The Wisdom and Persuadability of Threads

Robin Engelhardt,¹ Vincent F. Hendricks,¹ and Jacob Stærk-Østergaard¹

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

Online discussion threads are important means for individual decision-making and for aggregating collective judgments, e.g. the ‘wisdom of crowds’. Empirical investigations of the wisdom of crowds are currently ambivalent about the role played by social information. While some findings suggest that social information undermines crowd accuracy due to correlated judgment errors, others show that accuracy improves. We investigate experimentally the accuracy of threads in which participants make magnitude estimates of varying difficulty while seeing a varying number of previous estimates. We demonstrate that, for difficult tasks, seeing preceding estimates aids the wisdom of crowds. If, however, participants only see extreme estimates, wisdom quickly turns into folly. Using a Gaussian Mixture Model, we assign a persuadability score to each participant and show that persuadability increases with task difficulty and with the amount of social information provided. In filtered threads, we see an increasing gap between highly persuadable participants and skeptics.

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 scroll through comments on social media platforms and blogs. These activities inform us and influence our decisions, but researchers still debate under what conditions these types of social information improve decision-making [1–4], lead us astray [5–9], or simply add to our confusion [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 [12–15]. Then again, confusion is not always randomly scattered around the truth. Systematic biases in individual perceptions may create measurable disruptions in the wisdom of crowds [16–18]. Social information may add to those biases and create cascades, echo chambers, bandwagoning and herding behavior [19–22]. 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 [24–28]. 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], or simply by plain misinformation [32], 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 (ex ante), and subsequently ask them to revise their estimate after having received information about other people’s estimates of that very object (ex post) [3, 4, 6, 34–35]. This setup presumes that people change their mind because of the social information they receive. Other studies, however, have shown that people routinely 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 [36–38]. Such a psychological mechanism - and perhaps others such as hedging strategies due to anticipated regrets [39], and/or disappointments [40] - make it difficult to differentiate between ‘inner’ samples and ‘outer’ influences. It is therefore desirable to develop an alternative framework that is able to infer the extend of individual bias and social influence from a single estimation task.

We propose to use a probabilistic Gaussian Mixture Model (GMM) as it has properties that are highly valuable in crowd aggregation research. First, a GMM associates a measure to each data point, describing the influence of social information on that particular participant without any prior knowledge about that participant. Second, a GMM is comprised of several Gaussians which fit well to the right-skewed and heavy-tailed distributions emerging from free response elicitations. Finally, as a statistical model, a GMM provides confidence bounds to the estimates, which further adds a measure of model uncertainty.

We also propose the experimental mechanism of dot-guessing games [41], where participants guess the number of dots in an image. While dot estimations have been used previously in numerosity experiments [42–44] they have only recently been proposed as useful ‘model organisms’ for crowd aggregation research [41–45] due to their advantages in terms of cultural neutrality, resistance to expertise and/or prior knowledge, and their qualities as captchas [46]. In addition, dot estimation tasks are easy to implement and easy to understand. Most importantly, they have an objective solution and are tunable, allowing for nearly-continuous difficulty levels and performance measures.
We collected a total of 11,748 estimates from 6,196 unique participants on Amazon Mechanical Turk. 5,990 estimates were collected from participants placed in 12 different threads where they successively estimated the number of dots, \( d \in \{55,148,403,1097\} \), in an image, while seeing the \( v \in \{1,3,9\} \) preceding estimates (historical threads). Another 3,934 estimates were collected from participants who were placed in 12 other threads and shown the same images while seeing the \( v \in \{1,3,9\} \) highest estimates made so far (manipulated threads). Finally, 1,824 estimates were collected from participants who were shown the same images, but with \( v = 0 \), corresponding to control conditions for each \( d \) containing no social information.

We interpret the number of dots, \( d \), as the task difficulty, while the number of visible previous estimates, \( v \), is interpreted as the degree of social information. Participants were placed randomly in one of the 28 threads \((2 \times 3 \times 4 + 4 \text{ controls})\) and made their estimate one after another. In order to keep the estimates in a somewhat realistic range, participants could not submit numbers below 10 and above 1.000.000. No participant who had seen a certain image would be able to participate in another thread containing the same image again. In addition to a participation fee and a variable waiting fee, all participants in all threads received a bonus of $1 if their estimate was within 10% of the true value. See the Methods section and the Supplementary Information for additional information about the experimental design.

Free response elicitation of absolute values is known to create right-skewed distributions with long tails, which inflate the means. While it is still debated which measure is best suited to aggregate such data \cite{31}, we follow the lead of Galton \cite{12} and focus on the median as it is easy to interpret, robust against outliers, and best expresses the opinion of the crowd in the sense that the majority deems every other estimate as too high or too low. For the statistical analysis of thread aggregates, we therefore compare the log-ratio of thread medians, \( y_{dv} = \log(M(d,v)/d) \), using a linear normal model to quantify differences between threads in terms of \( \log(d) \) and \( v \), the latter as a categorical variable. The effects of social information on individual decision-making are analysed using a GMM on the log-ratio of individual estimates \( Y_i \) with the weighted geometric mean of the social information as the explanatory variable. The model assigns a persuadability score, \( \beta^w \), to each participant, which is high when people are highly influenced by the social information they can see, and small when people are not influenced, or when they are skeptical about the information (see Methods for details).

**Results**

In accordance with previous findings, participants do well in tasks without social information, especially when estimating small numbers. For higher difficulties estimates vary widely and biases become substantial \cite{16,18,48,44}. The median tends to underestimate the true value and the mean tends to overestimate the true value.

**Analysis of thread performance**

The relationship between the observed median log-ratio \( y_{dv} \) and the number of dots \( d \) are found suitable, according to quantile-quantile plots (see Supplementary Infor-
FIG. 2: GMM results: The left hand side shows three plots of a historical thread with \(d = 1097\) and \(v = 1\), and the right hand side shows three plots of a manipulated thread with \(d = 1097\) and \(v = 9\). Top plots show the log-ratio estimates over time (observation no.), with the weighted geometric mean of the social information shown by a dotted line. Bottom left plots show the log-ratio of the estimates as a function of the log-ratio of the social information, indicating how differently participants use their social information. Bottom right plots show the cumulative distribution of individual persuadability scores with 95% intervals derived from the fitted models.

In manipulated threads, Fig. 2 (right), the manipulation results in a large positive bias for \(d = 3\), \(v = 2\), and \(v = 3\) that becomes more pronounced as the task difficulty increases. This is due to the fact that the estimates are not very effective. This resonates well with the findings in [3] which shows that providing a moderate amount of incorrect information may counterbalance underestimation bias and improve collective performance.

Analysis of Persuadability

In Fig. 2 two exemplary threads show how the GMM model framework reveals interesting features of the data. Each thread is represented by three figures. The top figures show the log-ratio of estimates as the thread evolves over time with the weighted geometric mean of the social information shown by a dotted line. Each estimate has an associated color, given by the persuadability score \(\beta_w\), which measures the influence of social information on each participant. A high value \((\beta_w > 0.6)\) in dark green and blue colors implies a large social influence effect, a low value \((\beta_w ≈ 0)\) in red implies a small or no effect, and a medium value \((\beta_w ≈ 0.4)\) in brown and light green colors suggests a compromise between the two extremes. The RGB color scale (right) is the same for all threads and plots. The bottom right plots show the cumulative distribution of the persuadability scores with 95% confidence intervals derived from the fitted model. They clearly show how people may be categorized into those ‘sleeping dogs’ or ‘lost causes’ in red that do not take other peo-
FIG. 3: Thread dynamics: Comparison of typical historical threads (left) and their corresponding manipulated versions (right). Colors correspond to estimated $\beta^w$ values. The color scale is fixed across threads.

... the ‘persuadables’ that tend to follow others, and a large group of green ‘compromisers’, who try to strike a balance between what they see others have guessed, and what they believe themselves. Of course, such labels are only proximate. Given a participant has a personal estimate much in line with the social information seen, this participant may be labelled as a persuadable or as a compromiser, when in fact this is only partially true. In general, however, the distributions are remarkable stable across threads, which suggests that...
people indeed may be partitioned into such overlapping response types.

**Thread dynamics**

What becomes abundantly clear when examining the results from the GMMs, is a sharp contrast between how people act in pristine historical threads, and how people act in manipulated, filtered threads. This is shown in Figure 3, where we compare typical historical threads (left) with their corresponding manipulated treatments (right). When the task difficulty is low and $v = 1$, participants are not significantly influenced, no matter whether they see the preceding or highest estimate in their thread (top row in Figure 3). But as soon as the amount of social information increases (second row), more people tend to follow – or at least to compromise. This indicates that there is a bandwagon effect at work $[20, 50]$, making it more likely for people to follow others, the more people have done so already.

If we increase the difficulty of the task from $d = 55$ to $d = 148$ and keep $v$ constant at $v = 3$, as shown in the third row of Figure 3, the dynamics between the historical thread on the left diverges strongly from the dynamics of the manipulated thread on the right. While the majority of participants in the historical thread are easier to persuade (green colors), people in the manipulated thread become more diverse and slowly start to split between those, who are highly persuadable (blue) and those who are not (red). Increasing difficulty even further to $d = 403$, and also increasing the amount of social information to $v = 9$, as shown in the fourth row of Figure 3 makes this split in behaviour even more visible. For $d = 1097$ and $v = 9$, as shown in the bottom row, the difference is very clear: In the historical thread, a large majority of participants are medium to strongly influenced, having a persuadability score around 0.5, probably because most of the social information they see is regarded as more or less reasonable. However, as can be seen in the $\beta^w$-distribution, there is a high uncertainty in the scores, because we only model the weighted geometric mean of the nine preceding estimates, while people presumably are much more diverse in the way they process these estimates.

**Discussion**

Returning to the initial question of whether social information helps crowd wisdom, or whether it does more harm than good, our answer is both. Social information in online estimation threads does help when people have difficult questions to answer and when the information available to them is pristine and therefore representative. However, if the social information is filtered and not representative of the thread population, it becomes manipulative and may fool a substantial proportion of people, which, of course, interferes with any aspiration to harness the powers of collective intelligence.

Moving from the aggregate to the individual level, we find that people can be assigned a persuadability score by using a Gaussian Mixture Model. Persuadability generally increases with task difficulty and with the amount of social information provided. In the case of strong manipulation, we may see a split between a minority of persuadables and a majority of compromisers and skeptics. We do not know how much these influencing effects are transferable to other domains. However, due to the general nature of dot estimations, we suspect the effects to be substantial in other settings as well, and also for other questions – especially for those that are more emotional, subjective and political in nature. In computational social science and decision-making research, there is an acute need to further investigate the types of ‘disturbances’ in online threads and crowds, such as filters, rankings, likes, and recommendations, as well as the types of reaction to those disturbances. To the end of designing future collective intelligence systems, one needs to be vigilant about the way social information is gathered and framed. Crowd knowledge and thread wisdom is fragile, and can only be maintained with strong controls upon the way it is cultivated and recovered.

**Methods**

All experiments were coded in otree 2.1 $[51]$. The code itself is designed along the same lines as the classical information cascade experiments by Anderson and Holt $[19]$. A participant makes an estimate, and the next one receives the information about the estimate of the previous participant(s).

We obtained a total of 11.748 estimates from 6.196 unique participants. Any dot-image was only seen once by a participant, i.e. we had a total of 3.157 participants seeing only one image, 1.259 participants seeing two images, 1.047 participants seeing three images, and 733 participants seeing all four images (either in the unmanipulated or manipulated condition and with $v \in \{0, 1, 3, 9\}$). After 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 $v \in \{0, 1, 3, 9\}$ previous estimates. 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 Supplementary Information for screenshot and detailed design descriptions.

**Analysis of thread medians**

For the observed medians $M_{dv}, d \in \{55, 148, 403, 1097\}, v \in \{0, 1, 3, 9\}$ 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, the model structure for the medians was $\mu_{dv} = \alpha_v + \beta_v \log d$. Models were fitted separately to historical and manipulated threads to allow for different variance estimates $\sigma^2$ between these groups. Goodness of fit was assessed by quantile-quantile (QQ) plots of the residuals (see Supplementary Information, Figure 4). Although some deviation is present, with the available number of observations we do not discard the models based on these plots. The low number of observations also implies wide confidence bands, hence conclusions from the models can be viewed as conservative.
**GMM model for social influence**

A Gaussian Mixture Model (GMM) was used to cope with the heavy tails and skewness present in the data. The GMM employs a fixed number of states to fit a weighted sum of Gaussian distributions to each observation, hence the mixture labeling.

Let \( X_i = \{1, \ldots, k\} \) denote the state variable and \( Y_i \) the observation for participant \( i = 1, \ldots, n \). Then, conditional on \( X_i = j, j = 1, \ldots, k \), \( Y_i \) follows a normal distribution

\[
Y_i | (X_i = j) \sim \mathcal{N}(\mu_j, \sigma_j^2), j = 1, \ldots, k,
\]

hence the unconditional distribution of \( Y_i \) is obtained by integrating (i.e. summing since \( X_i \) is discrete) over \( X_i \)

\[
P(Y_i) = \sum_{j=1}^{k} P(X_i = j)P(Y_i | X_i = j).
\]

Including social information as an observed variable in the model, \( Y_i \) depends on both \( X_i \) and \( S_i \), if \( v > 0 \). Furthermore, if \( v > 0 \), then \( Y_i \) influences \( S_{i+1} \) as the next participant \( Y_{i+1} \) will see the estimate of \( Y_i \) (assuming the history provided is the previous views). A graphical presentation of model dependencies (with preceding views available) are shown in Figure 4.

**FIG. 4:** Gaussian Mixture Model dependencies with the observed estimate \( Y_i \) by participant \( i \), the latent state variable \( X_i \), and the social information \( S_i \) when consisting of preceding estimates. Dashed lines are connections that depend on the number of views \( v \). When \( v = 0 \), \( S_i \) is omitted from the model, when \( v = 1 \) then \( S_i \rightarrow Y_i \) and \( Y_i \rightarrow S_{i+1} \). For \( v > 1 \) then \( S_i \rightarrow S_{i+1} \) is added.

Thus, given \( k \) independent states and using \( \delta_{ij} = P(X_i = j) \), then \( Y_i \) is normally distributed with parameters

\[
\mu_i = \sum_{j=1}^{k} \delta_{ij} \mu_j \\
\sigma_i^2 = \sum_{j=1}^{k} (\delta_{ij} \sigma_j^2)
\]

where the mean \( \mu_i \) is modeled with social information \( s_i \) as

\[
\mu_i = \alpha_i + \beta_i s_i, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k.
\]

Note that \( \mu_i^w \) and \( \sigma_i^w \) refer to the weighted estimates, unique to each observation \( y_{i,v} \), \( i = 1, \ldots, n \). This implies, that the effect of social information becomes a weighted sum of \( \beta_j \) estimates

\[
\beta_i^w = \sum_{j=1}^{k} \delta_{ij} \beta_j, \quad i = 1, \ldots, n.
\]

Similarly for the \( \alpha_i^w \) parameter, modeling the accompanying intercept.

It should be emphasized that the weighted \( \alpha_i^w, \beta_i^w \) are unique for each observation due to the weights \( \delta_{ij} \), contrary to a standard regression model where all observations are assumed to adhere to the same effects. 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 \( \delta_{ij} \), but weighted among \( k \) general dimensions.

For each model, the number of states were set to \( k = 2, 3, 4, 5 \) and the model with the lowest Bayesian Information Criteria (BIC) was chosen as the preferable one. The BIC was used rather than the more usual Akaike 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. In our case AIC: \( 6k - 2\ln(L) \) and BIC: \( 3k\ln(n) - 2\ln(L) \), where \( L \) refers to the likelihood value. Note that both criteria are scaled with \( 3k \) parameters for \( k \) states. The BIC penalizes the number of parameters weighted by the (log) number of observations \( n \), as \( 3k \), corresponding to \( (\delta_{ij}, \mu_j, \sigma_j^2) \) for each state, whereas AIC only penalizes by a factor 2. Using objective criteria to choose a number of states/clusters is often prone to simply increase \( k \) continuously and as such the AIC yielded much larger \( k \) values \( (k \gg 5) \), practically implying that \( k = 5 \) for all models, due to us capping this parameter at this value. This effect was not seen for the BIC measure for \( 2 \leq k \leq 5 \), further supporting this choice of objective measure.

The goodness of fit for the GMM was assessed by calculating standardized residuals

\[
\hat{e}_i = \frac{y_i - \hat{\mu}_i^w}{\sigma_i^w} = (y_i - \hat{\alpha}_i^w - \hat{\beta}_i^w x_i)/(\hat{\sigma}_i^w)
\]

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

After removing the 5\% most extreme observations, i.e. estimates below the 2.5\% and above the 97.5\% thresholds, Fig. 7 in the Supplementary Information shows the QQ-plots of the fitted models. Fig. 8 in the SI-text shows the estimated \( \hat{\beta}_i^w \) distributions, with 95\% confidence bounds provided by the models. Fig. 9 in the Supplementary Information shows the observed threads, colored by the individual \( \hat{\beta}_i^w \) values, and the available social information. Fig. 10 shows the observed estimates against the available social information, again colored by their individual \( \hat{\beta}_i^w \) values.

**Modeled values**

The observations in the model were transformed to achieve a better fit. The estimates were standardized with the relevant number of dots and then log-transformed. Hence, for an observed estimate \( c_i \) on \( d \) dots

\[
y_i = \log(c_i/d), \quad i = 1, \ldots, n.
\]

The available social information, \( s_i \), available for participant \( i \) in a given thread was modeled as an aggregated value of the available previous estimates. If we let \( z_i(1, \ldots, v) \) denote these \( v \), then for the un-manipulated threads \( \hat{\beta}_i^w(1, \ldots, v) = \)
\{ e_{i-1}, \ldots, e_{i+v} \} \text{ and for the manipulated threads } \{ \hat{e}_{i}^{(1)}, \ldots, \hat{e}_{i}^{(m)} \} \text{ where } e_{i}^{(1)} \geq e_{i}^{(2)} \geq \cdots \geq e_{i}^{(m)} \geq e_{i} \text{ denote the } (\text{decreasingly}) \text{ ordered estimates } e_{i} \text{ among } e_{1}, \ldots, e_{i}, \text{ for } v < l \leq j < i. \text{ The superscripts } h, m \text{ refer to the type of information available, either historical } (h) \text{ or manipulated } (m). \text{ Hence, the historical information are the preceding } v \text{ estimates, whereas the manipulated information are the } v \text{ largest estimates among all preceding estimates. Following [4] the aggregation of these estimates was a weighted geometric mean. However, contrary to [4], the actual estimates were available to the participants, hence in this context } s_{i} \text{ becomes a proxy for the available social information. Letting } z_{ij} \text{ denote the } j^{th} \text{ visible estimate for the } i^{th} \text{ participant in either type of threads, then}

\begin{equation}
    s_{i} = \log \left( \frac{\prod_{j=1}^{v} w_{j}^{z_{ij}}}{d} \right) = \sum_{j=1}^{v} w_{j} \log \left( \frac{z_{ij}}{d} \right),
\end{equation}

where } \sum_{j=1}^{v} w_{j} = 1. \text{ Thus, the log-transform implies that the modeled values (log-normalized estimates) are aggregated as a simple weighted mean. The pair } (y_{i}, s_{i}), i = 1, \ldots, n \text{ were then used in the GMM framework. The weights } w_{j}, j = 1, \ldots, v \text{ were determined from density estimates based on the control threads of } v = 0. \text{ Hence, for } d \text{ dots, a density was estimated from the } v = 0 \text{ thread and then used as a proxy for how individuals in a thread with } v > 0 \text{ would estimate without any social information. Using this, an individual in a thread with } v > 0 \text{ is then assumed to assess the available previous estimates based on this density. This way extreme estimates are naturally filtered and a participants information is therefore almost exclusively weighted among sensible estimates. Practically this implied that estimates within boundaries considered extreme, were almost equally weighted, however extreme inputs, either by deliberate efforts to mislead the thread or otherwise just to be considered nuisances were weighted almost at } 0, \text{ yielding a more reasonable aggregated social information input } s_{i}. \text{ For manipulated threads, or threads with } v = 1, \text{ the social information available show these extreme guesses. This is simply what a participant have seen and will thus be discarded by the parameter estimation, fitting a low } \beta^{w} \text{ value for this participant (assuming that the current participant submits a realistic estimate and not taking the extreme info into account). In the case of manipulated threads, due to the removal of most extreme estimates to filter out deliberate extremities, the social information available can be higher than the cutoff limit of the 97.5% quantile. However, as in the case of } v = 1, \text{ this will simply be the available information that a participant can follow or disregard. The model will thus fit a suitable } \beta^{w} \text{ value based on the current participants use of the info available. The R package } \text{depmixS4} \text{ was used to fit the model, using an EM algorithm to maximize the log-likelihood function of the model.}

\textbf{Acknowledgments}

The experiments were implemented by Robin Engelhardt and Mikkel Birkegaard Andersen. Server infrastructure and development 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.

\textbf{Author contributions}

Robin Engelhardt designed the experiment, analysed the data, and contributed as a lead author; Vincent F. Hendricks contributed as the project manager and co-author; Jacob Stærk-Ostergaard designed the statistical methods, analysed the data and contributed as a lead-author.

\begin{thebibliography}{99}
\bibitem{Woolley2010} 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. \textit{Science} \textbf{330}, 686–688 (2010).
\bibitem{Gürçay2015} Gürçay, B., Mellers, B. A. & Baron, J. The power of social influence on estimation accuracy. \textit{J Behav Decis Making} \textbf{28}, 250–261 (2015).
\bibitem{Becker2017} Becker, J., Brackbill, D. & Centola, D. Network dynamics of social influence in the wisdom of crowds. \textit{P Natl Acad Sci USA} \textbf{114}, E5070-E5076 (2017).
\bibitem{Jayles2017} Jayles, B. \textit{et al.} How social information can improve estimation accuracy in human groups. \textit{P Natl Acad Sci USA} \textbf{114}, 12620–12625 (2017).
\bibitem{Caplan2011} Caplan, B. \textit{The Myth of the Rational Voter: Why Democracies Choose Bad Policies-New Edition} (Princeton University Press, 2011).
\bibitem{Lorenz2011} Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. \textit{P Natl Acad Sci USA} \textbf{108}, 9020–9025 (2011).
\bibitem{Minson2008} Minson, J. A. & Mueller, J. S. The cost of collaboration: Why joint decision making exacerbates rejection of outside information. \textit{Psychol Sci} \textbf{23}, 219–224 (2012).
\bibitem{King2011} King, A. J., Cheng, L., Starke, S. D. & Myatt, J. P. Is the true ‘wisdom of the crowd’ to copy successful individuals? \textit{Biol Lett} \textbf{8}, 197–200 (2011).
\bibitem{LeMens2011} Le Mens, G., Kovács, B., Ábrahám, J. & Kareev, Y. How endogenous crowd formation undermines the wisdom of the crowd in online ratings. \textit{Psychol Sci} \textbf{29}, 1475–1490 (2018).
\bibitem{Salganik2006} Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. \textit{Science} \textbf{311}, 854–856 (2006).
\bibitem{Salganik2009} Salganik, M. J. & Watts, D. J. Web-based experiments for the study of collective social dynamics in cultural markets. \textit{Top Cogn Sci} \textbf{1}, 439–468 (2009).
\bibitem{Galton1907} Galton, F. \textit{Vox populi (the wisdom of crowds)}. \textit{Anchor}, 2005.
\bibitem{Hong2011} Hong, L. & Page, S. E. Some microfoundations of collective wisdom, 56–71 (2008).
\bibitem{Izard2008} Izard, V. & Dehaene, S. Calibrating the mental number line. \textit{Cognition} \textbf{106}, 1221–1247 (2008).
\end{thebibliography}
Controlled lab 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 designed our experiment along the same lines as the classical information cascade laboratory experiments by Anderson and Holt [19]. While this design is not feasible in normal laboratory conditions, it is well suited for online labor markets and crowdsourcing platforms such as Amazon Mechanical Turk (Mturk), an online labor markets and crowdsourcing platform that has become a highly valued tool for social scientists who wish to conduct experimental research on the real time dynamics of large groups. Mturk has repeatedly been shown to meet or exceed the standards set by data collection methods using other means [1, 2]. The platform provides an integrated participant compensation system, has a large participant pool, and has been shown to be reliable, replicable, and significantly more diverse than typical American college samples [3–7].

Our experimental setup on Mturk is simple. After accepting our ‘HIT’ (Mturk acronym for a ‘human intelligence task’) and providing informed consent, participants are asked to wait in a ‘waiting room’ until the ‘choice room’ becomes available. After entering the choice room, participants are asked to take a look at the image and make an estimate (see screenshot in Figure 5). Depending on the view condition, participants can see \( v \in \{0, 1, 3, 9\} \) preceding estimates. We chose to present the ‘oldest’ previous estimate on the top of the list and the last estimate made on the bottom. When dealing with news, or financial data, users typically want to see the most recent activity first (think tweets, online banking transactions, news updates). With conversations it is different because there is a context to consider of whatever message came before and after the one you are looking at (think blogs or Facebook comments). We have chosen to use the conversation thread design (oldest on top) because there is no particular news criteria when guessing the weight of an ox or estimating the number of dots on a screen. Participants have one minute to think about the image and make their estimate. This might sound as a severe time constraint, but exploratory trials had shown that participants in general use less than a minute when performing this task. After submission, participants are thanked for their participation and the experiment ends. Waiting times are compensated with $0.20 per minute (maximally 5 minutes), participation fee is $0.10, and a bonus of $1 is paid if an estimate is within 10% of the true number.

FIG. 5: Screendump of the choice-page in the dot-experiment with \( d = 403 \) and \( v = 9 \).
Mturk settings and data quality

When working with Mturk it is important to consider the right settings in order to obtain the best data quality possible [8]. Fair wage, attrition rates, removal of duplicate participants and informative feedback are some of the most important issues to address.

Average wage for participants in our experiments was $\sim 12$ per hour, which is considered good according to Mturk guidelines and certainly above the estimated average of $6$ per hour when excluding un-submitted and rejected work [9].

Quitting a study before completing it is prevalent on Mturk, and varies systemically across experimental conditions [10]. On average 20 participants accepted our HITs within the first minute; after 10 minutes the average acceptance rate had dropped to 2-3 participants per minute and after an hour to less than one participant per minute. Our scripts were coded in such a way that participants were automatically assigned to a ‘waiting room’ in which they were asked to wait for maximally five minutes before entering the choice room. This meant that a lot of participants waited in vain. Due to such a high attrition rate in the first couple of experiments we changed the script slightly later on: Now the waiting room could contain a maximum of five participants, and when the waiting room was full, participants were told to come back and finish the HIT at a later point in time. This reduced the attrition rate to 6.5% on average.

All participants automatically received an image-specific qualification when accepting a HIT. This qualification ensured that participants could not accept any other HITs that use the same image. Further data inspection showed that 32 participants somehow managed to accept two HITs with the same image anyway. The reason may be that the time interval between accepting two HITs with the same image was too short for the qualification to register in the Mturk interface. All 32 duplicate participants were removed before data analysis. In addition, we set the qualification that participants should have completed at least 100 HITs and have an accepted HIT rate of 98% or above. This ensured that we would get only experienced and qualified participants.

Mturk participant attention was expected to be equal to or better than undergraduate participant’s attention [11], while various forms of dishonesty (practical joking or telling others about the true value offline or on an Mturk participants web page) was expected to be rare. Our screening of data files before data analyses revealed that a small fraction of participants submitted ridiculously high estimates across images seen, thereby skewing thread averages substantially. These estimates were included in the calculations of the geometric mean of the social information. However, they were given a low weight determined from density estimates based on the control threads with $v = 0$ (see section Modeled values above).

During our experiments, participants had easy access to our email for questions and possible bug reports. Apart from some minor difficulties when typing from a mobile device (less than 1%) participants had few comments or complaints.

Dataset

Anonymized data set of all dots-experiments can be found the the file dots.xlsx. Parameters: task = type of experiment; $d$ = number of dots in image; $v$ = number of visible preceding estimates; session = thread name; hashed_turker = anonymized participant id; decision order = order in which participants enter the queue; hist = list of guesses seen by participant; guess = estimate by participant

[1] Berinsky, A. J., Huber, G. A. & Lenz, G. S. Evaluating online labor markets for experimental research: Amazon. com’s mechanical turk. _Polit Anal_ 20, 351–368 (2012).
[2] Buhrmester, M. D., Talaifar, S. & Gosling, S. D. An evaluation of amazon’s mechanical turk, its rapid rise, and its effective use. _Perspect Psychol Sci_ 13, 149–154 (2018).
[3] Mason, W. & Watts, D. J. Financial incentives and the performance of crowds. In _Proceedings of the ACM SIGKDD workshop on human computation, _77–85 (ACM, 2009).
[4] Buhrmester, M., Kwang, T. & Gosling, S. D. Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data? _Perspect Psychol Sci_ 6, 3–5 (2011).
[5] Crump, M. J., McDonnell, J. V. & Gureckis, T. M. Evaluating amazon’s mechanical turk as a tool for experimental behavioral research. _PLoS One_ 8, e57410 (2013).
[6] Rand, D. G. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. _J Theor Biol_ 299, 172–179 (2012).
[7] Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: Conducting experiments in a real labor market. _Exper Econ_ 14, 399–425 (2011).
[8] Chandler, J. & Shapiro, D. Conducting clinical research using crowdsourced convenience samples. _Annu Rev Clin Psycho_ 12, 53–81 (2016).
[9] Hara, K. et al. A data-driven analysis of workers’ earnings on amazon mechanical turk. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 449 (ACM, 2018).

[10] Zhou, H. & Fishbach, A. The pitfalls of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. J Pers Soc Psychol 111, 493 (2016).

[11] Hauser, D. J. & Schwarz, N. Attentive Turkers: Mturk participants perform better on online attention checks than do subject pool participants. Behavior research methods 48, 400–407 (2016).
FIG. 6: Model diagnostic plots for median analysis. Top left: QQ plot for the residuals from the historical thread medians analysis. Top right: the same residuals plotted against session groups. Bottom left: QQ plot for the residuals from the manipulated thread medians analysis. Bottom right: the same residuals plotted against session groups.
FIG. 7: QQ-plots for the 29 unique threads from Amazon Mechanical Turk experiments. With some exceptions, the models fit fairly well. Some series with very few observations are discarded in the analysis.
FIG. 8: Distributions of estimated $\beta$’s for each thread. Color scale is fixed across all threads. Red colors indicate skeptics, blue colors indicate persuadables and green colors indicate compromisers.
FIG. 9: Individual guesses for each thread, colored by their estimated $\beta^w$ values. Social information available at each timepoint is shown as dashed black lines. The $\beta^w$ color scale is fixed across all threads. Red colors indicate skeptics, blue colors indicate persuadables and green colors indicate compromisers.
FIG. 10: Actual estimates by each participant against the given social information. Colored by the estimated $\beta_w$ values. Color scale is fixed across all threads. Red colors indicate skeptics, blue colors indicate persuadables and green colors indicate compromisers.
| method | d  | v  | thread id  | N  | median | mean  | SD       | CV     | skew | kurt | bonus (%) |
|--------|----|----|------------|----|--------|-------|----------|--------|------|------|-----------|
| history | 55 | 0  | fscmakcz   | 464| 55     | 2360.57| 42038.6 | 17.81  | 21.01| 445.08| 60.99     |
| history | 55 | 1  | hs6bovtz   | 477| 57     | 1682.13| 29381.7 | 17.47  | 21.18| 453.74| 57.02     |
| history | 55 | 3  | 3cda8qpn   | 405| 56     | 8320.59| 73524.4 | 8.84   | 9.59 | 93.58 | 66.91     |
| history | 55 | 9  | bvpj37io   | 476| 56     | 1239.85| 25435.5 | 20.51  | 21.74| 470.87| 56.09     |
| history | 148| 0  | dwnj09mb   | 464| 140    | 1349.14| 12891   | 9.55   | 12.83| 172.18| 22.41     |
| history | 148| 1  | ehaxc7lu   | 435| 150    | 2250.86| 27802.5 | 12.35  | 15.66| 261.32| 25.75     |
| history | 148| 3  | ck291lk5   | 466| 147.5  | 2523.51| 38071.3 | 15.09  | 20.42| 427.38| 23.18     |
| history | 148| 9  | 2hxe3g0w   | 473| 153    | 3182.51| 38085.7 | 11.97  | 13.85| 196.79| 24.52     |
| history | 403| 0  | hqpv0r75   | 473| 300    | 4816.02| 49624.9 | 10.3   | 14.16| 220.03| 7.82      |
| history | 403| 1  | 9bbkjlje   | 469| 320    | 2791.59| 31667.6 | 11.34  | 15.23| 248.43| 8.1       |
| history | 403| 3  | 5du4txa7   | 455| 350    | 2430   | 29944.3 | 12.32  | 15.25| 233.19| 11.65     |
| history | 403| 9  | spw8qgcd   | 470| 400    | 3035.65| 36078.8 | 11.89  | 15.32| 235.27| 10.43     |
| history | 1097| 0 | hal3jdlo   | 423| 657    | 8855.13| 76379.1 | 8.63   | 11.46| 138.21| 9.69      |
| history | 1097| 1 | huuyghto   | 461| 750    | 2924.95| 25012.2 | 8.55   | 17.91| 344.84| 9.76      |
| history | 1097| 1 | z0v5h02v   | 473| 650    | 8091.24| 68728.9 | 8.49   | 11.74| 144.62| 8.67      |
| history | 1097| 3 | hbb0if6e   | 470| 812.5  | 4811.97| 53329.9 | 11.08  | 15.61| 259.91| 20.64     |
| history | 1097| 9 | wv4xuqg7   | 460| 999    | 2748   | 31206.8 | 11.36  | 21.24| 451.09| 13.7      |
| max    | 55 | 1  | aebicytb   | 384| 57     | 68.7   | 35.77   | 0.52   | 2.82 | 8.76  | 54.17     |
| max    | 55 | 3  | 094p61xp   | 340| 60     | 101.05 | 413.25  | 4.09   | 18   | 326.23| 43.82     |
| max    | 55 | 9  | 8c80yycl   | 418| 60     | 75.03  | 45.69   | 0.61   | 4.72 | 34.23 | 51.44     |
| max    | 148| 1  | 7swuh7l    | 317| 152    | 271.06 | 323.75  | 1.19   | 3.84 | 19.06 | 17.03     |
| max    | 148| 3  | u2x0m02p   | 418| 220    | 366.6  | 596     | 1.63   | 11.6 | 179.79| 9.57      |
| max    | 148| 9  | mf57hnwb   | 412| 210    | 266.74 | 168.39  | 0.63   | 1.08 | 0.2  | 12.38     |
| max    | 403| 1  | 1r17post   | 25 | 400    | 549    | 402.64  | 0.73   | 1.83 | 4.14  | 4         |
| max    | 403| 3  | 6c4s02ki   | 16 | 90     | 452.5  | 614.18  | 1.36   | 2.26 | 4.85  | 0         |
| max    | 403| 1  | e8vw9575   | 64 | 600    | 734.38 | 506.71  | 0.69   | 0.48 | -1    | 6.25      |
| max    | 403| 3  | ua2230ux   | 315| 600    | 3215.08| 9635.18 | 3      | 4.84 | 22.79| 8.25      |
| max    | 403| 9  | 1xyev3dj   | 422| 887.5  | 4758.88| 25397.5 | 5.34   | 17.68| 339.78| 8.29      |
| max    | 1097| 1 | q5bhrghz   | 76 | 1062.5 | 3310.3 | 4690.72 | 1.42   | 2.13 | 4.19  | 5.26      |
| max    | 1097| 3 | 2b6km84z   | 117| 2500   | 7593.38| 9341.25 | 1.23   | 1.29 | 0.35  | 13.68     |
| max    | 1097| 3 | ud5vo371   | 78 | 2950   | 5167.06| 12588.5 | 2.44   | 7.43 | 58.71 | 3.85      |
| max    | 1097| 9 | lh7twb36v  | 416| 3410   | 13024.2| 38664.1 | 2.97   | 11.9 | 183.47| 8.17      |