Incentivising Monitoring in Open Normative Systems

Natasha Alechina
University of Nottingham
Nottingham, UK
nza@cs.nott.ac.uk

Joseph Y. Halpern*
Cornell University
Ithaca, USA
halpern@cornell.edu

Ian A. Kash
Microsoft Research
Cambridge, UK
iankash@microsoft.com

Brian Logan
University of Nottingham
Nottingham, UK
bsl@cs.nott.ac.uk

Abstract

We present an approach to incentivising monitoring for norm violations in open multi-agent systems such as Wikipedia. In such systems, there is no crisp definition of a norm violation; rather, it is a matter of judgement whether an agent's behaviour conforms to generally accepted standards of behaviour. Agents may legitimately disagree about borderline cases. Using ideas from scrip systems and peer prediction, we show how to design a mechanism that incentivises agents to monitor each other's behaviour for norm violations. The mechanism keeps the probability of undetected violations (submissions that the majority of the community would consider not conforming to standards) low, and is robust against collusion by the monitoring agents.

Introduction

Consider a setting in which agents contribute to and collectively maintain a shared resource. Such settings arise frequently. Examples include a collectively authored encyclopedia (such as Wikipedia), a community-authored street map (such as OpenStreetMap), an open source software project, and a discussion forum in which people comment on news stories. Agents can make changes to the resource, for example, by writing a new article in the encyclopedia or by making a series of edits to an existing article. We assume that there are generally accepted community standards governing what constitutes acceptable behaviour. These standards may, for example, regulate the content of articles that can be posted or require that articles must be balanced.[1] The community in general benefits if the standards are upheld. However, the actions of some agents may occasionally violate the community standards. Such violations may be inadvertent (the agent does not realise his behaviour is non-compliant) or malicious (the agent knows his behaviour is non-compliant).

*Supported in part by NSF grants IIS-0534064, IIS-0812045, IIS-0911036, and CCF-1214844, and by AFOSR grants FA9550-08-1-0438, FA9550-09-1-0266, and FA9550-12-1-0040, and ARO grant W911NF-09-1-0281.

[1]Examples of standards applying to Wikimedia projects can be found at `meta.wikimedia.org/wiki/Wikimedia_principles`.

In this paper, we focus on the question of how to ensure the community standards are applied and offending behaviour is detected. We do not consider *enforcement* here, that is, what should be done about someone who violates the community standards; our focus is only on detection of a violation. A key problem is that the community standards that govern the behaviour of agents are typically not hard and fast rules—they require interpretation in each particular case to determine whether a violation has occurred. We assume that the standards are broadly accepted, but agents may legitimately disagree about borderline cases.

The problem of incentivising agents to provide accurate answers in settings where there is no ground truth has been extensively studied in the literature on *peer prediction* (see, e.g., (Miller *et al.* 2005)). The key idea is that, while there is no certain way to tell whether an agent has answered correctly, agents doing their best to make a determination will tend to agree with each other much of the time, so we can use agreements with other agents as a measure of how good a job an agent is doing. Recently, these techniques have seen use in crowdsourcing settings as a way to reward people for quality work and deter behaviours such as spamming (Kamar and Horvitz 2012).

Prior work on peer prediction has studied settings where agents are incentivised with money; these techniques are not immediately applicable to settings where monetary incentives cannot be used (as is typically the case for collectively maintained resources). Alechina et al. (2016) showed how incentives can be supplied for detecting violated norms without the use of money in settings where (1) if a violation is detected, there is an easily-checkable incontestable 'witness' to the violation, that is, an easy way to convince others that there was indeed a violation (e.g., an inappropriate piece of text) and (2) that the violation would be found if an agent checked (even though checking might take time). In this paper, we show how the approach of Alechina et al. can be extended to allow us to use peer prediction without reliance on money. While, for concreteness, we focus on the specific problem of determining whether a norm has been violated, our techniques are equally applicable to other tasks for which agents have no access to ground truth, a common occurrence in crowdsourcing settings (e.g. labelling data).

In (Alechina *et al.* 2016), agents are incentivised to monitor by payments in scrip, which the agents use in turn to pay

for their own submissions in the future (which are monitored for norm violations). Each norm violation was assumed to have an easily checkable witness and was guaranteed to be found by a motivated checker, it sufficed to have a single motivated agent check for a violation. However, we consider standards that are a matter of interpretation, so we cannot simply choose an agent to check for a violation. A checking agent may, for example, have an incentive to falsely claim that that behaviour violates the community standards if she disagrees with the material or wishes to suppress a particular point of view. Similarly, a checking agent may prefer not to report what she suspects most people would view as a violation. We do assume that the community standards are sufficiently widespread that most people will agree on what constitutes a violation and that subject expertise is not needed to detect a violation. Moreover, we allow for the possibility that there is a small sub-community with different standards or that may deliberately want to violate the norm (for example, by defacing a Wikipedia page).

Thus, rather than choosing a single agent to monitor a contribution, we choose (at random) a small committee to do the monitoring. Roughly speaking, we take there to be a violation if a majority of the committee believes that there is a violation. Of course, we must set up the incentives so that agents cannot gain by collusion. Once we consider such committees, there are further problems. If we reward a committee for detecting a violation (i.e., if a majority of the committee says that there is a violation), it seems that an agent has an incentive to always declare there to be a violation. With some probability, her declaring a violation might result in a violation being declared where otherwise that would not have been the case, so she gets rewarded. We deal with this by punishing agents slightly for declaring a violation when other agents do not. As in (Alechina *et al.* 2016), the 'punishments' and 'rewards' are in terms of scrip. The connection between scrip and utility is somewhat indirect: scrip is needed to post a submission, which is what gives utility. Although what we have is essentially a *scoring rule*, of which many are known in the literature (Gneiting and Raftery 2007), the fact that we are dealing with both scrip and utility introduces subtleties. In any case, setting the penalties appropriately suffices to give a Nash equilibrium.

Preliminaries

As in (Alechina *et al.* 2016), we consider two settings. In the first setting, while agents may submit content that most community members would judge as not conforming to the community standards, they do so unintentionally, intuitively, because they viewed it as acceptable even though most community members would not. We shall refer to such submissions as 'bad' although we stress that we do not presuppose the existence of ground truth. Suppose that, in the case of unintentional violations, the probability of unintentionally submitting bad content is b. In the second setting, we assume that whether to make a bad submission is a strategic decision (e.g., someone may want to vandalize a Wikipedia page). In this second scenario, we will also be particularly interested in coalitions, since it may well be the case that there is a sub-community that desires to vandalize a page. (Dealing with coalitions is also important in the unintentional case.)

Roughly speaking, in the setting of Alechina et al. (2016), where an agent who found a violation could provide unimpeachable evidence of the violation, the idea was that a random monitor was chosen for each post. If the monitor detected a violation, then he was paid (in scrip). The payment was sufficient that, given the probability of violations, the expected amount of scrip received for monitoring was sufficient compensation for the loss of utility due to monitoring (because the scrip was needed to allow further posts, which increased utility). Thus, agents were incentivised to volunteer to be monitors.

We cannot just adopt the same approach here, since the two assumptions made by Alechina et al. do not hold: there is no 'witness' to a violation nor is a motivated monitor guaranteed to find a violation (since there may not be agreement on whether a violation even occurred). For modelling purposes, we assume that there are two types of submissions, which we call *good* or *bad*. A community member may view a good submission as unacceptable (this is a false negative), but this is unlikely; similarly, a community member may view a bad submission as acceptable (this is a false positive), but this is also unlikely. We assume for simplicity that there is a probability μ for both false positives and false negatives.

Unintentional Setting

Adapting the model of Alechina et al. (2016), the game in the unintentional violations case is described by the following parameters:

- a finite set of n agents $1, \ldots, n$;
- the time between rounds is $1/n$;[2]
- at each round t an agent is picked at random to submit (we implicitly assume that agents always have something that they want to submit);
- utility of submission (to agent doing the submitting): 1;
- (dis)utility of monitoring (to agent doing the monitoring, provided the agent actually does the checking): $-\alpha$ (where $0 < \alpha < 1$);
- probability of a bad submission: b;
- probability of a false positive or a false negative: μ;
- maximum acceptable probability of the system making an error: μ';
- discount rate: $\delta \in (0, 1)$.

The game runs forever. As is standard in the literature, we assume that agents discount future payoffs. This captures the intuition that a util now is worth more than a util tomorrow, and allows us to compute the total utility derived by an agent in the infinite game. We have assumed for simplicity that the system is homogeneous: all agents get the same utility for submitting (1), the same disutility for monitoring ($-\alpha$),

[2]The assumption that the time between rounds is $1/n$, originally due to Friedman et al. (2006), makes the analysis easier. It guarantees that, on average, each agent wants to make one submission per time unit, independent of the total number of agents.

have the same probability of being chosen to submit something ($1/n$), and have the same discount factor (δ). Using ideas from (Kash *et al.* 2012), we can extend the approach discussed here to deal with different *types* of agents, characterised by different parameters. Some agents may want to submit more often; other agents may be less patient (so have a smaller discount rate); and so on.

We need additional notation to describe what happens:

- $p^t \in \{1, \ldots, n\}$ is the agent chosen to submit in round t;
- $V^t \subseteq \{1, \ldots, n\} \setminus \{p^t\}$ is the set of agents chosen to monitor in round t;
- v_i^t: the vote of agent i if $i \in V^t$; $v_i^t = 1$ if the monitor 'votes' for a violation, and 0 otherwise.

The utility of an agent i in round t is:

$$u_i^t = \begin{cases} 1 & \text{if } i = p^t \text{ and } i\text{'s submission is posted;} \\ -\alpha & \text{if } i \in V^t \text{ and monitors honestly;} \\ 0 & \text{otherwise.} \end{cases}$$

Given the discount factor δ, the total utility U_i for agent i is $\Sigma_{t=0}^{\infty} \delta^{t/n}\, u_i^t$.

Mechanism for the Unintentional Setting

We would like to incentivise the agents to do the monitoring themselves, but we cannot pay them in utility because there is no source of additional utility in the system. In particular, we cannot levy fines for violations and pay monitors in fines because the system is open; instead of paying fines, agents can just leave and re-join under different identity. We address this problem using techniques from *scrip systems* (Friedman *et al.* 2006). We can think of scrip as 'virtual money' or 'tokens'. Performing an action costs tokens and monitoring is rewarded with tokens. The main difference between our setting and that of (Friedman *et al.* 2006) is that instead of always transferring a token from the agent who needs the work done to the agent who does the work, in the monitoring setting the agents who do the work (monitors) are not always rewarded. This requires a non-trivial adaptation of the techniques developed in (Friedman *et al.* 2006).

A fundamental difference from (Alechina *et al.* 2016) and the contexts in which prior scrip systems have been applied is that since monitors do not produce a witness (because they cannot), there is no way to check whether the work which is being rewarded by scrip tokens has actually been done. Nor can a monitor's verdict be checked in an objective way. To address this problem, we use techniques from *peer prediction* (Miller *et al.* 2005), which has been used for incentivising agents to honestly rate products.

At a high level, the idea of the mechanism is that every time an agent submits a request, a set of m agents is chosen to monitor, where m is even. If more than $m/2$ volunteers say that the submission is acceptable, it is accepted; otherwise it is rejected. The size m of this set is chosen to make the probability that the majority vote of the set leading to a false positive or false negative, namely, $\sum_{j=m/2+1}^{m} C(m,j)\mu^j(1-\mu)^{m-j}$, where $C(m,j)$ is m choose j, at most μ' (the desired bound on the system making an error). Payments to agents are calculated using a peer prediction rule to ensure that they maximise their expected number of tokens by truthfully monitoring. We also make sure that the volunteers actually monitor (rather than just pretending to do so) by ensuring that volunteers who simply report an answer without monitoring receive a negative token payment in expectation.

Our interpretation of how bad submissions occur with probability b is that there is a true underlying probability β that a submission is actually acceptable, but agents cannot access this ground truth directly. However, after generating a tentative submission, they can check if it is acceptable. They will post the submission only if it passes the check. However, with probability μ, an unacceptable submission will pass the check. Since the probability that a submission is unacceptable and passes the check is just the probability that it passes the check given that it is unacceptable multiplied by the probability that it is unacceptable, we have that $b = (1-\beta)\mu$.

As the idea of using peer prediction for norm monitoring where there is no witness for a norm violation is of interest even without the rest of our machinery for incentivising monitoring in the absence of money (or other methods of providing rewards and punishments), we state it as a separate lemma. For reasons the proof of the lemma makes clear, the specific payment rule that we use is that an agent who says 'acceptable' and the committee agrees receives no payment, an agent who says 'acceptable' and the committee disagrees must pay a token to the submitter, an agent who says 'unacceptable' and the committee disagrees must also pay a token to the submitter, and an agent who says 'unacceptable' and the committee agrees receives a payment of $\frac{1-b}{b} - \frac{(1-b)^2\mu''}{b^2(1-\mu'')+b(1-b)\mu''}$ from the submitter, where $\mu'' = \sum_{j=m/2+1}^{m-1} C(m-1,j)\mu^j(1-\mu)^{m-1-j}$. (Recall that b is the probability of a bad submission.) This is an example of a scoring rule, as in the peer-prediction literature. Other rules could also be used, as long as they have the required properties (as outlined in Lemma 1).

Lemma 1: *With the payments described, there is an equilibrium of the monitoring game where all agents who volunteer to monitor will exert effort and report their true belief. Furthermore, agents who deviate by not exerting effort receive a negative payoff in expectation while those who follow the equilibrium receive a positive payoff.*

Proof: First, we show that agents who do not exert effort and instead just select a report have a negative expected payment. If the agent just says 'acceptable', he gets a payoff of 0 if the majority says 'acceptable', and a payoff of -1 if the majority says 'unacceptable', so clearly the agent is worse off by saying 'acceptable' (without monitoring) than not volunteering at all. A monitor who deviates by saying 'unacceptable' without monitoring receives a payoff of -1 with probability $(1-b)(1-\mu'')+b\mu''$ (since this is the probability that the majority of the $m-1$ non-deviating monitors say 'acceptable'), and gets $\frac{1-b}{b} - \frac{(1-b)^2\mu''}{b^2(1-\mu'')+b(1-b)\mu''}$ with probability $b(1-\mu'')+(1-b)\mu''$. Note that $(\frac{1-b}{b} - \frac{(1-b)^2\mu''}{b^2(1-\mu'')+b(1-b)\mu''})(b(1-\mu'')+(1-b)\mu'') = (1-b)(1-\mu'')$

(which is exactly why that particular payment was chosen), so the deviator loses $b\mu''$ tokens in expectation with this deviation.

For agents who do actually monitor, it is easily verified that the payments constitute a peer-prediction scheme with correct incentives, so they maximize their expected payoff by monitoring. All that remains is to verify that this expected payment is non-negative, and a straightforward calculation shows that it is approximately $1 - b$. ∎

The lemma shows that correct behaviour is an equilibrium, but in peer-prediction settings, it is well known that there are other equilibria. For example, all monitors always reporting 'acceptable' is an equilibrium, as is all monitors always reporting 'unacceptable' (in the latter case, no agent will ever attempt to post anything). However, since the system is largely composed of agents who wish the norm to be enforced correctly, we expect that the 'cooperative' equilibrium, where all agents do in fact monitor correctly, to be the one that arises in practice.

By setting appropriate parameters and assuming the system has sufficiently many agents who are all sufficiently patient, the mechanism can ensure that the probability of undetected violations is arbitrarily low (with a caveat) and honest participation is arbitrarily close to an equilibrium. The caveat is that m monitors are required and monitoring costs each monitor α while the value of a submission is 1. So if $\alpha m \geq 1$ participation is irrational. This imposes an upper bound on m. Say that μ' is *achievable* if there exists $m \leq 1/\alpha$ such that $\sum_{j=m/2+1}^{m} C(m,j)\mu^j(1-\mu)^{m-j} \leq \mu'$.

Theorem 1: *For all achievable probabilities μ' and all $\epsilon > 0$, there exist a δ sufficiently close to 1, a committee size m, and an n sufficiently large such that if all n agents use a discount factor $\delta' \geq \delta$, then there exists a k such that the mechanism above with all agents volunteering to monitor iff they have fewer than k tokens is an ϵ-Nash equilibrium and the probability of undetected violations is at most μ'.*

Proof: [Sketch] The basic structure of the analysis is the same as that in (Alechina *et al.* 2016; Friedman *et al.* 2006), but the details are more complex. Intuitively, an agent does not want to run out of tokens, since then he will not be able to make a submission. Thus, if he starts running low on tokens, he will volunteer to be a monitor. What is the appropriate threshold at which an agent should volunteer to monitor? That depends in part on how much competition he will have when volunteering, which in turn depends on the other agents' thresholds for volunteering. It can be shown that, if we fix a threshold strategy for each agent, then the system quickly reaches a steady state where we can completely characterize the distribution of tokens (i.e., what fraction of agents will have k' tokens for each value of k'). The analysis involves a consideration of a Markov chain whose states are the number of tokens each agent has. The analysis is a little more complex in our setting than in earlier papers because more than one agent is needed as a monitor. Such systems have been studied by (Humbert *et al.* 2011); while they result in a more complex steady-state distribution of wealth than was the case in the earlier papers, the same basic analysis goes through. The key result is that there is a distribution d of tokens (i.e. the fraction of agents with each number of tokens) such that, for n sufficiently large, the probability of the distribution of tokens in the system being close to d is close to 1.

The rest of the argument is the same as that of (Friedman *et al.* 2006). Clearly, if μ' is achievable, then we can choose a group size m so that the error rate is at most μ'. It can be shown that if the system stayed exactly at the steady-state distribution then the optimal strategy is a threshold strategy. If n and δ are sufficiently large, the potential gains from deviating due to the distribution not staying exactly at the steady state distribution is less than ϵ. ∎

There are two problems with the mechanism above. The first is that we need to explain what to do if an agent has 0 (or a negative number of) tokens. Clearly the agent cannot make a submission, because he will not be able to pay the monitors if they find a problem. But he also cannot monitor, because he may not be able to pay the submitter if his evaluation of the submission does not agree with the majority. We can deal with this problem by allowing agents to monitor even if they have 0 or fewer tokens (although they still cannot make a submission if they do not have enough tokens to pay for it). Now it is possible that monitors will keep getting more and more in debt (we will need to keep track of how much is owed; we ignore how this is done here), but this is extremely unlikely, since, in expectation, agents gain tokens every time they monitor. Another concern is that an agent who is sufficiently in debt will just drop out of the system and re-enter with a new id. We expect that, in practice, for agents who are only slightly in debt, the overhead of re-registering with the system will not make it worth dropping out. For agents who are significantly in debt, it may well be worth dropping out, but this will happen with extremely low probability. This means that the true expected cost of a submission is a little larger than we have stated, because it must take into account that a submitter will not be paid (with extremely low probability) by an agent who owed him money and then drops out of the system. This small change in expectation does not affect the analysis. Friedman and Resnick (2001) discuss this issue and possible mitigations in further detail.

The second problem with this mechanism is that we need m monitors at each step, and m may be large. We can do much better by taking a two-step monitoring approach. We briefly sketch the details here. Suppose, for example, that $\mu = .1$ and we want $\mu' = .0005$. Then, with the mechanism above, we must take $m = 10$. With 10 monitors, we can ensure that the rate of false positives (a bad submission being posted) and the rate of false negatives (a good submission being rejected) is less than $.0005$. Suppose that we use a two-step approach: we take a group of 5 monitors, and accept if at least 4 of the 5 find it acceptable and reject if at least 4 of the 5 find it unacceptable. If neither of these two events occur, then we use 5 more monitors, and accept if greater than 5 find it acceptable and reject otherwise (i.e., we go back to our original scheme). The probability of a false positive and false negative is still less than $.0005$. However, now the ex-

pected number of monitors needed is significantly less. It is straightforward to check that, with probability greater than .9, a submission will be accepted or rejected using just 5 monitors. Thus, the expected number of monitors needed is $< .9(5) + .1(10) = 5.5$, which is significantly less than 10. In general, if m^* is the expected number of monitors needed to achieve a false positive/negative rate of μ', we must have $m^* < 1/\alpha$ for Theorem 1 to hold. The key point is that m^* may be significantly less than the m that was needed for the earlier mechanism.

We can get an even smaller expected number of monitors by starting with a committee of 4 and then increasing the committee one by one up to 10, stopping as soon the decisions are such that the rate of false positives and false negatives is guaranteed to be less than μ'. But, in practice, there might be some overhead in having too many stages in the committee process. While it is not uncommon to have a two-step process, it is rare to go beyond that. For example, while we might have a group of three reviewers for a paper, and then call in additional reviewers if the three reviewers cannot reach consensus, it is rare that we call in additional reviewers one at a time for a potentially extended sequence of steps.

Strategic Violations

We now turn to strategic violations. In this scenario, when an agent is chosen to submit a post, he can choose whether to submit something he believes to be good or something he believes to be bad. (We implicitly assume that agents have a collection of submissions that they believe good or bad available for posting.) Of course, as in the unintentional setting, the agent might be mistaken. Let γ_b be the probability that an intended submission he believes to be bad is bad and let γ_g be the probability that an intended submission he believes to be good actually is good. For simplicity, we take $\gamma_b = \gamma_g = \gamma$, and assume that $\gamma > 1/2$. The other parameters are the same as in the previous section, except that b, the probability that a submission is bad, is no longer determined exogenously, but is the result of the agent's strategic decision. Further, the utility of a bad posting is no longer 1, but $\kappa > 1$. (We must assume $\kappa > 1$ here, otherwise no agent would ever post anything bad: the utility of doing so is no higher than that of posting something good, and the violation may be detected.)

We now show that essentially the same mechanism used in the case of inadvertent bad postings can be used to control what b will be in equilibrium, by choosing the committee size appropriately; larger committees lead to smaller values of b and vice versa. The idea is the following: we proceed as in the previous section, using exactly the same mechanism except that all occurrences of b in the payments are replaced by $1-\gamma$ ($1-\gamma$ plays essentially the same role in the analysis as b), and choose a committee so that the probability that a good submission is accepted is at least $1-\mu'$ and, similarly, the probability that a bad submission is rejected is at least $1-\mu'$. Thus, an agent's expected utility if he posts something he believes to be good is at least $\gamma(1-\mu') + (1-\gamma)\mu'$ (since $\gamma > 1/2$), while the agent's expected utility if he posts something that he believes to be bad is at most $\gamma\mu'\kappa + (1-\gamma)\kappa(1-\mu')$ (here we are assuming that if the submission that the agent believed to be bad is accepted by the committee (so is likely to be good) and is posted, then the agent still gets utility κ). So, as long as we can choose μ' such that

$$\gamma(1-\mu') + (1-\gamma)\mu' > \gamma\mu'\kappa + (1-\gamma)\kappa(1-\mu'), \quad (1)$$

then posting only submissions that the agent believes to be good is better than posting a submission that the agent believes to be bad. Note that (1) will be easy to satisfy if $\gamma > (1-\gamma)\kappa$, which should be the case in practice (we would expect γ to be close to 1).

Note also that, as far as expected number of tokens goes, it is also better to post only submissions believed to be good. Roughly speaking, if an agent posts something that he believes to be good, then in the typical case (where the committee views it as acceptable), he pays 0 to all the agents who agree with the committee and receives 1 token from agents who disagree. On the other hand, if an agent posts something that he believes to be bad, then in the typical case (where the committee views it as unacceptable), then he must pay $\frac{\gamma}{1-\gamma} - \frac{\gamma^2\mu'}{(1-\gamma)^2(1-\mu')+\gamma(1-\gamma)\mu'}$ to each monitor who agrees with the decision. While a detailed calculation must take all the other possibilities into account (e.g., that the post is good but is viewed as unacceptable), the conclusion remains the same: in terms of expected number of tokens, the agent is better off submitting a post he believes to be good than submitting one he believes to be bad.

Thus, we get the following result in the strategic setting:

Theorem 2: *For all achievable probabilities μ' satisfying (1) and all $\epsilon > 0$, there exist a δ sufficiently close to 1, a committee size m, and an n sufficiently large such that if all n agents use a discount factor $\delta' \geq \delta$, then there exists a k such that the mechanism above with all agents using a threshold of k is an ϵ-Nash equilibrium and the probability of undetected violations is at most μ'.*

The proof is omitted because it involves only minor changes to the proof of Theorem 1.

It is interesting to compare this result with the corresponding theorem of Alechina et al. (2016) in the strategic setting. Because, in the setting of Alechina et al., a violation would certainly be found if a bad submission was checked, in equilibrium, the normative organization deliberately does *not* monitor some small fraction of submissions. (Roughly speaking, this is because, in the mechanism proposed by Alechina et al., monitors get paid only if they find a violation. If all posts are monitored, then no bad submissions are made, so monitors have no incentive to monitor.) Furthermore, their equilibrium result in agents *intentionally* making bad submissions, while in our equilibrium all bad submissions still arise inadvertently. On the other hand, in the setting of Alechina et al., the normative organization can get the probability of a bad post to be arbitrarily low. Here, the probability of an undetected violation is at most μ', but μ' must be achievable and satisfy (1).

In the simplified model we used for the analysis in this section, we assumed that all submissions are equally difficult to judge. In practice, some submissions may be obviously good or bad, while others may be borderline cases.

Moreover, malicious agents may have some knowledge of how difficult a submission will be to judge, and exploit this knowledge. Allowing for such additional knowledge would not substantively change our model, but would make the equations messier. For simplicity, we took the probability of a false positive and false negative to be the same, but the analysis would work equally well if they were different. Borderline cases would result in an increased probability of a false negative (and perhaps also decrease γ, since such agents also seem more likely to accidentally choose a non-violating submission). With these changes, essentially the same analysis would go through.

Coalition Resilience

Our previous analysis used peer-prediction techniques to ensure correct incentives for monitors and acceptable error rates. However, in a large open system there may be a small group who work together to either achieve norm violations or otherwise subvert the system (perhaps by simply reducing the effort they need to exert earning tokens).

It follows from the equilibrium analysis sketched in the proof of Theorem 1 that, in equilibrium, for any given submission, a significant fraction of the agents are willing to volunteer to monitor, so the odds of collusion being possible are actually quite low. For example, if there are 500 volunteers per submission and a coalition of size 10 for a 10 person committee, the probability that 2 or more coalition members is less than 0.015. To correct for this probability, we can simply decrease the value of μ' that the algorithm uses slightly, and still end up with a quite robust algorithm.

Related work

As we said, our approach uses techniques from peer prediction (Miller *et al.* 2005). The reward mechanism is very similar, but the crucial difference is that we use scrip to pay the monitors, rather than real utility.

Our analysis of the behaviour and incentives of the token economy draws heavily on prior work on scrip systems by Kash et al. (2006; 2015). We adopt many of their techniques, but extend their analysis to a variant model that applies to our setting. Other work has shown that changing the random volunteer procedure can improve welfare (Johnson *et al.* 2014) and that this approach still works if more than one agent must be hired to perform work (Humbert *et al.* 2011). Work from the systems community has looked at practical details such as the efficient implementation of a token bank (Vishnumurthy *et al.* 2003).

A number of norm-enforcement mechanisms with good incentive properties have been analysed (Kandori 1992; Ellison 1994). De Pinnick et al. (2010) proposed a distributed norm enforcement mechanism that uses ostracism as punishment, and showed both analytically and experimentally that it provides an upper bound on the number of norm violations.

There is a significant amount of work in the MAS literature on infrastructures for implementing normative organisations and on monitoring for norm violations. One common approach involves the use of additional components or agents to implement the normative organisation. For example, Boella and van der Torre (2003) propose 'defender agents' that detect and punish norm violations. Esteva et al. (2004) propose the use of 'governors' to monitor and regiment message exchanges between agents; each agent is associated with a governor, and all interactions with other agents are filtered by the governor to ensure compliance with norms. Grizard et al. (2007) propose an approach in which a separate system of 'controller agents' monitor norm violations and apply reputational sanctions to 'application agents' in a MAS; application agents avoid interactions with other application agents that have low reputation, hence eventually excluding bad agents from the system. Modgil et al. (2009) propose a two-layer approach, in which 'trusted observers' relay observations of states of interest referenced by norms to 'monitor agents' responsible for determining whether a norm has been violated (a similar approach is described by Criado et al. (2012)). Hübner et al. (2010) describe an approach in which 'organizational agents' monitor interactions between agents mediated by 'organizational artifacts'. Balke et al. (2013) used simulation to investigate the effectiveness and costs of paying 'enforcement agents' to monitor norm violations in a wireless mobile grid scenario; the mechanism they propose for rewarding enforcement agents results in a cost to the MAS (in their setting, the telecommunications company), and they assume sanction-based enforcement (agents who violate the norm are punished by the telecommunications company). Testerink et al. (2014) consider the problem of monitoring and enforcement by a network of normative organisations in which each normative organisation has only partial information about the actions of the agents and is capable of only local enforcement (by sanctioning).

In general, these approaches assume that a single component or agent can be used to reliably detect a norm violation. Moreover, the cost of monitoring is borne by the MAS organisation, either in the cost of running additional system components which monitor and regulate interactions (e.g., (Boella and van der Torre 2003; Esteva *et al.* 2004)) or by paying some agents (in utility) to monitor the rest (e.g., (Balke *et al.* 2013)). Fagundes et al. (2014) have explored the tradeoff between the efficiency and cost of norm enforcement in stochastic environments, to identify scenarios in which monitoring can be funded by sanctions levied on violating agents while at the same time keeping the number of violations within a tolerable level. However, in an open multi-agent system, approaches in which norm enforcement is based on sanctioning (e.g., (Grizard *et al.* 2007; Testerink *et al.* 2014)), sanctioned agents may simply leave the system and rejoin under a different id. Because we do not use sanctioning, this issue does not arise under our approach. (As we pointed out, agents may want to leave the system and rejoin if they have a significant token deficit, but this does not really cause a problem for us.)

Conclusions

We have presented an approach to incentivising monitoring for norm violations in open multi-agent systems such as Wikipedia. In such systems, there is no crisp definition of

a norm violation, so we need to rely on potentially erroneous evaluations of individual agents, who may legitimately disagree about borderline cases. Using ideas from scrip systems and peer prediction, we show how to design a mechanism that incentivises agents to monitor each other's behaviour for norm violations. The mechanism keeps the probability of undetected violations low, and is robust against collusion by the monitoring agents. In contrast to prior work, in the presence of strategic decisions about whether to violate norms, our equilibrium does not result in agents intentionally violating the norm or the system sometimes deliberately not monitoring. The equilibrium is achieved by agents following history-independent strategies; this shows that approximately optimal norm enforcement is possible without keeping track of agents' past behaviour. Our techniques may also be of interest for settings other than norm enforcement that also have the features of a lack of both money and access to ground truth.

References

N. Alechina, J. Y. Halpern, I. A. Kash, and B. Logan. Decentralised norm monitoring in open multi-agent systems: (extended abstract). In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1399–1400. ACM, 2016.

T. Balke, M. De Vos, and J. Padget. Evaluating the cost of enforcement by agent-based simulation: A wireless mobile grid example. In *PRIMA 2013: Principles and Practice of Multi-Agent Systems*, volume 8291 of *LNCS*, pages 21–36. Springer Berlin Heidelberg, 2013.

G. Boella and L. W. N. van der Torre. Norm governed multiagent systems: The delegation of control to autonomous agents. In *2003 IEEE/WIC International Conference on Intelligent Agent Technology (IAT 2003)*, pages 329–335. IEEE Computer Society, 2003.

N. Criado, E. Argente, P. Noriega, and V. J. Botti. A distributed architecture for enforcing norms in open MAS. In *Advanced Agent Technology - AAMAS 2011 Workshops, AMPLE, AOSE, ARMS, DOCM3AS, ITMAS. Revised Selected Papers*, volume 7068 of *LNCS*, pages 457–471. Springer, 2012.

A. P. de Pinninck, C. Sierra, and W. M. Schorlemmer. A multiagent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems*, 21(3):397–424, 2010.

G. Ellison. Cooperation in the prisoner's dilemma with anonymous random matching. *Review of Economic Studies*, 61:567–588, 1994.

M. Esteva, B. Rosell, J. A. Rodríguez-Aguilar, and J. L. Arcos. AMELI: An agent-based middleware for electronic institutions. In *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, pages 236–243. IEEE Computer Society, 2004.

M. S. Fagundes, S. Ossowski, and F. Meneguzzi. Analyzing the tradeoff between efficiency and cost of norm enforcement in stochastic environments. In *21st European Conference on Artificial Intelligence (ECAI 2014)*, pages 1003–1004. IOS Press, 2014.

E. J. Friedman, J. Y. Halpern, and I. A. Kash. Efficiency and Nash equilibria in a scrip system for P2P networks. In *Proceedings 7th ACM Conference on Electronic Commerce (EC-2006)*, pages 140–149. ACM, 2006.

E. J. Friedman and P. Resnick. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2):173–199, 2001.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

A. Grizard, L. Vercouter, T. Stratulat, and G. Muller. A peer-to-peer normative system to achieve social order. In *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, pages 274–289. Springer Berlin Heidelberg, 2007.

J. F. Hübner, O. Boissier, R. Kitio, and A. Ricci. Instrumenting multi-agent organisations with organisational artifacts and agents. *Autonomous Agents and Multi-Agent Systems*, 20:369–400, 2010.

M. Humbert, H. Manshaei, and J.-P. Hubaux. One-to-n scrip systems for cooperative privacy-enhancing technologies. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 682–692, 2011.

K. Johnson, D. Simchi-Levi, and P. Sun. Analyzing scrip systems. *Operations Research*, 62(3):524–534, 2014.

E. Kamar and E. Horvitz. Incentives for truthful reporting in crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, Vol. 3*, pages 1329–1330. IFAAMAS, 2012.

M. Kandori. Social norms and community enforcement. *Review of Economic Studies*, 59:63–80, 1992.

I. A. Kash, E. J. Friedman, and J. Y. Halpern. Optimizing scrip systems: crashes, altruists, hoarders, sybils and collusion. *Distributed Computing*, 25(5):335–357, 2012.

I. A. Kash, E. J. Friedman, and J. Y. Halpern. An equilibrium analysis of scrip systems. *ACM Transactions on Economics and Computation*, 2015.

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

S. Modgil, N. Faci, F. Meneguzzi, N. Oren, S. Miles, and M. Luck. A framework for monitoring agent-based normative systems. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 153–160, IFAAMAS, 2009.

B. Testerink, M. Dastani, and J.-J. Meyer. Norms in distributed organizations. In *Coordination, Organizations, Institutions, and Norms in Agent Systems IX*, pages 120–135. Springer, 2014.

V. Vishnumurthy, S. Chandrakumar, and E. G. Sirer. KARMA: a secure economic framework for peer-to-peer resource sharing. In *First Workshop on Economics of Peer-to-Peer Systems (P2PECON)*, 2003.