Swissnoise: Online Polls with Game-Theoretic Incentives

Florent Garcin and **Boi Faltings**
Artificial Intelligence Lab
Ecole Polytechnique Fédérale de Lausanne
Switzerland
{firstname.lastname}@epfl.ch

Abstract

There is much interest in crowdsourcing information that is distributed among many individuals, such as the likelihood of future events, election outcomes, the quality of products, or the consequence of a decision. To obtain accurate outcomes, various game-theoretic incentive schemes have been proposed. However, only prediction markets have been tried in practice. In this paper, we describe an experimental platform, *swissnoise*, that compares prediction markets with peer prediction schemes developed in recent AI research. It shows that peer prediction schemes can achieve similar performance while being applicable to a much broader range of questions.

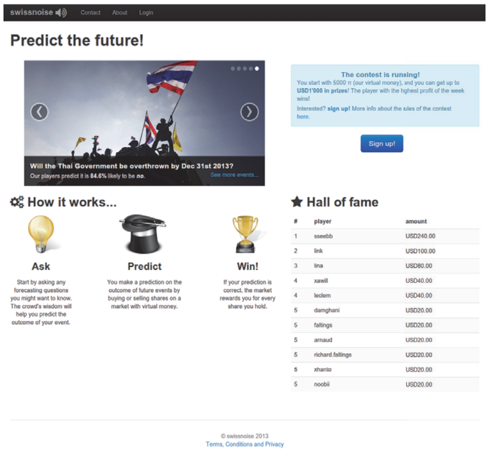
Introduction

The outcome of many important events depends on detailed information that is only known to certain individuals. For example, the outcome of a vote depends on the sympathies of voters for different options, the success of a project depends on a combination of details, and the sales of a new product are determined by how much an average consumer likes it.

Such questions are typically answered by polling a significant number of people who each provide a different judgement based on their perception of these details. Polls provide the best results when they are carried out on an unbiased sample. However, this requires that every member of the sample makes the effort to answer the questions, which is not easy to enforce. Thus, most online polls are based on voluntary participation or even self-selection, where people respond to a poll out of their own initiative. Responses are often given for ulterior motives, resulting in biased and questionable results. For example, in product review websites most reviews have either a positive or negative bias (Hu, Pavlou, and Zhang 2006; Jurca et al. 2010), so that it is not clear whether the average rating actually reflects the true quality (Garcin, Faltings, and Xia 2013).

One way to encourage participation by a broader sample of the population is to reward participants for their response. However, this raises a question of quality control: if random answers carry the same rewards as honest answers, why would anyone make an effort to give a correct answer?

Providing incentives for relevant and correct information has been addressed extensively in AI research but has not been applied very much in practice. In this paper, we report on an experimental platform, *swissnoise*[1] (Fig. 1), for conducting opinion polls on questions of public interest. Swissnoise experiments with two different models: prediction markets (Hanson 2003; 2007; Chen and Pennock 2010) and peer prediction (Miller, Resnick, and Zeckhauser 2005; Jurca and Faltings 2006; Witkowski and Parkes 2012a; 2012b; Radanovic and Faltings 2013). Peer prediction is a new scheme that can be applied more broadly than prediction markets. To our knowledge, ours is the first platform to implement a peer prediction scheme in a public opinion poll. The goal of our platform is to show that it can be practically implemented and achieve performance that is comparable to prediction markets.



Predict the future!

Will the Thai Government be overthrown by Dec 31st 2013?
Our players predict it is **84.6%** likely to be *no*. See more events...

The contest is running!
You start with 5000 π (our virtual money), and you can get up to **USD1'000 in prizes**! The player with the highest profit of the week wins!
Interested? **sign up!** More info about the rules of the contest here.

How it works...

Start by asking any forecasting questions you might want to know. The crowd's wisdom will help you predict the outcome of your event.

You make a prediction on the outcome of future events by buying or selling shares on a market with virtual money.

If your prediction is correct, the market rewards you for every share you hold.

Hall of fame

#	player	amount
1	sseebb	USD240.00
2	link	USD100.00
3	lina	USD80.00
4	xawill	USD40.00
4	leclem	USD40.00
5	damghani	USD20.00
5	faltings	USD20.00
5	arnaud	USD20.00
5	richard.faltings	USD20.00
5	xhanto	USD20.00
5	noobii	USD20.00

© swissnoise 2013
Terms, Conditions and Privacy

Figure 1: swissnoise's homepage.

Incentives for Online Polls

Rewarding participants to encourage accurate answers has been studied in game theory. If the correct answer becomes eventually known, as is the case in many prediction tasks, such incentives can be provided by *proper scoring rules* (Savage 1971). Agents submit a probability distribution $p(x)$ on their best estimate of the value of a variable x that is to be predicted. Once the true value $\bar{x}$ becomes

[1] http://go.swissnoise.ch

known, they get rewarded according to a scoring rule applied to the probability $p(\bar{x})$ they predicted for this true value. An example is the logarithmic scoring rule:

$$pay(\bar{x}, p) = a + b \log p(\bar{x}) \tag{1}$$

where a and b are constants with $b > 0$. It is also possible to use scoring rules to elicit averages, maxima and other functions of a set of measurements. Lambert and Shoham (2009) provide a complete characterization of the possibilities offered by scoring rules. This could be applied for example for rewarding participants in an unbiased sample.

Prediction Markets

In an open opinion poll, we would also like to encourage self-selection of those individuals that are the most knowledgeable about the subject, and pay more for information that makes the outcome of the poll more accurate rather than just confirms an already accurate poll. This is the idea behind prediction markets (Hanson 2003; 2007; Chen and Pennock 2010), where participants answer a poll with respect to the already known information. More precisely, participants in a prediction market trade *securities* linked to each outcome. When the outcome is decided, the securities for the correct outcome pay a reward of 1 whereas those for outcomes that did not materialize pay nothing. Thus, a participant can expect to gain by:

- buying securities at a price that is below the probability of the associated outcome, and
- selling securities at a price that is above this probability.

If all participants evaluate the outcomes in the same way, a prediction market is in equilibrium when the price of the securities is equal to the predicted probability of the outcome.

Another function provided by a prediction market is that of aggregating information. When participants do not agree on a single probability - and usually they will not - the aggregation is determined by how much money they are willing to risk on their prediction: as buying and selling securities changes the price, a participant may need to buy a large number of shares to move the price to her believed probability. If other participants have a strong opposite belief, they will readily sell their shares so that the price moves only very slowly. In practice, there are often not enough simultaneous participants, and thus this liquidity is simulated by an automated market maker. An automated market maker is based on a scoring rule and adjusts the price of securities so that the expected reward for changing the probability of an outcome is proportional to the difference of what a logarithmic scoring rule (Equation 1) would pay for the new probability and for the old probability. The scaling factor b that determines the actual amount is called the *liquidity parameter* and an important element of the design of the prediction market.

A major issue with practical deployment of prediction markets on public platforms is that at least when real money is used, many countries consider them a form of online gambling that is considered illegal. This is because participants have to place bets on particular outcomes that may or may not pay off. This could be overcome by just using scoring rules, so that payoff occurs only at the end, but this would mean that rewards are only paid much later and make the market less interesting.

Peer Prediction

Besides the legal issues, another problem that is common to both scoring rules and prediction markets is that they can only be applied when the predicted outcome can be verified with certainty. This makes it impossible to collect predictions for outcomes that will never be verified, such as product quality or appeal.

Peer prediction (Miller, Resnick, and Zeckhauser 2005) solves this issue. The idea is to consider the reports of other agents that observed the same variable, or at least a *stochastically relevant* variable, as the missing ground truth. A proper scoring rule is then used for the incentives. Provided that other agents truthfully report an unbiased observation of the variable, such a reward scheme makes it a best response to provide truthful and unbiased reports of the observations, and truthful reporting thus becomes a Nash equilibrium. Miller, Resnick and Zeckhauser (2005) describe such a mechanism and several variants, and Jurca and Faltings (2009) discuss further optimizations and variants.

An important limitation of peer prediction methods based on proper scoring rules is the need to know the agents' posterior probability distributions after each measurement. Zohar and Rosenschein (2006) investigate mechanisms that are robust to variations of these distributions, and show that this is only possible in very limited ways and leads to large increases in payments.

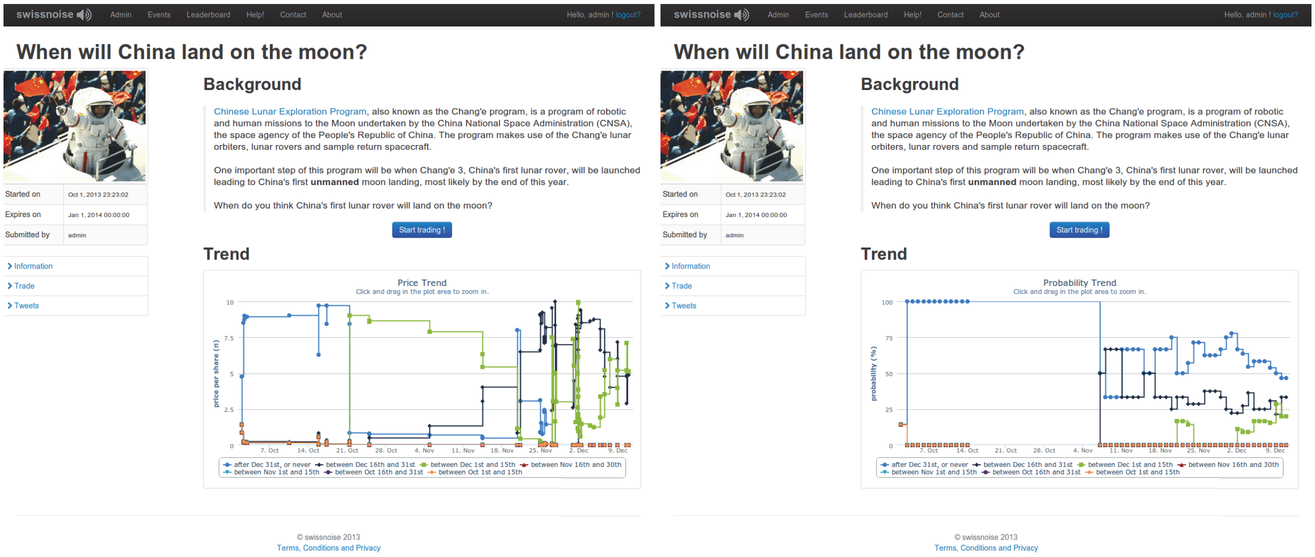
The Bayesian Truth Serum (Prelec 2004; Witkowski and Parkes 2012b; Radanovic and Faltings 2013) is a mechanism that elicits both the estimate itself as well as the beliefs about other's estimates. This elicitation of extra information eliminates the need to know the prior beliefs, but also requires participants to provide more information than just the answer to the question, which makes their task cognitively more difficult and too complex to implement.

We therefore took inspiration from prediction markets to implement a peer prediction scheme that assumes a common prior probability given by the current poll result. Similar to a prediction market, we display a current probability for each outcome. This probability is obtained by Bayesian updating from the reports received from different participants so far. Periodically, new reports are integrated into the current prediction to make it more accurate.

A reward is paid whenever the report matches a *peer report* that is randomly chosen among the reports that have been received in the same time period between updates of the public distribution. The amount of the reward is scaled so that it is inversely proportional to the currently estimated probability of this outcome: letting the probability estimate of an outcome x be $R(x)$, the reward for answer s is

$$f(R, s) = a + b/R(s) \tag{2}$$

where a and $b > 0$ are constants to scale the rewards. We call this scheme the *peer truth serum*: as the Bayesian Truth Serum, it does not require knowledge of prior probabilities. However, rather than requiring extra reports from participants, it takes this prior probability from the poll itself.



When will China land on the moon?

Started on	Oct 1, 2013 23:23:02
Expires on	Jan 1, 2014 00:00:00
Submitted by	admin

Background

Chinese Lunar Exploration Program, also known as the Chang'e program, is a program of robotic and human missions to the Moon undertaken by the China National Space Administration (CNSA), the space agency of the People's Republic of China. The program makes use of the Chang'e lunar orbiters, lunar rovers and sample return spacecraft.

One important step of this program will be when Chang'e 3, China's first lunar rover, will be launched leading to China's first **unmanned** moon landing, most likely by the end of this year.

When do you think China's first lunar rover will land on the moon?

Trend

(a) Prediction market.

swissnoise Admin Events Leaderboard Help! Contact About Hello, admin ! logout?

When will China land on the moon?

Started on Oct 1, 2013 23:23:02
Expires on Jan 1, 2014 00:00:00
Submitted by admin

Background

Chinese Lunar Exploration Program, also known as the Chang'e program, is a program of robotic and human missions to the Moon undertaken by the China National Space Administration (CNSA), the space agency of the People's Republic of China. The program makes use of the Chang'e lunar orbiters, lunar rovers and sample return spacecraft.

One important step of this program will be when Chang'e 3, China's first lunar rover, will be launched leading to China's first unmanned moon landing, most likely by the end of this year.

When do you think China's first lunar rover will land on the moon?

Trend

(b) Peer prediction.

Figure 2: swissnoise's event description panels.

This reward scheme will reward accurate reports whenever the participant a) believes that the current probabilities reflect the true prior distribution of other agents' reports, and b) believes that the true distribution of other agents' answers is actually shifted in the direction of its own opinion so that:

$$\frac{Pr_x(x)}{Pr(x)} > \frac{Pr_x(y)}{Pr(y)} \tag{3}$$

where $Pr(x)$ is the prior probability of the answer x while $Pr_x(x)$ and $Pr_x(y)$ are the posterior probabilities for the answers x respectively y, when the agent believes the true answer should be x. This condition is satisfied for example if the participant performs Bayesian updating to combine its own belief with the current poll outcome.

This can be shown easily by considering that the probability of answer x matching that of another randomly chosen participant is equal to $Pr_x(x)$, and the reward is equal to $R(x) = Pr(x)$ - the condition is then equal to the incentive-compatibility condition.

Design and Implementation

We designed a platform called *swissnoise*[2] with the goal of predicting results of Swiss ballots. However, swissnoise contains now more diverse events ranging from sports, entertainments or politics. Swissnoise was open to the public on April 22nd 2013. As of Jan 27th 2014, the platform had more than 200 active users with a total of 132 events (20 are currently open). It is free to signup and use. The platform has a virtual currency called π. Each user starts with π5000. Every week, we assign one gift card of CHF20 to the user who achieves the highest profit during that week.

Swissnoise implements two mechanisms to elicit information from the crowd: prediction markets and peer prediction. Since we are interested in comparing these two schemes, for a given event each user is assigned randomly to one of them, so that we get unbiased samples of even size.

Prediction Markets

We implemented the logarithmic market scoring rule (Hanson 2003; 2007). We determined the liquidity parameter $b = 100$ empirically in such a way that it allows newcomers to still be able to influence markets although their starting amount is lower than advanced users. We also scaled up the payments (by 10) to make it more attractive to the users.

Due to the way we rewarded the users, strategic behaviours emerged. Some users clear their trading positions at the last moment, on Sunday night, right before we determine the weekly winner in order to cash their profit. Figure 2(a) illustrates these dramatic price swings for one event.

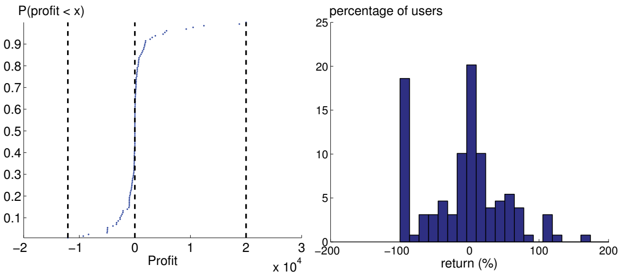
Peer Prediction

Since swissnoise is the first platform to implement a peer prediction scheme for public opinion polls, it was not clear a) how to implement the peer truth serum, and b) what are the best design choices.

In swissnoise, the peer truth serum is implemented as a "lottery". This analogy has the advantage of being well-understood among the majority of the users. However, we had to adapt it slightly to match the peer prediction scheme.

The key idea behind our implementation of the peer truth serum is that the user controls an agent that plays for her. Every day at midnight, we collect the statistics of the current day about an event and run a lottery. We randomly match a user's lottery ticket (report) to another user's ticket (peer report), and if their opinions are the same, the user is rewarded

[2] http://go.swissnoise.ch



(a) Cumulative distribution of profits. Dashed lines are min, median and max profit.

(b) Distribution of return across the user population.

Figure 3: Profits and returns.

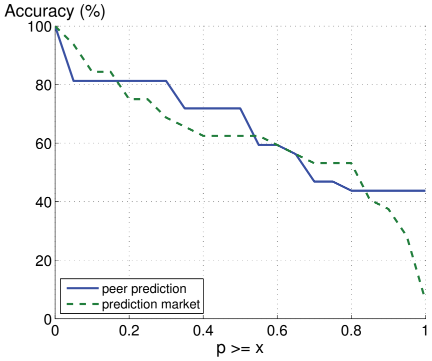


Figure 4: Accuracy of predictions.

according to Equation 2.

For a given event, the user selects first the number of days the agent is going to play, and buys lottery tickets accordingly. One lottery ticket is used per day. Then, the user selects the outcome she thinks is the best. She can update her choice at any time, but only the most updated information is taken into account for matching the tickets.

In the initial phase of the implementation, the reward of the lottery was very low, and it was making the peer prediction not interesting compared to the prediction market (Fig. 2(b)). We scaled the rewards to compensate for the risk-aversion of the users. The popularity of peer prediction events increased, but a risk-seeking behaviour emerged. Users started to choose the least likely outcome in order to get the highest reward. This behaviour resulted in daily oscillation around the 0.5 probability of the outcome. We observed this behaviour only for events with binary outcomes. Indeed, collusion/synchronization of behaviours among users on events with more than 2 outcomes is difficult because it requires to observe more than one signal.

To tackle this issue, we adjusted the reward of the peer prediction and, instead of taking the current statistics, we computed a running statistics over 5 days. This removed the oscillation effect.

Another issue was that some users did not update their votes, and the peer truth serum contained stale opinions. Thus, we decided to limit the number of possible tickets that a user can buy to 5. A user needs first to buy 5 tickets, and comes back after 5 days to buy again more tickets, and at the same time, updates her opinion if necessary.

The additional π earned on lottery events can be used to buy shares in prediction market events or tickets in peer prediction events. The profit made during one week with the peer prediction and prediction market determines whether she is the winner of the week or not.

Results

The performance of the users on the prediction market is depicted in Figure 3. The revenue is defined as the total amount received from selling shares and from payouts when a market is closed. The spending is the total amount spent in buying shares of events. The profit of a user is her revenue minus her spending, and her return is her profit over her spending. The profit and return indicates how good a user is on the platform. Note that the return has a lower bound at -100%, but the profit does not have a lower bound because users can get extra π by the means of a lottery. Figure 3(a) shows the cumulative distribution of the users' profits, with the minimum and maximum profit at -12051 and 20000, respectively. The median profit is slightly positive at 1.47 which shows that the median user improved the market's forecast accuracy. Figure 3(b) illustrates the distribution of users' return. The first peak at -100% corresponds to users whose predictions did not happen. The second peak on the right hand side of 0% is for users who slightly improved the market accuracy. A third smaller peak exists at around 50% showing that a small portion of users has improved the prediction more than the median user. Finally, another peak beyond 100% corresponds to risk-seeking users whose risky predictions actually paid off. These curves are similar to the ones reported recently by Dudik et al. (2013).

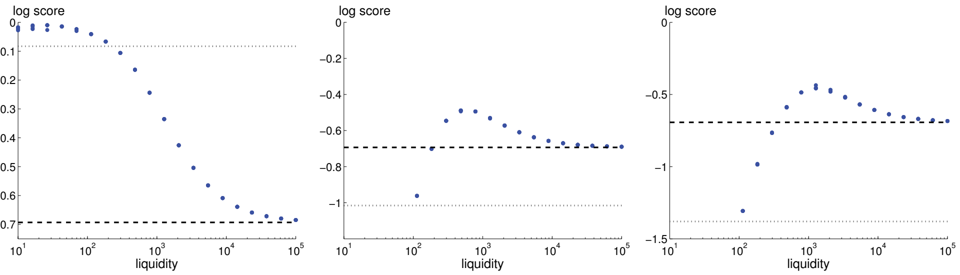
Figure 4 presents the accuracy of the two schemes as a function of the predicted probability. The accuracy is defined as the number of correctly predicted events over the total number of events (here 32 events). For instance, when an outcome has a predicted probability p of 0.5 or more, the prediction is correct for 72% of the events with the peer prediction and for 62% with prediction markets. We observe that the two schemes have similar performance.

To illustrate that peer prediction achieves a forecast accuracy comparable to the one by the prediction market, we focus our study on three events about the Swiss ballots because we can also compare our results with the ones from a traditional opinion poll company, named gfs.bern[3].

On November 24th 2013, Swiss citizens voted on 3 items[4]: two items where popular initiatives proposed by some eligible persons or a group of persons, and one was a mandatory referendum proposed by the Swiss Federal Assembly. At ballot stage, voters vote *yes* or *no*. The item comes into force if it is accepted by a double majority: majority of votes and majority of Swiss states. The 3 items are

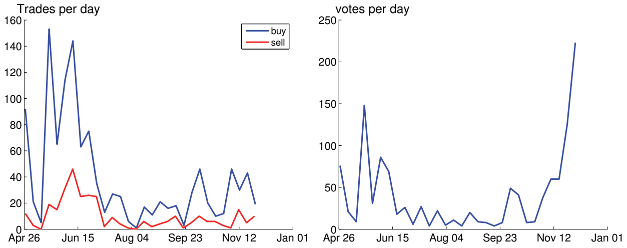
[3] www.gfsbern.ch

[4] More informations on the Swiss government portal https://www.ch.ch/en/federal-vote-of-the-24th-of-november-2013



(a) Initiative on fair wages. (b) Initiative for families. (c) Act on highway tax.

Figure 5: Forecast accuracy for different liquidity parameters. The dashed and dotted lines represent the log score for the initial probabilities and the actual market, respectively.



(a) Number of trades on prediction market events. (b) Number of votes on peer prediction events.

Figure 6: User activity on swissnoise.

- **(a) Initiative on fair wages.** This initiative asks that the highest salary paid in a company should not exceed twelve times the amount of the lowest salary. The question asked was *Are you in favour of the initiative on fair wages?* This item has been rejected at 65.5%.
- **(b) Initiative for families.** This initiative asks that parents who look after their children could deduct the same or a higher amount from their taxes as parents who pay for childcare. The question asked was *Are you in favour of the initiative on tax deductions for families?* This item has been rejected at 58.5%.
- **(c) Amendment on tax for highways.** This amendment aims at increasing the charge for using the highways from CHF40 to CHF100 a year, and introduce a two-month highway tax sticker costing CHF40. The extra revenue will be used to finance the running, maintenance and expansion of around 400km of roads. The question asked was *Are you in favour of the amendment on tax for highways?* This item has been rejected at 60.5%.

We posted these three items on swissnoise and asked what would be the outcome of the final vote.

We analyse the choice of the liquidity parameter by running a counterfactual simulation and following the same protocol as described by Dudik et al. (2013): we transform buy/sell transactions into a sequence of limit orders, and execute this sequence with different parameters. We then compute the accuracy as the log score, i.e. the log of the probability of the realized outcome.

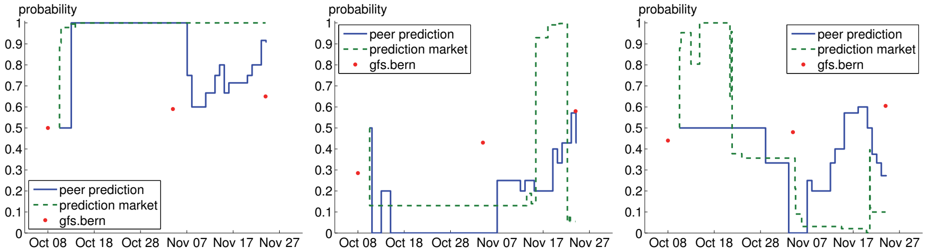
Figure 5 presents the forecast accuracy (average over 5 runs) for different liquidity parameters. Although the market accuracy for item (a) was very close to the optimal performance (Fig. 5(a)), the optimal liquidity is around 25 which indicates a low activity. Actually, the users knew at an early stage that item (a) would be rejected and it is also reflected by gfs.bern's opinion poll (Fig. 7(a)). Among the three items, the item (a) was the most certain to be rejected. On the other hand, for items (b) and (c), our choice of liquidity parameter was suboptimal. The optimal liquidities were around 480 and 1250 for items (b) and (c), respectively. This high liquidity reflects a higher activity for these two items. The gfs.bern's opinion polls (Fig. 7(b) and 7(c)) show that these two items were very uncertain.

Other factors influencing the user activity on the platform are a) the number of open events, b) the popularity of these events, and c) how close we are to determine the winner of the week. In addition, we see in Figure 6 a decrease in activity during the summer vacation.

The accuracy of both schemes are depicted in Fig. 7. For item (a), they both predicted correctly the outcome, while the opinion poll is more balanced (Fig. 7(a)). On Nov 7th, the forecast by peer prediction dropped to the same level as the opinion poll, and then increased until the end of the event, following the trend of the opinion poll. We believe that participants of the peer prediction scheme have been aware of this opinion poll and adapted their opinion, while users on the prediction market did not.

Regarding item (b), both schemes and the opinion poll predicted that it would be accepted. The closer we get to the realization of the event, the better we get to the actual prediction. One day before the event, the peer prediction touched the opinion poll. The price swings on the prediction market might be due to the strategic behaviours reported in the previous section. The peer prediction is better for item (c).

Both mechanisms have similar behaviour and accuracy, but it is difficult to compare them to traditional opinion polls for three reasons. First, the question asked to the users on swissnoise fundamentally differs from the one asked by the opinion poll company. Indeed, gfs.bern asked what *you* are going to vote, while on swissnoise we ask what you think the *outcome* is going to be.



(a) Initiative on fair wages (b) Initiative for families (c) Amendment on highway tax

Figure 7: Reject probability of the 3 items.

Second, it is not clear how gfs.bern samples the population and handles selection bias. Swissnoise's users might not be representative of the Swiss population. However, most users are related to Switzerland and follow local media. So they have a feeling of what could be the outcome.

Third, although anonymous, traditional opinion polls face the fact that people might still lie and do not reveal what they are going to vote. With this in mind and considering Switzerland's strong privacy awareness, we designed swissnoise in such a way that it is not possible for a user to check the current or past opinions of other users. Thus, contrary to most implementations of prediction market platforms, swissnoise preserves the privacy of its users.

Conclusion

Prediction markets have been applied with success for predicting events with a verifiable outcome. Recent research has developed the alternative technique of peer prediction which allows incentives without a verifiable final outcome. We have described how to adapt the peer prediction schemes developed in AI research to online opinion polls using the analogy of lotteries. This has been tested in the first experimental platform that implements peer prediction for online polls, called *swissnoise*. It shows that peer prediction has comparable performance to prediction markets, and thus constitutes a viable alternative. We are continuously collecting data about an increasing number of events. In future work, we would like to study users' behaviour about hypothetical questions when rewarded with the peer prediction mechanism, and explore the possibility of an adaptive liquidity parameter for prediction markets.

References

Chen, Y., and Pennock, D. 2010. Designing markets for prediction. *AI Magazine* 31(4):42–52.

Dudik, M.; Lahaie, S.; Pennock, D. M.; and Rothschild, D. 2013. A combinatorial prediction market for the us elections. In *Conference on Electronic Commerce*, 341–358.

Garcin, F.; Faltings, B.; and Xia, L. 2013. How aggregators influence human rater behavior? In *Workshop on Social Computing and User Generated Content*.

Hanson, R. 2003. Combinatorial information market design. *Information Systems Frontiers* 5(1):107–119.

Hanson, R. 2007. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets* 1(1):3–15.

Hu, N.; Pavlou, P. A.; and Zhang, J. 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *Conference on Electronic Commerce*, 324–330.

Jurca, R., and Faltings, B. 2006. Using chi-scores to reward honest feedback from repeated interactions. In *Conference on Autonomous Agents and Multiagent Systems*, 1233–1240.

Jurca, R., and Faltings, B. 2009. Mechanisms for making crowds truthful. *Journal of AI Research* 34(1):209.

Jurca, R.; Garcin, F.; Talwar, A.; and Faltings, B. 2010. Reporting incentives and biases in online review forums. *Transactions on the Web* 4(2):1–27.

Lambert, N., and Shoham, Y. 2009. Eliciting truthful answers to multiple-choice questions. In *Conference on Electronic Commerce*, 109–118.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51(9):1359–1373.

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 306(5695):462–466.

Radanovic, G., and Faltings, B. 2013. A robust bayesian truth serum for non-binary signals. In *AAAI Conf. on Artificial Intelligence*.

Savage, L. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66(336):783–801.

Witkowski, J., and Parkes, D. 2012a. Peer prediction without a common prior. In *Conference on Electronic Commerce*, 964–981.

Witkowski, J., and Parkes, D. 2012b. A robust bayesian truth serum for small populations. In *AAAI Conf. on Artificial Intelligence*.

Zohar, A., and Rosenschein, J. 2006. Robust mechanisms for information elicitation. In *Conference on Autonomous Agents and Multiagent Systems*, 1202–1204.