JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Naroditskiy V, Jennings NR, Van Hentenryck P, Cebrian M. 2014 Crowdsourcing contest dilemma. *J. R. Soc. Interface* **11**: 20140532.
http://dx.doi.org/10.1098/rsif.2014.0532

Received: 20 May 2014
Accepted: 24 July 2014

Subject Areas:
biocomplexity

Keywords:
crowdsourcing, game theory, Prisoner's Dilemma

Author for correspondence:
Victor Naroditskiy
e-mail: vn@ecs.soton.ac.uk

Royal Society Publishing

Crowdsourcing contest dilemma

Victor Naroditskiy[1], Nicholas R. Jennings[1,2], Pascal Van Hentenryck[3] and Manuel Cebrian[3]

[1]School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
[2]Department of Computing and Information Technology, King Abdulaziz University, Saudi Arabia
[3]National Information and Communications Technology Australia, Melbourne, Victoria 3003, Australia

Crowdsourcing offers unprecedented potential for solving tasks efficiently by tapping into the skills of large groups of people. A salient feature of crowdsourcing—its openness of entry—makes it vulnerable to malicious behaviour. Such behaviour took place in a number of recent popular crowdsourcing competitions. We provide game-theoretic analysis of a fundamental trade-off between the potential for increased productivity and the possibility of being set back by malicious behaviour. Our results show that in crowdsourcing competitions malicious behaviour is the norm, not the anomaly—a result contrary to the conventional wisdom in the area. Counterintuitively, making the attacks more costly does not deter them but leads to a less desirable outcome. These findings have cautionary implications for the design of crowdsourcing competitions.

1. Introduction

Numerous successful examples of the power of crowdsourcing to solve problems of extreme difficulty [1–14] have overshadowed important episodes where elaborate sabotage derailed or severely hindered collective efforts. The winning team in the DARPA Network Challenge obtained the locations of the 10 balloons after spending significant efforts filtering the majority of false submissions, including fabricated pictures containing individuals in disguise impersonating DARPA officials [15]. A team from the University of California at San Diego lost its lead in the DARPA Shredder Challenge after their progress was completely wiped out by a relentless number of coordinated overnight attacks [16,17]. The team that topped the US Department of State sponsored Tag Challenge had to withstand a smear campaign orchestrated in Twitter aimed at reducing its credibility [10]. Beyond crowdsourcing competitions, Ushahidi's collective conflict mapping for the Arab Spring had to be shut down for long periods of time owing to suspicions that it had been infiltrated by government officials [16].

These episodes have received a response in the form of emerging work at the intersection of the computer and economic sciences. Recent results in this area have elucidated that it is possible to design incentive structures and algorithmic strategies to modify the efficacy and quality of the crowdsourced solution, the vulnerability of crowdsourcing to malicious behaviour, and the cost of undertaking it [18–41]. These three factors can attain a wide variety of values depending on the particularities of the problem at hand, the economic incentives at stake as well as the algorithmic platform supporting the collective effort. For instance, the combinatorial nature of puzzle assembly in the DARPA Shredder Challenge makes the solution very vulnerable to an attack at a low cost—destroying puzzle progress is much easier than creating it. On the other hand, fabricating a false balloon sighting in the DARPA Balloon Challenge is costly, as it involves a certain degree of counterintelligence skills, more so when individuals posing in false submissions risk identifiability—while at the same time, a false submission does not affect the validity of the other submissions.

In this paper, we propose a formal analysis to explore the efficacy–vulnerability trade-off of crowdsourcing in competitive scenarios. We adopt a

scenario where two firms (players) compete against each other to obtain a better solution to a task which can be crowdsourced. Our goal is to understand whether malicious behaviour exemplified above is the norm or the anomaly, i.e. whether we should expect players to undertake attacks on the collective progress of competing players. We consider the cost of attacking a crowdsourcing strategy, and investigate the effect of higher costs on behaviour of the players. We also aim to quantify how this malicious behaviour affects the likelihood of using crowdsourcing as a strategy, and ultimately how it impacts social welfare under collective problem solving. Such an understanding will be helpful for the design of future crowdsourcing competitions, as well as support the decision process of institutions and firms considering crowdsourcing.

Our main finding is that making attacks more costly perhaps by making it more difficult to attack, does not deter the attackers and results in a more costly and less efficient equilibrium outcome. Paradoxically, making the cost of attacking zero is best for the players. This suggests that care should be used when attempting to discourage attacks by raising the cost. In situations where damage from attacking is high, the incentive to attack is very strong. Instead, crowdsourcing competitions are more suited for scenarios where damage inflicted by an attack is low.

In §2, we present a model for studying the trade-off between higher productivity offered by crowdsourcing and the increased vulnerability that comes with it. Then, we derive the equilibrium of the game defined by the model. We pay particular attention to the case where both players choose to crowdsource and investigate the players' incentives to attack. We also analyse how the cost of an attack and the damage from the attack affect the incentives of players to crowdsource.

2. The model

We study a non-cooperative situation where two players (or firms) compete to obtain a better solution to a given task. The firm with the better solution wins and receives a reward of R. Each firm can develop an in-house solution or crowdsource the task. The former is referred to as the closed strategy, the latter as open. We focus on scenarios where open strategies are likely to be more efficient even though the exact level of efficiency is not known until after a firm engages in crowdsourcing—this reflects a high level of uncertainty underlying engagement processes in social networks [42–49]. In addition, open strategies are susceptible to attacks: a firm using an open strategy can be attacked by the competing firm. The resulting damage from the attack impairs the other firm, and may let the attacking firm win.

To investigate the trade-off between higher efficiency of crowdsourcing and vulnerability to attacks, we propose a model that isolates these two factors. In our model, a firm decides whether or not to crowdsource, and whether or not to attack the competitor. The nature of a crowdsourcing strategy is that it is observable by everyone including the competitors. If the competitor chooses to crowdsource, the opponent would observe it, and can decide whether or not to attack. Note that the decisions are sequential: the decision about crowdsourcing is made first, whereas the decision about attacking is made second. We model this interaction as a sequential game.

Formally, we have a two-stage game. In the first stage, the firms decide whether to solve the task in-house S or to crowdsource it C. Then, the efficiencies of the firms that chose to crowdsource are observed. Owing to the open nature of crowdsourcing, the efficiency or *productivity*, of a crowdsourcing strategy used by player i becomes known not just to the crowdsourcing player, but also to the opponent. We use P_i for the random variable denoting the productivity and p_i for its realized value. The productivity of the in-house solution is fixed and normalized to zero. We assume that the productivity of a crowdsourcing strategy is uniformly distributed between 0 and 1: $P_i \sim U[0,1]$. In the absence of attacks, the firm with a higher productivity wins.

The assumption that the productivity of a crowdsourcing firm is publicly known is consistent with competitions such as the Shredder Challenge and the Tag Challenge. In the Shredder Challenge, the current progress of each competitor was publicly known. In the Tag Challenge, the teams announced the number of targets they had successfully identified as this increased the appeal of the team.

In the second stage, the players decide whether or not to attack the opponent (attacking is denoted by A and not attacking by N). An attack is costly, and the cost $q \in (0,1)$ is expressed as a fraction of the total reward R. This cost can represent a range of situations: the human effort in disrupting the opponent's solution; the complexity of creating multiple identities to carry out a Sybil attack; financial punishment received when the attack is detected in a crowdsourcing competition. The damage inflicted by the attack is denoted by $d \in (0,1)$, which determines how much productivity is taken away from the open strategy (equivalently, how much more productive the attacking firm becomes after 'stealing' the crowdsourced solution). The firm that has a higher productivity at the end of the second stage wins the prize R, which is normalized to be $R = 1$. The parameters q and d are publicly known. We characterize the equilibrium of this two-stage game as a function of these parameters.

3. Equilibrium analysis

We find the subgame perfect equilibrium of the two-stage game. Each pair of players' decisions made in the first stage (to crowdsource or not) result in a different second-stage game. We first analyse each second-stage game (to attack or not). Based on these, we then analyse how decisions are made in the first stage.

When both players use an in-house strategy S, there is no reason to attack, and they both choose N in the second stage. Each one is equally likely to win, and the expected utility of each is 1/2. We note this in the SS cell of the pay-off matrix in table 1.

When player 1 crowdsources and player 2 does not, player 1 cannot attack, but player 2 attacks if doing so puts her ahead of player 1. Player 2 attacks if the realized productivity of player 1 is less than the damage $p_1 < d$ player 2 can inflict. Recall that we normalized productivity of an in-house solution to zero. The ex-ante utility of player 1 before her productivity is realized is $1 - d$, i.e. she receives the pay-off of 1 when her productivity is high enough to not be attacked, which, for uniformly distributed productivity, happens with probability $1 - d$. The ex-ante utility of player 2 before the productivity of player 1 is realized is $d(1 - q)$, i.e. the productivity

Table 1. Expected pay-off matrix for the crowdsourcing game.

	C	*S*
C	$1/2 - (d - (d^2/2))q,\ 1/2 - (d - (d^2/2))q$	$1 - d,\ d(1 - q)1 - q,\ 0$
S	$d(1 - q),\ 1 - d$	1/2, 1/2

of player 1 is low enough to be overtaken after an attack (which happens with probability d), so player 2 attacks and receives the reward minus the cost: $1 - q$. Note that player 2 attacks for any cost of an attack $q \in [0, 1)$: attacking brings a positive utility while not attacking results in zero utility. The case when player 2 crowdsources, and player 1 does not is symmetric. We summarize this in the *CS* and *SC* cells of the pay-off matrix in table 1.

The most interesting case is when both players crowdsource. Let p_1 and p_2 denote their productivities, which are known before they decide on attacking. Consider the case when $p_1 > p_2$. The other case is symmetric.

If the difference in productivities $p_1 - d > p_2$ of the players is so large that the attack by player 2 does not let her reach player 1, then attacking does not change the outcome and neither player attacks. In this case, player 1 receives the utility of 1, whereas player 2 receives the utility of 0.

3.1. Crowdsourcing and attacking

We analyse the game when both players crowdsource and a unilateral attack by the weak player (i.e. player 2) will bring her ahead of the strong player (i.e. player 1). This is the case when $p_1 - d < p_2$. In this case, player 2 would like to attack if player 1 does not. At the same time, player 1 would like to attack in order to keep its lead only if player 2 is attacking. Consequently, there is no pure strategy equilibrium of this game. The pay-off matrix showing utilities for each player appears in the following table.

	A	*N*
A	$1 - q,\ -q$	$1 - q,\ 0$
N	$0,\ 1 - q$	$1,\ 0$

The game possesses a unique mixed equilibrium. Let λ_1,λ_2 denote the probabilities that, respectively, players 1 and 2 attack. For player 1, attacking results in the expected utility of

$$\lambda_2(1 - q) + (1 - \lambda_2)(1 - q) = 1 - q, \tag{3.1}$$

whereas non-attacking gives the expected utility of

$$\lambda_2 \cdot 0 + 1 \cdot (1 - \lambda_2) = 1 - \lambda_2. \tag{3.2}$$

In a mixed equilibrium, a player's expected utility from choosing either action must be the same. This is satisfied for player 1 when

$$1 - q = 1 - \lambda_2, \tag{3.3}$$

that is for

$$\lambda_2 = q. \tag{3.4}$$

Similarly, for player 2, the expected utility from attacking is

$$\lambda_1(-q) + (1 - \lambda_1)(1 - q) = 1 - q - \lambda_1. \tag{3.5}$$

The expected utility from not attacking is 0, yielding the equilibrium condition

$$\lambda_1 = 1 - q. \tag{3.6}$$

Thus, in the mixed equilibrium, player 1 attacks with probability $\lambda_1 = 1 - q$, and player 2 attacks with probability $\lambda_2 = q$ leading to the following observation.

Finding 1. *The higher the cost, the more likely the weak player* (i.e. player 2) will attack and the less likely the strong player will attack.

This behaviour contradicts the intuition that making attacks more costly helps prevent them, that is, the costlier it is to attack, the less either player should attack. We explain why this does not hold by looking into the reasons for players' attacks. Player 1 attacks only to counteract a possible attack of player 2. Player 2 attacks in the hope that the attack is not counteracted by player 1 allowing player 2 to get ahead. Crucially, the incentive of player 1 to attack *increases* in the likelihood that player 2 attacks, whereas the incentive of player 2 to attack *decreases* in the likelihood that player 1 attacks. This relationship becomes clear when we observe that

$$\lambda_1 = 1 - \lambda_2.$$

The cost of attacks *increases* the likelihood that player 2 attacks, because player 1 attacks less when the cost is higher. Whenever both players attack, player 1 wins with certainty, and player 2 would have preferred to avoid the useless and costly attack. Thus, the less player 1 is likely to attack (i.e. the higher the cost), the more eager player 2 is to attack. For example, when attacking is free, player 1 always attacks and always wins obtaining the utility of 1. When the cost is the same as the prize, $q = 1$, player 2 always attacks taking the prize away from player 1 who does not attack resulting in zero utility for both. Formally, player 2 wins when he attacks and player 1 does not, which occurs with probability

$$\lambda_2(1 - \lambda_1) = q(1 - (1 - q)) = q^2.$$

Consequently, player 1 wins with probability

$$1 - q^2.$$

The higher the cost of an attack, the more likely the weak player is to win!

Finding 2. *Instead of protecting the better crowdsourcing strategy, a higher cost of an attack increases the probability of the weaker crowdsourcing strategy winning.*

This is a striking observation. Making attacks more costly should help the society by ensuring the stronger strategy wins. This does not occur owing to a higher likelihood that the weak player attacks unilaterally, which would lead her to victory. Furthermore, the players spend more resources on attacking each

other: making attacks more costly results in a lower total utility of the players. To show this, we first note that the expected number of attacks does not change with the cost of an attack. Raising the cost results in a lower probability of attack by the strong player and in a higher probability of attack by the weak player: the weak player attacks with probability q and the strong player attacks with probability $1-q$.

Finding 3. *The expected number of attacks*[1] *is one, regardless of the cost of an attack* q.

A direct consequence is the following:

Finding 4. *Increasing the cost of an attack decreases* (*not increases*) *the total utility of the players*.

We now look at the competition from the point of view of what is socially optimal. Benefit to the society is proportional to the final quality of the solution as well as to the total utility of the players. We intentionally leave the exact definition of *social welfare* open. Combining findings 2 and 4, we conclude that

Finding 5. *Increasing the cost of an attack results in lower social welfare*.

This finding suggests that increasing the cost of an attack is damaging rather than helping. Setting the costs to zero would be optimal from the social welfare point of view: the strong player would attack and win with probability 1, the weak player would not attack, and the cost of attacking would be zero. Without an attack of the weak player, the final solution—that of the strong player—would be of the highest possible quality.

Finding 6. *Free attacks maximize social welfare*.

3.1.1. Uncertain attacks

We assumed that an attack inflicts a known damage of d. In this section, we extend the results to uncertain attacks. Let s denote the probability that an attack is successful at inflicting damage d. Otherwise, no damage is inflicted. Setting $s=1$ results in the game discussed above. At the end of the section, we interpret the results when s is proportional to the difference in the productivities of the players.

The pay-off matrix given the probability of success s is

	A	N
A	$1-q-(s-s^2), (s-s^2)-q$	$1-q, 0$
N	$1-s, s-q$	1, 0

When both players attack, the weak player wins if his attack is successful, and the attack of the strong player is unsuccessful. This occurs with probability $s(1-s)=s-s^2$. The utility of the strong player is

$$(1-(s-s^2))-q=(1-q)-(s-s^2).$$

Other entries in the pay-off matrix are computed in a similar manner.

We derive the equilibria of this generalized game. As before, for player 1, attacking results in the expected utility of

$$\begin{aligned}&\lambda_2(1-q-s+s^2)+(1-\lambda_2)(1-q)\\&\quad=1-q-\lambda_2 s+\lambda_2 s^2,\end{aligned}\tag{3.7}$$

whereas non-attacking gives the expected utility of

$$\lambda_2(1-s)+(1-\lambda_2)=1-\lambda_2 s.\tag{3.8}$$

In a mixed equilibrium, a player's expected utility from choosing either action must be the same. This is satisfied for player 1 when

$$1-q-\lambda_2 s+\lambda_2 s^2=1-\lambda_2 s,\tag{3.9}$$

that is for

$$\lambda_2=\frac{q}{s^2}.\tag{3.10}$$

For player 2, the expected utility from attacking is

$$\begin{aligned}&\lambda_1(s-s^2-q)+(1-\lambda_1)(s-q)=-\lambda_1 q+\lambda_1 s-\lambda_1 s^2\\&+s-q-\lambda_1 s+\lambda_1 q=-\lambda_1 s^2+s-q.\end{aligned}$$

The expected utility from not attacking is 0, yielding the equilibrium condition

$$\lambda_1=\frac{s-q}{s^2}=\frac{1}{s}-\frac{q}{s^2}=\frac{1}{s}-\lambda_2.$$

Our main conclusion regarding uncertain attacks is that the mixed equilibrium has the same structure as with certain attacks, and the lessons learned for certain attacks continue to hold. Specifically, the structure of the mixed equilibrium in both cases is $\lambda_1=\text{constant}\ -\lambda_2$. Interestingly, a higher probability of success makes, the weak player attack less often.[2] This finding is similar to finding 2: making attacks more costly or less certain does not help the strong player.

The expected number of attacks[3] is $1/s$.

Finding 7. *The lower the probability of the success of an attack, the more attacks occur*.

The equilibrium utility of the weak player is zero, whereas from equations (3.9) and (3.10), the equilibrium utility of the strong player is $1-(q/s)$. Therefore, the *more* effective the attacks are, the *higher* the utility of the strong player. This is a counterintuitive result similar to finding 5.

Finding 8. *The equilibrium utility of the strong player* (*and the total utility of the players*) *increases in* s.

The mixed equilibria derived above holds only for $0\le\lambda_1,\lambda_2\le 1$, i.e. when

$$\begin{aligned}q&\le s^2\\q&\ge s-s^2.\end{aligned}$$

This region is denoted by 'mixed' in figure 1. When these conditions do not hold, there is a unique pure strategy equilibrium. Specifically, for $q>s$, not attacking is the unique equilibrium for both players. For $q<s$ and $q>s^2$, the strong player does not attack, whereas the weak player does. For $q<s^2$ and $q<s-s^2$, both players attack.

In more detail, if the cost of an attack is above s, then no attacks occur—the best possible outcome from the social welfare point of view. Thus, higher q is beneficial, but only if it is above s. If q falls below s but remains above s^2, an undesirable equilibrium arises: the weak player attacks unilaterally, resulting in a high chance of the weak player winning. For relatively low values of q, both players choose to attack (see the AA region in figure 1). Observe, the case when $s=1$.

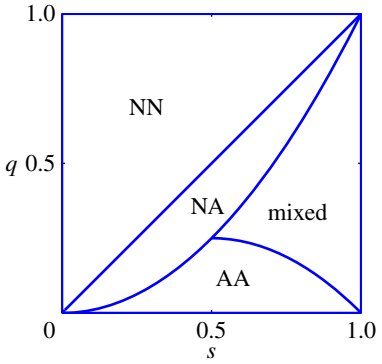

Figure 1. Equilibrium strategies in the second-stage game when both players crowdsource and $p_1 - d < p_2$. (Online version in colour.)

As we discussed in §3.1, for certain attacks, higher values of q do not increase social welfare (findings 2 and 5).

We characterized the equilibria from all values of s. We now explain how these results can be interpreted when the success of an attack is proportional to the distance between the players. An attack is successful from the weak player's point of view if it allows the weak player to overtake the strong player who does not attack. The closer the current solution qualities of the weak and the strong players, the easier it is for the weak player's to get ahead of the strong one. To model this, we set $s = 1-(p_1 - p_2/d)$ which denotes the probability that a unilateral attack takes the weak player ahead of the strong one. Note that s linearly increases from zero to one as the distance in productivities decreases from d to zero.

Knowing that the weak player may attack successfully, the strong player may want to attack to reduce the chances that the weak player gets ahead of the strong player. We use the same parameter s to denote the probability that the attack of the strong player is successful, i.e. that the strong player remains ahead of the weak player, thanks to its own attack.

The reader may find it counterintuitive that the success of the strong player's attack is inversely proportional to how much better his solution is relative to the weak player. However, the success here is defined as setting back the strategy of the weak player to prevent an otherwise successful attack of the weak player. It is easier to attack a solution that is of high quality. Indeed, at the extreme case of zero quality solution, an attack is not possible at all. The more the weak player has achieved in its solution, the more the strong player can take away from him. The parameter $s = 1-(p_1 - p_2/d)$ is consistent with this relationship.

The model where attacks are successful with probability s allows insight to be gained into a situation where a player does not know the exact productivity of the opponent. The probability s can be viewed as a player's estimate that an attack is going to let her overtake the opponent whose productivity she does not know.

The analysis and interpretations of this section point to the robustness of our main results to the assumptions of observability of the opponent's productivity and known damage of the attack.

3.2. To crowdsource or not

In §3.1, we computed expected utilities in the second-stage game when both players crowdsource and $p_1 - d < p_2$. We now take a step back and compute expected utilities when both players crowdsource, but before their productivities become known. To avoid confusion, we refer to these utilities as *ex-ante* utilities. We focus on the case of $s = 1$, i.e. the case where the mixed equilibrium is played for all values of q (figure 1). The ex-ante utility of player 1 (and symmetrically of player 2) is

$$u_1 = \Pr(P_2 < P_1 < P_2 + d)(1 - q) + \Pr(P_2 + d < P_1).$$

The first term corresponds to the utility of player 1 in the mixed equilibrium of the game described above, and the second term corresponds to player 1 winning with certainty when player 2 cannot reach her even after attacking. Under the assumption that P_1 and P_2 are uniformly distributed between zero and one, we derive

$$\begin{aligned}
\Pr(P_2<P_1<P_2+d) &= \int_0^1\int_0^1 1_{\{p_2<p_1<p_2+d\}}dp_2dp_1\\
&= \int_0^1\int_{p_2}^{\min(p_2+d,1)} 1dp_1dp_2\\
&= \int_0^1 (\min(p_2+d,1)-p_2)dp_2\\
&= \int_0^{1-d} d\,dp_2+\int_{1-d}^1 (1-p_2)dp_2\\
&= (1-d)d+\left(1-\frac{1}{2}\right)-\left(1-d-\frac{(1-d)^2}{2}\right)=d-\frac{d^2}{2}
\end{aligned}$$

$$\begin{aligned}
Pr(P_2+d<P_1) &= \int_0^1\int_0^1 1_{\{p_2+d<p_1\}}1dp_2dp_1\\
&= \int_d^1\int_0^{p_1-d} 1dp_2dp_1=\int_d^1 (p_1-d)dp_1\\
&= \frac{1}{2}-d-\left(\frac{d^2}{2}-d^2\right)=\frac{d^2}{2}-d+\frac{1}{2}
\end{aligned}$$

The ex-ante utility of either player is

$$\begin{aligned}
u_1 = u_2 &= \left(d-\frac{d^2}{2}\right)(1-q)+\left(\frac{d^2}{2}-d+\frac{1}{2}\right)1\\
&= \frac{1}{2}-\left(d-\frac{d^2}{2}\right)q.
\end{aligned}$$

The utility of the players decreases in both q and d. At the extreme case when both q and d are 1, the utility is zero. Whenever either of the parameters is at its minimum value of zero, the utility is at its maximum value of 1/2. Observe that the ex-ante utility decreases in the cost of an attack. This is a consequence of the number of attacks being independent of q when both $p_1 - d < p_2$ as we noted in finding 3 and a related finding 4.

In the Prisoner's Dilemma [50–53], the unique equilibrium for both players is to choose the action that hurts the other player, i.e. to defect. In the resulting equilibrium, both players are hurt. A similar situation (although in mixed strategies) arises in the crowdsourcing game. Both players choose to attack (in equilibrium, the expected number of attacks is 1), incurring unproductive costs. When both players attack, the outcome is the same as when neither player attacks except that each player incurs the cost of an attack. When only the strong player attacks, the outcome does not change, but the player incurs the cost of an attack. When only the weak player attacks, the outcome changes for a less efficient outcome (the weaker player wins), and the weak player incurs

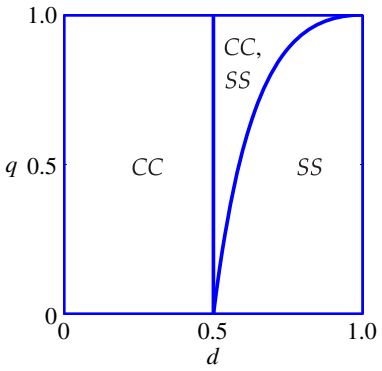

Figure 2. Equilibrium strategies. (Online version in colour.)

the cost of an attack. Only when neither player attacks, there is no loss to social welfare.

Having analysed the second-stage games, we can now describe the entire pay-off matrix for the game played in the first stage. Note that in the first stage, the players are unaware of their productivities, should they choose crowdsourcing. Thus, the pay-off matrix contains their ex-ante utilities, depicted in table 1.

We ask the question of how q and d affect the equilibria. For any choice of parameters, only *CC* and *SS* can be pure strategy equilibria. Higher damage from an attack and low cost of attacking corresponds to the crowdsourcing strategy being more risky. Indeed, as we detail below, for $d > 1/2$ and $q < ((2d-1)/d_2)$, *SS* is the only equilibrium strategy (figure 2). When the damage is low ($d < 1/2$), crowdsourcing is the only equilibrium strategy, regardless of the cost of attack. For the remaining range of parameters $d \geq 1/2$ and $q \geq ((2d-1)/d_2)$, both *CC* and *SS* are equilibrium strategies.

Crowdsourcing *CC* is a unique equilibrium when the damage inflicted by an attack is low $d < 1/2$. Indeed, when both firms use closed strategies, deviating to a crowdsourcing strategy provides a higher pay-off ($1-d > 1/2$); the productivity of a crowdsourcing strategy will make the crowdsourcing firm outside the reach of the in-house competitor even after being attacked. Crowdsourcing for one firm and not crowdsourcing for the other is not stable, as the non-crowdsourcing firm is better off switching to crowdsourcing $1/2 - (d - (d^2/2))q > d(1-q)$. This holds regardless of the cost of an attack $0 < q < 1$.

For $d \geq 1/2$, *CC* remains an equilibrium only if the cost of an attack is high enough: $q \geq ((2d-1)/d_2)$. Intuitively, players crowdsource when attacking costs a lot (higher q) and is not effective (lower d).

How does the cost of an attack influence the likelihood of players to crowdsource? Intuitively, we would expect that high costs prevent attacks and make crowdsourcing more appealing. Our model provides reasoning against this intuition. Figure 2 reveals that q has a limited effect on the choice of the equilibrium strategy. Cost of attacking matters only when the level of damage is high ($d \geq 1/2$). In this case, high costs enable *CC* to be an equilibrium; however, *SS* still remains an equilibrium.

The damage from an attack has a strong effect on the equilibrium. Low damage corresponds to fewer attacks on a crowdsourcing firm: when both players crowdsource, the likelihood that the strong player is within reach of the weak player is proportional to the level of damage, $\Pr(P_2 < P_1 < P_2 + d)$. Crowdsourcing is the unique equilibrium strategy for low levels of damage ($d < 1/2$). This leads to the conclusion that competitions where the maximal damage on a crowdsourcing opponent is limited are more likely to promote crowdsourcing than competitions where the designer attempts to make attacking more costly.

4. Discussion

Our results bear resemblance to the Prisoner's Dilemma but paint an even starker picture. When both players crowdsource (i.e. choose a more efficient way of performing the task) and are close to each other in terms of solution quality (specifically, within the damage inflicted by attacking, as was the case in the DARPA Shredder Challenge), the expected number of attacks is one, regardless of the cost of an attack. Increasing the cost of an attack offers no deterrence. Therefore, under our basic model, malicious behaviour is the expected behaviour, not the anomaly. Given this result and the examples of malicious behaviour in competitions, more emphasis should be given to the issue. There has been significant academic interest towards filtering misinformation; however, models of malicious behaviour in crowdsourcing scenarios have been absent until this work.

Our model applies more generally than malicious behaviour in competitive crowdsourcing. The same strategic considerations arise when instead of attacking the opponent a player can improve her own solution by d by investing q. The salient feature is that there is only one winner in the competition. Our results are for two competitors, however, they provide insight into strategies of two top competitors in a multi-competitor competition.

The finding that raising costs of attack is harmful for the players is striking and warrants further empirical investigation. Our model predicts that higher costs of attacks lead to more attacks by the weak player resulting in a higher probability that the weak player would win. Furthermore, the expected number of attacks remains the same, resulting in higher costs incurred by the players orchestrating attacks. We also find that making attacks less likely to succeed does not help prevent them (within the mixed equilibrium). On the contrary, a lower probability that an attack is successful results in a higher expected number of attacks. Confirming these findings in a laboratory or field experiment is a direction for future work.

We made a number of modelling choices: two firms, perfect observability of productivities, risk-neutrality of the players, and provided the analysis for a uniform distribution of productivities. These choices were guided by simplicity and the goal to isolate the factors relevant to the trade-off between higher productivity of open strategies and higher vulnerability. Owing to the simplicity of our model, we believe that our results capture the fundamental features of the trade-off between the productivity/vulnerability. Indeed, we showed that the results continue to hold when we allow attacks to be uncertain. Having said this, we acknowledge that our conclusions rely on the assumptions made, and that their broader applicability must be confirmed on a case-by-case basis.

We considered the relative performance of the firms while ignoring the absolute quality of the solution. This explains why the highest expected social welfare is obtained when neither firm crowdsources, and therefore, neither firm attacks

(see the pay-off of the (*S*,*S*) strategy in table 1). This is suitable for many competition settings where only relative performance is important (such as the aforementioned DARPA Network Challenge, DARPA Shredder Challenge or Tag Challenge). Requiring a certain minimum solution quality and modelling the cost of effort are interesting extensions for future work. For example, one could model effort as time required to find a solution, with the time being inversely related to productivity (e.g. $t = 1 - p$). Competitions may also mitigate aggression by using reward mechanisms where the reward received by the winner depends on the global progress of all teams—linking crowdsourcing games to public good games [54–57].

Repeated encounters in crowdsourcing competitions may provide opportunities for the emergence of a richer set of socially desirable strategies as in the iterated Prisoners Dilemma [58–60]. It would also be interesting to study how the presence of more than two players affects the behaviour displayed.

Our results emphasize that despite crowdsourcing being a more efficient way of accomplishing many tasks, it also a less secure approach. In scenarios of 'competitive' crowdsourcing, where there is an inherent desire to hurt the opponent, attacks on crowdsourcing strategies are essentially unavoidable. We expect these surprising results derived in our stylized model to hold in a variety of more complicated scenarios that exhibit the fundamental tension between openness, efficiency and vulnerability.

Acknowledgements. We thank Andres Abeliuk, Galen Pickard, Iyad Rahwan, Moshe Hoffman and Toby Walsh for discussions.

Funding statement. M.C. and P.V.H. acknowledge support from the National Science Foundation under grant no. 0905645, from DARPA/Lockheed Martin Guard Dog Programme under PO no. 4100149822, from the Army Research Office under grant no. W911NF-11-1-0363 and the Australian Government as represented by DBCDE and ARC through the ICT Centre of Excellence programme. N.R.J. and V.N. acknowledge funding from the UK Research Council for project ORCHID, grant no. EP/I011587/1.

End notes

[1]Note that in this one-shot game the number of attacks is limited to one per player. Therefore, the total number of attacks ranges from zero to two.

[2]A higher probability of attack makes the strong attack more often for $s > 2q$ and less often for $s < 2q$. In the former case, a higher s decreases the deterrence power of λ_1 on the weak player, and a higher probability of attack is needed. In the latter case, the reverse is true.

[3]The highest possible number of attacks is one per player, or two in total. Indeed, the relevant range of the parameter s is above $1/2$ as we will see in figure 1.

References

1. Howe J. 2006 The rise of crowdsourcing. *Wired Mag.* **14**, 1–4.
2. Hand E *et al*. 2010 Citizen science: people power. *Nature* **466**, 685. (doi:10.1038/466685a)
3. Von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. 2008 reCAPTCHA: human-based character recognition via web security measures. *Science* **321**, 1465–1468. (doi:10.1126/science.1160379)
4. Von Ahn L. 2006 Games with a purpose. *Computer* **39**, 92–94. (doi:10.1109/MC.2006.196)
5. Horowitz D, Kamvar SD. 2010 The anatomy of a large-scale social search engine. In *Proc. 19th ACM Int. Conf. on World Wide Web*, pp. 431–440. New York, NY: ACM. (doi:10.1145/1772690.1772735)
6. Huberman BA, Romero DM, Wu F. 2009 Crowdsourcing, attention and productivity. *J. Inf. Sci.* **35**, 758–765. (doi:10.1177/0165551509346786)
7. Cooper S *et al*. 2010 Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760. (doi:10.1038/nature09304)
8. Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A, Pentland A. 2011 Time-critical social mobilization. *Science* **334**, 509–512. (doi:10.1126/science.1205869)
9. Mason W, Suri S. 2012 Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* **44**, 1–23. (doi:10.3758/s13428-011-0124-6)
10. Rahwan I, D'Souza S, Rutherford A, Naroditskiy V, McInerney J, Venanzi M, Jennings NR, Cebrian M. 2013 Global manhunt pushes the limits of social mobilization. *Computer* **46**, 68–75. (doi:10.1109/MC.2012.295)
11. Barrington L, Turnbull D, Lanckriet G. 2012 Game-powered machine learning. *Proc. Natl Acad. Sci. USA* **109**, 6411–6416. (doi:10.1073/pnas.1014748109)
12. Hellerstein JM, Tennenhouse DL. 2011 Searching for Jim Gray: a technical overview. *Commun. ACM* **54**, 77–87. (doi:10.1145/1965724.1965744)
13. Alstott J, Madnick S, Velu C. 2014 Homophily and the speed of social mobilization: the effect of acquired and ascribed traits. *PLoS ONE* **9**, e95140. (doi:10.1371/journal.pone.0095140)
14. Zhang H, Horvitz E, Chen Y, Parkes DC. 2012 Task routing for prediction tasks. In *Proc. 11th Int. Conf. Autonomous Agents and Multiagent Systems*, vol. 2, pp. 889–896. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
15. Tang JC, Cebrian M, Giacobe NA, Kim H-W, Kim T, Wickert DB. 2011 Reflecting on the DARPA Red Balloon Challenge. *Commun. ACM* **54**, 78–85. (doi:10.1145/1924421.1924441)
16. Watts D, Cebrian M, Elliot M. 2013 Dynamics of social media. In *Public response to alerts and warnings using social media: report of a workshop on current knowledge and research gaps* (ed. National Research Council), pp. 22–33. Washington, DC: The National Academies Press.
17. Palmer C. 2011 UC San Diego team's effort in DARPA's Shredder Challenge derailed by sabotage. California Institute for Telecommunications and Information Technology Press Release. See http://www.calit2.net/newsroom/article.php?id=1938.
18. Kleinberg J, Raghavan P. 2005 Query incentive networks. In *46th Annual IEEE Symp. on Foundations of Computer Science, 2005 (FOCS 2005)*, pp. 132–141. Piscataway, NJ: IEEE.
19. Kittur A, Nickerson JV, Bernstein M, Gerber E, Shaw A, Zimmerman J, Lease M, Horton J. 2013 The future of crowd work. In *Proc. 2013 Conf. on Computer Supported Cooperative Work*, pp. 1301–1318. New York, NY: ACM.
20. Zhang H, Horvitz E, Miller RC, Parkes DC. 2011 Crowdsourcing general computation. In *ACM CHI 2011 Workshop on Crowdsourcing and Human Computation*. New York, NY: ACM.
21. Mao A, Parkes DC, Procaccia AD, Zhang H. 2011 Human computation and multiagent systems: an algorithmic perspective. In *Proc. Twenty-Fifth AAAI Conf. on Artificial Intelligence*. Menlo Park, CA: AAAI.
22. Karger DR, Oh S, Shah D. 2011 Budget-optimal task allocation for reliable crowdsourcing systems. See http://www.arxiv.org/abs/1110.3564.
23. Bernstein MS, Karger DR, Miller RC, Brandt J. 2012 Analytic methods for optimizing realtime crowdsourcing. See http://www.arxiv.org/abs/1204.2995.
24. Ipeirotis P. 2011 Crowdsourcing using mechanical turk: quality management and scalability. In *Proc. 8th Int. Workshop on Information Integration on the Web: in conjunction with WWW 2011*, p. 1. New York, NY: ACM.
25. Kamar E, Horvitz E. 2012 Incentives for truthful reporting in crowd-sourcing. In *Proc. 11th Int. Conf. on Autonomous Agents and Multiagent Systems*, pp. 1329–1330. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
26. Ipeirotis PG, Paritosh PK. 2011 Managing crowdsourced human computation: a tutorial. In *Proc. 20th Int. Conf. Companion on World Wide Web*, pp. 287–288. New York, NY: ACM.

27. Lorenz J, Rauhut H, Schweitzer F, Helbing D. 2011 How social influence can undermine the wisdom of crowd effect. *Proc. Natl Acad. Sci. USA* **108**, 9020–9025. (doi:10.1073/pnas.1008636108)
28. Pfeiffer T, Gao XA, Mao A, Chen Y, Rand DG. 2012 Adaptive polling for information aggregation. In *Twenty-Sixth AAAI Conf. on Artificial Intelligence*. Menlo Park, CA: AAAI.
29. Mason W, Watts DJ. 2012 Collaborative learning in networks. *Proc. Natl Acad. Sci. USA* **109**, 764–769. (doi:10.1073/pnas.1110069108)
30. Naroditskiy V, Rahwan I, Cebrian M, Jennings NR. 2012 Verification in referral-based crowdsourcing. *PLoS ONE* **7**, e45924. (doi:10.1371/journal.pone.0045924)
31. Cebrian M, Coviello L, Vattani A, Voulgaris P. 2012 Finding red balloons with split contracts: robustness to individuals' selfishness. In *Proc. 44th Symp. on Theory of Computing*, pp. 775–788. New York, NY: ACM.
32. Tran-Thanh L, Venanzi M, Rogers A, Jennings NR. 2013 Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *12th Int. Conf. on Autonomous Agents and Multi-Agent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
33. Venanzi M, Rogers A, Jennings NR. 2013 Trust-based fusion of untrustworthy information in crowdsourcing applications. In *12th Int. Conf. on Autonomous Agents and Multi-Agent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
34. Babaioff M, Dobzinski S, Oren S, Zohar A. 2012 On bitcoin and red balloons. In *Proc. 13th ACM Conf. on Electronic Commerce*, pp. 56–73. New York, NY: ACM.
35. Rutherford A, Cebrian M, Dsouza S, Moro E, Pentland A, Rahwan I. 2013 Limits of social mobilization. *Proc. Natl Acad. Sci. USA* **110**, 6281–6286. (doi:10.1073/pnas.1216338110)
36. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. 2010 Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688. (doi:10.1126/science.1193147)
37. Pentland A. 2012 The new science of building great teams. *Harvard Bus. Rev.* **90**, 60–69.
38. Drucker FA, Fleischer LK. 2012 Simpler Sybil-proof mechanisms for multi-level marketing. In *Proc. 13th ACM Conf. on Electronic commerce*, pp. 441–458. New York, NY: ACM.
39. Anderson A, Huttenlocher D, Kleinberg J, Leskovec J. 2013 Steering user behavior with badges. In *Proc. ACM Int. Conf. on World Wide Web*. New York, NY: ACM.
40. Nath S, Dayama P, Garg D, Narahari Y, Zou J. 2012 Threats and trade-offs in resource critical crowdsourcing tasks over networks. In *Twenty-Sixth AAAI Conf. on Artificial Intelligence*. Menlo Park, CA: AAAI.
41. Chitnis R, Hajiaghayi MT, Katz J, Mukherjee K. 2012 A game-theoretic model motivated by the DARPA network challenge. See http://www.arxiv.org/abs/1204.6552.
42. Centola D, Macy M. 2007 Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734. (doi:10.1086/521848)
43. Liben-Nowell D, Kleinberg J. 2008 Tracing information flow on a global scale using internet chain-letter data. *Proc. Natl Acad. Sci. USA* **105**, 4633–4638. (doi:10.1073/pnas.0708471105)
44. Golub B, Jackson MO. 2010 Using selection bias to explain the observed structure of internet diffusions. *Proc. Natl Acad. Sci. USA* **107**, 10 833–10 836. (doi:10.1073/pnas.1000814107)
45. Bakshy E, Hofman JM, Mason WA, Watts DJ. 2011 Everyone's an influencer: quantifying influence on twitter. In *Proc. fourth ACM Int. Conf. on Web Search and Data Mining*, pp. 65–74. New York, NY: ACM.
46. Iribarren JL, Moro E. 2009 Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.* **103**, 038702. (doi:10.1103/PhysRevLett.103.038702)
47. Goyal S, Kearns M. 2012 Competitive contagion in networks. In *Proc. 44th symposium on Theory of Computing*, pp. 759–774. New York, NY: ACM.
48. Toole JL, Cha M, González MC. 2012 Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS ONE* **7**, e29528. (doi:10.1371/journal.pone.0029528)
49. Ugander J, Backstrom L, Marlow C, Kleinberg J. 2012 Structural diversity in social contagion. *Proc. Natl Acad. Sci. USA* **109**, 5962–5966. (doi:10.1073/pnas.1116502109)
50. Rapoport A, Chammah AM. 1965 *Prisioner's Dilemma: a study in conflict and cooperation*. Ann Arbor, MI: University of Michigan Press.
51. Trivers RL. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
52. Aumann RJ. 1981 Survey of repeated games. In *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern*, vol. 4 of Gesellschaft, Recht, Wirtschaft, Wissenschaftsverlag (ed. V Bohm), pp. 11–42. Mannheim, Germany: Bibliographisches Institut.
53. Axelrod R, Hamilton WD. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)
54. Fehr E, Schmidt KM. 1999 A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868. (doi:10.1162/003355399556151)
55. Rand DG, Dreber A, Ellingsen T, Fudenberg D, Nowak MA. 2009 Positive interactions promote public cooperation. *Science* **325**, 1272–1275. (doi:10.1126/science.1177418)
56. Fowler JH. 2005 Altruistic punishment and the origin of cooperation. *Proc. Natl Acad. Sci. USA* **102**, 7047–7049. (doi:10.1073/pnas.0500938102)
57. Fowler JH, Christakis NA. 2010 Cooperative behavior cascades in human social networks. *Proc. Natl Acad. Sci. USA* **107**, 5334–5338. (doi:10.1073/pnas.0913149107)
58. Fudenberg D, Maskin E. 1986 The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533–554. (doi:10.2307/1911307)
59. Fundenberg D, Maskin E. 1990 Evolution and cooperation in noisy repeated games. *Am. Econ. Rev.* **80**, 274–279.
60. Nowak M, Sigmund K. 1993 A strategy of win–stay, lose–shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58. (doi:10.1038/364056a0)