The Wisdom and Manipulability of Threads

Robin Engelhardt[1], Jacob Stærk-Østergaard[1] and Vincent F. Hendricks[1]

[1] *Center for Information and Bubble Studies, Department of Communication, University of Copenhagen, Karen Blixens Plads 8, DK-2300 Copenhagen S.*

Abstract

Social decision-making is increasingly relying on digitized aggregates of people's opinions and judgments. These aggregates are frequently created and maintained as threads, i.e. as sequences of posts on a website. While it has been shown that knowledge of thread aggregates can distort individual decision-making, it is unknown how the ability to see preceding posts in a thread may influence collective accuracy, i.e. the wisdom of threads. We therefore investigate experimentally the accuracy of threads in which people make numerosity estimations of varying difficulty and varying degrees of social information in the form of visible preceeding estimates. We find a significant increase in collective accuracy for high numerosities and high social information in non-manipulated threads, while in manipulated threads collective accuracy declines quickly under the same conditions. Our result suggests people tend to rely on social information more when tasks are more complex, and that pristine threads without manipulations indeed can improve collective decision-making. Using gaussian mixture models we gain additional insights into how people use the social information provided and show how variably participants respond to social information.

1 Introduction

Social information in the form of opinions and judgments by other people is sampled sequentially. We read the news, hear rumors, listen to debates on TV, and flip through comments on social media platforms and blogs. These activities inform us and influence our decisions, but researchers still debate the conditions under which these types of social information help us make better decisions [1, 2, 3, 4], lead us astray [5, 6, 7, 8, 9], or just make us confused at a higher level [10, 11]. Collective estimates of a diverse group of people can outperform the majority of its members because any random confusion at the individual level is likely to average out and let the most accurate estimate prevail [12, 13, 14, 15]. Then again, confusion is not always randomly scattered around the truth. Systematic biases in individual perception may create measurable disruptions in the wisdom of crowds [16, 17, 18]. Social information can add to those biases and create cascades, echo chambers, bandwagoning and herd behavior [19, 20, 21, 22]. Partially sampled social information may lead to rich-get-richer dynamics [23] and to belief misattributions, which uphold harmful social practices despite being rejected by a majority of people [24, 25, 26, 27, 28]. Social information may also have been intentionally filtered or manipulated in various ways, for instance through group pressure [29], algorithmic filtering [30], false cues [10, 31, 32], or simply by plain misinformation [33], often with highly detrimental consequences for our economy and our health.

Observational data of decision-making processes is acutely sensitive to the social context in which people find themselves. Thus, researchers find it difficult to separate observational data into its social and individual components. How may we know how much weight an individual puts on her own 'independent' estimate relative to the weight put on the estimates by others? Randomized experimental studies have attempted to solve this problem by first letting participants make a magnitude estimate of an object without social information, and subsequently ask them to revise their estimate after having received information about other people's estimates of that object [3, 4, 6, 34, 35]. This two-stage information paradigm presumes that people change their mind because of the social information they have received. Other studies, however, have shown that people routinely can change their mind all by themselves, and that it may be more correct to assume an 'inner crowd' in the sense that people sample randomly from a probability distribution in their own mind [36, 37, 38]. This may make it difficult to differentiate between 'inner' samplings and 'outer' influences, and it would be preferable to have methods that can infer the degree of individual bias and/or social influence from a single estimate. In real life people rarely estimate twice and rarely have access to all the relevant social information. So if we wish to make somewhat realistic experiments, we should create situations in which people see only a small fraction of the social information out there. In addition, the social information is rarely sampled at random. People typically collect social information from certain people, in certain places and in certain time intervals, for instance via a discussion thread, in user reviews, or in similar successive pronouncements after which they make up their mind. The effects of the clustered, sparse, and sequential nature of social information on decision-making have not been investigated systematically before and turn out to have a substantial impact on both individual decision-making and on the accuracy of the crowd. In particular, we will show that seeing preceding estimates in a thread improves collective accuracy when

a) the task is difficult, b) the number of visible previous estimates is high, and c) the thread is pristine and unmanipulated in the sense that there is no filtering or rearrangement of preceding estimates. If, on the other hand, previous estimates are manipulated in such a way that people see only the highest (or lowest [1]) estimates done so far, accuracy decreases dramatically. We will also show that it is possible to reliably infer the degrees of individual bias and social influence from a single estimate and its associated social information by using gaussian mixed models fitted by an expectaion-maximization (EM) algorithm. Such models give very good fits by allowing for a more complex mean/variance relationships, and provide insights into the highly variabe use of social information within each thread.

2 Experimental Design

We created a number of artificial online threads on Amazon Mechanical Turk, where a total of 10,348 participants could make magnitude estimations of varying difficulty and with a varying number of visible preceding estimates. Participants were asked to either estimate the number of dots in one of four images each showing a certain number of dots, or, honoring Francis Galton [12], to estimate the weight of an ox on a photo together with information about the height and weight of a man standing next to it (see SI Appendix for screenshots of the experimental design). Dot estimation tasks have a long tradition in numerosity experiments [39, 40, 41] and have only recently been adopted as a useful 'model organism' for crowd aggregation research [42, 43]. The true numbers to be estimated by the participants may be interpreted as varying 'task complexity', which in the following will be expressed by the categorical variable $d \in \{55$ dots, 148 dots, 403 dots , 1097 dots, 1233 kilo$\}$ (with an expected ordering for the dots-experiments, but not necessarily with respect to the ox-experiment), while the visible estimates by preceding participants in a thread, $v \in \{0, 1, 3, 9\}$, may be interpreted as the degree of 'social information' available. Participants were placed randomly in one of the 20 thread configurations ($d \times v$ treatments) and made their estimate one after another. Treatments with $v = 0$ thus correspond to a control condition for each d that contain no social information.

In order to test for the manipulability of threads, we ran 5×3 additional treatments with 4,522 participants that saw the same five images, but instead of seeing the v preceding estimates, saw the v *highest* estimates made so far. This is a very heavy-handed type of thread filtering which we presumed would nudge participants to make ever higher estimates.[2] No participant who had seen a certain image would be able to participate in another treatment containing the same image again. In addition to a participation fee and a variable waiting fee, all participants in all treatments received a bonus of \$1 if their estimate was within 10% of the true value. See the Materials and Methods section and the Appendix for additional information about the experimental design.

[1] Data not reported here

[2] In order to keep the v observed estimates in a somewhat realistic range, participants were not able to see estimates made that were above 100.000 and 10.000 in the dots- and ox-experiments, respectively.

3 Methods

3.1 Analysis of group medians

For a given thread, let d and v denote the number of dots and available previous estimates, respectively. For observed medians $M_{dv}, d = 55, 148, 403, 1097, v = 0, 1, 3, 9$ the log ratios $y_{dv} = \log(M_{dv}/d)$ were modeled using a linear normal model. For a given thread, $\log d$ was used as a quantitative variable whereas v was used as a categorical factor. Hence,

$$y_{dv} = \alpha + \beta_d \log d * \sum_{l \in \{0,1,3,9\}} \beta_l \mathbf{1}(v = l) + \varepsilon_{dv}, \quad (1)$$

where $*$ denotes interaction between d and v in the model, and $\mathbf{1}(\cdot)$ denotes the indicator function

$$\mathbf{1}(v = l) = \begin{cases} 1 & \text{if } v = l \\ 0 & \text{otherwise.} \end{cases}$$

The goodness of fit was assessed by analyzing the residuals $\varepsilon_{dv} \sim \mathcal{N}(0, \sigma^2)$.

3.2 Effect of social influence

As in the analysis of group medians, the individual estimates were transformed as the log-ratio, relative to the true number of dots. Thus, the modeled response variable based on n observations $e_1, \ldots, e_n$ from a thread with d dots was $y_i = \log(e_i/d), i = 1, \ldots, n$. The threads were modeled individually, thus the only explanatory variable included was the social information available. However, in the experiments the participants were given either $v = 1, 3$ or 9 previous estimates. Following the idea of [4] the geometric mean of these estimates was used as a measure of the social information available. As with the observed estimates, the social information was transformed as the log-ratio relative to the true number of dots d. If we let $z_{i,\{1,\ldots,v\}}$ denote the v estimates available for the i'th participant, then for the un-manipulated threads

$$z^h_{i,\{1,\ldots,v\}} = \{e_{i-1}, \ldots, e_{i-v}\}$$

and for the manipulated threads

$$z^m_{i,\{1,\ldots,v\}} = \{e^*_{(1)}, \ldots, e^*_{(v)} | e^*_j = e_j, j < i\},$$

where $e^*_{(1)} \geq e^*_{(2)} \geq \cdots \geq e^*_{(v)} \geq e^*_{(l)}, v < l \leq j < i$ denote the (decreasingly) ordered estimates e^*_j among $e_1, \ldots, e_j$. The superscripts h, m refer to the type of information available, either historical (h) or manipulated (m). Hence, the historical information are the preceding v estimates, whereas the manipulated information are the v largest estimates among all preceding estimates. The quantified social information then be-

comes

$$x_i^h = \left(\prod_{l=1}^{v} e_{i-l} \right)$$
$$x_i^m = \left(\prod_{l=1}^{v} e_{(j_l)} \right).$$

However, contrary to [4], the actual estimates were available to the participants, hence in this context $x_i^{h,m}$ becomes a proxy for the available social information. As such, the model used was

$$\mu_i = \alpha + \beta_{\text{info}} x_i \quad , i = 1, \ldots, n \tag{2}$$

where x_i is the social information available for observation i.

To cope with heavy tails present in the data, a Gaussian Mixture Model (GMM) was used to model the effects of social influence. Given k states, the model fits k independent Gaussian distributions with parameters $\mu_j, \sigma_j^2, j = 1, \ldots, k$ along with weights $\delta_{ij}, i = 1, \ldots, n, j = 1, \ldots, k$ for each state. Note that the weight of each state is intrinsic to each of the n observations. Thus, every observation is modeled as a weighted sum of (independent) Gaussian distributions, which is itself a Gaussian distribution with parameters

$$\begin{aligned} \mu_{w,i} &= \textstyle\sum_{j=1}^{k} \delta_{ij} \mu_j \\ \sigma_{w,i}^2 &= \left(\textstyle\sum_{j=1}^{k} \delta_{ij} \sigma_j \right)^2 \end{aligned} \quad , i = 1, \ldots, n \tag{3}$$

where $\mu_{w,i}$ and $\sigma_{w,i}^2$ refer to the weighted parameter estimates for observation i. Applying (3) to (2), the effect of social information thus becomes a weighted sum of β estimates

$$\beta_{w,i} = \sum_{j=1}^{k} \delta_{ij} \beta_{\text{info},j}, \quad i = 1, \ldots, n, j = 1, \ldots, k. \tag{4}$$

It should be emphasized that the weighted $\beta_{w,i}$ is unique for each observation due to the weights δ_{ij}, contrary to a standard regression model where all observations are assumed to adhere to the same β effect. For the GMM this can be interpreted as each participant is modeled as a weighted average of k strategies. Hence, the personal strategy is unique, due to δ_{ij}, but weighted among k general axes. The goodness of fit for the GMM was assessed by calculating residuals

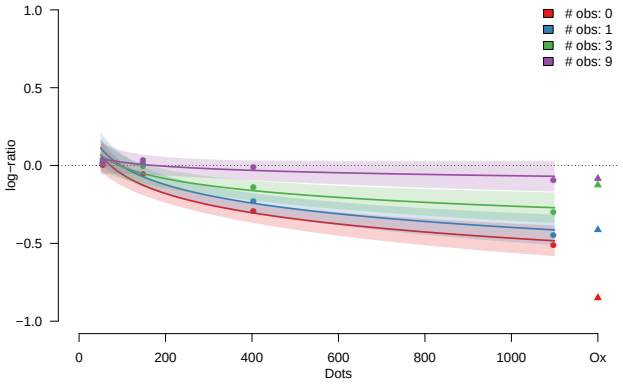
$$\hat{\varepsilon}_i = \frac{y_i - \hat{\mu}_{w,i}}{\hat{\sigma}_{w,i}^2} = \begin{cases} (y_i - \hat{\alpha} - \hat{\beta}_{w,i} x_i) / \hat{\sigma}_{w,i}^2 \\ (y_i - \hat{\alpha} - \hat{\beta}_{w,i}^{\text{info}} x_i - \hat{\beta}_{w,i}^{\text{pre}} y_i^0) / \hat{\sigma}_{w,i}^2 \end{cases} \tag{5}$$

and assessing the empirical distribution of $\hat{\varepsilon}_i, i = 1, \ldots, n$ against a standard $\mathcal{N}(0, 1)$ distribution.

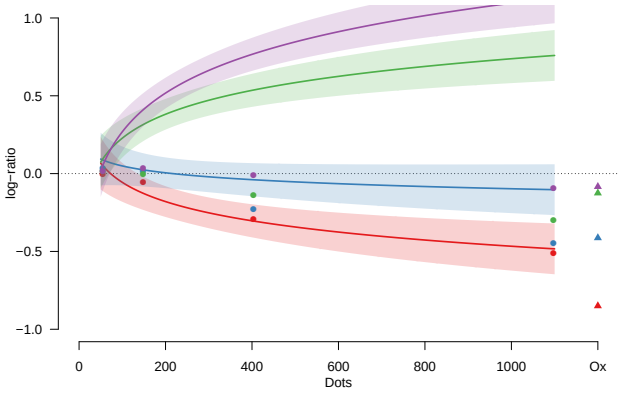
4 Results

4.1 Analysis of group medians

The two types of threads: historical (h) and manipulated (m) were modeled individually. This was to take into account the very different standard deviations of the two thread types. The relationship between the observed median ratio y_{dv} and the number of dots d were found to be optimal for both types, according to quantile-quantile plots, when modeling y_{dv} against $\log d$.



(a) Historic threads



(b) Manipulated threads

Figure 1: Relationship between median log-ratio and number of dots (lines) with 95% confidence bounds (shaded areas). The colors represent the four settings on number of visible estimates. There is a clear relation between number of dots and number of visible estimates in both thread types, but it is more pronounced for manipulated series. The ox thread (triangles) is added for reference.

Figure (1) presents the relationship between the number of dots d, visible estimates v and the log-ratio of the estimates against the true value y_{dv}. The historic threads, Figure (1a) show that for $v = 9$ there is no significant relationship between dots and the log-ratio, and that the log-ratio is not significantly different from 0. However, for $v < 9$, the log-ratio is significantly different from zero and decreasing with the number of dots, implying that there is a negative bias in thread performance and that this becomes worse when the task becomes more difficult, i.e. with more dots. The overlap among these groups also reveal that the thread performances, in terms of the median log-ratio, is comparable among $v = 0, 1, 3$. For the manipulated series, Figure (1b) all threads are significantly different. For $v = 1$ there is still a negative trend in the

bias, meaning that the manipulation is not effective. It is, however significantly different from $v = 0$, so some manipulation does occur with $v = 1$. For $v = 3, 9$ the manipulation is very clear, for both cases showing a large positive bias that increases with the number of dots. These findings imply that as a task becomes demanding, larger d, the amount of social information available, v, has a significant impact on thread performance. This is especially evident when threads are manipulated. In the two type of ox threads, it is also noticeable that the bias decrease with the number of observations, these observations where not included in the model, due to the fact that the task is qualitatively different compared to the dots scale, hence no confidence bounds are available to claim any significance on these observations.

4.2 Effect of social influence

A GMM (3) was fitted to all threads, including the ox threads. To choose a number of states k, each thread was fitted with 2, 3 and 4 states and the Bayesian Information Criteria (BIC) was evaluated to settle on the final k states. The BIC was used rather than the more usual Akaikes Information Criteria (AIC), since the BIC penalized the number of parameters more heavily. The aim was to settle with a decent model fit, with fewer parameters to avoid overfitting. Using objective criteria to choose a number of states/clusters is often prone to simply let this number increase (ref: Tibshirani, Elements of Statistical Learning section 14.3.11) and as such the AIC yielded much larger k values ($k \gg 4$). Hence, the BIC was chosen to simplify the models more so than the AIC. For each model, the standardized residuals (5) were evaluated visually using quantile-quantile (QQ) plots against a standard normal $\mathcal{N}(0, 1)$. These plots (see Figure **??**), revealed that generally the models using $k = 2, 3, 4$ states fitted the data quite well, with only a few models displaying a less adequate fit. This could indicate that either a larger k would be needed or that these threads simply contained very irrational responses from the models perspective. In the latter case a different modeling strategy might remedy this, however we will not pursue this idea in this paper, since the overall picture was that the GMM could cope with most of the observations in the various threads. Below a few of the threads are analyzed more in depth to show how the model framework can reveal interesting features of the data. Some of the data from [4] is also analyzed to give some perspective on the model as well as validating conclusions from the model based on only social information as explanatory, i.e. not including the pre-estimate of individuals. To better compare results from our experiments with [4], threads with 1 and 3 views were chosen as well as 1097 and 403 dots, to analyze threads of higher difficulty. This difficulty is closer to the questionnaire used in [4]. However, we also present the analysis of the manipulated thread with 1097 dots and 9 views, as this example clearly reveals how the manipulation dynamics unfold.

Each thread is presented in three figures, the top figure shows the log-ratio estimates as the thread evolves over time, bottom left show the log-ration of the social information (x-axis) against the log-ratio estimates (y-axis) and bottom right shows the distribution of the individually β_w's (4) with 95% intervals derived from the fitted models. These plots also show the interquartiles (25%, 50%, 75%) of the β_w distribution to give an impression of the relative group sizes. The middle plot displays a color scale that is the same among all threads. The green color is fixed around 0.4, indicating approximately average β_w's corresponding to participants that use social information to some extend (Compromisers). The red color indicates low β_w values < 0.2, corresponding to participants using little or no information, or simply contest the available information (Keepers/Contrarians). Finally the blue color indicates high β_w values > 0.6 corresponding to participants that are highly influenced by social information (Followers). Note that the interpretation of the participants is of course only to some degree. Given a participant has some personal estimate much in line with the social information available, this participant will be seen as an Adopter, when in fact this is only partially true. For Compromisers, this group also consist of participants, whos estimates are simply more difficult to categorize. Group biases are negative, i.e. people tend to underestimate the true values. Thus, as seen below the Compromiser group also consist of this majority, yielding both social information and estimates that lie somewhere below the true value (here corresponding to negative values).

4.2.1 Historical thread: 1097 dots and 1 view

This thread consist of rather difficult task with only a little social information available. As such, participants are prone to use whatever information is given, indicating a group of Followers with very high β_w values (in blue). The group of Keepers/Contrarians is also quite significant, indicating that despite the difficulty participants still weight their own opinion to a large extend. The interquartiles show that $\approx 25\%$ are Keepers/Contrarians, where as $\approx 15\%$ can be interpreted as Followers.

4.2.2 Historical thread: 403 dots and 3 views

Compared to the results in Figure (2), the task in this thread is also semi-difficult. With more social information given, the middle group of Compromisers (green) is larger and Keepers/Contrarians (red) is rather small, which is also the case for the Followers (blue). Evidently a small group of herders is present (streak of blue estimates). Compare to Figure (2) it appears that with more social information participants are more prone to compromise between their own opinion and the information available.

4.2.3 Manipulated thread: 1097 dots and 9 views

This thread is very different from either one presented in Figures (2) and (3) due to manipulation. And interesting phenomenon is clearly displayed in

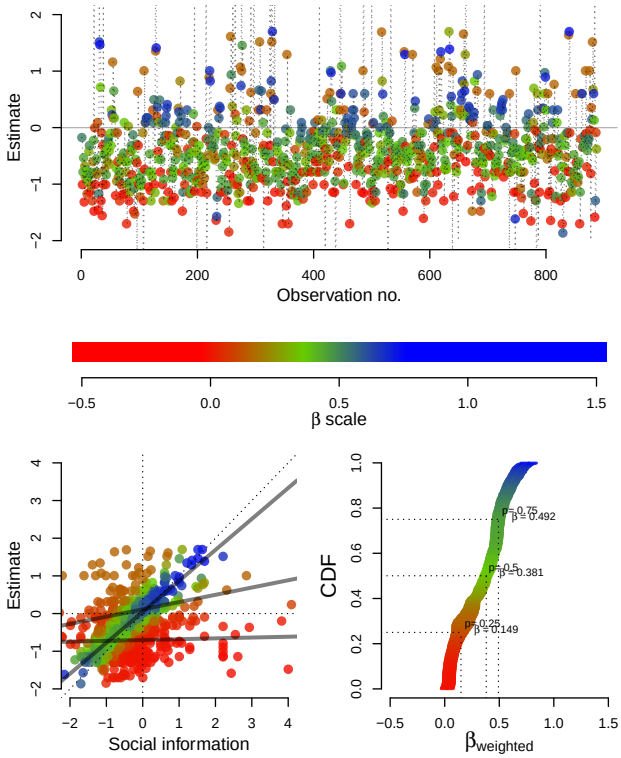


Figure 2: Historical thread from the AMT experiments with 1097 dots and 1 view, implying a difficult task with only little social information available.

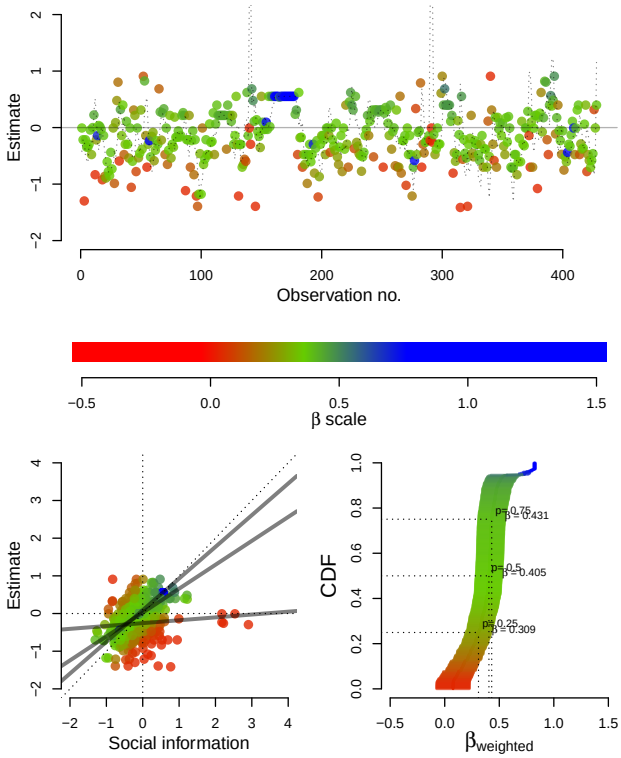


Figure 3: Historical thread from the AMT experiments with 403 dots and 3 views, implying a semi-difficult task with some social information available.

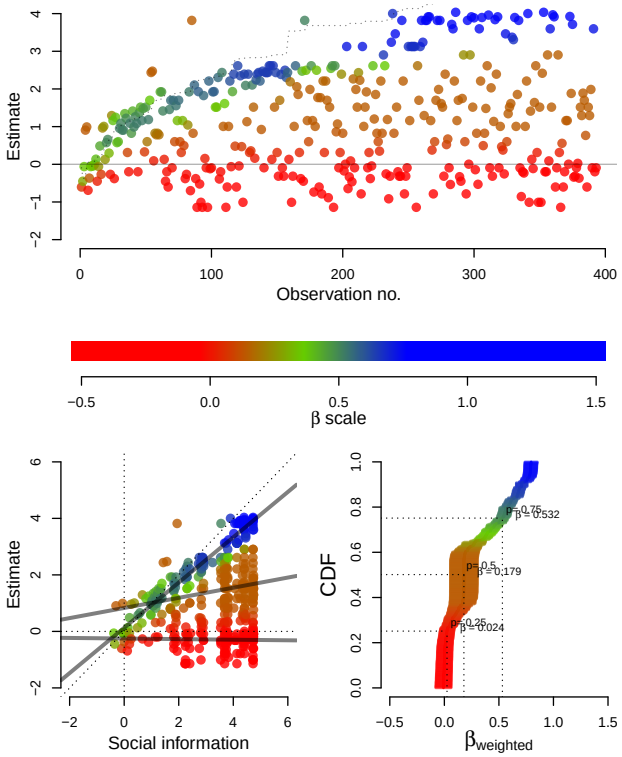


Figure 4: Manipulated thread from the AMT experiments with 1097 dots and 9 views, implying a difficult task with a lot of social information available.

the evolution of the thread. Initially the manipulation is not as pronounced, thus most participants are Compromisers (green), especially due to the large amount of social information presented. However, as the thread evolve (around 80-100 estimates in), the group clearly splits into a small group of Followers (blue) that drive the manipulation and a large group of Keepers/Contrarians clearly responding skeptically to the information presented with some participants being "suspicious" Compromisers (red-brown) with lower values $\beta_w \approx 0.18$. This group clearly adopts some information, but to a much lower degree than in Figure (3). A prototypical kink is visible in the β_w distribution, around the 65% quantile. This indicates the split between the more doubtful group and group of Followers. Note that the majority of green Compromisers is more or less gone after 80-100 estimates. Hence, this part of the β_w distribution is mostly from the initial part of the thread. Hence, further into the thread, mostly the Followers are in the blue region, indicating a high level of confidence in the social information.

4.2.4 Jayles data

The data from [4] is similar to the AMT in the sense that participants are given difficult estimation tasks. However, in this case, both a pre-estimate and their estimates after given social information were recorded. In addition, some manipulation took place in the sense that "expert" estimates (precise answers) were mixed into the geometric average of the social information. The questions were constructed so it would be near

impossible to know precise answers on all questions. Yet the difficulty of questions ranged from easy to very difficult, in the sense that participants had more/less intuition regarding the answer and their median answer was close to the true value or far apart, see Figure **??**. The models were all fitted using $k = 3$ states, this presented a decent fit (not shown) as evaluated by QQ-plots, as in the case of the AMT data.

Questions 1, 2 were of very similar type, referring to population sizes of two major Asian cities. The results in Figure 5-6 reveal that the participants attitude was similar in the use of social information. Roughly a third were Keepers/Contrarians (also found in Jayles??) 10-20% were Followers and the rest (majority group) consisted of Compromisers to some extend. Questions 8 and 10, Figures 7-8 where very different types, although the answer was numerically comparable to Questions 1 and 2. In the case of Figure 7, the group medians all had a negative bias, although less so with more experts included. This indicates that participants underestimated the answer to a large extend, but also weighed social information heavily. This is evident in Figure 7 where the almost complete lack of a Keeper/Contration group indicates that participants weighed their own opinion less. Regarding Question 10, Figure 8 reveals that the upper 50% of the β_w distribution is comparable to the one in Figure 7 on Question 8. However, clearly a significant group of participants weighed their own opinion highly due to the clear presence of a Keeper/Contrarian group. All the questions were presented to French students, which could explain this difference among Questions 8 and 10, since Question 8 dealt with the American Congress Library where as Question 10 dealt with sales of mobile phones in France. Finally, Question 23, Figure 9, is very different from the others in both type and numerical answer. Yet this Question seemed most difficult, according to Figure **??**. The Figure also shows that this question benefitted the most from the introduction of expert estimates as the severe negative group bias is mitigated strongly by these experts corrections of the social information. There are a few Keepers/Contrarians among the participants on this question, however Figure 9 shows that the estimated β_w's of this group are very imprecise, as the confidence bounds on the distribution of β_w is very wide in the bottom end. Contrary the Followers group displays much narrower confidence bounds indicating a good estimate of the upper 50% of the distribution. Hence, Figure 9 also indicates that Question 23 is difficult as most of the participants weigh the social information heavily, and those few who doesn't are quite difficult to categorize as the model imply they should be put somewhere between highly Contrarians ($\beta_w < 0.5$) to rather heavy Compromisers ($\beta_w \approx 0.5$).

"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas

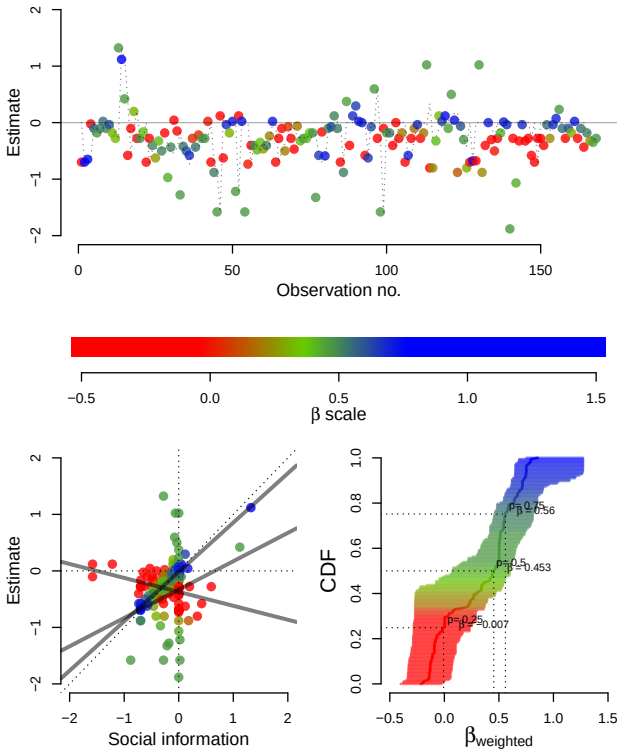


Figure 5: Jayles question 1: What is the population of Tokyo and its agglomeration? True answer: 38,000,000.

sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"

"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"

"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, to-

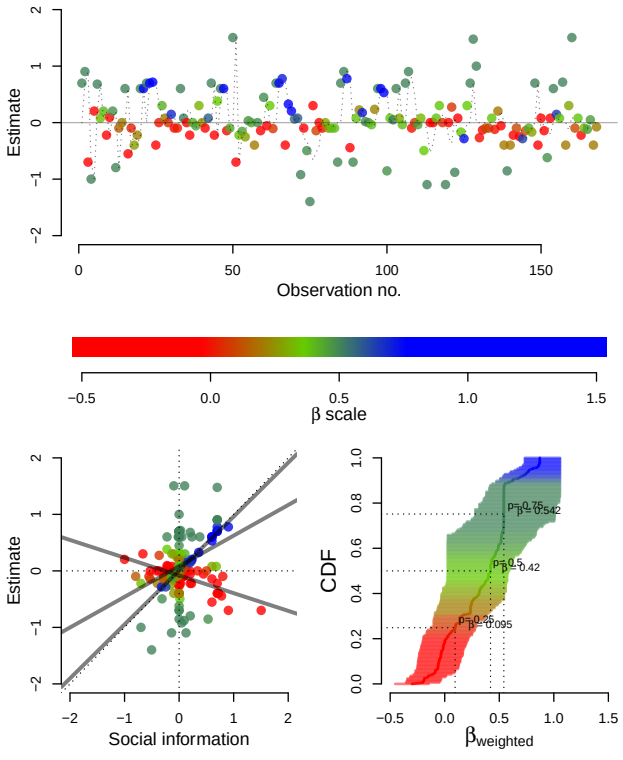


Figure 6: Jayles question 2: What is the population of Shanghai and its agglomeration? True answer: 25,000,000.

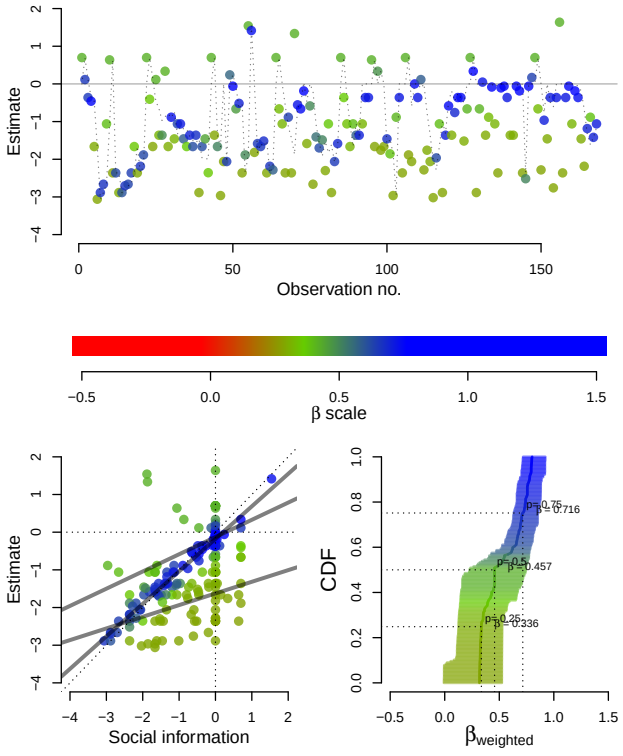


Figure 7: Jayles question 8: How many books does the American Congress library hold? True answer: 23,000,000.

tam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"

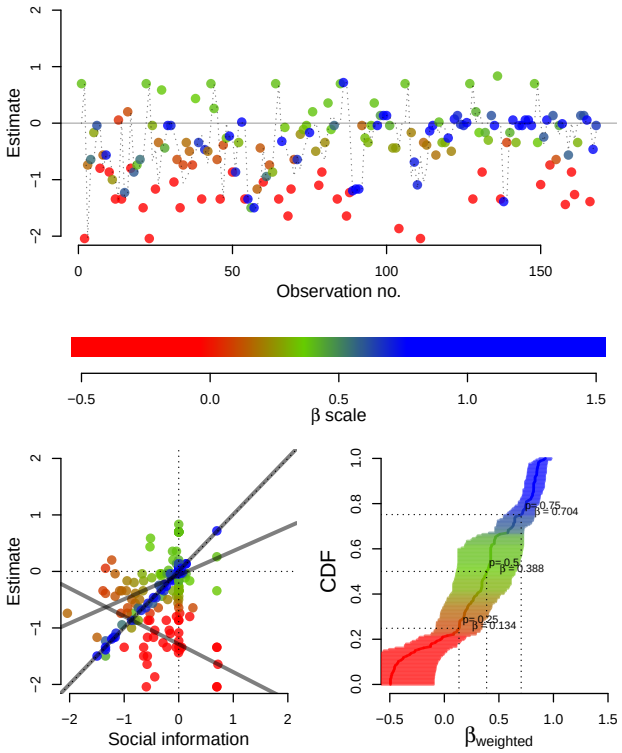


Figure 8: Jayles question 10: How many cell phones are sold in France every year? True answer: 22,000,000.

"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum

fugiat quo voluptas nulla pariatur?"

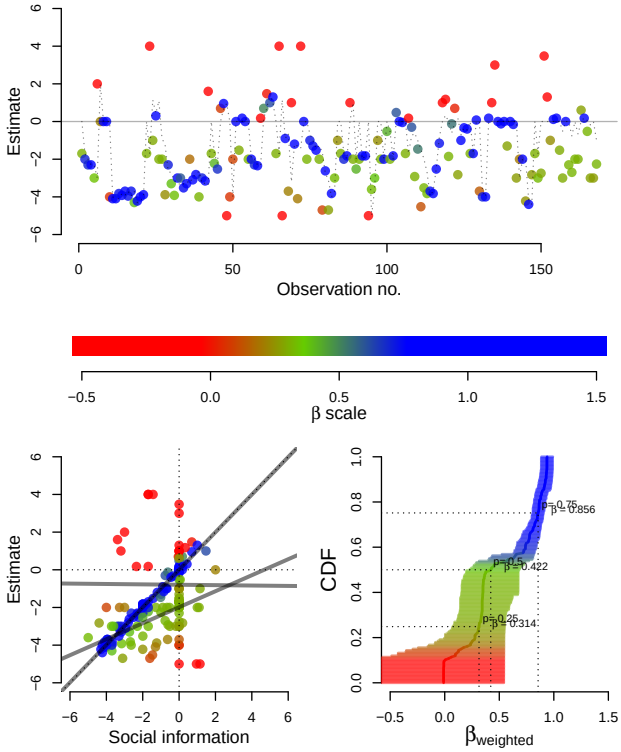
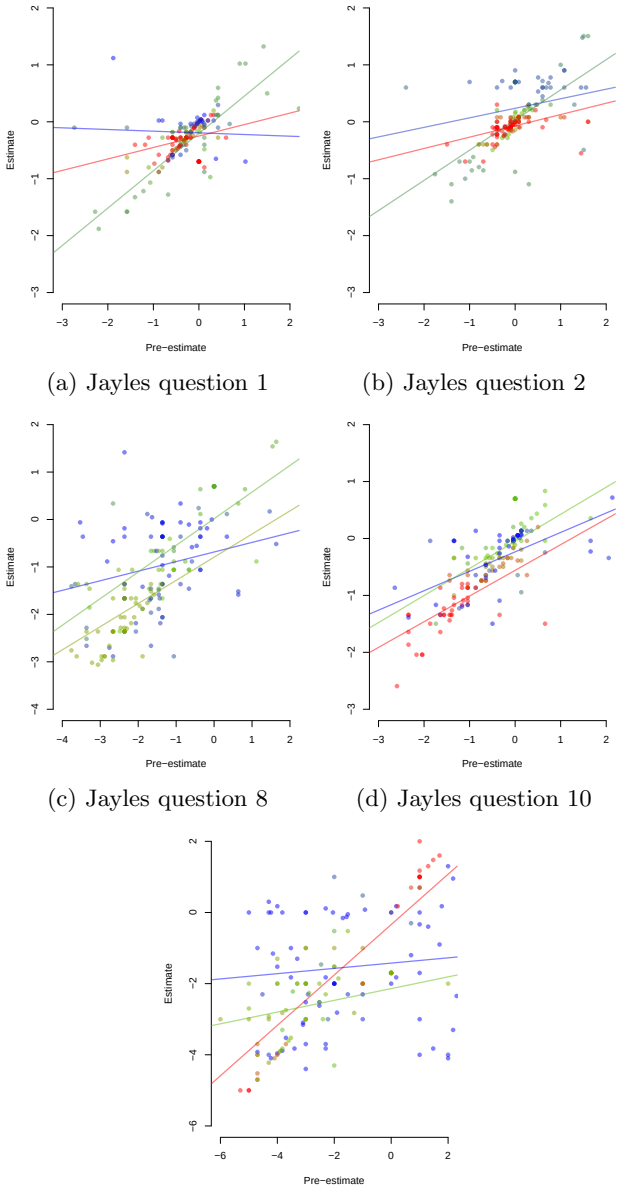


Figure 9: Jayles question 23: How many galaxies does the visible universe hold (in millions of galaxies)? True Answer: 100,000.

4.2.5 Jayles data, including pre-estimates

"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"

"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.



(a) Jayles question 1 (b) Jayles question 2

(c) Jayles question 8 (d) Jayles question 10

(e) Jayles question 23

Figure 10: Pre-estimates y_i^0 against final estimates y_i, with weighted regression lines. Colors are identical to the previous figures, and refer to the β_w estimate. The regression lines are colored according to the same scheme, using their estimates slopes.

Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?" "Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"

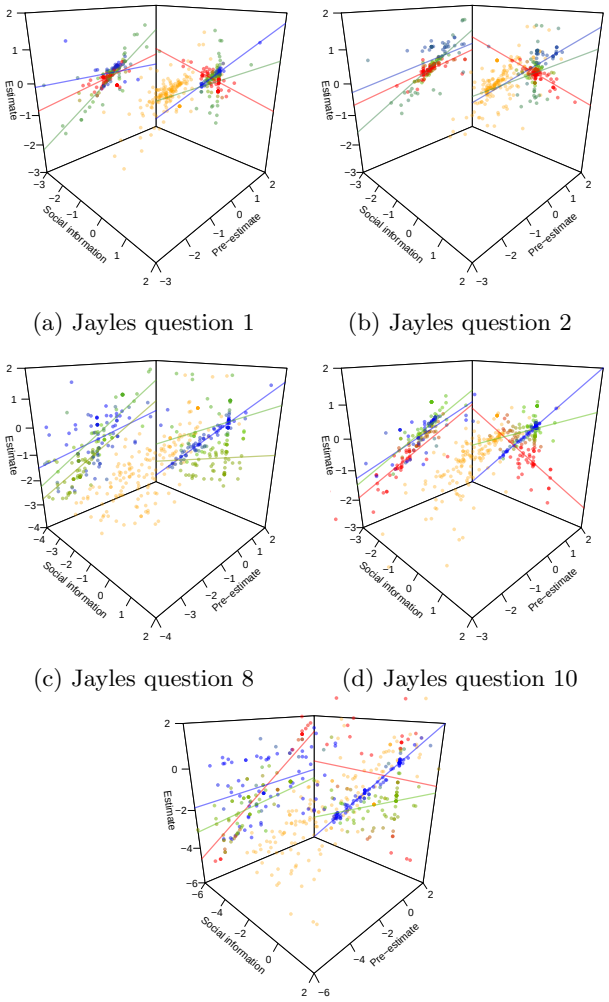
"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"

4.3 Aggregate Data

We model the log-error, $log_n(md_{(d,v)}/T)$, of thread medians, $md_{(d,v)}$, as a function of d using a generalized linear regression model. Since d is the dependent variable, we exclude the ox-data from the regression model as it is categorically different from the dots-experiments, and report only their final value.

4.4 Individual Threads

For each thread we use a mixture model to divide datapoints into states with corresponding independent gaussian components. Let Y denote the observation and $X \in \{1, 2, \ldots, k\}$ denote the state. Then the



(a) Jayles question 1 (b) Jayles question 2

(c) Jayles question 8 (d) Jayles question 10

(e) Jayles question 23

Figure 11: Pre-estimates y_i^0 against final estimates y_i and social information m_i, with weighted regression lines. Colors are identical to the previous figures, and refer to the β_w estimate. The regression lines are colored according to the same scheme, using their estimates slopes.

conditional variable $Y|X$ is gaussian with parameters depending on the state X. As an example, let $X = \{1, 2, 3\}$, i.e. we assume a model with $k = 3$ possible states. Then $Y|X = i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, 3$. Using notation $\delta = (\delta_1, \ldots, \delta_k)$ as the estimated weight of each state, with $\delta_i > 0$ and $\sum_{i=1}^{k} \delta_i = 1$, then the unconditional distribution of Y becomes

$$P(Y) = \sum_{i=1}^{k} \delta_i P(Y|X = i),$$

that is a weighted sum of gaussian probabilities, hence the name *gaussian mixture model* (GMM). This model structure can compensate for the heavy tails by adapting the different states to the more/less extreme observations. The number of states can be determined either by pre-specifying k or use an information criteria such as AIC: $6k - 2\log L$ or BIC: $3k\log(n) - 2\log L$ (note that both criteria are scaled with $3k$ parameters for k states), to pick a suitable number of states. Here BIC penalizes the number of parameters weighted by the (log) number of states as $3k$, corresponding to $(\delta_i, \mu_i, \sigma_i^2)$ for each state, whereas AIC only penalizes by a factor 2.

To keep things tractable we opt for a mix of specifying and using BIC where we set the number of states to $k \in \{2, 3, 4\}$ and then pick the number of states with the lowest BIC, conditional on convergence in the fitting routine. The model is fitted by the expectaion-maximization (EM) algorithm using the 'depmixS4' package in 'R'.

In order to account for other variables, the μ_i parameter can be extended, as in standard a linear regression model, with linear coefficients $\beta_{ij}, i = 1, \ldots, k, \quad j = 1, \ldots, p$. The specification of the mean μ_i is very flexible and can be adapted to each specific case. Here we opt for a model that can account for the social information present and in a special case a previous estimate for a given question, given prior to seeing social information.

4.4.1 Specific model for data analysis

Given N estimates $e_l, l = 1, \ldots, N$ from participants, we model the log-error $y_l = \log e_l/T$, where T denotes the true answer. Hence y_l is comparable among different questions, but obviously the location and deviation depends on the difficulty of the question. Harder questions tend to exhibit more devation from the true value and y_l tend to display a negative bias, often more pronounced when the true answer is of a high numerical value.

Each observation is associated with a piece of social information, i.e. previous estimates. In the setting of our experiments this amounts to a varying number of estimates: 0,1,3 or 9. In order to compare, we aggregate these into a single number m_l, namely the geometric average of the past v observations: $m_l = \left(\prod_{s=1}^{v} e_{l-s}\right)^{1/v}$. Of course in the case of 0 observations, this variable is excluded as no information is available. In this case the model only report a bias value. We also analyze the dataset from Jayles, where a prior estimate e_{l0} is given before any social information is revealed. Here we further include this estimate as explanatory. Hence we end up with 3 qualitatively different models for μ_i, here we omit the state and observation depedent subscripts i, l to ease readability

Note that all of these models of the mean are state dependent.

4.4.2 Model diagnostics

In order to evaluate a model fit, we calculate residuals. Here we exploit the fact that a sum of (independent) gaussian components is again a gaussian. Each observation y_l has an associated weighting of each state. These weights are positive and sum to 1 and can thus be interpreted as probabilities. That is, given k states, the model fits k gaussians with mean and variance $\hat{\mu}_i, \hat{\sigma}_i^2, i = 1, \ldots, k$. For the l'th observation, we let $\hat{\delta}_{il}$ denote the assigned probability of this observation belonging to state i. Note that the number of states is fixed when fitting a model, hence it does not represent a true number, but merely an assigned number of states. However, given this particular number of states, the model optimizes the fit as of which state the given observation should belong to. Hence, for each observation y_l, we can calculate the residuals first by substracting the weighted mean using $\hat{\delta}_{il}, i = 1, \ldots, k$. By scaling with the weighted variance analogously, we obtain centered and scaled residuals that should correspond to a standard normal $\mathcal{N}(0, 1)$. So we obtain the following residuals (we use the subscript l on the parameters to emphasize the possible dependency on expanatory variables associated with observation l)

$$\hat{\epsilon}_l = \frac{y_l - \sum_{i=1}^{k} \hat{\delta}_{il}\hat{\mu}_{il}}{\sum_{i=1}^{k} \hat{\delta}_{il}\hat{\sigma}_{il}} \sim \mathcal{N}(0, 1) \quad l = 1, \ldots, N$$

5 Results

In accordance with previous findings, participants do well when estimating small numbers independently. For higher numerosities, estimates vary widely and errors become substantial [16, 41, 44]. The median tends to underestimate the true value and the mean tends to overestimate the true value [18]. In comparison to Galton for instance, who found only a 0.8% difference between the median estimate and the true weight of the slaughtered ox [12], we find the median estimate to be less than half the true weight of the ox, as can be seen in the control condition $v = 0$ in the top left panel of fig. 12.

5.1 Unmanipulated threads

As soon as participants can see preceding estimates, accuracy improves substantially in difficult tasks, see the horizontal boxplots on the left hand side of figure 12. Increasing the view count v does not significantly improve collective accuracy in images with 55 and 148 dots. But as soon as the number of dots increases, social information starts to improve accuracy. And the larger the view count v, the more accurate the median (circles) becomes. Arithmetic means (triangles)

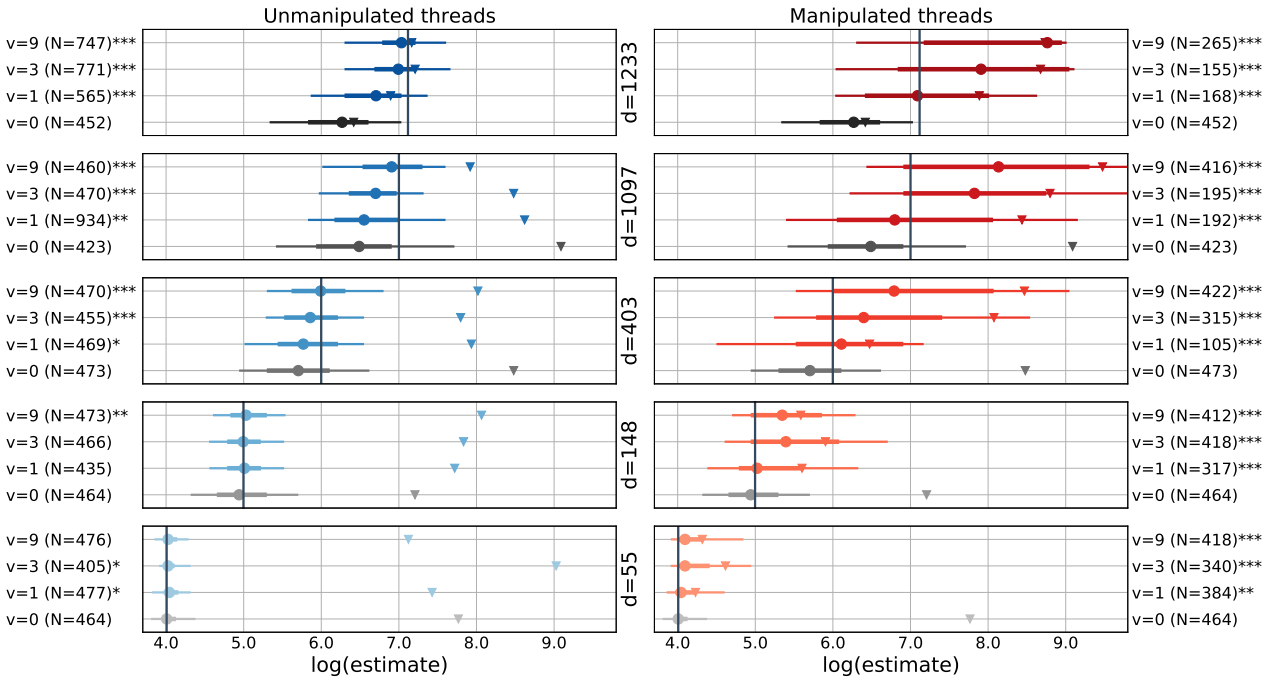


Figure 12: **Left:** Summary statistics of $d \times v$ treatments with a total of 10.348 magnitude estimates of either the number of dots in an image, $d \in \{55, 148, 403, 1097\}$, or the weight of an ox, $d \in \{1233\}$, while participants are able to see $v \in \{0, 1, 3, 9\}$ preceding estimates. Greys are the control treatments with $v = 0$. Estimates are log-transformed. Large circles indicate medians, triangles indicate arithmetic means, thick lines show interquartile ranges, thin lines show interdecile ranges, vertical black lines show the true value, and stars indicate significance levels compared to the control treatment (two-sided Wilcoxon-Mann-Whitney test). No outliers were removed, making the arithmetic means strongly right skewed. **Right:** Summary statistics of $d \times v$ treatments with a total of 4.522 additional magnitude estimates, where participants do not see the preceding estimates but the $v \in \{1, 3, 9\}$ *highest* estimates made so far. The controls, $v = 0$, are the same as on the left hand side.

quickly tend to overestimate the true value because free response elicitation of absolute values creates right-skewed, approximately log-normal distributions with a long tail, which inflate the means.[3] The interquartile and interdecile ranges show how high difficulty, d, leads to higher diversity (i.e. variation) in the dots-experiments, while a higher view-count tends to reduce variation when d is fixed, although not always.

5.2 Manipulated threads

The right hand side of figure 12 shows the equivalent results for manipulated threads. While participants are relatively unaffected by social information when difficulty is low in the unmanipulated threads, participants in the manipulated threads quick start to overestimate, even with $v = 1$. For higher d's and v's, estimates inflate ever more. In fact, we were able to create threads in which a majority of participants estimated the ox ($d = 1233$) to weigth more than 6000 kilo. Clearly, in the begining of a thread, only a few of the observed estimates are high numbers, making it not so probable for new participants to get influenced by them. But when threads become longer, still more of the observed estimates are very high numbers, making it more difficult to resist their influence. At a certain point, all visible estimates may be so improbably high, that participants may suspect foul play and start to ignore them. In some threads this effect can be observed by estimates branching off into two directions: those estimates showing herd behaviour by following the extremely high estimates seen, and those estimates ignoring them, see appendix figs 6 & 7 for examples. This may contribute to the high variance of the manipulated threads. Also note that for $v = 1$, the medians of the manipulated threads are often much closer to the true value than it is the case in the corresponding controls. This is so because the size of the manipulation (in this case showing only the single highest estimate in the thread) is just about enough to compensate for the naturally occuring underestimations in the controls.

5.3 Collective and individual performance

Figure 13 shows the benefits of social information for collective estimation accuracy (subplot **a**) and for individual estimation accuracy (subplot **c**) in unmanipulated threads. The effect is strong and robust in com-

[3] Due to the outliers, means become highly uninformative. Even when defining a cut-off for outliers, such as an error rate of 10, would still make the arithmetic means right skewed. Galton disliked the use of the mean for this very reason as it "would give a voting power to 'cranks' in proportion to their crankiness" [12]. While it is still debated [18] which measure is best suited to aggregate social information, we focus on the median: The median is easy to interpret, is robust against outliers, and best expresses the opinion of the crowd since the majority of participants in the thread deems every other estimate as too high or too low. See also the Appendix for a discussion of outliers.

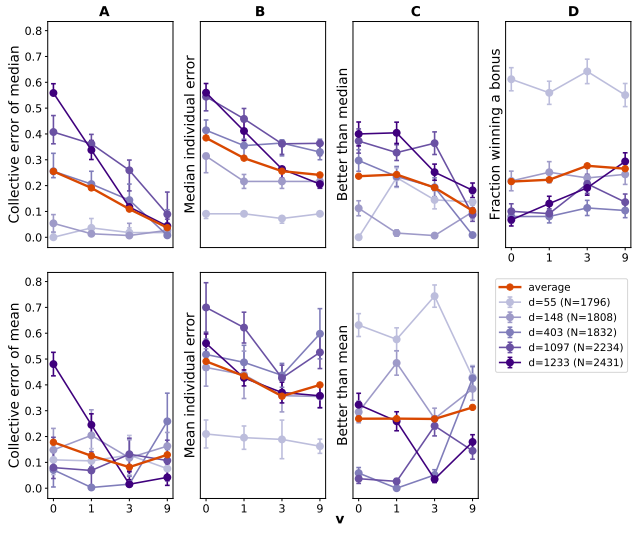


Figure 13: Collective and individual performances across d and v. In unmanipulated threads (blues), collective and individual errors decrease. In manipulated threads (reds), errors increase. **a**: In unmanipulated threads, the collective error, RE_C - i.e. the error of the median ($|truth - median|/truth$), drops substantially for high d and with higher v. Confidence intervals (95%) are show in error bars. **b**: In manipulated threads, collective errors increase dramatically for higher d and $v > 1$. The large error bar for $v = 3$ indicates that a large minority makes very high estimates. **c**: In unmanipulated threads, the median individual error, $MdRE$ - i.e. the median of $RE = |truth - estimates|/truth$, drops substantially for higher d and v, while for low d it is small and independent of v. **d**: In manipulated threads, the median individual error increases strongly for high d and high v, while again being independent of v when d is low. **e**: The WOC-index, W, measured as the fraction of participants who are worse than the error of the crowd median, RE_C, increases in general for higher v in unmanipulated threads, demonstrating an increasing wisdom of crowds-effect. **f**: In manipulated threads, the crowd becomes less wise, except for $v = 1$, which is due to the cancelling out of two opposing forces: a general underestimation bias and the influence of one observed high estimate. N shows the total number of estimates. No outliers were removed.

plex tasks with $d = \{403, 1097, 1233\}$. When threads are manipulated, however, the same reliance on social information creates very inaccurate collective as well as individual estimates (subplots **b** and **d**). For example: While the collective error of the median in the control with $d = 1097$ is 0.4, the error decreases to 0.1 when participants can see nine preceding estimates ($v = 9$), witnessing a highly beneficial effect on collective performance. Thus, a main result of the experiments is that participants tend to rely on social information more when the task is more complex, and that this reliance aids their decision-making as well as the wisdom of crowds in unmanipulated threads. The same reliance, however, leads to a very high collective and median individual error in complex tasks when the *highest* estimates are shown. While participants are able to calibrate their own estimation towards a more accurate value when they see preceding guesses, they are led astray as soon as these estimates have been purposefully changed. In a truthful world credulity is fine and well, but in a fraudulent one it is fatal.

A simple wisdom-of-crowds measure, W, is the fraction of participants who are less accurate than the median of the thread, i.e.

$$W = \frac{1}{N}\sum_{i}^{N} x_i,\ x_i = \begin{cases} 1 & \text{if } RE_i > RE_C \\ 0 & \text{otherwise} \end{cases}$$

where $RE_i = |e_i - T|/T$ is the error rate of estimate e_i by participant i, $RE_C = |md - T|/T$ is the error rate of the thread median, T is the true value, and N is the number of estimates. Figure 13**e** shows that W increases with increasing social information in the majority of cases, which means that less and less participants are better than the crowd median. Thus, increasing social information magnifies the wisdom of crowds effect in complex tasks. In contrast to [6] and in concert with [3, 45] this supports the claim that crowds indeed may become wise under social influence, even though wiser individuals give the aggregate wisdom of crowds measure a tougher baseline to compete against.

In manipulated threads the WOC-index decreases as expected (figure 13**f**), but is surprisingly high when participants see only the highest estimate in their thread at $v = 1$. This resonates well with the findings in [4], who showed that providing a moderate amount of incorrect information can counterbalance underestimation bias and improve collective performance. It is striking, however, that this manipulation works equally well in all threads (except in the easiest dot-experiment with $d = 55$), and confirms that human numerosity underestimation bias indeed is systematic, predictable and 'repairable' [4, 18, 40, 46].

5.4 Herd behavior

Visual inspection of thread dynamics (see minimal spanning trees in the Appendix, figs. S4-S7) reveals that most threads contain multiple occasions of herd behavior in the sense that participants frequently copy one of their preceding estimates. Prolonged herd behavior is rare in the unmanipulated dots-experiments (typically 1-5 instances in a thread of length 4-500), but common in the ox-experiments. Herd behavior drastically changes the visual appearance of the tree-like structure of the threads (see figures in appendix): Unmanipulated dot-threads typically show themselves as plump and well-formed trees with bushy branches, while ox-threads exhibit much sparser and fastigate

(i.e. vertically aligned) branches, indicating that herd behavior is much more pronounced when people estimate the weight of an ox.

As a simple point estimate of herd behavior we can use the proportion of participants who copy one of the preceding estimates they can see:

$$I(v) = \frac{1}{(N-v)} \sum_{i=v}^{N} x_i,\ x_i = \begin{cases} 1 & \text{if } e_i \in \{e_{i-1}, .., e_{i-v}\} \\ 0 & \text{otherwise} \end{cases}$$

where v is the number of visible estimates, and N is the total number of estimates in a thread. Of course, some people may become copycats only by accident, for instance due to the limited number of good guesses available in less complex tasks, or due to mental rounding, focal points [49], or simply by chance. The above proportion is therefore only informative when we compare it to the naturally occuring proportion of imitators in threads without social information, $I_0(v)$, which we estimate by counting how many times an estimate has occured by chance in v preceding steps in the control experiments with $v = 0$ (see table A1 in the appendix).

The relationship between the two proportions $I(v)$ and $I_0(v)$ is shown in figure 14a-d. Figures 14a and b show their difference, $\Delta_v = I(v) - I_0(v)$, giving an estimate of the overall magnitude of imitative behavior when social information is present. In contrast, figures 14c and d show the ratio between the two proportions $\Theta_v = \frac{I(v)}{I_0(v)}$, which estimates how much more (or less) people tend to copy each other when they have social information compared to when they have no such information. So if $\Theta_v \approx 1$ (i.e. $\Delta_v \approx 0$), the probability of imitation is no different from what would happen by chance. If $\Theta_v > 1$, imitation is on average Θ_v times higher than what would occur by chance, and if $\Theta_v < 1$, imitation is less than what would occur by chance, which means that people look at preceding estimates mostly in disbelief.

Figures 14a-d reveal that a majority of the complex unmanipulated threads have a positive probability difference ($\Delta_v > 0$, $\Theta_v > 1$), demonstrating that participants rely on social information more when the task is more complex. This relationship, however, is nontrivial: First, herd behavior is consistently higher in the ox-experiments (dark blues and reds) than in the dots-experiments ($p < 0.05$ in 5 out of 6 ox-threads, but only in 10 out of 24 dots-experiments). Second, the willingness to copy previous estimates is much higher than chance when $v = 3$ ($p < .05$ in 8 out of 10 threads) rather than when $v = 1$ ($p < .05$ in 3/10) or $v = 9$ ($p < .05$ in 4/10), which suggests that there may exist an optimal level of social information at which influence is maximized. Third, disbelief is not uncommon in low-complexity tasks, especially in the manipulated threads. This makes intuitively sense, because participants often see a bunch of highly inflated estimates in the manipulated threads, which may prompt them to make up their own mind instead.

Since the imitation proportions count only those people who make an exact copy of a predecessor estimate, we complement the analysis in the appendix with a more encompasing approach and include participants who make an estimate that is 'close' to one or more predecessor estimates. Figure SXX in the appendix shows the cumulative distribution of such copycats in terms of their percentual distance, p, to one or more preceding estimates in unmanipulated threads with the highest complexity and the highest amount of social information. The extend of herd behavior beyond simple copying is again much more pronounced in the ox-threads, indicating that people use different strategies when estimating different things.

—

blablabla..

6 Discussion

In the original wisdom of crowds experiment by Francis Galton, people estimated the weight of a "fat" slaughtered ox [12]. A multitude of formal and informal studies have subsequently tried to replicate Galton's findings by using various other items, from coins and jelly beans to stock prizes and best movie nominations. This has happened even though researchers were quick to point out that the social context of such experiments, the attributes of the items such as their numerical magnitude [16, 41], their extremeness and complexity [17, 50], as well as the expertise of the participants [51], are important factors to consider.

The differences in the results of our dot-experiments and our experiment with an even stouter ox show that item attributes indeed matter. Estimating dots on a screen may or may not require as much expertise as estimating the stoutness of cattle, but the observed differences in the levels of herd behavior at least indicate that participants use different strategies when estimating different things. What does not differ, however, is a systematic underestimation bias of large numbers and a rapid improvement in individual as well as in collective accuracy when the number of visible preceding estimates is increased.

This means that the social information contained in a thread in fact helps people to calibrate their own decision-making process, even if the information may be wrong. Concerns about gullibility and correlation of judgment errors may thus be less of an evil compared to the positive effects of seeing honest opinions from other people in a thread.

One possible reason for frequent herd behavior in the ox-threads is that participants are much more uncertain about their ability to estimate cattle weight compared to their ability to estimate dots on a screen. A high degree of uncertainty - or even an acceptance of cluelessness - may prompt participants to copy previous estimates rather than to make independent estimates by themselves [52].

Experimental design and data collection

Controlled experiments of thread dynamics are rare in the research literature due to the difficulties in keeping a large number of participants in a queue. We

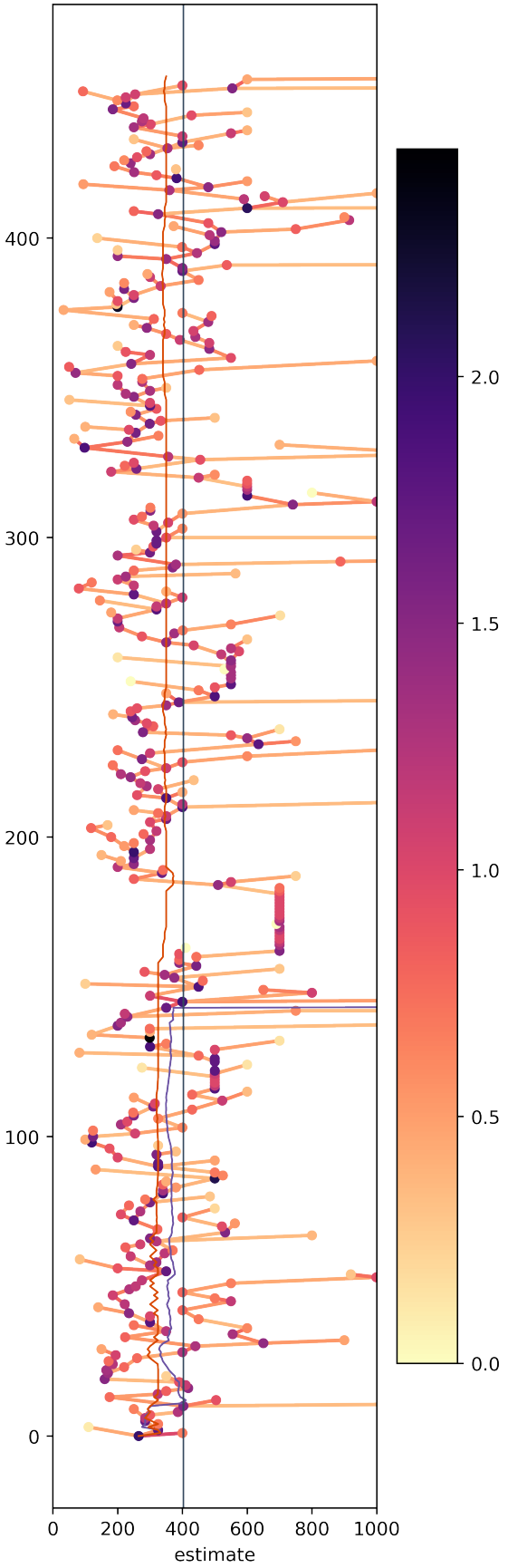


Figure 14: Herd behavior across tasks complexities and view counts. (**a, b**): The probability difference between proportions, $\Delta_v = I(v) - I_0(v)$, is a simple measure of herd behavior under social influence. In unmanipulated threads (**a**), only the ox-experiment sees a continuous increase in herd behavior across all v, while most dot-threads have a peak at $v = 3$. In the manipulated threads (**b**), Δ_v becomes negative in many threads, indicating that many participants do not belief the estimates they can see. Wald two-sided confidence intervals (95%) are given in error bars. (**c, d**): The ratio of the proportions, $\Theta_v = \frac{I(v)}{I_0(v)}$, is in the majority of cases higher than unity, and significantly so in most ox-experiments (darkest blue and darkest red), and in most threads with $v = 3$. As confindence intervals (95%) we use the recommended [47] Koopman asymptotic score interval as given by Nam [48]. **e,f**: The ratio of proportions of bonuses given when being an imitator and being not an imitator..... Color codings and sample sizes are the same as in figure 13.

designed our experiment along the same lines as the classical information cascade experiments by Anderson and Holt [19]. While such a design is not very feasible in normal laboratory conditions (at least for threads with several hundred participants), it is well suited for online labor markets and crowdsourcing platforms such as Amazon Mechanical Turk (AMT, see SI Appendix for further discussions of the pro and cons of AMT). We thus recruited participants from AMT to make a total of 10.808 magnitude estimations in various threads containing either an image of a certain number of dots ($d \in \{55, 148, 403, 1097\}$) or an image of an ox ($d = 1233$). After accepting our 'HIT' ('human intelligence task') and providing informed consent, participants waited in a 'waiting room' until the 'choice room' became available. When entering the choice room participants could see an image d together with $v \in \{0, 1, 3, 9\}$ estimates made by previous participants. After making an estimate, participants were thanked and paid a participation fee of \$0.10 and bonus of \$1 if their estimate was within 10% of the true number. The average time used was less than two minutes, see SI Appendix for screenshots and more detailed descriptions of the design and setup.

Acknowledgments

The experiments were implemented by Robin Engelhardt and Mikkel Birkegaard Andersen. Server infrastructure and devops was handled by Mikkel Birkegaard Andersen. The authors wish to thank Philipp Chapkovski for help with the cascade design, Rasmus Rendsvig for modeling discussions, and Ulrik Nash and Peter Norman Sørensen for their comments on an earlier draft of this paper. This research was approved by the Institutional Review Board at the University of Copenhagen and included informed consent by all participants in the study. The authors gratefully acknowledge the support provided by The Carlsberg Foundation under grant number CF 15-0212.

References

[1] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).

[2] Gürçay, B., Mellers, B. A. & Baron, J. The power of social influence on estimation accuracy. *J Behav Decis Making* **28**, 250–261 (2015).

[3] Becker, J., Brackbill, D. & Centola, D. Network dynamics of social influence in the wisdom of crowds. *P Natl Acad Sci Usa* **114**, E5070–E5076 (2017).

[4] Jayles, B. *et al.* How social information can improve estimation accuracy in human groups. *P Natl Acad Sci Usa* **114**, 12620–12625 (2017).

[5] Caplan, B. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies-New Edition* (Princeton University Press, 2011).

[6] Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *P Natl Acad Sci Usa* **108**, 9020–9025 (2011).

[7] Minson, J. A. & Mueller, J. S. The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychol Sci* **23**, 219–224 (2012).

[8] King, A. J., Cheng, L., Starke, S. D. & Myatt, J. P. Is the true 'wisdom of the crowd'to copy successful individuals? *Biol Lett* **8**, 197–200 (2011).

[9] Le Mens, G., Kovács, B., Avrahami, J. & Kareev, Y. How endogenous crowd formation undermines the wisdom of the crowd in online ratings. *Psychol Sci* **29**, 1475–1490 (2018).

[10] Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).

[11] Salganik, M. J. & Watts, D. J. Web-based experiments for the study of collective social dynamics in cultural markets. *Top Cogn Sci* **1**, 439–468 (2009).

[12] Galton, F. Vox populi (the wisdom of crowds). *Cah Rev The* **75**, 450–451 (1907).

[13] Muth, J. F. Rational expectations and the theory of price movements. *Econometrica* 315–335 (1961).

[14] Surowiecki, J. *The wisdom of crowds* (Anchor, 2005).

[15] Hong, L. & Page, S. E. *Some microfoundations of collective wisdom*, 56–71 (2008).

[16] Izard, V. & Dehaene, S. Calibrating the mental number line. *Cognition* **106**, 1221–1247 (2008).

[17] Nash, U. W. The curious anomaly of skewed judgment distributions and systematic error in the wisdom of crowds. *PLoS One* **9**, e112386 (2014).

[18] Kao, A. B. *et al.* Counteracting estimation bias and social influence to improve the wisdom of crowds. *J R Soc Interface* **15**, 20180130 (2018).

[19] Anderson, L. R. & Holt, C. A. Information cascades in the laboratory. *Amer Econ Rev* 847–862 (1997).

[20] Bikhchandani, S., Hirshleifer, D. & Welch, I. A theory of fads, fashion, custom, and cultural change as informational cascades. *J Polit Economy* **100**, 992–1026 (1992).

[21] Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on facebook. *Science* **348**, 1130–1132 (2015).

[22] Banerjee, A. V. A simple model of herd behavior. *Quart J Econ* **107**, 797–817 (1992).

[23] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[24] Katz, D., Allport, F. H. & Jenness, M. B. Students' attitudes; a report of the syracuse university reaction study. Tech. Rep., Oxford (1931).

[25] Darley, J. M. & Latané, B. Bystander intervention in emergencies: Diffusion of responsibility. *J Pers Soc Psychol* **8**, 377 (1968).

[26] Ross, L., Greene, D. & House, P. The "false consensus effect": An egocentric bias in social perception and attribution processes. *J Exp Soc Psychol* **13**, 279–301 (1977).

[27] Noelle-Neumann, E. The spiral of silence a theory of public opinion. *J Commun* **24**, 43–51 (1974).

[28] Lee, E. *et al.* Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour* 1–10 (2019).

[29] Asch, S. E. & Guetzkow, H. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men* 222–236 (1951).

[30] Pariser, E. *The filter bubble: What the Internet is hiding from you* (Penguin UK, 2011).

[31] Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: A randomized experiment. *Science* **341**, 647–651 (2013).

[32] Hanson, W. A. & Putler, D. S. Hits and misses: Herd behavior and online product popularity. *Market Lett* **7**, 297–305 (1996).

[33] Hendricks, V. F. & Vestergaard, M. *Reality Lost: Markets of Attention, Misinformation and Manipulation* (Springer, 2018).

[34] Sniezek, J. A. & Buckley, T. Cueing and cognitive conflict in judge-advisor decision making. *Organ Behav Hum Decis Process* **62**, 159–174 (1995).

[35] Mavrodiev, P., Tessone, C. J. & Schweitzer, F. Quantifying the effects of social influence. *Sci Rep* **3**, 1360 (2013).

[36] Vul, E. & Pashler, H. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science* **19**, 645–647 (2008).

[37] Herzog, S. M. & Hertwig, R. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science* **20**, 231–237 (2009).

[38] Herzog, S. M. & Hertwig, R. Harnessing the wisdom of the inner crowd. *Trends in cognitive sciences* **18**, 504–506 (2014).

[39] Minturn, A. & Reese, T. The effect of differential reinforcement on the discrimination of visual number. *J Psychol* **31**, 201–231 (1951).

[40] Indow, T. & Ida, M. Scaling of dot numerosity. *Percept Psychophys* **22**, 265–276 (1977).

[41] Krueger, L. E. Single judgments of numerosity. *Percept Psychophys* **31**, 175–182 (1982).

[42] Horton, J. J. The dot-guessing game: A 'fruit fly'for human computation research. *SSRN 1600372* (2010). URL `http://dx.doi.org/10.2139/ssrn.1600372`.

[43] Ugander, J., Drapeau, R. & Guestrin, C. The wisdom of multiple guesses. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 643–660 (ACM, 2015). URL `https://doi.org/10.1145/2764468.2764529`.

[44] Krueger, L. E. Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics* **35**, 536–542 (1984).

[45] Farrell, S. Social influence benefits the wisdom of individuals in the crowd. *P Natl Acad Sci Usa* **108**, E625–E625 (2011). URL `https://doi.org/10.1073/pnas.1109947108`.

[46] Nash, U. W. Sequential sampling, magnitude estimation, and the wisdom of crowds. *Journal of Mathematical Psychology* **77**, 165–179 (2017).

[47] Fagerland, M. W., Lydersen, S. & Laake, P. Recommended confidence intervals for two independent binomial proportions. *Statistical methods in medical research* **24**, 224–254 (2015).

[48] Nam, J.-M. Confidence limits for the ratio of two binomial proportions based on likelihood scores: non-iterative method. *Biometrical Journal* **37**, 375–379 (1995).

[49] Schelling, T. C. *The strategy of conflict* (Harvard university press, 1980).

[50] Taleb, N. N. Errors, robustness, and the fourth quadrant. *Int J Forecasting* **25**, 744–759 (2009).

[51] Perry-Coste, F. The ballot-box. *Cah Rev The* **75**, 509 (1907).

[52] Navajas, J., Armand, O., Bahrami, B. & Deroy, O. Diversity of opinions promotes herding in uncertain crowds. *PsyArXiv* (2018). URL `http://dx.doi.org/10.31234/osf.io/mvy25`.