Strategic Behavior in Contests with Ability Heterogeneous Agents: Evidence from Field Data*

Christoph Riedl[†‡] Tom Grad[§] Christopher Lettl[‡]

June 23, 2019

Abstract

We investigate two forms of strategic behaviors of agents with heterogeneous ability in large contests: sabotage and self-promotion. We test predictions from a theoretical model in a large dataset of more than 38 million observations of peer-ratings by around 75,000 individuals from 511 real world contests over a ten year period. We find that strategic behaviors might influence the outcome of up to 12% of all contests and 25% of close contests. We find the use of strategic behavior is ability-dependent: While self-promotion is the dominant form of strategic behavior of low ability agents, high ability agents are both culprits and targets of sabotage. We test additional predictions concerning the prevalence of strategic behavior in contests of different sizes. We find more sabotage in larger contests, but more self-promotion in smaller contests. We leverage a natural experiment to rule out that self-promotion is the result of overconfidence and a second natural experiment to support that sabotage ratings are strategic rather than a result of endogenous entry into contests.
JEL Codes: J22, J3, J24, D82.

Keywords: *Contests, tournaments, sabotage, self-promotion, heterogeneous agents, diff-in-diff, natural experiments*

*We gratefully acknowledge helpful comments by Blair Davey, Imke Reimers, Ulrich Berger, Ben Greiner, and participants at Digital Innovation Workshop 2019 at Boston College. This research was funded in part by the National Science Foundation [Grant IIS-1514283].

[†]Northeastern University
[‡]Harvard University
[§]Vienna University of Economics and Business

1 Introduction

Despite the widespread use of contests to spur innovation, R&D, industry development, employee effort, academic progress, and performance in arts and sports (c.f., Vojnović, 2016), the causes and consequences of strategic behavior of contestants are much better understood in principle (e.g., Dixit, 1987; Lazear, 1989; Chen, 2003) than in practice (Carpenter et al., 2010). However, incentive designers are also aware of a severe potential drawback of contests: they encourage non-productive effort in the form of strategic behavior like sabotage and self-promotion (Lazear, 1989; Prendergast and Topel, 1993). Given actors tendency to hide and obscure strategic behavior, only a few empirical studies of strategic behavior in contests outside of artificial lab settings exist. The empirical setting of these studies is either in sports (e.g., Balafoutas et al., 2012; Deutscher et al., 2013; Garicano et al., 2005) or manufacturing (Drago and Garvey, 1998).[1]

In sharp contrast to the empirical settings of prior work, today contests are predominantly taking place in online environments. This has two important implications: On the one hand, the digitalization of contests allows a large number of contestants to participate as entry costs are low and individuals can participate anonymously (at least as far as other contestants are concerned). This openness of entry means that contests with a large number of contestants are the norm rather than the exception. However, it also makes these contests vulnerable to malicious behavior through sabotage (Naroditskiy et al., 2014; Stefanovitch et al., 2014), as has become apparent for example in large contests sponsored by the US federal government (Rahwan, 2014).

Low entry costs also make it attractive for amateurs to participate (Boudreau, 2018), a phenomenon which has been coined "the rise of the amateur" (Howe, 2008). Consequently, in today's online contests low ability amateurs compete with high ability contestants which means that contestants' heterogeneity in ability is high. As a result, the little empirical insight we have on strategic behavior in contests may not fully reflect the settings in which the majority contests are used today. Furthermore, we can now leverage digital trace data to observe and measure strategic behavior in clean, disaggregated, and unobtrusive ways[2] at the individual level. This is crucial as contestants typically try to hide their strategic behavior due to its association of being illegal or immoral (Charness and Levine, 2004). The ability to observe strategic behavior outside of artificial lab settings is important as *"sabotage in the lab is almost always diffuse and blunt"* (Carpenter et al., 2010, p. 504).

Our paper extends prior literature in at least two important ways. First, we study strategic behavior in a context that is representative for most of today's contests that is more natural

[1] Two important papers that investigate sabotage in lab experiments are Charness et al. (2014) and Harbring and Irlenbusch (2011). They find that the possibility of sabotage reduces effort, that competitors are willing to incur costs to sabotage and artificially increase their own performance, and that sabotage increases with higher prize spreads.

[2] While the behavior is recorded by digital platforms, it is still likely to be hidden from other contestants.

compared to laboratory studies. This is possible given digital trace data that allows us to examine this behavior with clean, disaggregated, and unobtrusive measures on the individual level. Our study therefore complements prior lab-based work on strategic behavior in contests. Second, and no less important, our theoretical model and empirical setting enable us to study strategic behavior in contests more comprehensively from a theoretical perspective. Theory predicts that strategic behaviors are contingent on the ability of contestants (e.g., Schotter and Weigelt, 1992; Dixit, 1987; Chen, 2003). Our setting gives us the opportunity to study contestants' heterogeneity in ability in a fine-grained manner and how this relates to both the culprits and targets of strategic behavior. As a consequence we expand and test existing theoretical models by studying how heterogeneity in ability affects the decision to sabotage and who to target. Prior empirical studies have considered heterogeneity of contestants (Harbring et al., 2007; Vandegrift and Yavas, 2010; Balafoutas et al., 2012; Deutscher et al., 2013, see, e.g.,) but the results of these studies have been inconclusive. While there is some convergence in the findings of these studies that high ability contestants are more likely to be the targets of sabotage compared to low ability contestants, there is divergence with respect to the culprits: while Balafoutas et al. (2012), Carpenter et al. (2010), and Deutscher et al. (2013) find that low ability contestants are more likely to be the culprits of sabotage, Harbring et al. (2007) find evidence for "a battle of the giants", i.e., that sabotage is concentrated among high ability contestants. Our paper also develops and tests predictions concerning the prevalence of strategic behavior in contests of different sizes that have not been studied in detail before.

Our setting allows us to study two forms of strategic behavior simultaneously, namely sabotage and self-promotion,[3] and how contestants' heterogeneity in ability influences different uses of these two forms of strategic behavior. While some prior studies have exclusively looked at sabotage (Harbring and Irlenbusch, 2005, 2011; Harbring et al., 2007; Falk et al., 2008; Carpenter et al., 2010), other studies have focused entirely on self-promotion (Milgrom, 1988; Milgrom and Roberts, 1988a; Schweitzer et al., 2004; Edelman and Larkin, 2015). However, considerations of both forms of strategic behavior in combination are rare.[4] Investigating sabotage and self-promotion jointly rather than in isolation is important as the boundaries between performing and evaluating are increasingly blurred (Edwards Mark R. and Ewen, 1996; Ghorpade, 2000). This is prevalent in many of today's innovation contests where contestants participate in both the creation and evaluation of new technologies, new products or other types of innovation (Boudreau et al., 2016c). It is this particular blurring of boundaries that induces additional opportunities for using strategic behaviors. We explore how contestants of different ability use these different forms of strategic behavior to extend theory when more than one form of strategic behavior is available. In conclusion, our study extends prior work by investigating heterogeneity

[3]Other terms used in the economic literature for self-promotion are "influencing activities" (see e.g., Milgrom, 1988) or "directly unproductive, profit-seeking activities" (Bhagwati, 1982). However, we prefer the term self-promotion as used in, e.g., O'Reilly and Wade (1993) as it reflects the self-directed nature of this activity opposed to the outwards-directed nature of sabotage.

[4]The only exception is Arbatskaya and Mialon (2010) who theoretically look at multi-activity contests but only in contests with two players, in which sabotage acts differently as its externalities are only apparent in contests with at least three players (c.f. Konrad, 2000).

in ability when two forms of strategic behavior are available, and how it varies across contests of different size.

Our empirical setting is a large online community which comprises all the features of contests in today's digitalized world as outlined above. A large number of contestants compete in weekly innovation contests and participate in the contests' peer-rating mechanism. Contestants and outsiders (e.g., designers who did not enter the contest in that week) can then rate the submissions, and the host organization chooses a winner among the highest rated submissions and awards a monetary award. To study strategic behavior we analyze the rating behavior of designers in a week in which they entered their own submission in a contest compared to their rating behavior in a week when they did not enter their own submission in the contest. Our data consists of an longitudinal panel of 74,525 individuals participating in 511 weekly contests—spanning a 10-year period—and casting 38 million ratings on 154,086 contest entries (submissions). We model the effects of participating in the same contest on rating behavior while controlling for both individual-level and submission-level differences. We leverage two natural experiments—an incentive change and a change to rules governing the rating mechanism—to rule out endogenous contest entry and overconfidence as alternative explanations to strategic behavior.

Contrary to the prediction of Konrad (2000) we find that strategic behaviors matter even in contests with many participants: the contest winner would change in 12% of contests, and 25% of close contests, if (potentially) strategic ratings were not allowed. We also find that self-promotion is prevalent while sabotage is only a minor phenomenon overall. However, when considering heterogeneity of contestants with respect to their ability level, we find sabotage to be prevalent and concentrated among the high ability contestants who are both the culprits and victims of sabotage. Furthermore, we find more sabotage in *larger* contests, but more self-promotion in *smaller* contests, which is consistent with our model predictions. We find that allowing self-promotion levels the playing field among agents of different ability levels, which reduces sabotage. While our findings are consistent with prior research on who is targeted by sabotage (e.g., Carpenter et al., 2010; Münster, 2007) our findings regarding culprits is in contrast to the findings of Carpenter et al. (2010) and Balafoutas et al. (2012) who find that especially low ability contestants are likely to engage in sabotage targeted towards high ability contestants. Results from a natural experiment provide complementary evidence from the field for the theoretical prediction of Lazear (1989) and the laboratory experimental findings of Harbring and Irlenbusch (2011), i.e. that sabotage activities increase with an increase in prize spread. In our natural experiment we find that a doubling in prize spread increases the probability of sabotage by 2.2 percentage points.

The remainder of the paper is organized as follows. First, we present a simple model of contestants' behavior that incorporates sabotage and self-promotion as well as contestants of heterogeneous ability from which we derive formal propositions and comparative statics to guide our empirical analysis. After that, we describe the empirical setting and identification strategy,

followed by the presentation of results of our empirical analyses and two natural experiments. Finally, we discuss our results and the implications of several mechanisms contest organizers might employ to combat self-promotion and sabotage such as lifting anonymity or forbidding ratings on one's own submission in the light of our model and empirical findings.

2 Theoretical Model

We develop predictions of strategic behavior in contest where contest winners are determined through a community and peer evaluation mechanism that allows both sabotage and self-promotion. We build on the simple and tractable model of one-shot tournaments developed in Konrad (2000) and Harbring and Irlenbusch (2011), which incorporates a single winner prize, sabotage, and competitors of heterogeneous abilities. We extend the model by allowing votes of neutral outsiders and promotion.

2.1 Model Setup

The contest consists of two types of agents: a set H of h high ability agents, and a set L of l low ability agents. We refer to these as high and low types, respectively. Furthermore, the contests involves a set N of n of outsiders. We assume that $h < l < n$, which should be satisfied in most realistic contest settings. Following Moldovanu and Sela (2001) and Boudreau et al. (2016b) we do not consider the effort choice of contestants directly but their choice of bid quality b_i based on their ability level a_i. Low and high types produce a contest submission of low and high quality, respectively, which they enter into the contest. Outsiders do not enter the contest and thus remain neutral. The submissions of all agents are rated by each outsider and by each high and low type agent on a bounded quality scale in $\mathbb{R}_{\geq 0}$ with $[r_{min}, r_{max}]$. Without loss of generality, we normalize the rating scale to the unit interval so that $r_{min} = 0$ and $r_{max} = 1$. We assume that the quality bid of low types' submissions is b_l, whereas the quality bid of high types' submissions is b_h with $0 < b_l < b_h < 1$. Agents assume they can change their utility through their ratings (i.e., they are the marginal voter).

Rating can be sincere (i.e., b_l or b_h, respectively), but an agent can also sabotage any other agent, rating their submission at $r_{min} = 0$; or promote them by rating their product at $r_{max} = 1$. Let $\Delta s_h = b_h$ denote the damage done by sabotaging a high type and $\Delta s_l = b_l$ the damage done by sabotaging a low type. Since $b_l < b_h$ it follows that $\Delta s_l < \Delta s_h$. Further, let $\Delta p_h = 1 - b_h$ denote the benefit gained from promoting a high type and $\Delta p_l = 1 - b_l$ the benefit gained from promoting a low type. Since $b_l < b_h$ it follows that $\Delta p_l > \Delta p_h$. Figure 1 illustrates the rating scheme.

We assume that while rating sincerely is costless, sabotaging another agent costs c_s and promoting any agent (also includes oneself) agent costs c_p. Costs associated with promotion and sabotage might arise from the costs of identifying targets to sabotage (Harbring et al.,

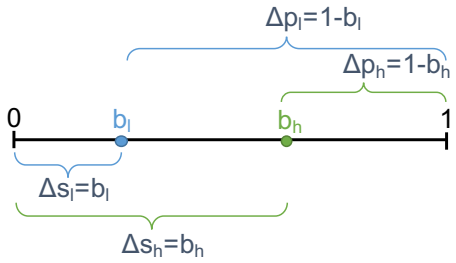


Figure 1: Rating interval.

2007; Münster, 2007), the moral costs associated with lying (Abeler et al., 2018; Atanasov and Dana, 2011; Gneezy et al., 2018) or violating social norms (Elster, 1989; Fehr and Fischbacher, 2004), and reputation costs (Engelmann and Fischbacher, 2009; Nowak and Sigmund, 2005). As a consequence all N outsiders rate contestants h and l sincerely. This implies that the size n of the set of outsiders N affects the effectiveness of any type of strategic behavior performed by the set of H and L agents. Effectively, n can be considered a modifier in our model that governs how effective strategic behavior is. Note also that n are not necessarily outsiders who never compete—they simply do not compete in the current contest. We will use this feature to identify strategic behavior in our empirical analysis where contestants do not enter every contest: they rate as neutral outsiders in some weeks and rate as competitors with stakes in the contest in others.

The rating given by agent i to agent j is denoted by v_{ij}. The value v_j of agent j's submission for $j \in L \cup H$ is the sum of all the realized ratings $v_j = \sum_{i \in N \cup L \cup H} v_{ij}$. Throughout the remainder, let v_i denote the value of some arbitrary high type (she), and v_k the value of some arbitrary low type (he). As an example, in the case of sincere rating by everyone (no promotion and no sabotage), the value of a high type $v_i = b_h(n + l + h)$ and that of a low type $v_k = b_l(n + l + h)$. It is important to note that we set up the model within the peer-rating context in online contests to better integrate theory and empirics. The basic premises and predictions of the model are, however, more general and apply to other contests with promotion and sabotage. The key assumption—a fixed rating scale and resulting limits to potential gain from promotion and damage conferred by sabotage—are, by no means, specific to contests in this setting.

While the probability to win is deterministic up to now, we assume that on top of this community evaluation a stochastic component reflecting idiosyncratic preferences of the contest organizer is added, such that the probability of winning the contest is strictly positive even for low types and strictly increasing in an agent's value (v_i and v_k, respectively). For simplicity, we assume that this results in a Tullock contest (Tullock, 1980) with the evaluations of submissions v_i as inputs. Each agent i's probability of winning the contest is $p_i = \frac{v_i}{\sum_{j \in L \cup H} v_j}$. For simplicity, let $S = \sum_{j \in L \cup H} v_j$ denote the total value of all contestants and we write the Tullock contest success function as $p_i = \frac{v_i}{S}$. The winner of the contest receives a prize M which is without loss of generality normalized to 1. Given the equilibrium values for productive effort and the resulting quality bid in the ordinary Tullock lottery contest, agents maximize their expected winning

probability minus their total costs of sabotage and cost of promotion. The utility function for agents strategic contest behavior can be then be written as:

$$E\Pi_i = Mp_i - c_s(\sum_j sab_{ij}) - c_p(\sum_j prom_{ij}), \tag{1}$$

where M is the winner prize, p_i is agent i probability of winning the prize, sab_{ij} ($prom_{ij}$) is 1 if agent i sabotages (promotes) agent j and 0 otherwise, and c_s and c_p are the cost of sabotage and promotion, respectively.

Formally a strategy of agent i is a list (r_{ij}) of rating behavior towards all agents $j \in L \cup H$, where $r_{ij} \in \{\text{sincere}, \text{sabotage}, \text{promote}\}$. Since sabotaging oneself and promoting any agent other than oneself are clearly strictly dominated and payoffs are symmetric with respect to actions towards different other agents of the same type, we only look at equivalence classes of strategies of the form $s_i = (hsab_i, lsab_i, sprom_i)$, where $hsab_i$ is the number of high types i sabotages, $lsab_i$ the number of low types i sabotages, and $sprom_i$ takes the value 1 or 0, indicating whether or not i self-promotes. We now investigate the strategic behavior for self-promotion and sabotage.

2.2 Self-Promotion

Self-promotion increases agent i's value v_i and total contest output S by the same amount and hence increases agent i's chance of winning. The size of this increase of course depends on agent i's type, the number of other high and low types in the contest, and the number of outsiders.

Lemma 1. *In contests with a sufficiently large performance gap g between low and high types, the expected gain from self-promotion is higher for a low type than a high type.*

Proof. See appendix. □

The proof introduces the notion of a performance gap g between high and low types. A sufficiently large performance gap of $g \geq 1$ ensures that the high type values v_i are higher than the low type values v_k even if the high types are being sabotaged by all H_{-i} and L agents, while low types are not sabotaged at all (i.e., a worst-case scenario for high types).

A high type will of course self-promote if the gain from self-promotion is higher than its costs. More specifically:

Lemma 2. *If, for a high type with value v_i and a performance gap $g \geq 1$, $c_p < \frac{\Delta p_h (S - v_i)}{S(S - \Delta p_h)}$ in equilibrium, then all agents self-promote.*

Proof. If the high type did not self-promote, her winning probability would fall by $\frac{v_i}{S} - \frac{v_i - \Delta p_h}{S - \Delta p_h} = \frac{\Delta p_h (S - v_i)}{S(S - \Delta p_h)}$. By assumption, this is larger than c_p, so her net gain would be negative. She will therefore stick to self-promotion. By Lemma 1, this continues to hold for the low types if $g \geq 1$. Alternatively, low types will self-promote if $c_p < \frac{\Delta p_l (S - v_k)}{S(S - \Delta p_l)}$. □

2.3 Sabotage

If an agent sabotages another agent, she decreases that agent's output v_i and thus decreases S which increases her probability of winning the contest. This decrease in S benefits all other agents as well and thus sabotage has—contrary to self-promotion—an important externality (Konrad, 2000). We first show that if an agents sabotages another agent, the agent sabotages all agents of that type.

Lemma 3. *The expected marginal gain in winning probability is increasing in the number of other agents (of the same type) being sabotaged.*

Proof. Consider a high type of value v_i and sincere rating by everyone. Her winning probability is $\frac{v_i}{S}$. If she sabotages one more high type, her winning probability increases to $\frac{v_i}{S - b_h}$. The marginal gain in probability from this sabotage action is $\frac{v_i}{S - b_h} - \frac{v_i}{S} = \frac{b_h v_i}{S(S - b_h)}$. We can then express the marginal gain of sabotage as a function of the number of other high types x_i that i sabotages: $f(x_i) = \frac{x_i b_h v_i}{S(S - x_i b_h)}$. The first derivative of which is $f'(x_i) = \frac{b_h v_h}{(S - x_i b_h)^2}$ which is always positive for positive x_i and given that $S > n b_h > h b_h > x_i b_h$ holds since by design $n > h$ and $h > x_i$ since i does not sabotage herself and at most sabotages $h - 1$ high types. Thus, the marginal gain of sabotaging another high type is increasing in the number of high types already sabotaged. If the costs of sabotage are smaller than the (maximum) marginal gain from sabotaging the last other high type, a rational high type will therefore sabotage all other high types. An analogous argument shows that the marginal gain for a high type from sabotaging one more low type is increasing in the number of low types already sabotaged. Exactly the same arguments also hold for low types sabotaging other agents. □

We can use Lemma 3 to compute the bounds of sabotage costs c_s when agents will engage in strategic contest behavior. For example, consider a high type of value v_i not currently sabotaging anyone and sincere voting by everyone else so that $v_i = b_h(n + l + h)$ and $v_k = b_l(n + l + h)$, and $S = h v_i + l v_k$. If she sabotages all other high types (remember she will not sabotage herself) her winning probability is $\frac{v_i}{S - b_h(h-1)}$. If she were to sabotage only $h - 2$ high types her probability of winning is $\frac{v_i}{S - b_h(h-2)}$. Consequently, her maximal marginal increase in winning probability from sabotaging other high types results from sabotaging all other $h - 1$ instead of just $h - 2$ high types and is of size

$$\frac{v_i}{S - b_h(h-1)} - \frac{v_i}{S - b_h(h-2)} = \frac{v_i b_h}{(S - b_h(h-1))(S - b_h(h-2))} \tag{2}$$

If the cost of sabotage c_s are smaller than the maximal marginal gain in Eq. 2, a rational high type will therefore sabotage all other high types. In a similar fashion we can compute the bounds for all other sabotages (high types sabotaging low types, low types sabotaging high types etc.). We show all bounds in the appendix. Inspecting the bounds, the following relationships between the bounds are apparent.

Lemma 4. *Both high and low types have more to gain from self-promotion than from sabotaging any other types.*

Proof. This follows directly from the externality associated with sabotage. □

Lemma 5. *High types have more to gain from sabotaging other agents of a given type (i.e., high types), than low types have to gain from sabotaging those agents.*

Proof. See appendix. □

Lemma 6. *Both high and low types have more to gain from sabotaging high types than low types.*

Proof. See appendix. □

2.4 Nash Equilibria

This contest has many Nash equilibria. Which equilibrium materializes depends on the sizes of the groups n, l, and h, the costs c_s and c_p, and the performance gap g between high and low types.

Proposition 1. *In contests with a positive performance gap $g \geq 1$ between high and low types, the following seven states are the set of possible Nash equilibria.*

(NE1) $s_i = (0, 0, 0)$ for $i \in H$, and $s_k = (0, 0, 0)$ for $k \in L$. No one self-promotes, and no one sabotages anyone.

(NE2) $s_i = (0, 0, 0)$ for $i \in H$, and $s_k = (0, 0, 1)$ for $k \in L$. Low types self-promote, and no one sabotages anyone.

(NE3) $s_i = (0, 0, 1)$ for $i \in H$, and $s_k = (0, 0, 1)$ for $k \in L$. All agents self-promote, and no one sabotages anyone.

(NE4) $s_i = (h-1, 0, 1)$ for $i \in H$, and $s_k = (0, 0, 1)$ for $k \in L$. All agents self-promote, high types sabotage each other but no low types, and low types do not sabotage.

(NE5) $s_i = (h-1, 0, 1)$ for $i \in H$, and $s_k = (h, 0, 1)$ for $k \in L$. All agents self-promote, high types sabotage each other but no low types, and low types sabotage only high types.

(NE6) $s_i = (h-1, l, 1)$ for $i \in H$, and $s_k = (h, 0, 1)$ for $k \in L$. All agents self-promote, high types sabotage everyone, and low types sabotage only high types.

(NE7) $s_i = (h-1, l, 1)$ for $i \in H$, and $s_k = (h, l-1, 1)$ for $k \in L$. All agents self-promote, and all agents sabotage all other agents of both types.

Proof. This follows directly from Lemma 1 (low types will self-promote before high types do), Lemmas 4-6 which give an oder of the bounds, and the additional proof that the benefit for low types to sabotage high types is larger than the benefit of high types to sabotage low types (see appendix). □

2.5 Revisit Performance Gap g

It is now interesting to revisit the condition of $g \geq 1$ for self-promotion. It is clear that the closer b_h is to 1 the less high types have to gain from self-promotion. Consequently, they will only self-promote if the costs c_s for doing so are increasingly smaller. In the equilibrium condition NE4, for example, in which high types sabotage each other and low types do not sabotage anyone, then $g \geq 1$ is violated if $(n + h + l)b_l > (n + l + 1)b_h$ and the resulting gain from self-promotion would be lower for low types than high types. This would then lead to the interesting case where low types do not self-promote while high types do. Considering both self-promotion and sabotage, it is interesting to note the following opposing effect of the performance gap: The larger the performance gap, the more attractive *self-promotion* becomes for *low types* and the more attractive *sabotage* becomes for *high types*. As a result, in contests with a large performance gap, the ability to use strategic behavior levels the playing field, thus making the contest more attractive to low types.

From this, we can draw another observation: Self-promotion acts as an equalizing mechanism that reduces the performance difference between high and low ability agents and thus reduces incentives for high ability agents to sabotage (this can be seen by calculating and comparing Equation 2 in our model with and without self-promotion).

2.6 Example and Additional Predictions

Let the prize be $M = \$5,000$ and let $h = 10$, $l = 30$, and $n = 100$. Further, let $b_h = 0.8$ and $b_l = 0.2$ (this corresponds to a common case of a finite rating scale from 0-stars to 5-stars where low types produce contest entries of 1-star quality and high type agents of 4-stars quality). The

resulting performance gap between low and high types is then $g = \frac{b_h}{b_l}\frac{n+1}{n+l+h} = \frac{0.8}{0.2}\frac{100+1}{100+30+10} > 1$ which means that high types always have higher values than low types, even when being sabotaged. Figure 3 shows agent utility (Mp) as a function of cost of sabotage c_s. The figure illustrates the transitions between equilibria and their relative sizes in terms of the range of c_s values spanned. The figure suggests that there is one equilibrium in particular that has a large basin of attraction that spans a wide range of c_s values that seem plausible in online contest ($0.06 to $0.23): all agents self-promote and high types sabotage other high types while low types do not sabotage (NE4).

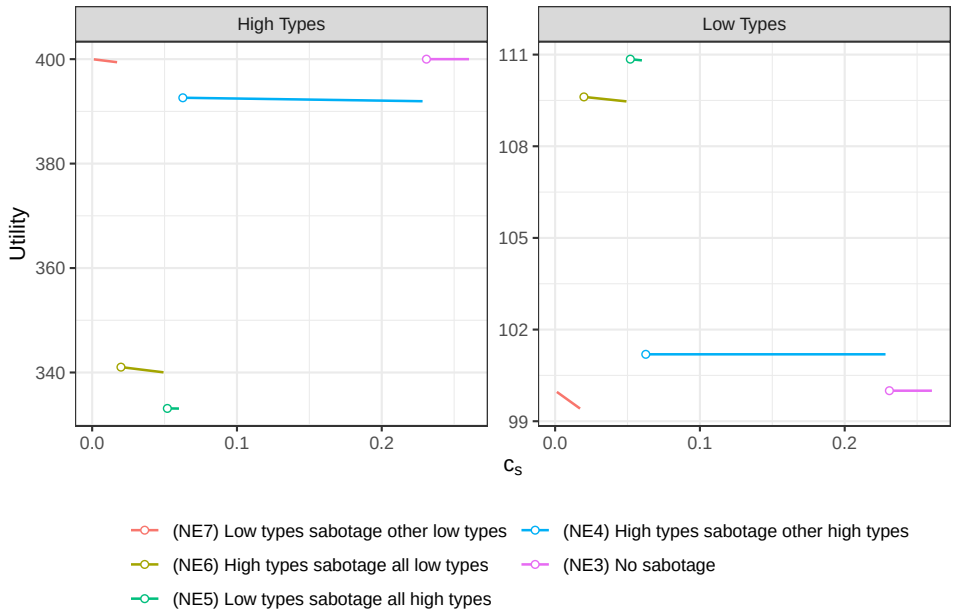


Figure 2: Equilibria and utilities.

From the equilibrium predictions above, we develop additional predictions regarding the relationship between strategic behavior (self-promotion and sabotage) and the number of competitors (contest size). The panel in Figure 3 shows some key predictions for changing contest sizes. Regarding self-promotion, comparative statics suggest decreasing marginal utility for this form of strategic behavior. We predict that the response to increased competition is a reduction in self-promotion for both high and low ability agents. That is, we predict self-promotion to be less prevalent in more competitive contests.

For sabotage, we predict a broad band of c_s values in which high ability agents sabotage other high ability agents. Unless the actual cost for sabotage that an agent faces is exactly at the upper bound, that agent will continue to sabotage all other agents even if one additional high type agent were to join the contest. That is, only if the cost faced by an agent is exactly the upper limit, will that agent move from sabotaging all other agents to not sabotaging anyone. Thus, increasing the contest by one additional high type, will lead to $2h$ additional acts of sabotage (every of the h agents in the contest sabotages the additional high type, and the additional high type sabotages all h already in the contest). We predict that the response to

increased competition is an increase in sabotage by all agents.

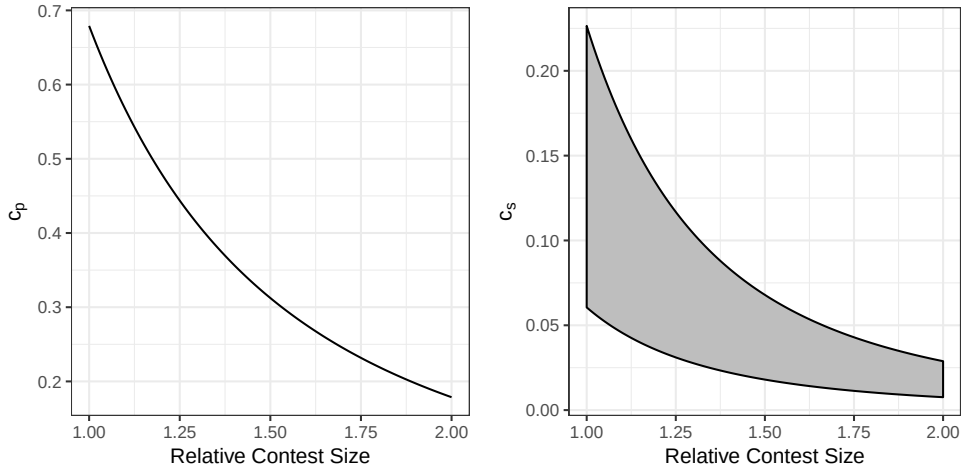


Figure 3: Equilibrium strategies for high types. (a) Marginal utility of self-promotion in contest of changing size. (b) Upper and lower bound of c_s in contests of changing size. If cost c_s fall below this lower bound low types will also sabotage high types.

2.7 Summary of Model Predictions

In this section we have developed a model of strategic behavior in contest where contest winners are determined through a community and peer evaluation mechanism. Our theoretical model and the subsequent comparative statics in our example suggests that one equilibrium in particular has a large basin of attraction (NE4: all agents self-promote, high types sabotage other high types, low types do not sabotage). Based on this, we derive several predictions to frame our subsequent empirical analyses.

i. *High ability agents are the most likely to sabotage* (Lemma 5).

ii. *High ability agents sabotage each other* (Lemma 5 & 6).

iii. *Self-promotion is prevalent: Most agents, including low ability agents, self-promote* (Lemma 2 & 4).

Based on the comparative statics and example in Section 2.6 we can make the following secondary predictions:

iv. *Self-promotion is more prevalent in smaller contests and less common in large contests* (Section 2.6).

v. *Sabotage is more prevalent in larger contests* (Section 2.6).

3 Empirical Setting

The data for our study come from a longitudinal panel of weekly design contests on the Threadless website (`www.threadless.com`). Threadless is a crowdsourcing and e-commerce platform that hosts weekly t-shirt design contests (Nickell and Kalmikoff, 2010). Started in 2001, the site has developed into a leading crowdsourcing platform pioneering a community-based business model. The company involves its community of 1.5 million in nearly all aspects of the innovation and product development process and employs no in-house designers (Lakhani and Kanji, 2008). The platform draws on the creative talent of designers from across the world and has a distinct focus on building and nurturing an online community of creative individuals (Riedl and Seidel, 2018). For example, they attempt to make most profits available to designers (Nickell and Kalmikoff, 2010).

The site hosts weekly design contests to which anyone is free to submit. Threadless provides a designer kit and template, and submissions from any standard software program or digitized drawings are accepted, thus offering exceptionally low barriers to entry. Submitted designs are then posted publicly on the website for peer rating, using a scale from 0 to 5 stars.[5] Ratings are anonymous (to the community, but not to us the researchers) and no average rating is shown at the time of rating so as to avoid social influence and herding.[6] Our empirical setting thus resembles key characteristics of modern contest environments including low barriers of entry, a high number of contestants, high heterogeneity of ability among contestants, and anonymity.

Among the submissions that received the highest average rating, Threadless typically picked three to six designs per week as contest winners. We show the rating percentile of contest winners in Figure 7. On average, contest winners scored in the 95^{th} percentile of all designs submitted to the same competition in the same week. The median rank of printed designs is the 98^{th} percentile. These contest winners would then be printed and subsequently sold through the Threadless e-commerce site. Designers receive a cash prize and additional store credit if their design was selected for printing. Threadless generated $30 million in revenue in 2012 and gave out over $775,000 in prizes over the observation period. We focus on the regular weekly contests, excluding from our analysis special and themed contests that Threadless also hosts from time to time. Our data consists of a unbalanced cross-panel of 74,525 individuals participating in 511 contests and casting over 38 million ratings on 154,086 contest entries (Table 1). As a proxy for ability, we use the highest average rating a submission by a designer received in the past. The ability distribution is right skewed (Figure 4(a)) and thus, our empirical setting matches our model which assumes fewer high types than low types. The average contest size is 407 submissions and Figure 4(b) shows the entire distribution.[7]

[5] The platform switched to a 1 to 5 stars scale recently but during the period observed in our data rating was on a 0 to 5 scale.

[6] The site does show a counter of how many ratings have already been submitted.

[7] The bimodal distribution of contest sizes is an artifact of doubling contest sizes in the early years (2001-2005) and a smaller increase in the later years.

	Rate own Submission	
Submitted to same contest	No	Yes
No	30,246,070 (13.58%)	
Yes	7,772,772 (10.73%)	114,914 (74.58%)

Table 1: Summary of rating behavior. Parentheses show probability of rating conditional on submitting.

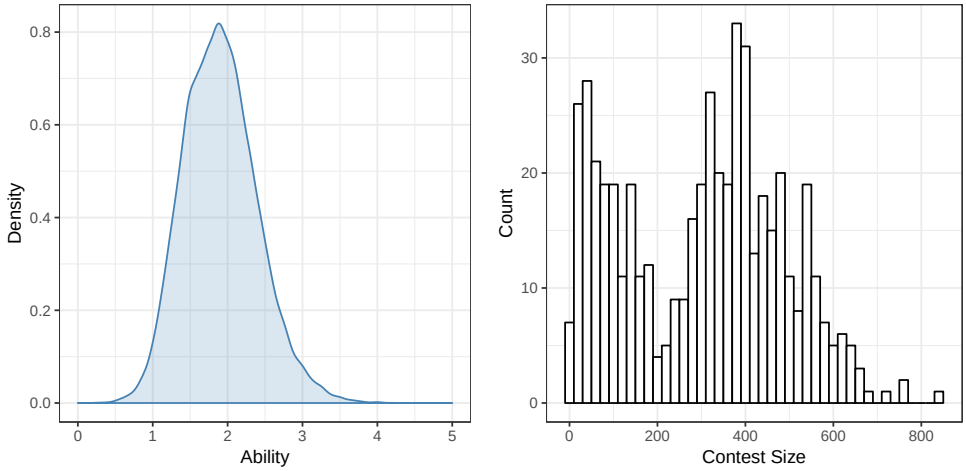


Figure 4: (a) We compute ability as the highest average rating a submission by a designer received in the past. The ability distribution is right skewed (skewness: 0.37). (b) Distribution of contest sizes (i.e., number of contest entries; mean: 407).

3.1 Identification Strategy

To empirically identify strategic contest behavior, we leverage the fact that individuals actively participate in the rating process even when not competing. We model the effects of participating in the same contest on rating behavior using a difference-in-difference approach. That is, we model within individual differences between ratings cast on competitors versus non-competitors and within submissions differences between ratings cast by those who submitted to the contest and those that did not, thus controlling for submission quality. Ratings cast by an individual in a contest that the individual did not compete in him-/herself serve as control for those contests in which he/she is competing in (i.e., in a week that a contestant submitted a contest entry, she is a (high/low type) agent rating her competitors, in a week that a contestant did not submit a contest entry she is a neutral outsider). Simultaneously, ratings on the same submission from individuals who are not themselves competing in the contest serve as controls for individuals who are competing in the contest.

We model sabotage as the change in probability that individual i submits a 0-stars rating—the lowest possible rating—on submission j when having submitted to the same contest. Conversely, we model self-promotion as the change in probability of rating ones own submission with 5-stars—the highest possible rating—when rating one's own submission. The two key explanatory variables are (a) a dummy indicating whether an individual is a participant in the

contest (i.e., whether i has submitted to the same contest c and is in the running for the prize) or not; and (b) a dummy indicating if the individual is rating his or her own submission (this is only possible if the other dummy is also 1). That is, we estimate the following equations:

$$\text{0-Stars Rating}_{ij} = \beta_{11}\text{Submitted to same contest}_{ij} + \beta_{12}\text{Rate own submission}_{ij} + \alpha_i + \alpha_s + \epsilon_{ij} \tag{3}$$

$$\text{5-Stars Rating}_{ij} = \beta_{21}\text{Submitted to same contest}_{ij} + \beta_{22}\text{Rate own submission}_{ij} + \alpha_i + \alpha_s + \epsilon_{ij} \tag{4}$$

where

$\text{0-Stars Rating}_{ij}$ is an indicator that is 1 if the rating submitted by individual i on the submission j is a 0-stars rating

$\text{5-Stars Rating}_{ij}$ is an indicator that is 1 if the rating submitted by individual i on the submission j is a 5-stars rating

α_i are individual-level fixed effects

α_s are submission-level fixed effects

ϵ_{ij} are error terms.

We use this approach to separately identify the two types of strategic contest behavior outlined in the theoretical model above:

- **Self-promotion:** if a 5-star vote is assigned to one's own submission this indicates self-promotion (β_{22} in Eq. 3). We provide robustness tests using a natural experiment to address concerns that this would not reflect strategic behavior but rather the result of overconfidence or an increased preference fit in section 4.3.2.
- **Sabotage:** Sabotage is indicated if 0-star votes are assigned more liberally when competing in the same contest versus being an outsider who has no stakes in the contest (β_{11} in Eq. 3). We provide results from a second natural experiment to provide evidence that low ratings are strategic rather than a result of endogenous entry into contests in section 4.3.1.

We estimate our models as linear probability models for ease of interpretation (Aldrich and Nelson, 1984; Angrist, 2001; Greene, 2012). A discussion of the usefulness of the approach appears in Angrist (2001) and Moffitt (2001). Furthermore, the limitations of the approach notwithstanding, efficient estimation approaches for OLS exist for large datasets that support demeaning of multiple fixed effects. We estimate linear probability model using the lfe package

for R (Gaure, 2013) which supports demeaning of multiple fixed effects. This algorithmic approach is mandatory as our dataset is extremely large with more than 38 million rows, 74,525 individual fixed effects, and 154,086 submission fixed effects (or alternatively 511 contest fixed effects).[8]

To investigate how sabotage and self-promotion vary across ability levels we compute the time lagged ability for each individual who is casting a rating (we term this the *source*) or who submitted the contest entry receiving a rating (we term this the *target*). As a proxy of ability, we use the highest average rating a submission by a designer received in the past. Specifically, we compute lagged ability as the maximum average rating that a contest submission by that individual received prior to casting the current rating (source ability) or to making the current submission (target ability).[9] We then estimate versions of the models in Equation 4 and 3 in which we include source and target ability and their interactions with the *Submitted to same contest* and *Rate own submission* dummies. Since target ability is time-invariant at the submission level, we use contest-level fixed effects instead of submission-level effects. To investigate how the use of strategic behavior varies across contests of different sizes, we also compute the *Contest size* as the number of submissions to a contest. To ease interpretation, we normalize contest size to the unit interval. We show descriptive statistics and correlations of main variables in Table 2.

	Mean	SD	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)
Rating (1)	1.93	1.53	0.00	5.00						
Submitted to same contest (2)	0.71	0.41	0.50	1.50	0.03					
Rate own submission (3)	0.50	0.05	0.50	1.50	0.11	0.11				
Contest size (4)	406.74	131.88	2.00	848.00	0.07	0.11	0.00			
Average score in contest (5)	1.93	0.23	1.26	2.75	0.15	0.04	0.00	0.50		
Source ability (6)	2.26	0.61	0.00	5.00	0.02	0.05	0.00	0.17	0.29	
Target ability (7)	2.33	0.61	0.00	5.00	0.23	0.00	0.00	0.14	0.30	0.10

Table 2: Descriptive statistics and correlations of main variables. All correlations are significant at $p < 0.001$.

4 Results

4.1 Strategic Behavior

Table 3 shows baseline estimates of a difference-in-difference model estimated using OLS (Model 1). When looking at scores overall, we do not find signs for sabotage. That is, scores are actually significantly higher when contestants have submitted to the same contest ($\beta = .024$; $p < .001$), indicating more lenient ratings from other contestants. However, we find strong signs for self-promotion ($\beta = 2.820$; $p < .001$). When contestants rate their own design they assign much

[8]The naive approach of estimating the model with dummies for the two fixed effects is computationally intractable even using advanced high-performance computing.

[9]We use the average rating across all prior designs as a robustness check which is constistent with our results.

higher scores compared to designs of others.

We next estimate a linear probability model to estimate the effect of being a participant in a contest on the likelihood of submitting low evaluations in Model 2 (0-Stars Rating). Consistent with the OLS model, we find a significant negative effect ($\beta = -.008$; $p < .001$) indicating that likelihood of submitting low ratings actually diminishes when rating ones competition. That is, we find no overall sabotage when averaging across all ability levels. To investigate effects of sabotage across contests of different sizes, we introduce an interaction term between the *Submitted to same contest* dummy and the *Contest size* variable. We find that the probability for low ratings increases significantly in larger contests (Model 3: $\beta = .008$; $p < .001$).

	OLS	Linear Probability			
Dependent Variable:	*Rating*	0-Stars Rating		5-Stars Rating	
	(1)	(2)	(3)	(4)	(5)
Submitted to same contest: Yes	0.024***	−0.008***	−0.012***	−0.002***	−0.002***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)
Rate own submission: Yes	2.820***	−0.195***	−0.194***	0.855***	0.860***
	(0.003)	(0.001)	(0.001)	(0.001)	(0.002)
Contest Size					
× Submitted to same contest: Yes			0.008***		
			(0.001)		
× Rate own submission: Yes					−0.011**
					(0.004)
Individual	*Fixed*	*Fixed*	*Fixed*	*Fixed*	*Fixed*
Submission	*Fixed*	*Fixed*	*Fixed*	*Fixed*	*Fixed*
Adj. R^2	0.382	0.372	0.372	0.210	0.210
Num. obs.			38,102,880		

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Estimates for strategic behavior in contest. Standard errors are in parentheses, clustered at the submission level.

Next, we estimate a linear probability model on casting a 5-stars rating. We find significant level of self-promotion (Model 4: $\beta = .855$; $p < .001$). The effect of self-promotion is extremely strong. Investigating the data further, we find that self-promotion is very prevalent. 75% of individuals rate their own submissions and do so with the highest possible rating: 97% of self-votes are 5-stars. Given that the average rating of non-self-ratings is only 1.80, this implies agents are self-promoting with dramatically inflated ratings that are significantly above the actual quality of their submission. Finally, we investigate how self-promotion changes across contests of different size. We find that self-promotion decreases as contest size increases (Model 5: $beta = -.011$; $p < .001$).

4.2 Heterogeneous Contestant Ability

We now turn to investigate strategic contest behavior with regard to heterogeneous ability of contestants. To estimate heterogeneity with regard to ability among who is performing sabotage and who is targeted, we estimate a variation of Equation 4 which includes measures of ability

for the contestant casting the rating (source ability) and the contestant who submitted the contest entry being rated (target ability), an interaction between source and target ability, and interaction terms between the *Submitted to same contest* dummy and source and target ability, respectively. This analysis relies on a reduced sample sizes as we have to exclude (a) all ratings that an individual casts before making the first submission (to compute source ability) and (b) all ratings on every contestant's first submission (to compute target ability). Note that the main effects for source ability and target ability capture the variance in ability over time due to the inclusion of individual fixed effects. As before, we estimate the equation as a linear probability model with individual and submission fixed effects (Table 4). We find significant heterogeneity in sabotaging behavior (Model 1: 0-Stars Rating). We find that more able individuals are more likely to be the target of sabotage ($\beta = .002$; $p < .001$) and more likely to be the source of sabotage ($\beta = .012$; $p < .001$). Turning to self-promotion (Model 2: 5-Stars Rating), we estimate a similar model using the *Rate own submission* dummy but include only the interaction with source ability, since source and target are the same in case a contestant is voting on herself. We find that contestants of higher ability are significantly less likely to self-promote ($\beta = -.06$; $p < .001$) compared to lower ability contestants.

To better interpret the heterogeneity across ability levels, we compute predicted values for the expected *change* in the likelihood to rate 0-stars when competing versus not competing (Figure 5). The heat map in Panel A shows that high ability contestants target other high ability contestants with an up to 4.5% increase in assigning a 0-stars rating when competing compared to their baseline when not competing. The null-line marking the start of sabotaging behavior is not perfectly symmetric: The highest ability contestants are sabotaged by contestants of all other ability levels. This again matches our theoretical predictions that low types sabotaging high types is the second most lucrative form of sabotage (after high types sabotaging high types). Interestingly, on the opposite end of the ability spectrum we find that low ability contestants appear to be more lenient toward other low ability contestants with their ratings when competing.

To facilitate the interpretation of our results of self-promotion, we similarly compute the relative *change* in probability of rating a submission with 5-stars if that submission is the contestants own submission compared to a submission of another contestant of the same ability (since source and target ability are the same in case of self-rating, this analysis is effectively the diagonal of the heatmap in Panel A). Panel B in Figure 5) shows that the likelihood that contestants self-promote decreases with their own ability. This finding is in line with our the predictions from our theoretical model that self-promotion is less prevalent among high types.

4.3 Alternative Explanations

Our primary identification is a difference-in-difference approach contrasting contestants' evaluation behavior in contests in which they are competing—and may be strategically motivated—

Dependent Variable:	Linear Probability	
	0-Stars Rating	5-Stars Rating
	(1)	(2)
Submitted to same contest: Yes	−0.055***	−0.000
	(0.004)	(0.001)
Rate own submission: Yes	−0.179***	0.986***
	(0.004)	(0.007)
Target ability	−0.066***	−0.001
	(0.002)	(0.001)
Source ability	−0.019***	−0.084***
	(0.004)	(0.003)
Target ability × Source ability	0.002	0.027***
	(0.001)	(0.001)
Submitted to same contest: Yes × Target ability	0.012***	
	(0.001)	
Submitted to same contest: Yes × Source ability	0.008***	
	(0.001)	
Rate own submission: Yes × Source ability		−0.060***
		(0.003)
Individual	*Fixed*	*Fixed*
Contest	*Fixed*	*Fixed*
Adj. R^2	0.360	0.211
Num. obs.	18,787,584	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Estimates heterogeneous effects by ability level. Contest-level fixed effects are used instead of submission-level as agents ability is time invariant at the submission level. Standard errors are in parentheses, clustered at the contest level.

from contests in which they have no stakes—and should consequently not be strategically motivated. In this section we address two potential concerns and alternative explanations regarding (a) low ratings being a result of endogenous entry into contests rather than sabotage and (b) an high ratings for ones own submission reflecting overconfidence in ones ability or an increased preference fit for one's own creative work rather than self-promotion. For both cases we provide evidence that suggests that contestants' behavior is indeed strategically motivated from two independent natural experiments.

4.3.1 Sabotage vs. Endogenous Entry

We model the strategic behavior of sabotage as the shift in probability to assign low rating when competing compared to when not competing. An alternative explanation for the observed negative rating behavior could be seen in endogenous decision to enter contests (c.f., Bockstedt et al., 2016; Gradstein, 1995; Morgan et al., 2012). Contestants may submit entries to contests where they expect that competition is weak and, therefore, deserve lower ratings. To rule out this alternative explanation we make use of a natural experiment that occurred on the platform within our window of observation: a change in prize money awarded to contest winners. The prize for winning the weekly contest was raised from $500 to $1,000 in 2005. This doubling of the contest prize was announced on the company's blog on the day it became effective and consequently contestants had no prior knowledge of it and were not able to withhold submissions

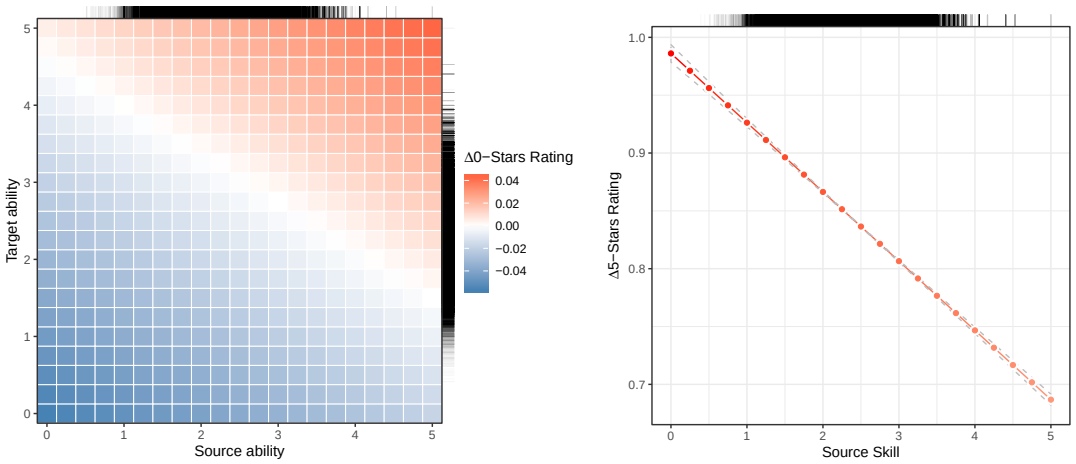


Figure 5: Strategic behavior by competitors of heterogeneous ability levels. **Panel A (Sabotage).** Relative change in probability of rating 0-stars when competing compared to not competing across ability levels. Outer rugs show distribution of data. There are over 15,000 (10,000) observations for source (target) ability greater than 4.5. **Panel B (Self-promotion).** Relative change in probability of rating 5-stars when rating own submission compared to submissions by others of same ability (error band is 95% confidence interval).

or otherwise alter their behavior. The prior literature on sabotage has stressed the fact that sabotage activities should increase with an increase in prize (or, more precicely the prize spread between the contest winner and the contest loser(s); e.g., Lazear, 1989; Harbring and Irlenbusch, 2011). Thus, following the interpretation of sabotage as strategically motivated, we would expect an increase in the probability to assign 0-stars after the incentive change. Following the argumentation of endogenous entry, however, we would expect either no change in probabilities as agents should not be able to strategically time their contest entry. Moreover, if rating is not strategically motivated, we would expect a decrease in the probability to rate 0-stars after the incentive change submission quality increases as a result of the incentive effect leads to increased effort (Morgan et al., 2012).

We use a regression discontinuity design to identify the causal effects of the incentive changes (see Hahn et al., 2001; Hausman and Rapson, 2018; Imbens and Lemieux, 2008). We use a six months time window before and after the incentive changes. We estimate the same linear probablility model for 0-Stars ratings with individual and submission fixed effects as before (Table 5). We find a significant 2.2% increase in the probability to rate 0-stars when competing after the incentive changes (Model 1: $\beta = .022$; $p < .001$). The effect is stronger (2.3%) in the first quarter after the incentive change and slightly weaker (2.0%) in the second quarter (Model 2: $\beta = .023, \beta = .020$; both $p < .001$). A placebo tests in which we chose a fake date four months before the actual incentive change for the discontinuity provides further evidence for the causal impact of the incentive change on the rating behavior (Model 3: $\beta = -.009$; $p < .001$) as it rather indicates a negative time trend. These findings are in line with our explanation that

Dependent Variable:	Linear Probability		
	0-Stars Rating		
	Base	Persistence	Placebo Test
	(1)	(2)	(3)
Submitted to same contest: Yes	−0.037***	−0.037***	−0.035***
	(0.000)	(0.000)	(0.000)
Rate own submission: Yes	−0.179***	−0.179***	−0.179***
	(0.000)	(0.000)	(0.000)
Submitted to same contest: Yes			
× After	0.022***		0.028***
	(0.000)		(0.000)
× 1st Quarter After		0.023***	
		(0.000)	
× 2nd Quarter After		0.020***	
		(0.000)	
× Fake After			−0.009***
			(0.000)
Individual	*Fixed*	*Fixed*	*Fixed*
Submission	*Fixed*	*Fixed*	*Fixed*
Adj. R^2	0.390	0.390	0.390
Num. obs.		825,504	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: Natural experiments of sabotage behavior after an incentive changes from \$500 to \$1,000 prize money using ±6 months observation windows. Standard errors in parentheses, clustered at the submission level.

the increase in probability to assign 0-stars when contestants rate their competition is indeed strategically motivated sabotage and not a result of endogenous entry.

4.3.2 Self-promotion vs. Overconfidence

We model the strategic behavior of self-promotion as an shift in probability to rate ones own contest entry with 5-stars compared to how likely one is to rate the entries by others with 5-stars. That is, to increase their own probability of winning contestants strategically assign the highest rating possible to their own output. Possible alternative explanations for this behavior might be overconfidence of contestants in their own abilities (Benabou, R., & Tirole, 2002; Camerer and Lovallo, 1999; Clark and Friessen, 2009) or an increased preference fit of ones' own creative work compared to the work of others (e.g., Berg, 2016; Franke et al., 2010). That is, contestants might simply perceive their own designs as better than those of others due to biased perception. The analysis of the contest size (Table 3) provides first indications that the observed patterns cannot be explained solely by overconfidence or an increased preference fit as the perception of one's own design should not be influences by the number of other designs in the contest. However, to further address these concerns we provide another robustness test which leverages another natural experiment.

In 2005 the platform owner made a change to the scoring mechanism that we use to

strengthen our claim that rating one's own contest entry is strategically motivated.[10] Before the change, every contest entry was posted on the website for the full seven day rating period. After the change, contest entries that had not received ratings from 100 different people or the average rating was less than 1.5 stars were dropped from the rating process before the seven day rating period was complete. This rule change was introduced to help the platform owner to focus on a smaller set of contest entries, dropping low quality submissions from the contest earlier. The change was announced and immediately implemented without any pre-announcement. If rating ones' own submission is not intended strategically but only reflect biased perception we should not see any change in rating behavior around this rule change. If, however, rating on ones' own submission is intended strategically, such rating should happen earlier to maximize chances to make it past the 100 vote / 1.5 stars cutoff.

We analyze the sequence of ratings for each contest entry and test when in the sequence the self-promoting rating was cast (Figure 6). We focus our analysis on the 20 weeks period around the rating rule change (±10 weeks before/after the rule change). We find that before the rule change, self-rating happens roughly in the middle of the rating sequence (43rd percentile). That is, the self-rating was cast roughly at a random day during the rating period as there was no incentive to rate early. After the rule change, the self-rating is cast significantly earlier, falling in the 25th percentile of the rating sequence. Our analysis shows that contestants changed their self-rating behavior, shifting it earlier, after the rule change. Thus, our analysis of this natural experiment suggests that rating on ones own design is used as a strategic device to self-promote ones own work.

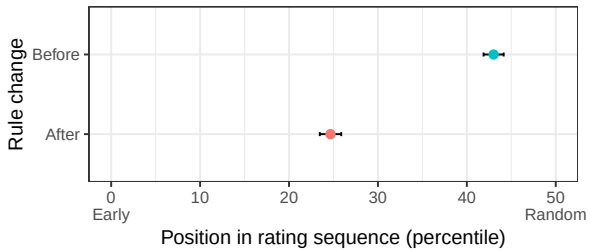


Figure 6: Analysis of rating sequence. Self-rating happens significantly earlier after the rule change.

4.4 Welfare Analysis

We first look at the question whether strategic rating behavior does matter in the generally large contests we observe. We first look at self-promotion. To gage the impact of self-promotion, we exclude all ratings from designers who rated their own designs and recompute average ratings and the resulting contest ranking. We find that, keeping everything else constant, in seven out of 511 contest (1.4%) the winner of the contest changes and in 28 of 511 contests (5.5%) there

[10] https://www.threadless.com/infoblog/1236/important_new_scoring_functionality

is a change in at least one of the top three ranks. To analyze the effect of sabotage we exclude all ratings from competitors as they might be strategically motivated.[11] We find that in 12% of the contests the winner of the contest changes and that in 48% of the contests there is a change in the top three ranks. The effect is especially pronounced in close contests[12] where the contest winner would change in 25% of cases and 65% would see a change in the top three. This can be considered an upper boundary for the effect of strategic behavior via the rating mechanism in our setting. Overall, given that Threadless has payed approximately $775,000 in prize money over the observation period we find that up to $93,000 may have been payed to the "wrong" contestant due to strategic behavior.

4.5 Discussion of Findings and Policy Implications

In our empirical analysis we find support for the prediction that high ability contestants are likely to be both victims and culprits of sabotage. While our findings are consistent with prior research on who is targeted by sabotage (e.g., Carpenter et al., 2010; Münster, 2007) our findings regarding culprits is in contrast to the findings of Carpenter et al. (2010) and Balafoutas et al. (2012) who find that especially low ability contestants are likely to engage in sabotage targeted towards high ability contestants. This difference can be explained by the arguments of Chen (2003) who differentiates conditions for culprits of sabotage, i.e., the relation of talent for constructive and destructive activities. In our empirical setting it can be assumed that the ability to produce productive outputs (i.e., create good designs) is correlated with the ability to identify other high ability contestants (i.e., recognize good designs; Riedl and Seidel, 2018). Interestingly, we also find significant evidence for lenient ratings among low-ability contestants. This is in line with the experimental evidence of "gift-giving" of Carpenter et al. (2010). This might be the result of empathy induced prosocial behavior (c.f., Decety and Jackson, 2004). That is, low ability designer who participate in the same contest tend to give fewer negative scores because they share the feeling of being judged by others. Furthermore, counter to the prediction of Konrad (2000) that sabotage should be a minor phenomenon in large contests due to the externality associated with it, our empirical observations across 511 contests provide evidence that sabotage increases in larger competitions. This indicates that in the field, counter to frequent assumptions of formal models, costs of sabotage are heterogeneous across contestants which underscores the importance of investigating strategic behavior in the field. Leveraging a natural experiment we are further able to provide field evidence for the theoretical prediction of Lazear (1989) and the laboratory experimental findings of Harbring and Irlenbusch (2011), i.e. that sabotage activities increase with an increase in prize spread. In our natural experiment we find that a doubling in prize spread increases sabotage by 2.2%.

Our empirical analyses show that self-promotion is ubiquitous. Contestants across all ability

[11] Note: This also includes votes from designers on their own designs, i.e., it contains the previous reported effects.

[12] We define a close contest as one where the standard deviation in ratings of the top ten placed submissions is 10th percentile.

levels engage in non-productive strategic activity (Bhagwati, 1982), that increases their own chances of winning the contest without increasing the output for the contest designer. This is underlined by the fact that our measure of self-promotion is rather conservative. Especially low ability contestants have more to gain from self-promotion and therefore change their behavior more drastically when competing compared to high ability contestants. Our findings are thus in accordance with prior research of Milgrom and Roberts (1988b) that predicts that agents in contest settings tend to shift their effort towards producing signals rather than productive effort. Our finding that self-promotion decreases with contest size supports our interpretation that ratings on one's own design are in fact used strategically because overconfidence and biased perception are less likely to be affected by changes in contest size. This is further underlined by our second natural experiment which shows that contestants strategically adjust their self-ratings in response to changes in the peer-rating mechanism in ways that cannot be explained by overconfidence (Camerer and Lovallo, 1999).

To tackle sabotage contest organizers could make the ratings not anonymous (c.f., Kim et al., 2019). This would force contestants to put their reputation on the line and thus would increase both the costs for self-promotion c_p and for sabotage c_s. Because reputation can be assumed to correlate with ability, it would disproportionately raise costs for high-ability contestants and level the playing field (c.f., Schotter and Weigelt, 1992). However, it can also be assumed that it in turn increases cooperation through indirect reciprocity (c.f., Nowak and Sigmund, 2005). Another mechanism would be to exclude contestants from rating their own submission. This would certainly reduce direct self-promotion, but our formal model also shows that self-promotion reduces the incentives for high ability contestants to engage in sabotage (c.f. Section 2.5). Self-promotion thus can be seen as a form of "opt-in" affirmative action (c.f., Schotter and Weigelt, 1992). Thus, eliminating self-promotion might be accompanied by a rise in sabotage from high ability contestants and that contestants would engage in indirect forms of self-promotion, e.g., by mobilizing their social network (Pickard et al., 2011).

As attempts to reduce only self-promotion might raise sabotage, contest organizers might be inclined to exclude contestants from the peer-rating altogether. This, however, may shift the problem to outsiders: As the incentives for contestants to engage in strategic behavior are still existent they might exert strategic behavior via contracts with otherwise neutral outsiders, i.e. they might buy outside ratings to increase their chances of winning. This would compromise the signal of the peer-rating by undermining the neutrality of outsiders N. Thus, contest organizers might think about refraining from peer-rating entierly and perform the task of evaluation themselves. However, there is ample research that suggest that internal evaluators are often biased (Boudreau et al., 2016a; Li, 2017; Li and Agha, 2015) especially under high workload (Criscuolo et al., 2017; Piezunka and Dahlander, 2015), and benefits of collective intelligence from large groups of evaluators (Blohm et al., 2016). However, rather than applying top-down approaches, a more subtle approach may be promising: contest organizers could think about fostering social norms of honesty among participants (Bauer et al., 2016; Elster, 1989). Once established, social norms would raise the moral costs of strategic behavior. Such an approach,

however, takes time and it is at least to a certain extent outside the control of the contest organizer.

5 Conclusion

In this paper, we analyze digital trace data of 38 million peer ratings of 74,525 individuals who compete in 511 contests to study strategic behaviors that contestants usually try to hide with unobtrusive measures and in a natural context. We study the behavioral patterns that arise when heterogeneous contestants have the possibility to engage in multiple strategic actions, in our case the use of self-promotion and sabotage. We generate primary and secondary predictions from a simple theoretical model of behavior of contestants in settings that reflect the characteristics of modern contests, i.e., high heterogeneity of contestants in ability, high number of participants, and multiplicity of strategic behaviors. The model also shows the interdependence of strategic behaviors: Because self-promotion acts as an equalizing mechanism diminishing the performance difference between high and low ability contestants (c.f., Section 2.5), it reduces incentives for high ability contestants to sabotage.

Our empirical analysis of more than 38 million observations reveals that strategic behavior has a significant impact on the final outcomes of contests in general and especially in close contests. This means that considerable amounts of prize money is awarded to the wrong contestants as a result of sabotage and self-promotion. We find support for our primary predictions that high ability contestants are likely to be both, victims and culprits of sabotage. This concentration of sabotage among high-ability contestants induces a relatively strong impact on the outcome of winner-take-all competitions. Our empirical analyses also provides support for the primary prediction of our formal model that self-promotion is ubiquitous. However, contestants are aware of their limited power to influence contest results as self-promotion decreases in larger contests, supporting one of our secondary predictions. Our model and findings also reveal that self-promotion disproportionately helps low-ability contestants. Finally, we also find support for the secondary prediction that sabotage increases with contest size.

Evidence from two natural experiments helps us to establish that our findings are actually driven by sabotage and self-promotion respectively and not by alternative explanations such as endogenous contest entry or overconfidence of contestants. Strategic behaviors like self-promotion and sabotage remain relevant in today's contest environments and, as our discussion of policy implications indicates, frugal ground of future research. While our findings certainly provide first insights on the highly relevant but sparsely research theme of malicious behavior in real world contests there is ample potential for future research leveraging new sources of data and unobtrusive measures.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2018). Preferences for truth-telling, forthcoming in. *Econometrica*.

Aldrich, J. H. and Nelson, F. D. (1984). *Linear Probability, Logit and Probit Models (Quantitative Applications in Social Sciences*. Sage Publications, Newbury Park.

Angrist, J. D. (2001). Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors. *Journal of Business & Economic Statistics*, 19(1):2–28.

Arbatskaya, M. and Mialon, H. M. (2010). Dynamic Multi-Activity Contests. *Economic Theory*, 43:23–43.

Atanasov, P. and Dana, J. (2011). Leveling the playing field: Dishonesty in the face of threat. *Journal of Economic Psychology*, 32(5):809–817.

Balafoutas, L., Lindner, F., and Sutter, M. (2012). Sabotage in Tournaments: Evidence from a Natural Experiment. *Kyklos*, 65(4):425–441.

Bauer, J., Franke, N., and Tuertscher, P. (2016). Intellectual Property Norms in Online Communities : How User-Organized Intellectual Property Regulation Supports Innovation. *Information Systems Research*, 27(4):724–750.

Benabou, R., & Tirole, J. (2002). Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Berg, J. M. (2016). Balancing on the Creative Highwire: Forecasting the Success of Novel Ideas in Organizations. *Administrative Science Quarterly*, 61(3):433–468.

Bhagwati, J. N. (1982). Directly Unproductive, Profit-Seeking (DUP) Activities. *Journal of Political Economy*, 90(5):988–1002.

Blohm, I., Riedl, C., Füller, J., and Leimeister, J. M. (2016). Rate or trade? identifying winning ideas in open idea sourcing. *Information Systems Research*, 27(1):27–48.

Bockstedt, J., Druehl, C., and Mishra, A. (2016). Heterogeneous Submission Behavior and its Implications for Success in Innovation Contests with Public Submissions. *Production and Operations Management*, 25(7):1157–1176.

Boudreau, K. J. (2018). Amateurs Crowds & Professional Entrepreneurs as Platform Complementors. *NBER Working Paper*.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., and Riedl, C. (2016a). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. *Management Science*, 62(10):2765–2783.

Boudreau, K. J., Lakhani, K. K. R., Menietti, M., Helfat, C. C. E., Lakhani, K. K. R., and Menietti, M. (2016b). Performance Responses to Competition across Skill-levels in Rank Order Tournaments: Field Evidence and Implications for Tournament Design. *RAND Journal of Economics*, 47(1):140–165.

Boudreau, K. J., Lakhani, K. R., and Menietti, M. (2016c). Performance Responses to Competition across Skill-levels in Rank Order Tournaments: Field Evidence and Implications for Tournament Design. *RAND Journal of Economics*, 47(1):140–165.

Camerer, C. and Lovallo, D. (1999). Overconfidence and Excess Entry: An Experimental Approach. *American Economic Review*, 89(1):306–318.

Carpenter, B. J., Matthews, P. H., and Schirm, J. (2010). American Economic Association Tournaments and Office Politics : Evidence from a Real Effort Experiment. *The American Economic Review*, 100(1):504–517.

Charness, G. and Levine, D. (2004). Sabotage! Survey Evidence of When it is Acceptable. *Center for Responsible Business, UC Berkeley Working paper*, 12:1–31.

Charness, G., Masclet, D., and Villeval, M. C. (2014). The Dark Side of Competition for Status. *Management Science*, 60(1):38–55.

Chen, K.-P. (2003). Sabotage in Promotion Tournaments. *Journal of Law, Economics, and Organization*, 19(1):119–140.

Clark, J. and Friessen, L. (2009). Rational expectations of own performance: An experimental study. *Economic Journal*, 119(2004):229–251.

Criscuolo, P., Dahlander, L., Grohsjean, T., and Salter, A. (2017). Evaluating Novelty : The Role of Panels in the Selection of R & D Projects. *Academy of Management Journal*, 60(2):433–460.

Decety, J. and Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Deutscher, C., Frick, B., Gürtler, O., and Prinz, J. (2013). Sabotage in Tournaments with Heterogeneous Contestants: Empirical Evidence from the Soccer Pitch. *Scandinavian Journal of Economics*, 115(4):1138–1157.

Dixit, A. (1987). Strategic Behavior in Contests. *The American Economic Review*, 77(5):891–898.

Drago, R. and Garvey, G. T. (1998). Incentives for Helping on the Job: Theory and Evidence. *Journal of Labor Economics*, 16(1):1–25.

Edelman, B. and Larkin, I. (2015). Social Comparisons and Deception Across Workplace Hierarchies: Field and Experimental Evidence. *Organization Science*, 26(1):78–98.

Edwards Mark R. and Ewen, A. J. (1996). How to manage performance and pay with 360-degree feedback. *Compensation and Benefits Review*, 28(3):41–46.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

Engelmann, D. and Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, 67(2):399–407.

Falk, A., Fehr, E., and Huffman, D. (2008). The power and limits of tournament incentives. *Working Paper*, pages 1–44.

Fehr, E. and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4):185–190.

Franke, N., Schreier, M., and Kaiser, U. (2010). The I Designed It Myself Effect in Mass Customization. *Management Science*, 56(1):125–140.

Garicano, L., Palacios-Huerta, I., and Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, 87(2):208–216.

Gaure, S. (2013). lfe: Linear group fixed effects. *The R Journal*, 5(2):104–116.

Ghorpade, J. (2000). Managing five paradoxes of 360-degree feedback. *Academy of Management Executive*, 14(1):140–150.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108(2):419–453.

Gradstein, M. (1995). Intensity of competition, entry and entry deterrence in rent seeking contests. *Economics & Politics*, 7(1):79–91.

Greene, W. W. H. . (2012). *Econometric analysis*. Prentice Hall, New Jersey, 6th edition.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209.

Harbring, C. and Irlenbusch, B. (2005). Incentives in Tournaments with Endogenous Prize Selection. *Journal of Institutional and Theoretical Economics*, 161(4):636–663.

Harbring, C. and Irlenbusch, B. (2011). Sabotage in Tournaments: Evidence from a Laboratory Experiment. *Management Science*, 57(4):611–627.

Harbring, C., Irlenbusch, B., Kräkel, M., and Selten, R. (2007). Sabotage in corporate contests - An experimental analysis. *International Journal of the Economics of Business*, 14(3):367–392.

Hausman, C. and Rapson, D. S. (2018). Regression Discontinuity in Time: Considerations for Empirical Applications. *Annual Review of Resource Economics*, 12(4):1–20.

Howe, J. (2008). *Crowdsourcing*. Random House, London.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Kim, K., Chung, K., and Lim, N. (2019). Third-party reviews and quality provision. *Management Science, Forthcoming*.

Konrad, K. A. (2000). Sabotage in Rent-Seeking Contests. *Journal of Law, Economics, and Organization*, 16(1):155–165.

Lakhani, K. R. and Kanji, Z. (2008). *Threadless: The Business of Community*. Harvard Business School Multimedia/Video Case.

Lazear, E. P. (1989). Pay Equality and Industrial Politics. *Journal of Political Economy*, 97(3):561.

Li, D. (2017). Expertise vs . Bias in Evaluation : Evidence from the NIH. 9(2):60–92.

Li, D. and Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, 348(6233):434–438.

Milgrom, P. and Roberts, J. (1988a). An Economic Approach to Influence Activities in Organizations. *American Journal of Sociology*, 94:154–179.

Milgrom, P. and Roberts, J. (1988b). An Economic Approach to Influence Activities in Organizations.

Milgrom, P. R. (1988). Employment Contracts, Influence Activities, and Efficient Organization Design. *Journal of Political Economy*, 96(1):42–60.

Moffitt, R. (2001). Policy interventions, low-level equilibria, and social interactions. *Social Dynamics*, 4(45-82):6–17.

Moldovanu, B. and Sela, A. (2001). The Optimal Allocation of Prizes in Contests. *The American Economic Review*, 91(3):542–558.

Morgan, J., Orzen, H., and Sefton, M. (2012). Endogenous entry in contests. *Economic Theory*, 51(2):435–463.

Münster, J. (2007). Selection tournaments, sabotage, and participation. *Journal of Economics and Management Strategy*, 16(4):943–970.

Naroditskiy, V., Jennings, N. R., Van Hentenryck, P., and Cebrian, M. (2014). Crowdsourcing contest dilemma. *Journal of The Royal Society Interface*, 11(99).

Nickell, J. and Kalmikoff, J. (2010). *Threadless: Teen Years of T-shirts from the World's Most Inspiring Online Design Community*.

Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291.

O'Reilly, C. A. and Wade, J. (1993). Top Executive Pay: Tournament or Teamwork? *Journal of Labor Economics*, 11(4):606–628.

Pickard, G., Pan, W., Rahwan, I., Cebrian, M., Crane, R., Madan, A., and Pentland, A. (2011). Time-Critical Social Mobilization. *Science*, 334(28):509–513.

Piezunka, H. and Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizations filtering of suggestions in crowdsourcing. *Academy of Management Journal*, 58(3):856–880.

Prendergast, C. and Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, 37(2-3):355–365.

Rahwan, I. (2014). How crowdsourcing turned on me. *Nautilus*.

Riedl, C. and Seidel, V. P. (2018). Learning from mixed signals in online innovation communities. *Organization Science*, 29(6):1010–1032.

Schotter, A. and Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *The Quarterly Journal of Economics*, 107(2):511–539.

Schweitzer, M. E., Ordóñez, L., and Douma, B. (2004). Goal setting as a motivator of unethical behavior. *Academy of Management Journal*, 47(3):422–432.

Stefanovitch, N., Alshamsi, A., Cebrian, M., and Rahwan, I. (2014). Error and attack tolerance of collective problem solving: The darpa shredder challenge. *EPJ Data Science*, 3(1):1–27.

Tullock, G. (1980). Toward a theory of the rent-seeking society. chapter Efficient. Texas A&M University Press, College Station.

Vandegrift, D. and Yavas, A. (2010). An Experimental Test of Sabotage in Tournaments. *Journal of Institutional and Theoretical Economics*, 166(2):259–285.

Vojnović, M. (2016). *Contest theory: Incentive mechanisms and ranking methods*. Cambridge University Press.

Appendix

A Model: Bounds and Proofs

In this section, we provide details for the calculation of bounds when agents of a specific type engage in strategic behavior, as well as proofs for ordering of these bounds so as to establish the full set of possible equilibria.

A.1 Lemma 1: Expected gain from self-promotion is higher for a low type than a high type

Proof. The expected gain from self-promotion for a currently non-self-promoting high type agent i is $\frac{v_i+\Delta p_h}{S+\Delta p_h} - \frac{v_i}{S} = \frac{\Delta p_h(S-v_i)}{S(S+\Delta p_h)}$, and $\frac{v_k+\Delta p_l}{S+\Delta p_l} - \frac{v_k}{S} = \frac{\Delta p_l(S-v_k)}{S(S+\Delta p_l)}$ for a currently non-self-promoting low type agent k. Considering the worst case value for a high ability agent gives us $v_i \geq b_h(n+1)$ (i.e., sincere rating by all n outsiders and i herself, all l and $h-1$ sabotage and rate 0). Conversely, the best case value for a low type gives us $v_k \leq b_l(n+l+h)$ (i.e., sincere rating by everyone).

In a contest where the performance gap between high and low types is large with respect to the proportion of neutral outsiders n (who vote truthfully) and the overall contest size (where l and h might potentially sabotage) we get $S > v_i > v_k$. Define g as the performance gap between high and low types so that $v_i = gv_k$. Expressing g as the inequality between the worst case value for a high type (receiving sabotage from all other high types and all low types) and the best case value for a low type (not being sabotaged at all)

$$g = \frac{b_h}{b_l}\frac{n+1}{n+l+h}. \tag{5}$$

This expression gives a lower bound on g for a contest with a positive performance gap. A positive performance gap ensures that the high type values v_i are higher than the low type values v_k even if the high types are being sabotaged by all H_{-i} and L agents. This ensures that the $b_l < b_h$ relationship also holds for the values $v_k < v_i$. Note that this expresses the necessary size of the performance gap as a product of the difference in production ability and the proportion of neutral outsiders to overall contest size. Considering the smallest possible contest will have $n = 3, l = 2, h = 1$ provides a lower bound for g so that

$$g = \frac{b_h}{b_l}\frac{n+1}{n+l+h} \geq \frac{b_h}{b_l}\frac{2}{3} \tag{6}$$

Now consider a contest with a significant performance gap between high and low types so

that $g \geq 1$ holds, then we can compare the expected gain from self-promotion between high and low types.

$$\frac{\Delta p_l(S - v_k)}{S(S + \Delta p_l)} - \frac{\Delta p_h(S - v_i)}{S(S + \Delta p_h)} = \frac{S((\Delta p_l - \Delta p_h)S + v_i - \Delta p_l v_l) + \Delta p_l(v_i - v_k)}{S(S + \Delta p_h)(S + \Delta p_l)}. \tag{7}$$

This expression is > 0 as can be seen by substituting gv_k for v_i and $1 - b_l$ for Δp_l:

$$\begin{aligned} S((\Delta p_l - \Delta p_h)S + v_i - \Delta p_l v_k) \geq \\ S[(\Delta p_l - \Delta p_h)S + gv_k - (1 - b_l)v_k] = \\ S[(\Delta p_l - \Delta p_h)S + (g - 1 + b_l)v_k] > 0 \end{aligned} \tag{8}$$

Thus, in contests where the performance gap between low and high types is large enough such that $g \geq 1$ is satisfied, the expected gain from self-promotion for a low type is larger than for a high type. □

A.2 High Types Start Sabotaging High Types

Consider a high type of value v_i not currently sabotaging anyone and sincere voting by everyone else so that $v_i = b_h(n+l+h)$ and $v_k = b_l(n+l+h)$, and $S = hv_i + lv_k$. High types will sabotage other high types if the cost of sabotage c_s is lower than the maximum marginal increase of sabotaging the last unit (i.e., going from sabotaging $h-2$ to $h-1$).

$$\frac{v_i}{S - b_h(h-1)} - \frac{v_i}{S - b_h(h-2)} \tag{9}$$

$$\Leftrightarrow \quad \frac{v_i(S - b_h(h-2)) - v_i(S - b_h(h-1))}{(S - b_h(h-1))(S - b_h(h-2))} \tag{10}$$

$$\Leftrightarrow \quad \frac{v_i S - v_i b_h(h-2) - v_i S + v_i b_h(h-1)}{(S - b_h(h-1))(S - b_h(h-2))} \tag{11}$$

$$\Leftrightarrow \quad \frac{v_i b_h}{(S - b_h(h-1))(S - b_h(h-2))} \tag{12}$$

A.3 High Types Start Sabotaging Low Types

There are two possible transition points into this equilibrium: High types may start sabotaging low types while high types already sabotage other high types, but no sabotage otherwise. Or

high types may start sabotaging low types, after low types have already started to sabotage high types. That is, the exact bound for high types to sabotage low types, will depend on whether it falls before or after the bound of low types sabotaging high types (in terms of c_s) as this determines the behavior of low types, which in turn affects the costs of sabotaging behavior for high types. We compute both bounds and then compare them to establish this sequence.

A.3.1 Option H1: High types sabotage all other high types and low types do not sabotage anyone

Consider a high type of value v_i not currently sabotaging anyone and sincere voting by everyone else so that $v_i = b_h(n+l+h)$ and $v_k = b_l(n+l+h)$, and $S = hv_i + lv_k$. In this equilibrium, h high types vote $-b_h$ on $h-1$ others.

$$\frac{v_i}{S - hb_h(h-1) - b_l l} - \frac{v_i}{S - hb_h(h-1) - b_l(l-1)} \tag{13}$$

$$\Leftrightarrow \quad \frac{v_i(S - hb_h(h-1) - b_l(l-1)) - v_i(S - hb_h(h-1) - b_l l)}{(S - hb_h(h-1) - b_l l)(S - hb_h(h-1) - b_l(l-1))} \tag{14}$$

$$\Leftrightarrow \quad \frac{v_i S - v_i hb_h(h-1) - v_i b_l(l-1) - v_i S + v_i hb_h(h-1) + v_i b_l l}{(S - hb_h(h-1) - b_l l)(S - hb_h(h-1) - b_l(l-1))} \tag{15}$$

$$\Leftrightarrow \quad \frac{v_i b_l}{(S - hb_h(h-1) - b_l l)(S - hb_h(h-1) - b_l(l-1))} \tag{16}$$

A.3.2 Option H2: High and low types sabotage all high types

Consider a high type of value v_i not currently sabotaging anyone and sincere voting by everyone else so that $v_i = b_h(n+l+h)$ and $v_k = b_l(n+l+h)$, and $S = hv_i + lv_k$. In this equilibrium, h high types vote $-b_h$ on $h-1$ others and l low types vote $-b_h$ on h high types. This turns out to be the relevant bound (see proof below).

$$\frac{v_i}{S - hb_h(h-1) - lb_h h - b_l l} - \frac{v_i}{S - hb_h(h-1) - lb_h h - b_l(l-1)} \tag{17}$$

$$\Leftrightarrow \quad \frac{v_i(S - hb_h(h-1) - lb_h h - b_l(l-1)) - v_i(S - hb_h(h-1) - lb_h h - b_l l)}{(S - hb_h(h-1) - lb_h h - b_l l)(S - hb_h(h-1) - lb_h h - b_l l)} \tag{18}$$

$$\Leftrightarrow \quad \frac{v_i S - v_i hb_h(h-1) - v_i lb_h h - v_i b_l(l-1) - v_i S + v_i hb_h(h-1) + v_i lb_h h + v_i b_l l}{(S - hb_h(h-1) - lb_h h - b_l l)(S - hb_h(h-1) - lb_h h - b_l l)} \tag{19}$$

$$\Leftrightarrow \quad \frac{v_i b_l}{(S - hb_h(h-1) - lb_h h - b_l l)(S - hb_h(h-1) - lb_h h - b_l l)} \tag{20}$$

$$\Leftrightarrow \quad \frac{v_i b_l}{(S - hb_h(h-l-1) - b_l l)(S - hb_h(h-l-1) - b_l l)} \tag{21}$$

A.4 Low Types Start Sabotaging High Types

Similarly, there are two possible bounds when low types start to sabotage high types: one in the equilibrium where high types high types sabotage low types and one where they do not.

A.4.1 Option L1: High types sabotage only other high types, but no low types

Consider a high type of value v_i not currently sabotaging anyone and sincere voting by everyone else so that $v_i = b_h(n+l+h)$ and $v_k = b_l(n+l+h)$, and $S = hv_i + lv_k$. In this equilibrium, h high types vote $-b_h$ on $h-1$ all others (high and low types) and low type start sabotaging h high types with $-b_h$. This turns out to be the relevant bound (see proof below).

$$\frac{v_k}{S - hb_h(h-1) - b_h h} - \frac{v_k}{S - hb_h(h-1) - b_h(h-1)} \tag{22}$$

$$\frac{v_k b_h}{(S - hb_h(h-1) - b_h h)(S - hb_h(h-1) - b_h(h-1))} \tag{23}$$

$$\frac{v_k b_h}{(S - hb_h h)(S - b_h h^2 + b_h)} \tag{24}$$

A.4.2 Option L2: High types sabotage all other high types and low types

Consider a high type of value v_i not currently sabotaging anyone and sincere voting by everyone else so that $v_i = b_h(n+l+h)$ and $v_k = b_l(n+l+h)$, and $S = hv_i + lv_k$. In this equilibrium, h high types vote $-b_h$ on $h-1$ other high types and vote $-b_l$ on l low types and low types start sabotaging h high types with $-b_h$.

$$\frac{v_k}{S - hb_h(h-1) - hb_l l - b_h h} - \frac{v_k}{S - hb_h(h-1) - hb_l l - b_h(h-1)} \tag{25}$$

$$\frac{v_k b_h}{(S - hb_h(h-1) - hb_l l - b_h h)(S - hb_h(h-1) - hb_l l - b_h(h-1))} \tag{26}$$

$$\frac{v_k b_h}{(S - hb_h h - hb_l l)(S + b_h - b_h h^2 - hb_l l)} \tag{27}$$

A.5 Proof: Order of Bounds

To establish the correct order of bounds, we show that maximum marginal gain for low types to sabotage high types is higher than the maximum marginal gain for high types to sabotage low types.

$$\frac{v_i b_l}{(S - hb_h(h-1) - b_l l)(S - hb_h(h-1) - b_l(l-1))} < \frac{v_k b_h}{(S - hb_h h)(S - b_h h^2 + b_h)} \tag{28}$$

Proof. This inequality holds given $v_k < v_i$ and $b_l < b_h$ which are true by construction. □

This shows that low types have more to gain from sabotaging high types than high types have to gain from sabotaging low types. Consequently, as the cost for sabotage decrease, low types will switch from not sabotaging anyone to sabotaging high types before high types will switch from sabotaging only other high types to sabotaging all other high types and low types. As a consequence, the relevant bound for low types switching to sabotaging high types is the one where only high types sabotage high types but no low types (option H2 above). Furthermore, the correct bound for high types to sabotage low types is one where everyone already sabotages high types (option L1 above). Given that we have now established a sequence of bounds, we establish that the set of possible Nash equilibria is NE = {no sab, h sab h, l sab h, h sab l, l sab l}

B Empirical Analysis

In this section we provide additional empirical analyses.

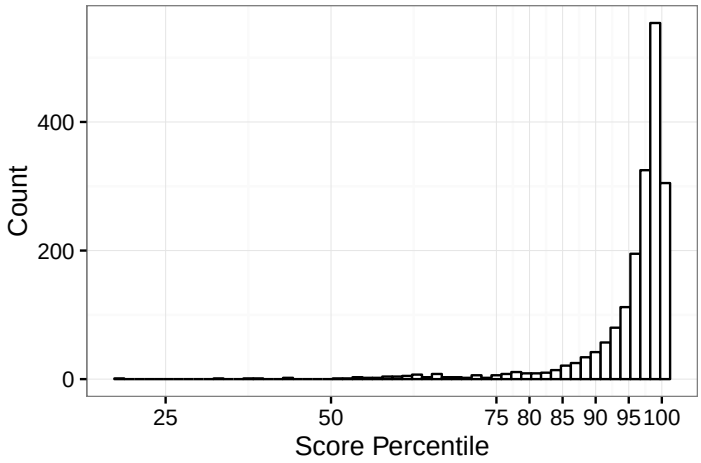


Figure 7: Percentile score of contest winners. Contest winners scored in the highest percentiles among all submissions in a given week. The median percentile score was 98th percentile.