Crowdsourcing of Formal Scientific Knowledge

Ahmed Abdelmeged
Northeastern University
Boston, MA 02115 USA
mohsen@ccs.neu.edu

Karl Lieberherr
Northeastern University
Boston, MA 02115 USA
lieber@ccs.neu.edu

WebSci'13, April 27 – May 2, 2013, Paris, France.
ACM 978-1-XXXX-XXXX-X/XX/XX.

Introduction

Popperian Scientific Method

The main contribution of this paper is a generic game, called the Scientific Community Game (SCG), for the Popperian Scientific Method [14]. Popper, one of the most influential philosophers of science of the previous century, promoted in "Conjectures and Refutations" the idea that each claim should have a description how to refute the claim. SCG is a game designed to encourage constructively-solvable disputes about a predefined set of claims. The reason we want to encourage those constructively-solvable disputes is that they help advance and focus the scientific discourse and learning.

We apply Popper's ideas to formal sciences as well as to formal models in other sciences. Formal sciences are not concerned with the validity of claims based on observations in the real world, but instead with the properties of formal systems based on definitions and rules. Examples of formal sciences are many: logic, mathematics, theoretical computer science, information theory, systems theory, decision theory, statistics. Formal models exist in sufficiently well understood sciences so that we have efficient simulation software that defines a sufficiently precise computable model. In biology, this is called an "in-silico" science.

We consider a lab that formulates claims about the science. We have several lab users, called scholars, who form their opinions about the claims of the lab. Then they meet in pairs on the web and engage in a scientific discourse to determine who has likely the correct opinion. This determination is done by playing several binary games in which the players assume the position they believe is true.

An example of a claim is SolarCell(R,t,s,f) which will be used in the context of a game

$$G(SolarCell(R,t,s,f),p1,p2,verifier,falsifier)$$

where p1 and p2 are the two players. We assume that p1 wants to be the verifier and p2 wants to be the falsifier. If the falsifier gives to the verifier raw materials, energy and equipment of kind and amount R, the verifier produces in time t a quadratic solar cell of area s and efficiency f. In the game the falsifier provides, all in silico, raw materials etc. of kind and amount K, and the verifier is given time t to apply its secret construction to produce a solar cell of the predicted size and efficiency. If the verifier does not produce what it predicted it is said to be in a contradiction and the verifier loses the game.

In our version of Popperian Science, each claim c has a two-person game

$$G(c,p1,p2,r1,r2)$$

attached, called the Semantic Game (SG), that defines what is needed to defend one's position of a claim in a specific lab. The claims are grouped into claim families that we call labs. The purpose of the game is to (1) ask the players to show a personal performance of the skills needed to defend the true claims in the lab (2) evaluate the players with respect to their skills relevant to the lab (3) bring at least one of the players into a "personalized" contradiction. This does not mean that the position of the player who got into the contradiction is false. But it is an indication that it might be false casting doubts on the skills of the player who took the position. The personalized contradiction has the flavor: you predict an outcome of an experiment connected to the claim and involving the two players but when the experiment is carried out the predicted outcome does not happen.

Crowdsourcing
Crowdsourcing has become an important problem solving approach that enables us to tackle large scale problems that require human intelligence to solve. Crowdsourcing has been successfully applied to several problems over the past decade. Including, labeling images indexed by Google on the web [15], discovering protein foldings [6], synthesizing proteins [3] and building the Wikipedia.

It is our goal to apply crowdsourcing to solve computational problems. To achieve this goal, there are four challenging questions that we need to address [7]:

1. What contributions can users make?
2. How to evaluate users and their contributions?
3. How to combine user contributions to solve the target problem?
4. How to recruit and retain users?

Thesis

Our thesis is that semantic games of interpreted logic statements provide a useful foundation for building *successful* crowdsourcing systems for building formal scientific knowledge.

Rationale and Limitations of Semantic Games
SGs of claims provide attractive answers to the four challenging questions of crowdsourcing systems. However, these answers are only valid in a limited context. Most notably, SGs define an interaction mechanism between two users only. A *successful* SG-based system must generalize SGs to a much wider context and improve on the way SGs address these four challenging questions, whenever possible.

User Contributions
During the course of playing an SG, users make two kinds of *formal* contributions: positions and supporting actions. Further details about the syntax can be found in [1].

Apart from playing SGs, users can still contribute by improving their own SG playing strategies. By doing so, players are able to spot more problems in the positions taken by their opponents in future games. Because users have to follow a well defined formal protocol **??** to play an SG, this enables users to *automate* the execution of their strategies into *avatars*. Algorithms used in avatars are themselves yet another potential formal contribution.

Evaluating Users
SGs provide *relative*, *objective*, and *self-sufficient* approach to assess the *relative strength* of users. Simply put, the winner of an SG is considered *stronger* than the loser. This approach is fundamentally different than the current evaluation schemes used in crowdsourcing systems such as: gold standards, trusted workers and probabilistic oracles, and disagreement-based schemes [9].

Disagreement-based schemes evaluate the *absolute strength* of users based on how often the user's contribution is "correct" where a "Correct" contribution is defined to be *similar* to the "majority vote". SG-based evaluation is independent of the "correctness" of user contributions. Instead SG-based evaluation can *objectively* judge one contribution to be "better" than the other. It is worth noting that the "better" contribution is not always necessarily *similar* to the "majority vote".

SG-based evaluation is said to be *self-sufficient* because, unlike gold standard evaluation, it is not based on a set of pre-populated test cases. Instead, the two users test each others.

An SG-based crowdsourcing system must somehow decide which SGs to be played. To decide on an SG to be played, the system must decide on a claim, a user to play the role of the verifier and a user to take the role of the falsifier. It is possible that a system can delegate some of these decisions to the users. The system must also utilize the outcome of a large number of SGs to evaluate users' strength. The naive approach of summing the number of SGs the user won is unlikely to be fair due to several concerns that gives one group of players an advantage over another group of users. A comprehensive list of these concerns given by:

1. Users can be at an advantage (or at a disadvantage) if the system chooses to force them to participate in more SGs where they are at an advantage (or at a disadvantage). A player is at an advantage (or at a disadvantage) in an SG if either the claim **(CONCERN 1.a)** or the position **(CONCERN 1.b)** is only forced on their adversary (or only forced on them).
2. Users can be at an advantage (or at a disadvantage) if the system chooses to forces them to participate in more (or fewer) than the average number of SGs played by their counterparts **(CONCERN 2)**.

3. Users can be at an advantage (or at a disadvantage) if the system chooses to force them to participate in more SGs against other weaker (or stronger) users **(CONCERN 3)**.

4. If a group of users can form a coalition with the goal of artificially increasing the strength of a particular user through losing against that user on purpose, then that user is at an advantage **(CONCERN 4)**.

One potential approach to address the first concern is to ensure that the system only chooses SGs in which neither of the players is at a disadvantage. However, this approach is ineffective because it make it impossible to get two users to play an SG because they have to hold two opposing positions on some claim that they independently come up with. Furthermore, it makes it harder to spot users holding the correct position on a particular claim but for the wrong reason. The second and third concerns can be addressed through either restricting the algorithm by which the system decides which SGs to be played, or through a more sophisticated approach to assess the user strength, or through both approaches. Anonymity can be used to defend against the fourth concern.

Evaluating User Contributions
By definition, the contributions of an SG winner are "better" than the contributions of the loser. However, we can not consider the winner contributions as *potentially correct* unless:

1. The position taken by the winner was not forced **(CONCERN 5)**.

2. There is a mechanism to discourage "cheating" (i.e. *knowingly* making "incorrect" contributions) either because their adversary is weak enough not to discover the "cheat", or to lose on purpose against their opponent **(CONCERN 6)**.

Again, anonymity can be used to discourage "cheating". It is also possible to hold the positions taken by users against themselves in future SGs.

Combining User Contributions
It is possible to collect the potentially correct contributions of all winners of SGs into a contribution database. The *crowd beliefs* about claims can be assessed from the contribution database. It is possible that "incorrect" contributions make it to the contribution database **(CONCERN 7)**. Therefore, it is necessary to have a mechanism to periodically clean the contribution database in order to enable more accurate assessment of the crowd beliefs.

Apart from estimating the crowd beliefs, SG losers get precise feedback on how they can improve their SG playing strategies. Furthermore, users can then build on the crowd beliefs. For example, suppose that the winners where mostly taking the verifier position on the claim $\forall k : divides(k, 3571) \Leftrightarrow k \in \{1, 3571\}$, then this likely-to-be-true claim can be used as a test case for factorization algorithms.

Recruiting and Retaining Users
Participating in an SG can provide users with an intrinsically rewarding experience. The exact intrinsic rewarding experience is user dependent. For example, some participants can find the act of game play against an adversary to be fun. Others can enjoy the educational (or collaborative) nature of SGs that comes from the fact that the winner of an SG gives the loser a very targeted

feedback.

We believe that the following three factors that could enhance the intrinsically rewarding experience that SGs provide to users:

1. Choosing claims that both players find interesting **(CONCERN 8)**.
2. Allowing users to choose their positions on claims **(CONCERN 9)**.
3. Matching players with similar levels of strength **(CONCERN 10)**.

Neither intrinsic nor extrinsic reward is absolutely superior [1] [2]. However, most certainly, a crowd would have users that prefer both kinds of rewards. Therefore, it is still useful to include other encouragement and retention schemes **(CONCERN 11)** such as: instant gratification, providing ways to establish, measure, and show different qualities of the users, establishing competitions, and providing ownership situations [7].

Initial Investigation

To support our thesis, we designed and partially implemented [2] a proof of concept SG-based crowdsourcing system. We briefly describe our system below. Further details about the current system be found in [1]. Details about earlier versions and their evolution can be found in [5], [4], [12].

[1] For example, consider using Amazon Mechanical Turk (AMT) to label all images indexed by Google. Would that be as cost effective as the ESP game? A second example is building the Wikipedia. Would it be as cost effective to build the Wikipedia using AMT?

[2] Extrinsic reward is believed to be superior in motivating automatic (motor) tasks, while intrinsic value would be superior in motivating intelligent (cognitive) tasks [13], [10], [8].

In a nutshell, our system uses first order logic to express claim families, and uses the semantical games of first order logic formulas defined by Hintikka's Game-Theoretic-Semantics [11].

To ensure that claims are never forced on users our systems uses labs. Labs define special interest groups of users. A lab is created by an owner (one kind of users) and consists of a family of claims. Scholars (another kind of users) choose to join the labs they find *interesting*. The system only allocates users to SGs of claims from the labs they joined. This enhances the users' experience while participating in SGs **(CONCERN 8)** and guarantees that users are never at a disadvantage regardless of the method used to chose the underlying claims for SGs **(CONCERN 1.a)**.

The main interaction mechanism of scholars with the system is called the *driver* mechanism. The driver mechanism decides on the SGs to be played, the scholars participating in every SG as well as their roles. However, rather than making scholars participate in SGs directly, the driver mechanism in our system makes user participate in Contradiction-Agreement Games (CAGs). Although CAGs are composed of SG, CAGs can be played by two player taking the same position on the underlying claims. This enhances the users' experience **(CONCERN 9)**. Furthermore, CAGs are specifically designed to provide fair evaluation **(CONCERN 1.b)** and to identify potentially correct contributions **(CONCERN 5)**. CAGs are described in [1]. Currently, our system has a per-lab driver mechanism. Lab owners are required to provide their driver mechanisms taking into account to match scholars with close enough strength. This is critical to enhance the users' experience **(CONCERN 10)** and fairness **(CONCERN 3)**.

Our system uses an algorithm to evaluate the users' strength as fairly as possible. Our algorithm is designed to address the fairness concerns **(CONCERN 2,3)**. The algorithm is described in [1]. To estimate crowd beliefs, our system uses a simple formula that is presented in [1]. To discourage "cheating" **(CONCERN 4,6)**, our system relies on anonymity. Currently, our system does not provide a mechanism for cleaning the contributions database **(CONCERN 7)** nor any encouragement and retention schemes **(CONCERN 11)** other than the fun that scholars get from participating in SGs.

References

[1] Website. http://www.ccs.neu.edu/home/mohsen/proposal-ahmed.pdf.

[2] Website. https://github.com/amohsen/fscp.

[3] EteRNA. Website, 2011. http://eterna.cmu.edu/.

[4] Abdelmeged, A., and Lieberherr, K. J. The Scientific Community Game. In *CCIS Technical Report NU-CCIS-2012-19* (October 2012). http://www.ccs.neu.edu/home/lieber/papers/SCG-definition/SCG-definition-NU-CCIS-2012.pdf.

[5] Abdelmeged, A., and Lieberherr, K. J. FSCP: A Platform for Crowdsourcing Formal Science. In *CCIS Technical Report* (February 2013). http://www.ccs.neu.edu/home/lieber/papers/SCG-crowdsourcing/websci2013_submission_FSCP.pdf.

[6] Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., and Popović, Z. The challenge of designing scientific discovery games. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, ACM (New York, NY, USA, 2010), 40–47.

[7] Doan, A., Ramakrishnan, R., and Halevy, A. Y. Crowdsourcing systems on the world-wide web. *Commun. ACM 54*, 4 (Apr. 2011), 86–96.

[8] Ipeirotis, P. G., and Paritosh, P. K. Managing crowdsourced human computation: a tutorial. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, ACM (New York, NY, USA, 2011), 287–288.

[9] Joglekar, M., Garcia-Molina, H., and Parameswaran, A. Evaluating the crowd with confidence. Technical report, Stanford University, August 2012.

[10] Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[11] Kulas, J., and Hintikka, J. *The Game of Language: Studies in Game-Theoretical Semantics and Its Applications*. Synthese Language Library. Springer, 1983.

[12] Lieberherr, K. J., Abdelmeged, A., and Chadwick, B. The Specker Challenge Game for Education and Innovation in Constructive Domains. In *Keynote paper at Bionetics 2010, Cambridge, MA, and CCIS Technical Report NU-CCIS-2010-19* (December 2010). http://www.ccs.neu.edu/home/lieber/evergreen/specker/paper/bionetics-2010.pdf.

[13] Pink, D. *Drive: The Surprising Truth About What Motivates Us*. Canongate Books, 2011.

[14] Popper, K. R. *Conjectures and refutations: the growth of scientific knowledge, by Karl R. Popper*. Routledge, London, 1969.

[15] von Ahn, L., and Dabbish, L. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, ACM (New York, NY, USA, 2004), 319–326.