Truthful Opinions from the Crowds

RADU JURCA
Google Inc.
and
BOI FALTINGS
Artificial Intelligence Lab, Ecole Polytechnique Fédérale de Lausanne (EPFL)

An increasing number of applications of artificial intelligence extract knowledge from large groups of agents, also termed the *wisdom of the crowds*. One example are online feedback forums (also known as reputation mechanisms) for obtaining information about the products or services. The testimonies of previous buyers disclose hidden product attributes such as quality, reliability, ease of use, etc., that can only be observed after the purchase. This previously unavailable information allows the buyers to make more efficient decisions, and eliminates some of the problems that would otherwise lead to the collapse of online markets [Akerlof 1970].

Recent studies, however, raise important questions regarding the ability of existing reputation mechanisms to reflect the real quality of a product. First, the absence of clear incentives drives only some of the users to voice their opinions. For example, most Amazon ratings for a book or CD are either very good, or very bad, while controlled experiments on the same items reveal normally distributed opinions [Hu et al. 2006]. Second, some users intentionally lie to distort the public reputation in their favor. Fake reviews can be seen on Amazon [Harmon 2004], TripAdivsor [Keates 2007], or in song charts [White 1999]. Although we still see high levels of altruistic (i.e., honest) reporting, the increasing awareness that gains can be made by manipulating online reputation will likely attract more dishonest reporting in the future.

Both problems can be solved by explicitly rewarding users for reporting feedback. The mechanism scales the *payments* (monetary or in kind) to the reporters such that (i) the expected reward is greater than the cost of reporting, and (ii) honest reporting becomes the optimal strategy. This technique is not limited to reputation mechanisms, but applies more generally to any setting where a private signal is inferred from reports by a crowd of self-interested rational agents.

[Miller et al. 2005] describe a framework for designing incentive-compatible payments for online feedback forums characterized by *pure adverse selection*. In such environments, buyers observe noisy signals about the underlying quality attributes of products or service providers, and the role of the reputation mechanism is to

Authors' addressses: radu.jurca@gmail.com, boi.faltings@epfl.ch

signal, or differentiate products of different quality. Examples of such a situation are product rating forums such as Amazon, ePinions or Bizrate, and most services that are provided through machines or networks in an anonymous fashion.

Intuitively, incentive-compatible payments exploit the correlation between the private signal observed by a buyer, and her beliefs regarding the feedback of another buyer (the reference reporter). Different observations change the private belief regarding the product's true quality, and hence generate different expectations regarding the reference report. The payment reflects the distance between the updated expected distribution of the reference report, and the reference report itself. Every buyer has the incentive to "align" her private belief regarding the reference report to the public one, which can be done by reporting the truth. Honest reporting becomes a Nash equilibrium.

As a first contribution we explored alternative algorithms for designing incentive-compatible rewards [Jurca and Faltings 2006]. We implemented the principle of *automated mechanism design* [Conitzer and Sandholm 2002], and defined the design problem as an optimization program that minimizes the total cost of the reputation mechanism. Other extensions include mechanisms that consider several reference reports, that filter reports likely to be false, or that can accommodate imperfect information [Jurca and Faltings 2007c]. All these additions increased significantly the performance and applicability of incentive-compatible reward mechanisms.

A second direction of our research addresses the problem of collusion. Incentive-compatible reward mechanisms generally have other equilibria as well [Jurca and Faltings 2005]. For example, a simple lying equilibrium is for all agents to always report the same, thus leading to perfect prediction of the reference reports. As any product or service in the real world will have occasional defects, truthful reporting will always be a noisy predictor of the reference report, and thus not be able to match the payoff of the lying strategy. Rational agents can exploit this fact and collude to extract payments from the mechanism.

A simple idea to combat collusion is to reward more the reports that predict the slight imperfections that inherently occur in every real product. For example, a report can be scored against a *set* of at least 4 reference reports, without considering their order. By giving a higher reward for matching all but one of the reference reports, it is possible to give a higher expected payoff to the truthful reporting equilibrium.

As it is difficult to see how to scale the rewards to obtain this characteristic, we use *automated mechanism design* to compute rewards that satisfy both incentive compatibility constraints, and additional criteria directed to combat collusion. We individually consider different collusion scenarios where:

—all or only some of the agents can become part of a lying coalition,

—colluders can coordinate or not on using different strategies (e.g., colluding strategies can be a symmetric or asymmetric),

—colluders can pay other colluders or not (e.g., settings with transferable vs. non-transferable utilities).

We also consider different degrees of resistance to collusion. Ideally, honest reporting would be the dominant strategy for the colluders, such that no matter

what other colluders do, honest reporting is optimal for every individual colluder. Clearly no rational agent should be expected to join a lying coalition under such circumstances.

When dominant truth-telling is not possible, the second most preferable option is to have honest reporting as the unique Nash Equilibrium. Any lying coalition would imply a non-equilibrium situation, where individual colluders would rather report differently than specified by the colluding strategy. Assuming non transferable utilities and the absence of effective punishments that a coalition can enforce on its members, any lying coalition would be unstable, and therefore, unlikely to form.

Finally, when honest reporting cannot be neither the dominant strategy, nor the unique Nash Equilibrium, collusion resistance can emerge if honesty is the Pareto-optimal Nash equilibrium. The intuition here is that any stable (i.e., equilibrium) lying coalition would make at least one of the colluders worse off than in the honest equilibrium, which hopefully prevents that agent from joining the coalition in the first place (assuming, again, non-transferable utilities).

The detailed results are described in [Jurca and Faltings 2007a] and [Jurca 2007], and are summarized by Table I. When all agents may collude, colluders may coordinate on asymmetric strategies and can redistribute the payments, collusion resistance is trivially impossible. Also, it does not make sense to look at settings where utilities are transferable, but collusion strategies are restricted to symmetric ones: assuming that colluders are sophisticated enough to make payments among themselves, they should also be able to coordinate on asymmetric strategies. For all of the remaining five cases, we obtain positive results.

For example, the lower right corner of the table addresses the scenario where only a fraction of the agents may collude, but colluders can redistribute the payments and may use asymmetric strategies. This setting closely models the *sybil attack* [Cheng and Friedman 2005] where the same strategic entity controls a number of fake online identities. Here, even if only one agent is assumed to report truthfully (without knowing the identity of the honest reporter) we could design payments that make it more profitable for the coalition as a whole to report honestly.

Another example is the second column of the table where utilities are non-transferable, and colluders may use asymmetric strategies. When all agents are assumed to be potential colluders, honest reporting may only be imposed as a Pareto-optimal equilibrium. If, however, at least one of the agents is assumed to report the truth (again, without knowing the identity of the honest reporter) there is a payment mechanism that makes honest reporting the unique equilibrium. Moreover, if the majority of the agents is assumed to be honest, there is a payment mechanism that makes truthtelling the dominant strategy for the colluders.

Finally, a third dimension of our research addresses settings with *moral hazard*. This exists in environments where the provider can vary the quality attributes for each particular buyer in a strategic manner. The role of the reputation mechanism is to spread information about seller misbehavior and increase its cost to make it unattractive.

We generalize the results obtained by [Dellarocas 2005] for binary settings, and show how to design efficient reputation mechanisms when sellers can exert an arbitrary number of effort levels, and buyers can observe an arbitrary number of

	Non-Transferable Utilities		Transferable Utilities	
	symmetric strategies	asymmetric strategies	symmetric strategies	asymmetric strategies
all agents collude	-unique honest NE; -Pareto-optimal honest NE	-Pareto-optimal honest NE	unreasonable assumption	impossible
some agents collude	-unique honest NE; -Pareto-optimal honest NE	-honesty as dominant strategy ($N_{col} < \frac{N}{2}$); -unique honest NE; -Pareto-optimal honest NE	unreasonable assumption	-(sybil attack), the coalition maximizes its revenue by reporting honestly;

Table I. Summary of collusion scenarios and mechanism results

feedback signals [Jurca 2007]. We also describe a first mechanism [Jurca and Faltings 2007b] that encourages the buyers to report the truth by allowing them to develop a reputation as honest reporters.

REFERENCES

Akerlof, G. A. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics 84,* 3, 488–500.

Cheng, A. and Friedman, E. 2005. Sybilproof reputation mechanisms. In *Proceeding of the Workshop on Economics of Peer-to-Peer Systems (P2PECON)*. 128–132.

Conitzer, V. and Sandholm, T. 2002. Complexity of mechanism design. In *Proceedings of the Uncertainty in Artificial Intelligence Conference (UAI)*.

Dellarocas, C. 2005. Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard. *Information Systems Research 16,* 2, 209–230.

Harmon, A. 2004. Amazon Glitch Unmasks War of Reviewers. *The New York Times*.

Hu, N., Pavlou, P., and Zhang, J. 2006. Can Online Reviews Reveal a Product's True Quality? In *Proceedings of ACM Conference on Electronic Commerce (EC 06)*.

Jurca, R. 2007. Truthful Reputation Mechanisms for Online Systems. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL).

Jurca, R. and Faltings, B. 2005. Enforcing Truthful Strategies in Incentive Compatible Reputation Mechanisms. In *Internet and Network Economics (WINE'05)*. LNCS, vol. 3828. Springer-Verlag, 268–277.

Jurca, R. and Faltings, B. 2006. Minimum Payments that Reward Honest Reputation Feedback. In *Proceedings of the ACM Conference on Electronic Commerce (EC'06)*. Ann Arbor, Michigan, USA, 190–199.

Jurca, R. and Faltings, B. 2007a. Collusion Resistant, Incentive Compatible Feedback Payments. In *Proceedings of the ACM Conference on Electronic Commerce (EC'07)*. San Diego, USA, 200–209.

Jurca, R. and Faltings, B. 2007b. Obtaining Reliable Feedback when Clients can Commit to Report Honestly. *Journal of Artificial Intelligence Research 29*.

Jurca, R. and Faltings, B. 2007c. Robust Incentive-Compatible Feedback Payments. In *Trust, Reputation and Security: Theories and Practice*, M. Fasli and O. Shehory, Eds. Vol. LNAI 4452. Springer-Verlag, Berlin Heidelberg, 204–218.

Keates, N. 2007. Deconstructing TripAdvisor. *The Wall Street Journal*, page W1.

Miller, N., Resnick, P., and Zeckhauser, R. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science 51*, 1359 –1373.

White, E. 1999. Chatting a Singer Up the Pop Charts. *The Wall Street Journal*.