Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Quantifying the Impact of Cognitive Biases in Question-Answering Systems

Keith Burghardt
University of California at Davis
kaburghardt@ucdavis.edu

Tad Hogg
Institute for Molecular Manufacturing
tadhogg@yahoo.com

Kristina Lerman
Information Sciences Institute
University of Southern California
lerman@isi.edu

Abstract

Crowdsourcing can identify high-quality solutions to problems; however, individual decisions are constrained by cognitive biases. We investigate some of these biases in an experimental model of a question-answering system. We observe a strong position bias in favor of answers appearing earlier in a list of choices. This effect is enhanced by three cognitive factors: the attention an answer receives, its perceived popularity, and cognitive load, measured by the number of choices a user has to process. While separately weak, these effects synergistically amplify position bias and decouple user choices of best answers from their intrinsic quality. We end our paper by discussing the novel ways we can apply these findings to substantially improve how high-quality answers are found in question-answering systems.

Introduction

According to the wisdom of crowds, a group can find a better solution to a problem than a typical individual (de Condorcet 1976; Galton 1908; Surowiecki 2005; Kaniovski and Zaigraev 2011). This effect has become the foundation of crowdsourcing, including systems for content creation (Kittur and Kraut 2008), product review (Lim and Van Der Heide 2015), peer recommendation (Stoddard 2015; Weninger, Johnston, and Glenski 2015), and question-answering (Q&A) (Adamic et al. 2008; Yao et al. 2015). In many cases, the crowd's solution aggregates many users' recommendations or votes as they are sequentially added. Recent work suggests that this should determine the best items (Celis, Krafft, and Kobe 2016; Krafft et al. 2016), and displaying item popularity makes high-quality items easier to find. However, individual decisions can be affected by cognitive biases, which may reduce the relation between wisdom (the quality of ideas) and crowds (popularity).

For example, recent research has demonstrated that social influence introduces correlations between decision makers, which can reduce the quality of collective solutions (Lorenz et al. 2011; Kaniovski and Zaigraev 2011) and make them less predictable (Muchnik, Aral, and Taylor 2013; Weninger, Johnston, and Glenski 2015; Salganik, Dodds, and Watts 2006). Empirical studies of crowdsourcing systems suggest that users' bounded rationality, and reliance on heuristics like item position (*position bias*), are even more important limiting factors in collective performance (Stoddard 2015; Burghardt et al. 2017).

This paper examines cognitive factors that affect crowd performance in Q&A systems in order to better correlate item popularity with quality. We define quality of an answer as how well it addresses the question or how well it is written, which are independent of where or how the answer is shown to users. The research questions we address are:

- **RQ1: What cognitive factors contribute to answer popularity?**
- **RQ2: How does popularity relate to quality?**

We explore these questions with two complementary methods: simulating Q&A experimentally and using empirical data. The experiment identifies why answers become popular while controlling potentially confounding variables, and empirical data checks the experiment's ecological validity. We focus on Stack Exchange (SE), a popular Q&A site covering a wide range of topics from cooking to computer codes. The design of SE, particularly the way the site displays answers, are similar to other major Q&A boards, such as Yahoo! Answers and Quora. Thus our study is relevant to a variety of Q&A sites.

Methods

Our experiment directs Amazon Mechanical Turk workers to a web page instructing them to "choose the most correct answer for each of ten questions" [1]. The page models the Stack Exchange *English Language Learners* (ELL) forum, from which the questions were selected. Each question had at least 8 answers. The workers are mostly from the US, Canada, or Britain (90% of IP addresses).

We assign each MT worker to one of two experimental conditions. In both the "random" and "social influence" conditions, workers see 2 or 8 oldest answers (the same answers a SE user would have seen) from the ELL website in a random order below the question (independently for each worker). In the "random" condition no score is shown. In the "social influence" condition, however, workers are told that "scores listed next to each answer denote the number of

[1] Full code and data is available at https://github.com/KeithBurghardt/CognitiveBiasesQuestion-AnsweringSystems

Table 1: Number of questions in each experiment trial.*

# Answers	Trial	# Questions (Random)	# Questions (Social Influence)
2 Answers	Trial 1	440	438
	Trial 2	473	174
8 Answers	Trial 1	410	412
	Trial 2	930	1256
	Trial 3	447	–

*270 & 228 workers are in the random and social influence conditions, respectively.

individuals who chose this answer in the past" (as on SE). Scores are randomly generated for each worker, and ordered such that the first answer receives the highest score, the second receives the next-highest score, and so on, which simulates how scores are ordered in SE. Scores were numbers between 0 and 100 in the 2-answer scenario and between 0 and 25 in the 8-answer scenario (such that the scores add up to 100 on average)

We recruit workers with an approval rate of over 95% and more than 1000 Human Intelligence Tasks (HITs). For Trial 1 in the random condition (shown in Table 1), we requested "Master" workers. Their voting behavior was statistically similar to that of other users. Because it takes much more time to find Master workers, we dropped this requirement in later experiments. Workers have up to one hour to complete an assignment (the median time is 8.0 minutes in the random condition, and 9.5 minutes in the social influence condition). Each worker is paid 50 cents upon completion. The equivalent hourly wage was less than half that originally designed because the tasks took unexpectedly long compared to initial tests in which the authors were subjects.

Once workers choose an answer, they click to advance to the next question. We perform multiple trials for each condition, as listed in Table 1. In the experiments, we record:

1. the question number
2. the number of answers for each question
3. the time a worker answers each question
4. the answer a worker chooses
5. the order answers are listed for each question
6. the times when workers scroll their computer mouse (or track pad) over an answer
7. each answer's score (if applicable), and
8. the start and end time for a worker to complete the task

In the first and last trials of the random condition, and trial 1 in the social influence condition (see Table 1), the number of answers are randomly chosen to be either 2 or 8 for each question. Trial 2 in both conditions always has 8 answers. There is, however, no significant change in user behavior between the two conditions when comparing the probability a worker chooses an answer versus its position between trials 1 & 2 (Kolmogorov-Smirnov test p-values $>$ 0.1).

Finally, to check ecological validity, we compared the probability to choose answers versus their position on SE. We collected all votes from boards defined on SE as "non-technical" from August 2008 until September 2014[2], and recorded the position of answers just before they were voted on if the following was true: 2 or 8 answers are visible to voters, an answer has not been "accepted" by an asker, and votes were made after August 2009. Accepting an answer automatically pushes that answer to the top regardless of their score, and before August 2009, answers with the same score were not sorted randomly, which could affect our results (Oktay, Taylor, and Jensen 2010). These data conditions allow for answers to be strictly ordered from highest-score to lowest-score, which closely matches our experiment. In total, we have 790K votes when 2 answers are visible and 43K votes when 8 answers are visible. We see similar behavior in data from other forums, e.g., Stack Overflow and other technical boards, but we believe non-technical boards most closely match ELL forum questions. The ELL forum alone had too little data to make an adequate comparison.

Results

Our experiment disentangles contributions to position bias from cognitive load, score, and attention, and determines how these factors affect decisions. Figure 1 shows the probability a worker chooses an answer as a function of answer's position in each experimental condition, and a null model (described below) where users choose answers based on the amount of attention they receive. To allow for the best agreement between this model and data, we removed cases in which users chose an answer that was moused over less than an arbitrary threshold of 0.01 seconds (0 (0%), 706 (40%), 5 (0.008%), and 641 (38%) of votes were removed from Figs. 1a–b, respectively). The trends shown in the figures are the same when including all votes in the dataset, and when the threshold is larger, such as 0.1 seconds.

In the random condition (Fig 1a inset), workers prefer the last answer when 2 answers are visible (p-value$< 10^{-6}$), but they prefer the first few answers when 8 answers are visible (Fig. 1a). Future experiments will be necessary to understand this reversal. Nevertheless, the trend in which top answers are increasingly preferred as the number of answers increases agrees with previous research (Burghardt et al. 2017). In the social influence condition, workers prefer the first answer most when both 2 and 8 answers are shown (Figs. 1b), and prefer the first few answers much more than in the random condition. Furthermore, the top half of the answers are more likely to be chosen as the number of answers increases (58% and 68% for 2 and 8 answers, respectively), in qualitative agreement with the random condition. Controlling for position, there is no statistically significant correlation between score and the probability an answer is chosen (all p-values $>$ 0.1). Initial research suggests that scores have little effect even when we randomize the posi-

[2] Raw data is at https://archive.org/details/stackexchange, and data we parsed is available at www.openicpsr.org/openicpsr/project/102420/version/V1/view/.

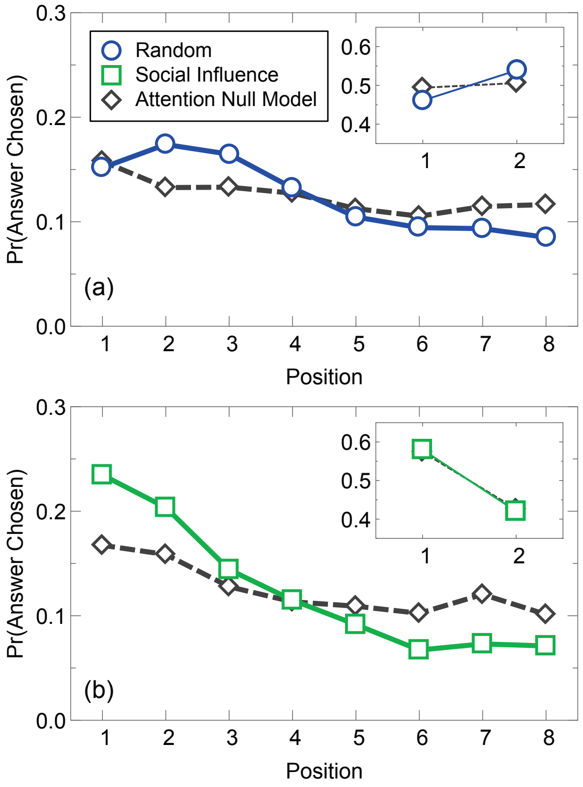

Figure 1: Probability to choose an answer versus its position in the experiment. (a) 8 answers visible in the random condition (inset: 2 answers visible) and (b) 8 answers are visible in the social influence condition (inset: 2 answers visible). Also shown is the null attention model, discussed in the main text. Error bars for all data are smaller than the plot markers.

tions of both answers and scores, therefore it is the ordering of scores (regardless of the score values) that amplifies the position bias.

We determine how attention contributes to position bias by using mouse movement, which correlates with eye tracking (Chu, Anderson, and Sohn 2001). We only record when the mouse is moving over or clicking on the answer, not when users scroll over it with their scroll wheel, for greater confidence that mouse movements were intentional. Although mouse tracking data is not perfect, it is a practical way to estimate attention. To check this, we compared the probability to choose an answer versus the fraction of time a worker mouses over it, which we call *time share*. We find that the probability strongly increases with time share (Cragg & Uhler's Pseudo R^2 values are between $0.44-0.54$ using logistic regression), in qualitative agreement with previous research (Krajbich, Armel, and Rangel 2010; Krajbich and Rangel 2011).

To test how well time share explains votes, we create a null model in which users choose an answer with probability proportional to the answer's time share. The dashed lines in Figure 1 compare this null model to experiments. We find that position bias is significantly stronger than the null model when 8 answers are visible. Namely, the trend is steeper than null model (p-value < 0.01) although all trends are negative. Thus the null model can partly–but not fully–explain position bias.

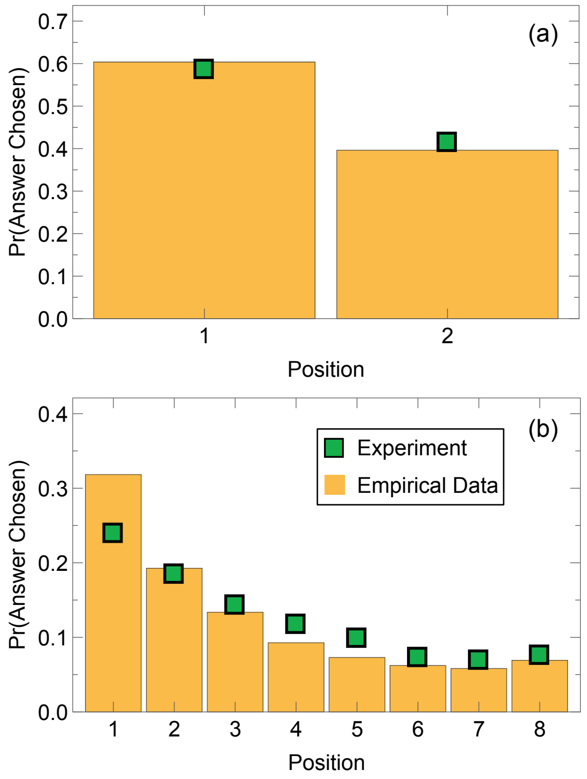

Figure 2: Comparison between experiment and empirical data. The probability SE users in non-technical boards (yellow bars) and MT workers (green squares) vote for an answer when scores are visible and (a) 2 or (b) 8 answers are visible. Error bars are smaller than the plot markers.

In summary, these observations lead us to the following conclusions.

1. *Cognitive load (number of answers visible) increases position bias*,
2. *Perceived popularity increases the position bias*,
3. *Perceived popularity, when corrected for position, is not a significant factor*, and
4. *Attention increases the position bias.*

Finally, to check the ecological validity of our experiments, we compared them with the vote data from all non-technical SE forums from August 2009 through September 2014. Specifically, Fig. 2 compares probability answers are voted on versus their position. Answers are ordered, by default, from highest to lowest score, which provides a direct comparison between our experiment results and the results from the data. The experiment agrees qualitatively with the observed user behavior on SE non-technical forums, even though SE answers are presumably ordered by from highest-to-lowest quality (Yang et al. 2014). Answer position may more strongly bias how answers are voted than their underlying quality.

Discussion

Our experiments elucidate factors affecting user choices of the best answers to questions. We find that an answer's position plays an important role in this decision and is strongly enhanced by perceived popularity, information load, and the attention given to top answers. We see strong agreement between our experimental model and empirical data, which demonstrates the experiments capture many aspects of real Q&A systems, despite differences in the populations, and the different motivations, of MT workers and SE users.

Because the design of SE is similar to other popular Q&A sites, we believe that our results are widely applicable to Q&A crowdsourcing systems. Furthermore, our observations are in line with recent work showing that position bias coupled with popularity-based ranking reduces collective ability to identify highest-quality items (Abeliuk et al. 2017).

How do we improve the wisdom of crowd in Q&A systems? We found that the position bias strongly affects whether an answer is voted on, even in empirical data, therefore we want to find ways to reduce the position bias. Our findings suggest two novel ways to do this: not showing answer scores and reducing the number of answers people see (e.g., removing older unpopular answers) but still ordering from highest-to-lowest score. Future work, however, is necessary to reduce the impact of older answers that accumulate votes due to their age and not their quality.

Acknowledgments

Our work is supported by the Army Research Office under contract W911NF-15-1-0142 and by the Defense Advanced Research Projects Agency under contract W911NF-18-C-0011.

References

Abeliuk, A.; Berbeglia, G.; Hentenryck, P. V.; Hogg, T.; and Lerman, K. 2017. Taming the unpredictability of cultural markets with social influence. In *WWW 2017*, 745–754.

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and Yahoo Answers: Everyone knows something. In *WWW 2008*, 665–674. New York, NY: ACM.

Burghardt, K.; Alsina, E. F.; Girvan, M.; Rand, W.; and Lerman, K. 2017. The myopia of crowds: A study of collective evaluation on Stack Exchange. *PLOS ONE* 12(3):e0173610.

Celis, L. E.; Krafft, P. M.; and Kobe, N. 2016. Sequential voting promotes collective discovery in social recommendation systems. In *ICWSM 2016*, 42–51. AAAI Press.

Chu, M.; Anderson, J. R.; and Sohn, M. H. 2001. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI EA '01 CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 281–282.

de Condorcet, M. 1976. *"Essay on the Application of Mathematics to the Theory of Decision-Making." Reprinted in Condorcet: Selected Writings*. Indianapolis, Indiana: Bobbs-Merrill,.

Galton, F. 1908. Vox populi. *Nature* 75:450–451.

Kaniovski, S., and Zaigraev, A. 2011. Optimal jury design for homogeneous juries with correlated votes. *Theory Dec.* 71:439–459.

Kittur, A., and Kraut, R. E. 2008. Harnessing the widom of crowds in Wikipedia: Quality through coordination. In *CSCW 2008*, 37–46.

Krafft, P. M.; Zheng, J.; Pan, W.; Penna, N. D.; Altshuler, Y.; Shmueli, E.; Tenenbaum, J. B.; and Pentland, A. 2016. Human collective intelligence as distributed Bayesian inference. *arXiv preprint:1608.01987*.

Krajbich, I., and Rangel, A. 2011. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *PNAS* 108(33):13852–13857.

Krajbich, I.; Armel, C.; and Rangel, A. 2010. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience* 13(10).

Lim, Y.-s., and Van Der Heide, B. 2015. Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on Yelp. *JCMC* 20(1):67–82.

Lorenz, J.; Rauhut, H.; Schweitzer, F.; and Helbing, D. 2011. How social influence can undermine the wisdom of crowd effect. *PNAS* 108(22):9020–9025.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science* 341:647–651.

Oktay, H.; Taylor, B. J.; and Jensen, D. D. 2010. Causal discovery in social media using quasi-experimental designs. In *SOMA 2010*, 1–9. ACM.

Salganik, M.; Dodds, P.; and Watts, D. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854–856.

Stoddard, G. 2015. Popularity dynamics and intrinsic quality in Reddit and Hacker News. In *ICWSM 2015*, 416–425.

Surowiecki, J. 2005. *The wisdom of crowds*. New York: Anchor.

Weninger, T.; Johnston, T. J.; and Glenski, M. 2015. Random voting effects in Social-Digital spaces: A case study of Reddit post submissions. In *HT 2015*, HT '15, 293–297. New York, NY, USA: ACM.

Yang, J.; Tao, K.; Bozzon, A.; and Houben, G. J. 2014. Sparrows and owls: Characterisation of expert behavior in Stack Overflow. In *UMAP*, 266–277.

Yao, Y.; Tong, H.; Xie, T.; Akoglu, L.; Xu, F.; and Lu, J. 2015. Detecting high-quality posts in community question answering sites. *Information Sciences* 302:70–82.