Wisdom of crowds in practice

JUHO SALMINEN, Lappeenranta University of Technology

1. INTRODUCTION

One of the premises of crowdsourcing is exploiting the wisdom of crowds. Under the right circumstances the aggregated judgment of a crowd can be closer to the truth than that of the best individuals in the crowd. To be wise, a crowd should be diverse, judgments of its members should be independent, and there should be a way to aggregate the judgments (Surowiecki 2005). Previous research has shown that diverse problem solvers can outperform the best problem solvers both in simulations (Hong and Page 2004) and experimental settings (Krause et al 2011). On the other hand even a minor social influence can decrease the accuracy of a crowd (Lorenz et al 2011). Crowds are also susceptible to self-fulfilling prophecies, in which perceived but false popularity can become real over time (Salganik and Watts 2008). Despite these shortcomings, it has been suggested that "in a well-designed setting, a collective evaluation can match the performance of experts on a given evaluation task" (Riedl et al 2010). But exactly how wise are the crowds in practical crowdsourcing settings? To my knowledge assessments of the wisdom of crowds on actual crowdsourcing sites have not been publicly reported. The question is important for the design of crowdsourcing applications. It would be useful to know to what extent the crowd's judgment can be relied upon and if additional decision making mechanisms are necessary. In this study the judgments of a crowd are compared to decisions of experts using data collected from crowdsourcing site Threadless (http://www.threadless.com/). The company describes itself as a creative community that makes, supports and buys great art. It is perhaps best known for producing and selling t-shirts designed by its user community. Threadless describes the selection process of designs as follows (Threadless 2013):

"All designs printed on Threadless are voted on and picked by the community. Users can submit designs to Threadless, which are then voted on for a 7 day period by other users in the community. Once the scoring period has ended, the design receives a score from 1 to 5. This is used as a gauge by Threadless to decide what gets made into a tee!"

The comparison of community voting scores to what Threadless actually prints reveals a statistically significant effect: when the average score given to a design by the crowd increases by one, the odds of Threadless printing that design increase by a factor of 15. Still, the accuracy of the crowd is not high enough to be relied upon alone in the decision making.

2. METHODS

Data was collected using a web crawler written in R programming language (R Core Team 2012). The algorithm was slowed down on purpose to prevent disrupting the website during the data collection. The dataset contains all designs accepted to the ongoing Threadless Challenge (http://www.threadless.com/threadless/) between 24 July 2012 and 1 July 2013. The collected data includes the name of the design, user, date of approval, average score of community votes (scale from 1 to 5), number of votes, number of ones, number of fives, and whether the design was eventually printed by Threadless. All this data is publicly available on the Threadless website, although the site does not link printed designs to submitted designs in a consistent way. Therefore multiple approaches were

used to assess the status of designs. A design is considered to have been printed if 1) the user who submitted the design has the URL of the submitted design on her list of printed designs, 2) the page displaying all designs printed by Threadless (www.threadless.com/all) contains the URL of the submitted design, or 3) a printed design at Threadless shop has a link to the submitted design. 70 % of the data was randomly selected to the training set and analyzed using logistic regression analysis (Field et al 2012) implemented in R environment (R Core Team 2012). This analysis produced estimates of the probabilities of submitted designs being printed based on the average score given by the crowd. The model fit was assessed by simulating decision-making 10 000 times on the test set and comparing the number of printed designs, their average scores and standard deviation of average scores to actually observed values. Simulations were also carried out using a baseline model, where each design was given an equal probability of being printed.

3. RESULTS

During the observation period Threadless printed 205 designs out of 15,592 submitted by 12,478 users to the Threadless Challenge. Mean score of all submitted designs was 2.75 (SD 0.395), while the printed designs scored 3.18 (SD 0.458) on average. The results of logistic regression analysis on the training set with odds ratios and their 95 % confidence intervals are presented in table 1.

			95 % CI for odds ratio		
	B (SE)	p	Lower	Odds ratio	Upper
Intercept	-12.467 (0.675)	0.000	1.00*10^-6	3.85*10^-6	1.41*10^-5
Average score	2.752 (0.211)	0.000	10.41	15.69	23.83

Table 1. Results of logistic regression analysis, odds ratios and their 95 % confidence intervals.

Estimated parameters are statistically highly significant. When the average score of a design rises by one unit on the scale from 1 to 5, the odds of that design getting printed by Threadless increase approximately by a factor of 15. Effects of number of scores, number of ones and number of fives were also analyzed, but none of them were statistically significant ($p = 0.19$, $p = 0.17$, $p = 0.31$). Simulated decision making with a model from table 1 showed that the observed data are within plausible range in terms of the expected number of printed designs, average score of selected designs and standard deviation of scores. The baseline model did not produce average scores matching the observations.

A) B)

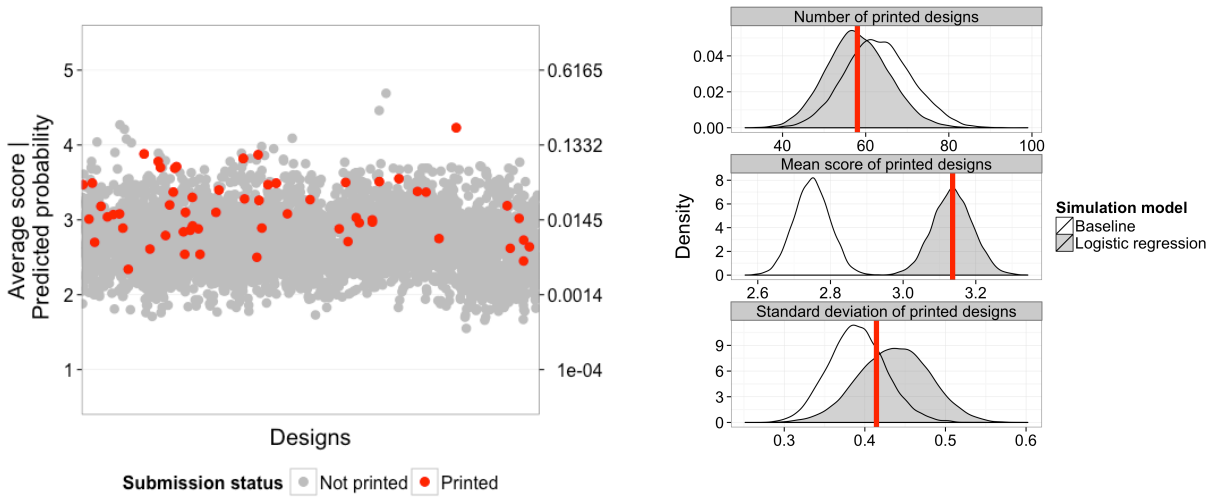

Fig. 1. A) Average scores, predicted probabilities and submission status of 4678 designs in the test set. B) Simulation results. Vertical lines represent observed values in the test set.

4. DISCUSSION AND CONCLUSIONS

The logistic regression model predicts that when the average score of a submitted design increases by one unit on the scale from 1 to 5, the odds of that design getting printed at Threadless increase by a factor of 15. The effect is statistically highly significant, but it is not strong enough to be the sole determinant of what gets printed at Threadless, as figure 1 clearly demonstrates. Simulations indicate a reasonable fit between the model and observed data. The results suggest that the wisdom of crowds in practical contexts may not be high enough to be relied upon in decision-making, although it may support it. Therefore the design of crowdsourcing sites should include secondary decision-making mechanisms in addition to community voting. Violations of the independence condition might decrease the accuracy of the crowd. Although users cannot see the evaluations of others before the voting has ended, the promotion of submitted designs to friends and relatives is encouraged, which might bias some of the evaluations. There are also known cases of outright cheating, where some users create multiple accounts to give high scores to their own designs and low scores for everyone else (known as downvoting at Threadless community forums). Sometimes this behavior even determines the top scoring designs in a challenge. Observations on the website suggest that the accuracy of evaluations is not the only use of community voting. According to Threadless Forum, many designers consider the average score given by the community to offer important feedback and help their development as graphic designers. Voting on submitted designs can also be a pleasurable activity and thus work as a means of engaging users to spend time on the site, browsing designs that might be on sale in the future. Further research might investigate whether there is a tradeoff between these different uses of community voting, which should be taken into account when designing crowdsourcing applications. Performance of a different aggregation mechanism could also be compared. Would some other aggregation mechanism, such as non-linear average used by honey bees for nest site selection (Salminen and Harmaakorpi 2012) be more accurate than the arithmetic mean Threadless currently uses?

ACKNOWLEDGEMENTS

The author wishes to thank the European Regional Development Fund and the Regional Council of Päijät-Häme for the opportunity of presenting his research at the CI2014 conference.

REFERENCES

Andy Field, Jeremy Miles and Zoë Field. 2012. *Discovering Statistics Using R*. Sage Publications, 957 pages. ISBN:978-1446200469

Lu Hong, and Scott Page. 2004. Groups of diverse problem-solvers can outperform groups of high-ability problem-solvers, *PNAS*, 101, 16385-16389. DOI:10.1073/pnas.0403723101.

Stefan Krause, Richard James, Jolyon Faria, Graeme Ruxton, Jens Krause. 2011. Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour,* 81, 941-948. DOI: 10.1016/j.anbehav.2010.12.018.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *PNAS*, 108, 9020-9025. 10.1073/pnas. DOI:1008636108.

Matthew Salganik and Duncan Watts. 2008. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly*, 71:4, 338-355. DOI:10.1177/019027250807100404.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Christoph Riedl, Ivo Blohm, Jan M. Leimeister and Helmut Krcmar. 2010. Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it right. In *Thirty First International Conference on Information Systems*, St Louis.

Juho Salminen and Vesa Harmaakorpi. 2012. Collective intelligence and practice-based innovation: An idea evaluation method based on collective intelligence. In *Helinä Melkas and Vesa Harmaakorpi (eds.) Practice-Based Innovation: Insights, Applications and Policy Implications*. Springer. ISBN: 978-3-642-21722.

James Surowiecki. 2005. *Wisdom of Crowds*. Anchor Books, 306 pages. ISBN: 978-0385721707.

Threadless. 2013. Frequently Answered Questions. Retrieved 7.10.2013.
URL: http://support.threadless.com/ics/support/default.asp?deptID=15140&_referrer=http://www.threadless.com/.