Economics Bulletin

Volume 37, Issue 1

Testing for Malice

Brishti Guha
Jawaharlal Nehru University

Abstract

Consider two parties disputing claims over an indivisible prize. A malicious claimant may or may not intrinsically value the prize for its own sake, but always derives pleasure "malice utility" from depriving the rival claimant. I devise a method for detecting malice in experimental settings. I derive a simple mechanism which allows third parties (such as experimenters) to distinguish whether (i) both claimants bear each other malice (two-sided malice) (ii) whether only one claimant bears the other malice (one-sided malice) and if so, the identity of this malicious claimant, and (iii) whether neither of the claimants are maliciously motivated. I show that, with slight modifications, this mechanism is applicable both to the case where the claimants know each other and to the case where they are strangers. I also discuss a method whereby the experimenter may infer an upper bound on the malice of the less malicious party in the case of two-sided malice.

I would like to thank the editor, Professor John Conley, and an anonymous referee.
Citation: Brishti Guha, (2017) "Testing for Malice", *Economics Bulletin*, Volume 37, Issue 1, pages 327-335
Contact: Brishti Guha - brishtiguha@gmail.com.
Submitted: November 26, 2016. **Published:** February 22, 2017.

1. Introduction

An individual exhibits "malice" (spite) when she gets some utility simply when another individual is deprived – without this outcome necessarily translating into a direct economic benefit to herself. This paper connects two strands of the literature, one being the behavioral/experimental literature on malice in decision-making, and the second being the game theoretic literature on King Solomon's problem. The experimental literature on the importance of malice and envy in decision-making, includes, among others, Beckman et al (2002), Bosman and van Winden (2002), Bosman et al (2006), Albert and Mertins (2008), Zizzo and Oswald (2001), Abbink and Sadrieh (2008), and Abbink and Herrmann (2011). These papers provide considerable evidence that malice matters. For example, Beckman et al find that 50% of their experimental subjects oppose Pareto improvements that make others better off without making them any worse off. Bosman and van Winden (2002) find, in a "power to take" experiment, that 21% of their subjects destroy their own earnings when told that a portion of these earnings would later go to another subject. Zizzo and Oswald (2001) find, in a "money burning" game, that two-thirds of their subjects were actually willing to pay in order to destroy others' earnings. Abbink and Herrmann (2011) use a one-shot "joy of destruction" game to show that 10-25% of their subjects destroyed others' endowments without any economic gain to themselves.

King Solomon's problem – inspired by the biblical story of two women who both petitioned King Solomon, claiming the same child – has been interpreted by economists as the problem of allocating an indivisible prize between a high-valuation and a low-valuation claimant, when the mechanism designer is unaware of which claimant has the higher valuation (Glazer and Ma 1989, Moore 1992, Perry and Reny 1999, Olszewski 2003, Bag and Sabourian 2005, Artemov 2006, Qin and Yang 2009, Mihara 2012). This game theoretic literature does not, however, discuss malice.

Guha (2014) incorporates malice into this traditional mechanism design problem. A malicious claimant is driven by the pleasure she gets, not from possessing the prize herself (though she may place a positive, albeit low, intrinsic valuation on the prize), but from depriving her rival. Thus, she obtains a "malice utility" whenever her rival does not get the prize, which includes contexts where the mechanism designer simply retains the prize without giving it to either claimant.[1] The paper derived a simple mechanism which allocates the prize to the claimant with the higher *intrinsic* valuation for the prize, at no cost. The mechanism designer was aware that there was malice involved, but did not know (in cases of one-sided malice) which claimant was malicious; nor did he know the actual valuations or the actual extent of malice. Relevant real-life applications included the case of separating spouses, one or both of whom may bear the other malice, contesting ownership of an asset; extended-family land ownership disputes; patent

[1]Guha (2014) also argued that the original problem of two mothers petitioning King Solomon involved one-sided malice with the false mother claiming the child simply out of spite for the true mother. This was indicated by the fact that she preferred that the child be cut in half rather than given to the true mother. She does not obtain a child in either case, but prefers the outcome in which the true mother is deprived.

trolling (patent trolls' main objective in applying for a patent is not to develop the product themselves, but to harm other companies), and child custody.

In the current paper, I examine whether it is possible to detect if malice is or is not present in a setting where two claimants are claiming the same prize. Is it also possible to detect whether – in case malice is present – it is one-sided or two sided? If it is one-sided, is it possible to detect the identity of the malicious claimant? I explore these issues in this paper, and show that the answer to all of these questions is yes. I thus devise a method which experimenters can use to detect whether malice is absent, one-sided or two-sided in a setting where two claimants both claim an indivisible prize (as well as uncovering the identity of the malicious claimant in one-sided malice). My method involves designing a game (or mechanism) which yields different outcomes in each of these cases; the outcome of the game can allow us to infer if malice is two-sided, one-sided or absent.

This mechanism is different from that in Guha (2014), where the outcome will not indicate (to the designer or other third parties) whether malice was present. The mechanism here also differs from the previous paper's, in that its only objective is to allow third parties such as experimenters to make deductions about malice; while in the previous paper the objective was to ensure that the claimant with the higher intrinsic valuation obtained the prize. Happily, the designer here does not require any information about the claimants (unlike in the previous paper). Moreover, the claimants themselves need less information. When the claimants know each other, they do know if the other claimant bears them malice (but do not know the extent or the other claimant's true valuation). However I also explicitly consider the case where the claimants are strangers to one another and therefore have no information even about whether the rival claimant is malicious. As a side contribution, I also show how, in the event that malice is two-sided, the experimenter may infer an upper bound on the malice of the less malicious party.

Thus, this paper devises a method of testing for malice (and identifying the malicious party in the case of one-sided malice) in any setting where two claimants are contesting for a single prize. It uses a game-theoretic approach to devising a method which can be of use to experimenters, thus linking the game theoretic and the experimental literatures.

The rest of the paper is organized as follows. Section 2 presents the main results, devising a mechanism which can be used to detect whether malice is absent, one-sided or two-sided, and from which the identity of the malicious party can also be inferred. I discuss both the case where the claimants know each other and the case where they are strangers. A subsidiary result discusses a method of obtaining an upper bound on the malice utility of the less malicious party in the event of two-sided malice. Section 3 concludes.

2. A Mechanism to Detect Malice

2.1 Detecting malice in non-anonymous interactions

In this sub-section I design a mechanism whose purpose is to make deductions about the absence or presence (and one or two sidedness) of malice in two-person interactions where both claimants are contesting over an indivisible prize and *know each other.* Here, malice may be a

product of the personal history of the two people, or may just be a product of ill nature. More precisely, the *information structure* is as follows.

2.1.1 Information structure and payoffs

1. Both players A and B know their own valuations (V for A and v for B, where V>v w.l.o.g) and whether they bear the other player malice. If they are malicious, they obtain a positive "malice utility" whenever the other party does not obtain the object. These malice utilities, if positive, are denoted by λ for A and κ for B; a non-malicious player has a malice utility of zero.
2. In addition, each player knows if the other player bears her malice, but has no information on the extent of this malice, and knows nothing about the other player's valuation.[2]
3. The mechanism designer (experimenter) knows nothing.[3] The designer has an indivisible object which he allocates using a mechanism to be specified below.

In cases where the mechanism involves repeated rounds of play, we assume that the designer starts off with a sufficient number of indivisible objects to repeat the experiment even if the earlier rounds result in one of the players being allocated an object. While, for convenience, we fix the players' valuations across rounds, nothing is lost if we allow these valuations to vary.

Besides deducing whether malice is absent, one-sided or two-sided, I will also show how the experimenter may (i) deduce the identity of the malicious party if malice is one-sided, and (ii) when malice is two-sided, estimate an upper bound on the malice of the less malicious party. In a later sub-section, I will also discuss how similar deductions can be made for the case where the two claimants are *strangers* (in which case malice would merely be a product of ill nature rather than history, and a claimant would not know beforehand if the other claimant was or was not malicious).

2.1.2 The Mechanisms

In the experiment below, the designer designs a two-stage game (a two-step mechanism). Depending on the outcome in Stage 1, the designer may or may not proceed with Stage 2. However, note that the participants are not informed of the possibility of a second stage. Thus, while entering Stage 1, they behave as if there would be no further play, enabling us to analyze their behavior in the two stages independently. Our solution concept is pure strategy Nash equilibrium.

*Definition 1.*Define a mechanism Γ such that, in Stage 1, the designer (experimenter) asks both claimants to state "mine" or "hers" *sequentially*, with the second claimant observing the first claimant's statement, and

[2] This is in contrast to Guha (2014), where the players might need to know an upper bound on the valuation of the low-valuation party, a lower bound on the valuation of the high-valuation party, and an upper bound on the malice of the more malicious party.

[3] Again, this is in contrast to Guha(2014), where the designer needs to have the information mentioned in the previous footnote.

(i) If one says "mine" and the other says "hers", allocates the prize to the one who said "mine".
(ii) If both say "mine", the designer retains the prize and charges both claimants an infinitesimally small fee of ε each.
(iii) If both say "hers", the designer conducts a lottery randomly assigning the prize to either claimant with probability half.

If both players stated "mine" in Stage 1, the designer does not implement a second stage. Otherwise, the designer proceeds to Stage 2, in which he again announces and implements Stage 1, but changes the identity of the first mover (that is, if A were the first mover in the previous step, A now moves second). As stated earlier, during Step 1, the players are not informed of the possibility of a second step.

We now proceed to our main result.

Proposition 1.*(i) Let the designer implement the mechanism* Γ. *Then, both disputants say "mine" and the mechanism stops after Stage 1 with the designer keeping the prize, if and only if two-sided malice is present. Otherwise, one says "mine" and the other says "hers" in Stage 1.*

(ii) If the mechanism proceeds to Stage 2, and the outcome in Stage 2 remains the same as that of Stage 1 regardless of which player moves first, there is one-sided malice with the malicious party saying "mine" while the non-malicious one says "hers". If the outcome changes when player order changes, with the first mover in either case obtaining the prize, then malice is absent.

Proof: (i) Suppose there is two-sided malice. Then, the payoff structure resulting from Stage 1(allowing A to be the first mover w.l.o.g) can be depicted in Figure 1.

Since ε is infinitesimally small, we have $\lambda,\kappa>\varepsilon$. It is then easy to see that saying "mine" is the dominant strategy for the second mover. Knowing this, the first mover says "mine" and the designer keeps the prize. This establishes the "if" component of part (i). (One can check that exactly the same logic holds if B were the first mover, instead of the case depicted in Figure 1).

To show that this outcome will not obtain if malice is one-sided or absent, first consider one-sided malice. There are two possibilities, (i) the first mover is malicious, and (ii) the second mover is malicious. Consider case (i) first, setting $\kappa = 0$ in the game tree above. Then, the first mover works out that the second mover will say "hers" if the first mover says "mine", and "mine" if the first mover says "hers". Since the former is better for the first mover, she says "mine" while the second mover says "hers" since $-\varepsilon < 0$. Now consider case (ii), setting $\lambda = 0$ in Figure 1. Then, the first mover knows that the second mover will always say "mine". The first mover then says "hers", since $-\varepsilon < 0$.

Next, consider absence of malice, setting $\lambda = \kappa = 0$ in Figure 1 above. Now, knowing that the second mover will say "hers" if the first mover says "mine", and "mine" if the first mover says "hers", the first mover says "mine". Thus, the NE involves the first mover saying "mine" and the second saying "hers".

This establishes the "only if" part; unless malice is two-sided, therefore, the outcome will be that one of the claimants says "mine" and the other says "hers" during Stage 1. Therefore, the mechanism proceeds to the second step if and only if two-sided malice is absent.

(ii) Suppose malice is one-sided with B being the malicious party (again w.l.o.g). Stage 1 is implemented; suppose, as in Figure 1, A is the first mover in Stage 1. This corresponds to case (ii) above; knowing that B will always say "mine", A says "hers" as $0 > -\varepsilon$. Thus, B is given the object. Now, let the experimenter proceed to Stage 2 (with another unit of the same object), this time giving B the first move. B realizes that A will say "mine" if B says "hers", and will say "hers" if B says "mine", as $0 > -\varepsilon$. Since the latter is better for B, she says "mine", while A says "hers". Thus, B again obtains the object. Therefore, changing the first mover does not change the outcome. Moreover, the malicious party is the one who said "mine", while the non-malicious one says "hers".

Now consider absence of malice. If A moves first (in either stage) she realizes that B's best response will be "hers" if A says "mine", and "mine" if A says "hers". Since the outcome where A says "mine" and B says "hers" is strictly better for A than the one where A says "hers" and B says "mine" (since $V>0$), A will say "mine". Playing her best response, B then says "hers" and A obtains the prize. If B moves first, then an exactly analogous argument ensures that B selects the outcome where B says "mine" and A, in response, says "hers", so that B obtains the prize. Thus, the first mover in either stage always obtains the prize if and only if[4] malice is absent; changing the player order changes the outcome. ***QED***

The intuition underlying Proposition 1 is straightforward. A malicious person is anxious that the other claimant not obtain the object, and is willing to pay the designer a small amount for the designer to retain the object; she always prefers this outcome to letting the rival claimant obtain the object. Therefore, if her rival says "mine", she would like to say "mine" as well. If her rival says "hers", then she will always say "mine", since by doing so she can obtain the prize while otherwise she will only get the prize with probability half. Therefore, saying "mine" is a dominant strategy for a malicious person, but not for a non-malicious person. The latter strictly prefers the outcome in which the rival obtains the object to the one where the designer keeps the object but imposes a tiny fee, as she does not obtain the prize in either outcome, and avoids the tiny fee in the former. This asymmetry can then be exploited in making the mechanism distinguish between cases where malice is two-sided, one-sided and absent. Varying the order of moves helps distinguish between one-sided malice and absence of malice; this also fixes the identity of the malicious party if there is one-sided malice.

Estimating the magnitude of malice

Corollary 1. If an experimenter establishes that two-sided malice exists, by implementing the mechanism of Proposition 1, he can then progressively increase ε in small increments. The level of ε at which one party switches her statement from "mine" to "hers" marks an upper bound on that party's malice utility.

[4]If Stage 2 is implemented for two-sided malice, the outcome would always be "mine, mine" regardless of player order.

Remark: The mechanism above does not involve subsidies made by the experimenters to the subjects. Therefore, it does not generate incentives to collude.

2.2 A test for strangers

What if the two disputants claiming the prize are unknown to each other? We discuss this possibility, in addition to the possibility where subjects are known to each other, because an experimenter may in practice be able to implement both types of experiments (the second type can be implemented even if the subjects actually know each other but the experimenter succeeds in concealing their identity from each other). On one hand, if subjects do not know whom they are competing against, there may be less basis for malice; however, people who are malicious or envious by nature may derive pleasure from depriving any rival, even if he or she is a stranger. Now, the test devised in Proposition 1 needs to be modified because of the fact that (in addition to the experimenter having no information), neither of the claimants knows if the other bears her malice.

2.2.1 Information structure and payoffs

1. Both players know their own valuations (V for A and v for B, where V>v w.l.o.g) and whether they bear the other player malice. These malice utilities are denoted by λ for A and κ for B as usual. They do not know if the other player is malicious.
2. The mechanism designer knows nothing.

Definition 2. Denote the normal-form version of Stage 1 of mechanism Γ by G. That is, the experimenter asks both players to state "mine" or "hers" *simultaneously*. He then follows the same allocation procedure as in Stage 1 of mechanism Γ, so that in case of two-sided malice, we have the following payoff matrix:

Table 1: The Payoff Matrix in mechanism G

B

	Mine	**Hers**
Mine	(λ-ε, κ-ε)	(λ+V, 0)
Hers	(0, κ+v)	((λ+V)/2, (κ+v)/2)

where A is the row player and B the column player.

2.2.2 The Mechanism

Proposition 2.*Suppose A and B do not know each other. Then, if the designer repeatedly implements G,*

(i) With two-sided malice, both A and B keep saying "mine" in all rounds.
(ii) With one-sided malice, one of the claimants – the malicious one – keeps saying "mine" in all rounds while the other switches to "hers" at least in the later rounds.
(iii) If neither party is malicious, neither consistently says "mine".

Proof: Since ε is infinitesimally small, we have λ,κ>ε with two-sided malice. It is then easy to see that saying "mine" is the dominant strategy for a malicious player so that in the DSE with

two-sided malice, both players say "mine" in every round of the repeated game. With one-sided malice, the malicious player says "mine" in every round for the same reason. The non-malicious party initially does not know that her opponent is malicious; however after a few rounds she learns that her opponent always says "mine" and therefore systematically begins saying "hers" which is her best response to "mine". If both A and B are non-malicious, however, they lack a dominant strategy and hence neither will consistently say "mine" (or "hers"). ***QED***

Thus, even in the case of strangers, it is possible for the experimenter to distinguish between cases where malice is two-sided, one-sided, or absent, and to deduce the identity of the malicious party under one-sided malice. Moreover, if malice is two-sided, Corollary 1 can still be applied to infer an upper bound on the malice of the less malicious party.

3. Conclusion

This paper devises game-theory based tests that third parties – primarily experimenters – can use to detect if malice is absent, one-sided or two-sided in a "King Solomon" like setting, that is a setting in which two individuals both want the same (indivisible) prize. I develop a new mechanism that does this (the one in Guha 2014 does not help make deductions about this) by resulting in different observable outcomes based on whether malice is absent, one-sided or two-sided. This mechanism also requires minimal information. It assumes no prior information on the part of the designer. When the subjects know each other, the mechanism does assume that they know if their rival claimant bears them malice (but no other knowledge is required). However, I also develop a mechanism for the case where the claimants do not know each other and therefore do not even know if the other claimant is malicious. Here, neither the designer nor the claimants have any information about individuals other than themselves. I also obtain a few subsidiary results. I show how to uncover the identity of the malicious party in cases of one-sided malice, and how to derive an upper bound on the malice of the less malicious party in cases of two-sided malice. Finally, as the new mechanism does not involve subsidies from the designer to the claimants (aside from allocation of the prize itself), it is not vulnerable to collusion.

References

Abbink, K and B. Herrmann (2011) "The Moral Costs of Nastiness" *Economic Inquiry* **49**, 631-633.

Abbink, K and A. Sadrieh (2008) "The Pleasure of Being Nasty" *Economics Letters* **105**, 306-308.

Albert, M and V. Mertins (2008) "Participation and decision making: a three-person power-to-take experiment" Joint Discussion Paper Series in Economics Working Paper No 05-2008.

Artemov, G. (2006) "Imminent Nash implementation as a solution to King Solomon's dilemma" *Economics Bulletin* **4**, 1-8.

Bag, P.K and H. Sabourian (2005) "Distributing Awards Efficiently: More on King Solomon's Problem" *Games and Economic Behavior* **53**, 43-58.

Beckman, S.R, J.P Formby, W. James Smith and B. Zheng (2002) "Envy, malice and Pareto efficiency: an experimental examination" *Social Choice and Welfare* **19**, 349-367.

Bosman, R and F. van Winden (2002) "Emotional hazard in a power-to-take experiment" *Economic Journal* **112**, 146-169.

Bosman, R, H. Hennig-Schmidt and F. van Winden (2006) "Exploring group decision-making in a power-to-take experiment" *Experimental Economics* **9**, 35-51.

Glazer, G and C.T.A Ma (1989) "Efficient Allocation of a "Prize" – King Solomon's Dilemma" *Games and Economic Behavior* **1**, 222-233.

Guha, B. (2014) "Reinterpreting King Solomon's Problem: Malice and Mechanism Design" *Journal of Economic Behavior and Organization* **98**, 125-132.

Mihara, H.R. (2012) "The second-price auction solves King Solomon's dilemma" *Japanese Economic Review* **63**, 420-429.

Moore, J. (1992) "Implementation, Contracts and Renegotiation in Environments with Complete Information" in Laffont, J.J. (ed) *Advances in Economic Theory: Sixth World Congress Volume 1* by J.J. Laffont, Ed., Cambridge University Press: Cambridge, 182-282.

Olszewski, W. (2003) "A simple and general solution to King Solomon's Problem" *Games and Economic Behavior* **42**, 315-318.

Perry, M and P.J. Reny (1999) "A general solution to King Solomon's Dilemma" *Games and Economic Behavior* **26**, 279-285.

Qin, C.Z and C.L. Yang (2009) "Make a guess: a Robust Mechanism for King Solomon's Dilemma" *Economic Theory* **39**, 259-268.

Zizzo, D.J and A.J. Oswald (2001) "Are people willing to pay to reduce others' incomes?" *Annales d' Economieet de Statistique* **63**, 39-65.

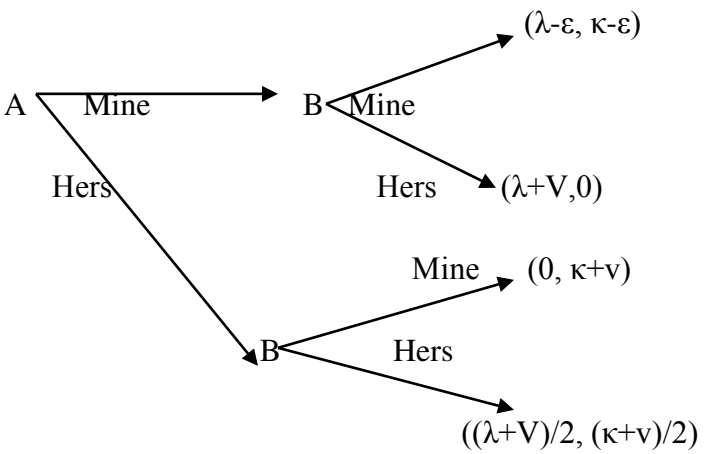

Figure 1: Stage 1 with two-sided Malice