An Idea Filtering Method for Open Innovations

ANA CRISTINA BICHARRA GARCIA, Universidade Federal Fluminense and MIT
MARK KLEIN, University of Zurich, MIT

1. INTRODUCTION

The adoption of crowdsourcing to boost innovation in organizations (Boudreau, Lacetera & Lakhani, 2011), at very low cost, has raised a new dilemma: how to evaluate and select ideas worthwhile trying to implement. Open innovation engagements tend to generate idea corpuses that are large, highly redundant, and of highly variable quality. Convening a group of experts to identify the best ideas, from these corpuses, can be prohibitively slow and expensive. To deal with the unexpected deluge of ideas for the 10 to the 100th project, for example, Google had to recruit 3000 Google employees to filter the ideas in a process that put them 9 months behind schedule. Organizations have thus turned to the wisdom of the crowds to not just generate ideas, but also filter them, so only a relatively small selection of the best ideas needs to be considered by the decision makers. It has in fact been shown that crowds, under the right circumstances, can solve such classification problems with accuracy equal to or even better than that of experts [(Surowiecki 2004). But this approach, in practice, has been no panacea. Rating and ranking systems, when applied to large idea corpuses, are prone to quickly locking, because of positive feedback loops, into fairly static and arbitrary idea rankings,: people do not have time to rate all ideas and thus tend to consider only ideas that have already received good ratings (Salganik, Dodds & Watts 2006). We propose a novel crowd-based idea evaluation technique called the "bag of stars", which produces crowd assessments of equal quality to that of standard rating systems, but in only a fraction of the time. In this paper, we will describe the approach, present an empirical evaluation conducted as part of a "real-world" open innovation engagement, and discuss the contributions and possible future directions for this work.

1.1 The Bag Of Stars/Lemons Approach

Our approach to this challenge is based on the following three key concepts:

- framing: the crowd is asked to predict the judgment of the individuals who will make the final idea selection, rather than simply evaluating the ideas based on their own criteria.
- incentives: participants are provided with financial incentives for making evaluations that align with those of the decision makers, as opposed to (for example) giving them a flat payment for entering their ratings. We explored both prize-based (giving people extra money for accuracy) and penalty-based (reducing people's payments for mistakes) payment mechanisms. The latter approach is particularly interesting because studies have shown that people are more motivated by avoiding loss than they are by achieving gains of equivalent value (de Meza & Webb 2007; Fryer, Levitt, List & Sadoff 2012; Holt & Laury 2005).
- budgets: rather than asking participants to rate *all* ideas, they are provided with a limited number of tokens ("stars or lemons") that they are asked to allocate to the ideas they consider the most (or least) promising. The more confident they feel about this judgment, the more stars they can allocate to that idea (within the limits of the overall token budget). The rationale for this is simple. Many ideas, we have observed, can be rejected quickly based on relatively shallow criteria. With the bag of stars approach, participants are encouraged to not waste time evaluating ideas that are unlikely to make the final cut, and can focus their efforts on the remaining ideas.

Our hypothesis is that these innovations will allow crowds to substantially compress the idea corpus from open innovation engagements, while retaining the best ideas, more quickly and effectively than existing idea filtering techniques. The design and results of our experiments to test this hypothesis are discussed in the sections below

Our basic rational is as follows. Tracing an advection pathway for a particle dropped in a flow field is a perceptual task that can be carried out with the aid of a visual representation of the flow. The task requires that an individual attempts to trace a continuous contour from some designated starting point in the flow until some terminating condition is realized. This terminating condition might be the edge of the flow field or the crossing of some designated boundary. If we can produce a neurologically plausible model of contour perception then this may be the basis of a rigorous theory of flow visualization efficiency.

1.2 Conditions

Our experimental evaluation consisted of two stages:

- an open innovation engagement to collect a corpus of productivity enhancement ideas from the lab members
- an idea filtering engagement which compared the bag of stars approach with a widely-used idea rating scheme - the five-level Likert Scale (Likert 1932)- as techniques for identifying the best ideas from within the idea corpus.

The open innovation engagement occurred in the context of a University R&D lab, in Brazil, that develops software solutions for complex problems in the petroleum exploration and production domain for which no solutions exist in the market. The lab has existed for almost 17 years and is growing in size and complexity, currently including 70 students, professors, managers, programmers and staff with expertise in such areas as computer science, engineering, statistics, physics, linguistics, and management. Typically, teams of about 5 people are formed to address client problems. Projects generally last from 2-3 years and cost about $500K.

The research lab's open innovation engagement was framed as a contest, with significant financial rewards offered for the three best ideas. The contest stayed opened for one month, and generated a total of 48 ideas. Each idea contained a title, a proposal description, sometimes with examples, and at least 2 advantages and 2 disadvantages. There were some similarities within the 48 ideas. For example, one idea suggested instant awards for employees responsible for some extraordinary event, while another suggested an annual reward indexed by the Lab performance as a whole. Although both ideas related to rewards, they were quite different in detail, and were not merged.

A committee of composed of 4 highly experienced research managers was asked to select the top three ideas based on the following criteria:

- cost for implementing the idea (lower is better)
- productivity benefit of the idea (higher is better)
- time needed to measure the benefit (lower is better)

A mediator applied a Delphi method to allow the committee reach an agreement without communication among them. It took four cycles to converge to the 3 winner ideas. These ideas were used to measure the quality of the crowd evaluation.

The ideas were given to the crowd that was submitted to 5 different evaluation conditions:

- Likert (group G1): Participants were asked to rate each idea using a 5-point Likert scale, ranging from 1 (very unlikely to have been selected as a winner by the committee) to 5 (highly likely to have been selected as a winner). Every participant was given a fixed payment for participating.
- Bag of stars (BOS) (groups G2 - G5): In these groups, the interface displayed the list of ideas, and allowed users to add or remove tokens (represented as gold icons) to these ideas, with the constraint that they can allocate no more than their total budget of tokens. Every user was given a budget of 10 tokens. The four bag of stars conditions differed solely in the payment mechanism used, as follows:

- BOS fixed: participants were given a fixed payment that was independent of their accuracy in predicting the expert committee's decisions.
- BOS penalty: participants paid a penalty for adding a star to an idea that the expert committee did not select as a winner.
- BOS bonus: participants were paid a bonus for each star they correctly placed on an idea that the expert committee selected as a winner.
- BOS eliminate: participants were paid a bonus for each star they correctly placed on an idea that the expert committee did not select as a winner. The focus was thus on eliminating bad ideas, rather than identifying good ones.

Every group was given one week to enter their idea scores. All the idea filtering engagements took place in parallel, and participants were asked to not discuss their evaluations with each other during the experiment.

1.3 Results

Our primary hypothesis was that the bag of stars approach will perform better at idea filtering than existing, Likert-scale based, methods, where performance is defined as including both compression (i.e. how much the idea corpus is pruned) and recall (i.e. how many of the top ideas remained in the filtered set). Clearly, an ideal filtering mechanism will have perfect compression and recall i.e. the final filtered set will include all and only the ideas that the expert committee would have selected if it saw the entire corpus. Our secondary hypothesis was that a "penalty for errors" approach will produce better performance than a "bonus for correctness" approach because of the loss aversion phenomena cited above. We present our key results below.

Figure 1 presents recall performance as a function of compression, for all groups. The vertical axis displays recall, where N% recall means that N percent of the ideas selected by the expert committee were included in the corpus filtered by the crowd. The horizontal axis displays the compression rate, for which M% compression rate means that M% of the crowd was eliminated from the corpus of ideas.

Each group is represented by a line representing the filtering performance as a function of the threshold used for including an idea in the final filtered set. The line for G1 (the Likert-scale group) shows the recall and compression values when we vary the minimum average rating an idea must exceed to be included in the filtered set. The lines for G2 through G5 (bag of stars) show the recall and compression values we get when we vary the minimum number of stars an idea should accrue before it is included in the final filtered set. The better the performance of a filtering approach, the greater the area its associated line covers in the figure.

We can immediately see that all conditions offered us a tradeoff: if we want a high compression rate (by setting a high threshold for including an idea in the final filtered set), we reduce recall, and vice versa. It is also clear that all groups offered roughly the same tradeoff between recall and compression. Group G1 (Likert scale) offered slightly better performance overall, but G3 (bag of stars with penalties for inaccuracy) was very close. Since each condition had just one group, we were unable in this study to assess the statistical significance of these differences. We thus also cannot conclusively prove, or disprove, our secondary hypothesis that basing a payment mechanism on loss aversion (G3) offers superior performance than flat or bonus-based schemes. The absolute magnitude of the effects is, in any case, clearly relatively small.

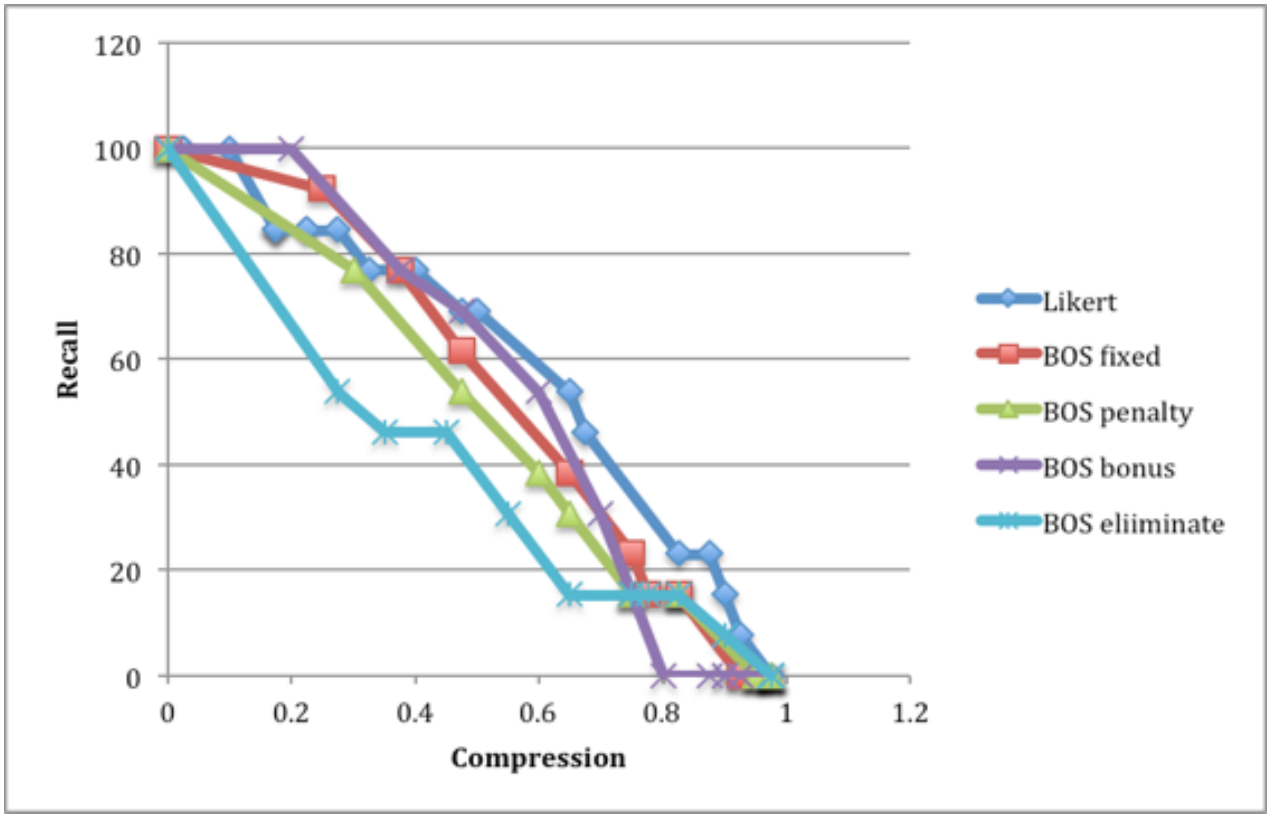

Fig 1: Compression and recall for the five idea filtering conditions.

We did, however, encounter dramatic differences in how long it took the participants to filter the ideas in the different groups (Table 1). The Likert-scale approach took participants over four times as long as the fastest bag of stars approach, and this effect was statistically significant (p < 0.001) for all the bag of stars groups. While the bag of stars approaches had substantial speed differences between them, with G3 being the fastest, the effects were statistically marginal due to the high variances

Table 1. Average duration of participants' activities, plus or minus 1 standard deviation.

Group	Time (minutes)
G1	83 +/- 16
G2	33+/- 37
G3	20 +/- 12
G4	35 +/- 15
G5	25 +/- 12

1.4 Lessons Learned

Our data suggest that the idea filtering performance for a bag of stars approach is equivalent, in terms of the critical recall/compression tradeoff, to that of the ubiquitous Likert-scale approaches, while requiring only a fraction of the participants' time. The reason for this, we believe, is simple: a bag of stars approach does not force participants to spend time assigning an exact score to every idea, allowing them to focus rather on deciding how much they want to bet on the best ideas in the corpus.

This work represents, we believe, a novel and important contribution to the literature in this field. The high level of participation the world has observed with social media systems has been shown to reply crucially on reducing the cost of participation (Benkler 2006). Our work points the way to how participation costs can be drastically reduced for the important problem of crowd-based idea filtering.

The closest analogue to our work, we believe, is the idea of prediction markets i.e. where participants are asked to buy and sell stocks that each represent a distinct prediction, with the understanding that they will receive a payoff, monetary or otherwise, based on how many of the stocks they own turned out to represent correct predictions (Arrow 1963; Wolfers & Zitzewitz 2004). Such markets have been used, with significant success, for purposes as diverse as predicting terrorist events, presidential elections, sports results, and Hollywood box office results (Berg & Rietz 2003; Berg et al. 2008; Gerstad 2004; Hankins & Lee 2011). Prediction markets have, however, significant weaknesses. They are prone, for example, to the same dysfunctions that financial stock markets face, in terms of stock prices being deeply influenced by short-term profit-seeking behavior rather than by the inherent value of the stocks. It can also be a challenge to encourage sufficient trading activity in prediction markets, since the benefits to the traders of getting the correct portfolio are usually too nominal to merit a substantial ongoing time investment. The bag of stars approach achieves many of the same benefits, by providing incentives for careful decision-making about which ideas to bet on, while requiring substantially less time investment and avoiding the problems of short-term profit taking and insufficient market liquidity.

ACKNOWLEDGMENTS

The work described in this paper was partially supported by the European Union FP7 CATALYST project.

REFERENCES

Arrow, K. J. (1963). Social choice and individual values. Wiley.

Benkler, Y. (2006). The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press.

Berg, J. and Rietz, T. (2003) Prediction Markets as Decision Support Systems. Information System Frontiers, 5(1), 79–93.

Berg, J. E., Forsythe, R., Nelson, F. and Rietz, T. A. (2008). Results From a Dozen Years of Election Futures Markets. In Charles R. Plott & Vernon L. Smith (ed.), Handbook of Experimental Economics Results, Elsevier, 1(5). 742-751.

Boudreau, K. J., Lacetera, N., & Lakhani, K. R. (2011). Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis. Management Science, 57(5)(5), 843-863.

de Meza, D. and Webb, D. C. (2007), Incentive Design under loss aversion. Journal of the European Economic Association, 5: 66–92

Fryer, R.G., Levitt, S.D., List, J. and Sadoff, L. (2012) Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. NBER Working Paper No. 18237. Issued in July 2012

Gjerstad, S. (2004) Risk Aversion, Beliefs, and Prediction Market Equilibrium. Steven Gjerstad. http://EconPapers.repec.org/RePEc:wpa:wuwpmi:0411002.

Hankins, R.A. and Lee, A. (2011) “Crowd Sourcing and Prediction Markets”. CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Holt, C. A., and Laury, S.K.. "Risk aversion and incentive effects: New data without order effects." The American economic review 95.3 (2005): 902-904.

Likert, R. (1932). A Technique for the Measurement of Attitudes. Archives of Psychology, 140, 1-55.

Morgan, J. and Wang, R. (2010) Tournament for Ideas. California Management Review, 52(2): 77-97

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science, 311(5762)(5762), 854-856.

Surowiecki James (2004), “The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, socie- ties and nations”, Doubleday, New York.

Wolfers, J. and Zitzewitz, J. 2004. Prediction Markets. Journal of Economic Perspectives, 18(8), 107-126.