The Wisdom of Threads

Protocol for WoT online experiments on Amazon Mechanical Turk

Robin Engelhardt

12 June 2018

Background

This experiment investigates how respondents are influenced in their estimate of a quantity when they have variable acces to previous estimates of the same quantity. A classical example of an estimation task without access to other estimates is that of people guessing the weight of an ox at a country fair, originally investigated by Francis Galton in 1906 (Galton, 1907). Galton showed that a pronounced 'Wisdoms of Crowds'- effect was present in the sample: While individual estimates were often wrong, their aggregate in terms of the mean or the median estimate, was remarkably close to the correct value.

Such collective accuracy is believed to be a statistical feature of noise. Individual estimates are subject to error, but when errors scatter equally in all directions around the truth, the mean and the median become an accurate measure of the truth. This phenomenon has been called 'Vox Populi' (Galton, 1907), 'Rational Expectations' (Muth, 1961), the 'Many Wrongs Principle' (Simons, 2004), and most recently the 'Wisdom of Crowds' (Surowiecki, 2005).

The main conditions for the Wisdom of Crowds-effect to materialize are believed to be relatively large crowds, independence of opinions, and a high diversity in expertise. Looking at the recent spread and omnipresence of prediction markets, crowdsourced information and recommmender systems, as well as peer-to-peer and distributed labour systems, Galton's findings may now be seen as a fundamental feature of aggrating information technology platforms and Big Data analysis.

Contrary to previous findings (Lorenz et al., 2011), recent research results have expanded the scope of the Wisdom of Crowds-effect by showing that social influence on decentralized information networks can in fact improve the accuracy of group estimates, even as the diversity of individual estimates becomes smaller (Becker et al., 2017). This is an important finding because most social information networks are decentralized, although not as regular as the network structure used in (Becker et al., 2017).

In addition to being decentralized, real life communication in online social media is typically ordered in threads: when reading through a facebook or a youtube thread, people read sequentially through a certain number of comments, depending on contextual cues and their specific psychological state. Typically people read the most recent comments first and stop after

reading only a few (ref?). When the thread is about estimating a quantity, research from the field of judgement and decision theory has shown that people seldomly calculate an average of other people's estimate before providing an estimate themselves. Instead, they tend to either dismiss other people's estimates in the belief that their own opinion "is as good as any other opinion" (Larrick, Soll, 2006) or they anchor their own estimate on a previous estimate which seems reasonable to them. In aggregate, people engange in what has been called 'egocentric discounting' of other's estimates (Harvey and Fischer, 1997; Yaniv, 2004) by shifting their own estimate in the direction of others by 20-40% on average (Rader, 2017).

In a sequential setting people pay more attention to items at the top of a web page or a list of items than those below them. A consequence of this herding-bias is a strong effect of presentation order on choices people make. For instance, presentation order affects which items in a list of search results users click on, and the answer they select when responding to a multiple choice question. Thus, a content provider can change how much attention items receive simply by changing their presentation order. (Lerman and Hogg, 2014) In finance it is also known that individuals herd more when they see others' forecasts (Chen & Jiang, 2005).

Research Questions

The standard belief is that social influence and herding can reduce the accuracy of a group's average answer. According to Surowiecki: "The more influence we exert on each other, the more likely it is that we will believe the same things and make the same mistakes. That means it's possible that we could become individually smarter but collectively dumber." (Surowiecki, 2005). But there are some pros: Additional information may improve the accuracy of private signals. E.g.: We become both individually smarter and collectively smarter. According to Becker et al. (Becker et al., 2017) this depends on the type of network.

So, our research questions would be: To which degree are respondents influenced by each other in such a sequential setting? Does this eventual influence reduce the accuracy of the aggregates (mean & median & geometric mean) across all respondents or not? Can this influence be described as herding, or is it of a intermittent and ephemeral type? Will the overall accuracy of the crowd will be reduced, while the average individual accuracy will increase?

Experimental Design and Setup

The experimental setup is very simple. After accepting the Amazon Mechanical Turk 'hit' all participants are asked to provide informed consent during the registration process. Upon being provided with a link to a UCPH-server at https://cibs.mef.sc.ku.dk and entering the experimental platform, participants wait until the choice room is available. A web page appears showing the following:

Time left to complete this page: **0:33**

Take a look at this ox

You are number 6 to guess. Previous players guessed the following:

player	guess (in kilograms, kg)	guess (in pounds, lb)
3. player	1100.0	2425.1
4. player	1234.0	2720.5
5. player	400.0	881.8

Please guess the weight of this ox (the man standing next weighs 174 lb (79 kg) and is 5'9'' (180 cm) tall).

If your guess is less than 10% away from the true weight, you will get a bonus of $0.10.

900

You submit your answer by choosing your unit of measurement for the weight:

pounds (lb) kilograms (kg)

The list under the image shows h previous guesses as determined by the treatment condition. 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's different because there is the context of whatever message came before and after the one you're looking at (think blog or facebook comments). We've chosen to use the conversation thread design (oldest on top) because there is no particular news criteria when guessing the weight of an ox.

After guessing once by choosing the appropriate unit, respondents see the following result-page:

Result

Player	guess (kg)	bonus
3. player	1100.0	$0.00
4. player	1234.0	$0.10
5. player	400.0	$0.00
6. player	900.0	$0.00

You were a bit off. The weight of the ox was 1233.0 kg.

Your guess was 900.0 kg, e.g. 27.01 % away from the true weight.

Your bonus is $0.00.

Your total payoff is your participation fee $0.10 + waiting time fee $0.20

Total = $0.30. Thank you for your participation.

Next

this ends the study and participants are propted back to their original Amazon Mechanical Turk page where they can redeem their payoff.

Study type

The experiment will be done online on the Amazon Mechanical Turk Platform. Participants will be from all over the world. The software platform is otree 2.1.

Inclusion and exclusion rules on Mturk

No player is allowed to play the game more than once which we will secure by giving each respondent a permanent qualification which excludes them from further participation.

Questions

1. (You see an image of an ox with a man standing next to it.) “How much does this ox weigh? The man weighs 174 pounds (79 kg) and is 71 inch (180 cm) tall.”
2. (You see an image with a lot of dots on it) “How many dots are there?” (70-100)

3) (You see an image with a lot of dots on it) “How many dots are there?” (150-200)

4) (You see an image with a lot of dots on it) “How many dots are there?” (400-800)

Alternativt: billede af en park med mange mennesker: hvor mange mennesker er der i parken?

Supplementary Questions

None.

Treatments / Condition

Treatment	Trials	# obs	endowment	bonuses	waiting time fee	total
normal						
0	1	500	$50	$100	$10	$70
1	1	500	$50	$100	$10	$70
3	1	500	$50	$100	$10	$70
9	1	500	$50	$100	$10	$70
highest (< 10000 kg)						
1	1	500	$50	$50	$10	$70
3	1	500	$50	$50	$10	$70
9	1	500	$50	$50	$10	$70
lowest (> 100 kg)						
1	1	500	$50	$50	$10	$70
3	1	500	$50	$50	$10	$70
9	1	500	$50	$50	$10	$70
total	**10**	**5000**	**$500**	**$700**	**$100**	**$1300**

Payoff Structure

Respondents receive a participation fee of $0.10 plus a bonus of $1 when guessing within 10% of the true value. I expect that respondents in the normal condition will be within 10% of the correct value 20% of the time, while in the highest and lowest condition only 10% of the time. In addition, respondents recieve a waiting time fee of $0.20 per minute with a maximum of 5 minutes waiting time.

Theoretical Approach and Methods

A time series is a sequence of measurements of the same variable(s) made over time. In the case of people making sequential estimates of a quantity, we have a special kind of time series: A

series of people (players) with a large variety of experiences and biases by which to make 'measurements' (estimates) of a thingy. In addition, players have access to a finite horizon of previous estimates. We can expect that players differ a lot in their opinions, and that the weights they put on the estimates of previous players also varies a lot. Therefore, our time series is quite noisy.

Now, let's imagine a player who is asked to make an estimate of a thingy X_t at time t. What probably happens is that she looks at the thingy and attaches a "prior" estimate β_t to it, based on her experiences and biases. Without loss of generality β_t can be expressed as a combination of the true value c plus an individual error ε_t, $\beta_t = c + \varepsilon_t$. Then she looks at the other previous estimates available and chooses to (or chooses not to) adjust her prior estimate accordingly.

This cognitive adjustment process is typically performed as a comparison (Yaniv, 2004), (Rader, 2017): "Oh, I think this guess is too high" is a typical thought, or "Ah, maybe I should go a bit lower" is another. This means that we might define a weight parameter ω telling us how much a player at time t is weightening all the available guesses, including her own prior:

$$X_t = \omega_{t0}(c + \varepsilon_t) + \sum_{i=1}^{q} \omega_{ti} X_{t-i} \qquad (1)$$

where the $\omega's$ are the parameters of the model (interpretable as a 'weight of influence'), c is a constant (interpretable as the 'true value'), and $\varepsilon_t \sim$ i.i.d $(0, \sigma^2)$ is white noise (in our case interpretable as the players abilities to guess correctly).

At least we can say that a simple autoregressivel model AR(p) will not do. The general expression for an AR(p) model is: $X_t = c + \varepsilon_t + \sum_{i=1}^{p} \omega_i X_{t-i}$ with ω_i only depending on i, e.g. on a (constant) weight of previous estimates. In our case ω_{ti} also depends on to what degree (and if at all) a player at time t choses to be informed by them. Some players are influenced by the previous guesses, others are not, e.g. $\omega_{ti} = 0$. Some players are heavily influenced by them, others only a little bit. Some players are influenced by the last guess only, and still others try to calculate the average of all visible previous guesses. So, in real life we should expect $\mathbf{\Omega}$ to be a not very well-behaved matrix of parameters.

If we asume that the weight parameters are random numbers, the model becomes something which at least resembles a 'random coefficient autoregressive' (RCA) model, investigated by (Nicholls and Quinn, 1980):

$$X_t = \omega_{t0}(c + \varepsilon_t) + \sum_{i=1}^{p} \omega_{ti} X_{t-i} \qquad (2)$$

where $\omega_{t1}, \dots, \omega_{tp}$ are independent and identically distributed random variables from a distribution with mean zero and unit variance, $\omega \sim$ i.i.d $(0, \sigma^2)$. But the problem is that $\mathbf{\Omega}$ probably is not very random.

Existing knowledge

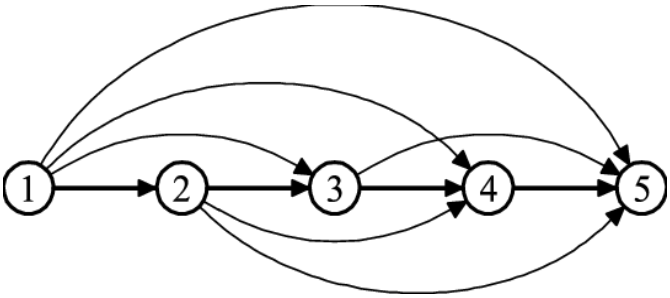
So, how should we expect the matrix $\boldsymbol{\Omega}$ to look? Browsing through the research literature, which as made extensive studies of how - and how much - people get influenced by others in estimation tasks, we find that people are very idiosyncratic in the way they incorporate information from others. Here some general findings:

1. $\boldsymbol{\Omega}$ is sparse: Both in the judgment and decision making (JDM) literature and in the observational learning literature it is well-known that people tend to stick to their guns, e.g. they do not follow others as much as they should. On average, 40% of all participants do not get influenced by others at all, and ignore the information available (Yaniv and Milyavsky, 2007)(Yaniv and Kleinberger, 2000). Equally in cascade experiments, (Weizsäcker, 2010) showed that the average player contradicts her own signal only if the empirical odds ratio of the own signal being wrong, conditional on all available information, is larger than 2:1, rather than 1:1 as would be implied by rational expectations.
2. Players shift their estimate in the direction of others by 20-40% on average (Harvey and Fischer, 1997), (Rader, 2017), (Yaniv, 2004). This is called 'egocentric discounting', e.g. the tendency to favor one's own opinions over those of others. The weight on advice (ω_{ti}) is typically measured as the proportional shift toward another's opinion. In one representative study, people chose to stay with their original belief 36.1% of the time, deferred entirely to one advisor 10.0% of the time, and took a simple average of the opinions only 17.5% of the time (Soll & Larrick, 2009).
3. The large the group of peers (number of guesses seen) the greater is the tendency to follow the crowd. But players are reluctant to take a simple average of opinions, for example, because it treats the opinions of experts and nonexperts alike, which in their eyes is a recipe for mediocrity (Larrick & Soll, 2006), (Mannes, et al., 2014).
4. Egocentric discounting operates in such a way that estimates distant from your own estimate are greatly discounted (Yaniv and Milyavsky, 2007), ($\sum^{i=1} \omega_{ti} \approx 0$). But if a player herself makes an outlier-guess, there is a strong tendency to conform ($\sum^{i=1} \omega_{ti} \approx 1$).
5. Primacy effects: Data that occurs early in a sequence influences players more than does later data (Peterson and DuCharme, 1967).
6. Position bias: People pay more attention to items at the top of a web page or a list of items than those below them. A consequence of this bias is the strong effect of presentation order on choices people make. For instance, presentation order affects which items in a list of search results users click on, and the answer they select when responding to a multiple choice question. Thus, a content provider can change how much attention items receive simply by changing their presentation order. (Lerman and Hogg, 2014)
7. Players often have unfavorable opinions about a whole-crowd strategy (averaging) and use it infrequently when paid to make accurate judgments. Instead, people are drawn to a best-member strategy, which performs admirably when there are large differences in ability and reliable cues for identifying a best member. But these conditions are often not met in practice, which means people will inevitably err in their choice of an imagined best member." (Mannes, et al., 2014)

We could try to be even more precise and assume that the $\omega's$ are a combination of three parameters: $\omega_{ti} = \mu_t(\rho_i + \eta_{it})$, where μ_t is a binary coefficient telling us whether the player at time t is influenced by the others or not, ρ_i is a position bias (which is present in all players, see (Payne, 1951), (Buscher, 2009) and (Lerman, Hogg, 2014), and η_{it} is the weight of influence a given estimate x_{t-i} has on the player at time t.

Other approaches

Information cascades are a special case of directed acyclic graphs (DAG's). The the case of h=1, e.g. when there is only one visible previous estimate, we have a chinese whisper game. For larger h we have a transitive acyclic tournament with h nodes:



Every guess at position n implies an endorsement of at leat one previous guess at position m < n, namely the guess which is numerically most close to the guess at node n (if there is a tie, take the oldest). This endorsement defines the hamiltonian path of "most popular" guesses, which might show some interesting properties.

Ethical considerations

Approval by the Institutional Review Board

All procedures in this study were approved by the Faculty of Humanities' Research Ethics Committee, which functions as an Institutional Review Board of the Faculty of Humanity, UCPH.

We provide participants with a consent page before the experiment starts and participants need to check a box in order to proceed. If not, they are not allowed to participate in the experiment.

Anonymity and data management

We will closely follow the faculty's guidelines in accordance with the Danish Data Protection Act, cf. requirements from the Danish Data Protection Agency and the faculty's Institutional Review Board.

Data will be stored in a separate database at https://cibs.mef.sc.ku.dk, where the experiment will take place. No data will be stored in the cloud.

Anonymity: There will be no need for anonymization because the game is played anonymously already. The only personal information that will be available to the researchers is what is publicly available on the MTurk participant profile and any information that participants choose to provide during the course of the study. This information will not be shared with any individuals who are not part of the research team. We will store data anonymously, and any publications resulting from the experiment will have no references to any personal information. The only thing we will report on is the participant's guesses.

Additional considerations

Time plan

The experiment is expected to finish within a week.

Total Costs and financing

We plan to ask 4 questions, each involving 5.000 respondents and costing around $1.200, excluding server and Mturk-fees.

References:

Beal, George M.; Bohlen, Joe M. "The Diffusion Process". Iowa State University of Science and Technology of Ames, Iowa. http://ageconsearch.umn.edu/bitstream/17351/1/ar560111.pdf

Becker, J., et al. (2017). "Network dynamics of social influence in the wisdom of crowds." Proceedings of the National Academy of Sciences: 201615978.

Buscher G, Cutrell E, Morris MR (2009) What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In: Proc. the 27th Int. Conf. on Human factors in computing systems. New York, NY, USA, 21–30.

Chen, Q. and W. Jiang (2005). "Analysts' weighting of private and public information." The Review of financial studies 19(1): 319-355.

Galton, F. (1907). "Vox populi (The wisdom of crowds)." Nature 75: 450-451.

Hoffman, R. M., et al. (2011). "Simultaneous versus sequential information processing." Economics Letters 112(1): 16-18.

Larrick, R. P. and J. B. Soll (2006). "Intuitions about combining opinions: Misappreciation of the averaging principle." Management Science 52(1): 111-127.

Lerman, K. and T. Hogg (2014). "Leveraging position bias to improve peer recommendation." PLoS One 9(6): e98914.

Lorenz, J., et al. (2011). “How social influence can undermine the wisdom of crowd effect.” Proceedings of the National Academy of Sciences 108(22): 9020-9025. Mannes, A. E., et al. (2014). “The wisdom of select crowds.” J Pers Soc Psychol 107(2): 276-299.

Muth, J. F. (1961). “Rational expectations and the theory of price movements.” Econometrica: Journal of the Econometric Society: 315-335.

Nicholls, D. F. and Quinn, B. G. (1980). The estimation of random coefficient autoregressive model. (I), f. Time Series Anal., 1, 37-46.

Payne SL (1951) The Art of Asking Questions. Princeton University Press.

Peterson, C. R. and W. M. DuCharme (1967). “A primacy effect in subjective probability revision.” Journal of experimental psychology 73(1): 61.

Rader, C. A., et al. (2017). “Advice as a form of social influence: Informational motives and the consequences for accuracy.” Social and Personality Psychology Compass 11(8).

Simons, A. M. (2004). “Many wrongs: the advantage of group navigation.” Trends Ecol Evol 19(9): 453-455.

Soll, J. B. and R. P. Larrick (2009). “Strategies for revising judgment: how (and how well) people use others’ opinions.” J Exp Psychol Learn Mem Cogn 35(3): 780-805.

Surowiecki, J. (2005). The wisdom of crowds, Anchor.

Walden, Eric; Glenn Browne (2002). “Information Cascades in the Adoption of New Technology”. ICIS Proceedings.

Weizsäcker, G. (2010). “Do we follow others when we should? A simple test of rational expectations.” The American Economic Review 100(5): 2340-2360.

Yaniv, I. and E. Kleinberger (2000). “Advice taking in decision making: Egocentric discounting and reputation formation.” Organizational behavior and human decision processes 83(2): 260-281.

Yaniv, I. (2004). “Receiving other people’s advice: Influence and benefit.” Organizational behavior and human decision processes 93(1): 1-13.

Yaniv, I. and M. Milyavsky (2007). “Using advice from multiple sources to revise and improve judgments.” Organizational behavior and human decision processes 103(1): 104-120.