RECOVERING TARNISHED REPUTATIONS AND SHARED UNDERSTANDINGS IN INTERNATIONAL POLITICS

April 4, 2008

Koji Kagotani

Robert F. Trager

Department of Political Science
University of California, Los Angeles

The material structure of the international system is usually not sufficient to determine a unique international social equilibrium. The dynamics of international conflicts are determined by the shared understandings that have developed over time about how other states react to each others' behavior. We show that how states believe reputation is recovered is a key to understanding international crisis behavior. We contrast the case where tarnished reputations recover naturally over time (Sartori 2005) or attach to leaders whose political demise is imminent to the case where states must take action to recover their bargaining reputations. In the prior case, war is most likely when states have not been caught bluffing, while in the latter case, the probability of war is not related to bluffing behavior. Surprisingly, the shared expectation that states must take action to recover reputation diminishes the value of diplomatic communication, and when states need to go to war in order to recover a lost reputation for being willing to go to war, war is less likely.

Prepared for Presentation at the 2008 meetings of the Midwest Political Science Association. Comments are very welcome, but *please do not cite without permission*. The authors can be contacted at kagotani@ucla.edu and rtrager@ucla.edu.

In April and May of 1981, Israel threatened to use military force against Syria if the latter did not moderate its demands against a Lebanese Christian militia known as the Lebanese Front, and remove missiles positioned around the Lebanese town of Zahle. Israeli Prime Minister, Menachim Begin, stated: "There are grounds to assume we will not be content with [Syrian] action." Syria resisted Israeli demands, however, and Israel did not follow through on its promise to use military force. Did Israeli officials worry the country's reputation for following through on commitments had been damaged in a way that would worsen its bargaining leverage in future negotiations? If Israel did have such fears, would it have been more likely to take actions in other contexts to demonstrate its willingness to follow through on commitments – actions like actually employing military force? In particular, was Israel's near coterminous decision to escalate its conflict with the Palestinian Liberation Organization partly a result of the outcome of the crisis with Syria?[1]

In this paper, we examine the implications for the dynamics of international crises of the belief, on the part of foreign policy elites, that reputations for following through on threats must (or can) be recovered by demonstrating a willingness to follow through. That is, we examine the implications of a shared understanding that bargaining reputations are recovered by demonstrating a willingness to go to war. We will show that such a belief significantly undermines the efficacy of diplomacy in crises. We also exploit a simple non-parametric test to argue that the empirical record is most consistent with a reduced view of the efficacy of diplomatic signals that rely on bargaining reputation.

The theoretical discussion employs the model developed by Sartori (2002, 2005). Two states are engaged in an international crisis and have the opportunity to send "costless" signals of intention. Sartori demonstrates the existence of an equilibrium in the model in which diplomacy is made meaningful by the shared expectation that dissemblers will not be listened to for some time to come.[2] As in many recent models of international politics, however, the structure of the strategic

[1] See Lewis and Schultz (2008), Evron (1987), and Kifner (1981).

[2] On the point that the punishment phase can be of any length, see Kurizaki (2007).

context allows for multiple social equilibria. We shall study another equilibrium of the model in which actors believe they must fight to recover their reputations.[3]

The paper addresses several issues in the study of international politics. First, the theoretical discussion helps to illuminate what the material structure of the crisis bargaining context leaves to agency, and what roles even the most minimal aspects of international society play in shaping international outcomes. We show that seemingly small variations in states' shared understandings have large implications for the dynamics of international crises, and that at least some of these effects can be captured in a game theoretic model. Second, with respect to reputational signals sent by state adversaries, we show that when states believe they must fight to recover bargaining reputations that are tarnished (a reasonable conjecture about the beliefs of foreign policy elites, and one which actors would not be disabused of by play on the equilibrium path), private diplomacy loses its effectiveness. Third, we employ a non-parametric test that offers divergent evidence from that presented in Sartori (2005). One implication of the equilibrium studied by Sartori is that war is more likely when states are believed by others to be honest, and thus when their diplomatic signals convey information, than when states are believed to be dishonest. We replicate Sartori's coding of international crises and find that the reverse is the case: states are much more likely to go to war when they have been caught in a bluff. When reputation is coded based on the model we discuss below, however, we find that the prediction of that model is born out: there is no significant difference between the probability of war when states have or have not been caught bluffing in the past.

The paper proceeds as follows. In the next section, we discuss the state of the literature on reputational signaling in crises as well as the role of shared understandings of actor behavior in influencing the course of events in international politics. We then present Sartori's (2005) model and contrast two of its equilibria. The last section analyzes a simple non-parametric test of the predictions of the two equilibria.

[3] See also Trager (2010, 2011, 2012, 2013, 2015a, 2015b).

Reputation and Shared Expectations

One of the paradoxes of the literature on reputation is that while there is general agreement that policy-makers sometimes take actions for the sake of their state's bargaining reputation, it is unclear whether states pay any attention to reputation (or past actions generally) when evaluating the likely future behavior of an adversary. On this latter question, analysts of international politics are divided. Studies of the effect of reputations on bargaining in crises have come to widely divergent conclusions. Press (2004) argues, for instance, that even in cases where we should expect a state's past actions to influence other states' evaluations of it future intentions, we find no such evidence. On the contrary, states evaluate each others' likely behavior based on a current calculus of capability and interest, no matter what has gone on before.[4] Sartori (2005), Jervis (1970, 1997), and Schelling (1966), by contrast, argue that diplomacy sometimes conveys information as a result of states' need to maintain a bargaining reputation.

Our view is that reputational signaling is sometimes effective. The conditions and contexts under which it works, however, are still not fully understood. For one, the efficiency and effect of these diplomatic signals depend on the shared understandings of the international system's functioning possessed by system actors. In fact, as Sartori demonstrated, shared understandings can cause reputational dynamics to emerge from agent interaction even when no characteristic or disposition of agents actually exists for which they are thought to have developed a reputation. Put differently, there are no honest or dishonest "types" in her analysis; rather, a "reputation" for honesty can be said to be entirely socially constructed in an important sense.

The essential role played by such shared understandings has been increasingly recognized in recent years. Wendt (1999), for example, distinguishes between "3 cultures of anarchy" that correspond to three very differently functioning international systems. These differences are driven in part by the ways actors are "socialized" into the different cultures, which affects what goals they seek after, and – importantly –

[4] See also Mercer (1996).

what their expectations are about how other actors in the system will behave. Shared experience in the system leads to shared understandings of the rules by which it functions. Actor expectations may thus converge to common knowledge of the rules of the game in a particular social equilibrium.

Similarly, formal work in international relations has demonstrated the possibility of multiple social equilibria with very different dynamics in a single strategic and material context. Examples include Slantchev (2003), Niou and Ordeshook (1990, 1991), and all cheap talk models, such as Sartori (2005) and Kydd (2003). In these models, the shared understanding of actors of which equilibrium they are in can mean the difference between peace and protracted war.

While most constructivists and modelers have theorized the differing effects of large scale changes in international culture, we shall examine the effects of slight changes in actor understandings and expectations. We show that even these slight changes have large scale effects.

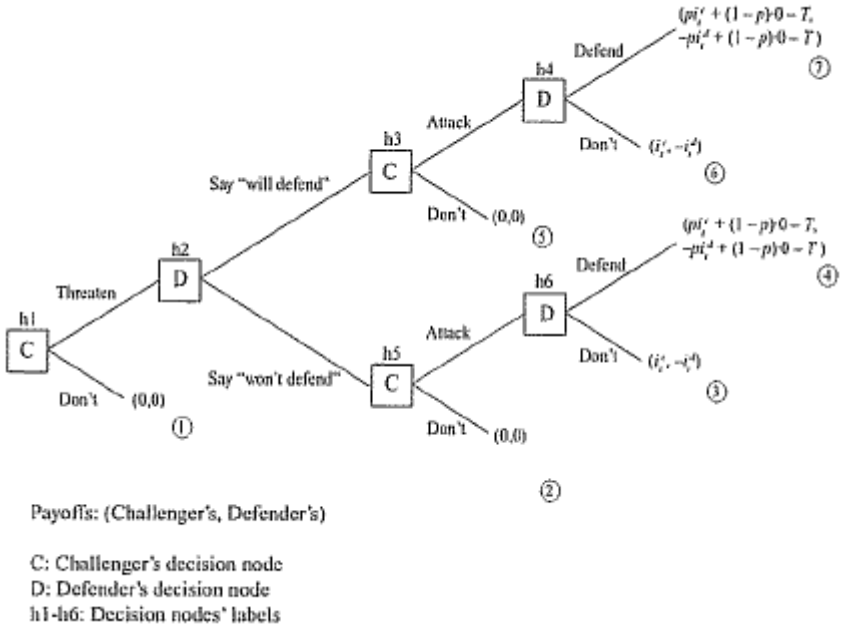
The Model

We use Sartori's (2005) model as a base-line to represent costless diplomatic signaling in an international crisis. A challenger and a defender declare their intention to fight or not over a particular issue and then decide whether or not to follow through. States care more about some issues than others, and do not know each other's subjective evaluations of issue importance. Defenders are involved in similar crises again and again, and make current decisions with the knowledge that these may effect future bargaining in addition to the outcome of the current crisis. We will investigate the defender's behavior over the long-term, including the conditions under which diplomacy is informative.

We formalize a series of crises as the infinitely iterated game where only the defender is repeatedly involved in crises. Each crisis between a randomly chosen challenger and the defender is described as a stage game of the entire game at the period t $(=1,2,\ldots,\infty)$. The stage game consists of both the use of diplomacy and the use of force. Figure 1 shows the sequence of moves in the stage game. The challenger decides to

threaten or not (at h1) against the defender. If the challenger chooses not to threaten, a crisis does not happen and the outcome is status quo. If the challenger threatens, the defender decides to say "will defend" or "won't defend" (at h2). Then, the challenger decides to attack or not after it hears the defender's message (at h3 or h4). If the challenger chooses not to attack, a crisis ends and the outcome is status quo. If the challenger attacks, the defender decides to defend or not (at h4 or h6). If the defender defends, a war occurs. Otherwise, the defender appeases to the challenger's demand.

Figure 1: Game Tree



Citation: Sartori (2005) p.132.

We allocate payoffs of both sides to each international outcome in the stage game. Suppose that the challenger demand for a unit of the pie from the defender's material possessions and that the size of the bargained pie is normalized to one. We also assume that the challenger and the defender individually place the values, i_t^c and i_t^d, on it. That is, these values represent resolve of both sides for the issue. When all the bargained pie is transferred from the defender to the challenger, this transfer of a

unit of the pie means gain of i_t^c for the challenger and loss of i_t^d for the defender. If the challenger does not threaten or attack, international outcomes remain the status quo and each state receives zero. If a war occurs due to the defender's resist, the challenger will win with probability p and lose with probability $1-p$. When the challenger wins, it takes all the pie under a dispute and receives i_t^c, and the defender receives $-i_t^d$. When the challenger loses, the defender keeps all the pie under a dispute and each state receives zero because nothing change from the status quo. Any war costs T on the both sides $(0<T<1)$. The challenger's and defender's expected payoffs for war are $pi_t^c+(1-p)\cdot 0-T$ and $-pi_t^d+(1-p)\cdot 0-T$, respectively. If the defender appease to the challenger's demand, the challenger takes all the pie under a dispute for sure, and the challenger's and the defender's payoffs are i_t^c and $-i_t^d$.

We also assume that each state knows its own the value on a given issue but is unsure of the value the other state place on the issue. In other words, each state has the private information about its value on the issue, i_t, which we call the player's "type." At the beginning of each stage game, the value of i_t^c and i_t^d is determined by the nature (i_t^c, i_t^d ~unif [0,1]). Each state can observe only its own the value on the issue. This also implies that each state cannot update its belief about the other's resolve over stage games and that the values on the issue change over time.

The stage game above addresses a signaling problem in a crisis. The existing literature shows that strongly resolved states engage in costly behavior such as mobilization and tying-hands to distinguish themselves from weakly resolved states.[5] However, we do not focus on costly behavior as a signaling device in a crisis here. The stage game captures *costless* behavior such as a diplomatic talk to reveal the information about the defender's type. In the stage game, any defender's message does not change the subsequent structure of the game, the sequence of the moves and payoffs. This means that sending any message does not cost on the defender. We imply that the challenger cannot learn the defender's type from the subsequent changes in the sequence of the moves and payoffs that would result from the defender's behavior. A diplomatic message can change only the challenger's belief about the

[5] Fearon (1997).

defender's type. Thus, a diplomatic message is cheap and seems to be noisy because every type can easily use it as an information transmission device. The game is thus one of costless signaling or "cheap talk".[6]

Now we consider the situation where the stage game is infinitely iterated. One of our concerns is when and how the defender's message is informative in the long-term perspective. We assume that the defender is always challenged by different challengers at every stage game and the issue changes over time periods. This means that each challenger cannot update its belief about the defender's type from stage game to stage game and that each challenger only knows what the defender did in previous periods.

When the challenger recognized the defender as honest in the past, the challenger uses the defender's message to update his belief about the defender's resolve. However, if the defender's bluffing was caught in the past, the challenger does not do so. The defender's message and behavior at the current stage affects the prospective challengers' cognitive bias that will lead to behavioral changes. Thus, the defender's current behavior not only affects the prospective challenger's cognitive processes but also change the behavior of both sides at the future stage games, resulting in the change of expected payoffs. The defender cares not only about current payoffs, but also about her future payoff stream.

We shall examine whether equilibria exist in which a called bluff on the part of the defender results in a reputation for dishonesty in one or more future periods. Sartori studies an equilibrium in which the defender's tarnished reputation can be gradually recovered over time. We shall contrast this equilibrium to one in which states believe that defenders can recover their reputation for following through on a commitment to go to war only by following through on such a commitment. That is, recovering a reputation for following through (i.e. for honesty) requires that states demonstrate their willingness to go to war by going to war. If the defender with tarnished reputations does not fight with the challenger and a war is not observed, the prospective challenger will not update it's beliefs based on the defender's diplomatic

[6] For a clear discussion of cheap talk games, see Gibbons (1992) pp. 210-213.

message.

A key question we ask below is whether or not conflict behavior is affected by these two different processes for recovering tarnished reputations. If the two recovering processes of tarnished reputation generate different behavioral patterns in crisis, we can infer social norms, behind behavioral patterns in history, on how tarnished reputation is recovered. Given these assumptions, we will explore the conditions under which a "cheap" diplomatic talk is informative.

The Informative and Non-informative Equilibria

The game above has at least two equilibria, the informative and non-informative equilibria. The informative equilibrium indicates the possibility that the defender honestly tells its resolve and its message is valuable for the challenger to infer the likelihood of the defender's fighting more precisely. The other equilibrium describes the possibility that the challenger never listens to the defender's message because it is not helpful for revealing the likelihood of the defender's fighting. As far as the challenger believes the defender's message as informative, the defender's deterrent threat is sometimes effective and the defender can achieve its goal peacefully. For this reason, credible diplomacy has a value to the defender. Defenders therefore have incentive to use a diplomatic talk as an information-revealing tool, and challengers find it to their benefit to listen to the message in inferring their opponent's intentions.

The informative equilibrium consists of a set of strategies in the two stages, the honesty stage and the punishment stage. In the honesty stage, the challenger listens to the defender's message to infer its resolve and the defender's bluffing is sometimes observed. The defender loses its reputation for honesty when its bluffing is called: (1) the challenger makes a threat, (2) the defender says "will defend" (a counter threat), (3) the challenger attacks, and then (4) the defender do not defend. When the defender's bluffing is called in the honesty stage, the challenger will not listen to the defender's message in the several subsequent periods. However, the defender's tarnished reputation can be recovered in several ways. Sartori argues that a state

caught bluffing can use diplomacy again because the reputation for bluffing cannot last forever. This means that a tarnished reputation exists for some periods but a reputation for honesty is gradually recovered. She assumes that the reputation for bluffing continues over two periods and it will be recovered. In contrast, we suggest that a reputation for bluffing can be immediately recovered by fighting a war. Fighting a war is the opportunity for the defender with a tarnished reputation to show its resolve. This scheme indicates that the defender is not always bluffing and that potential challengers change their minds and start listening to the defender's message again. These recovering processes are the two different norms about how tarnished reputation is recovered. We start investigating whether or not these norms support the informative equilibrium.

Figure 2: The Informative Equilibrium Strategies

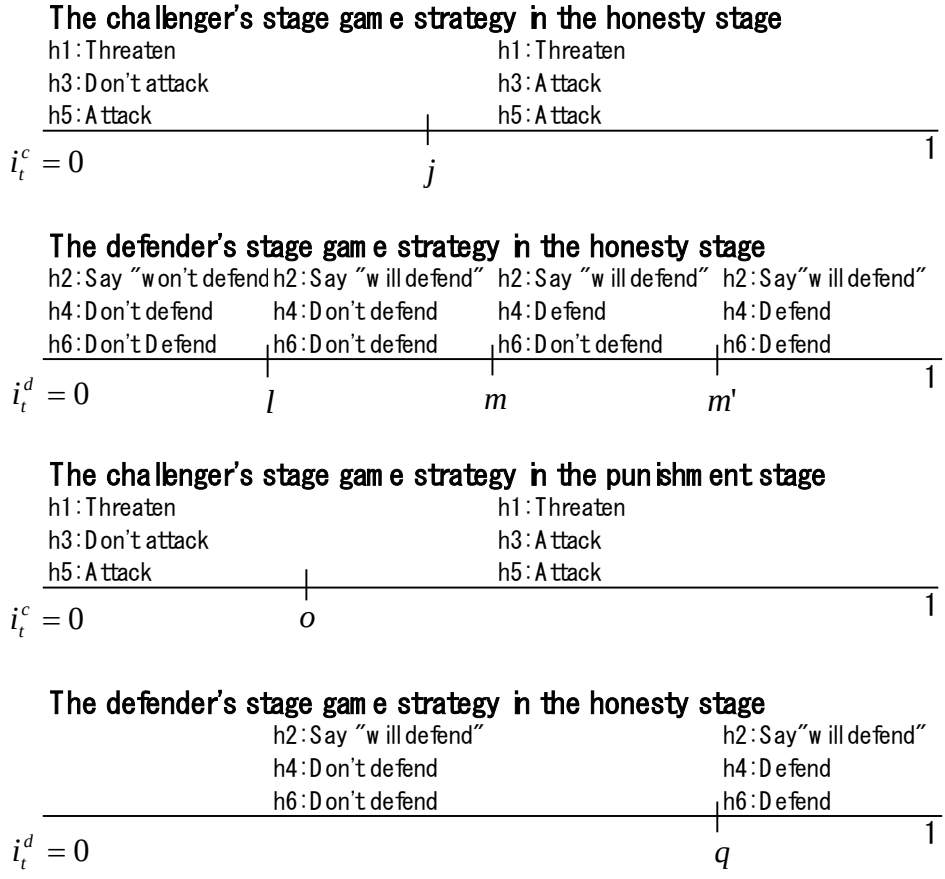


Figure 2 shows the set of strategies in the informative equilibrium.[7] The top half of figure 2 describes the honesty stage under which the defender was not caught bluffing in time $(t-1)$ and $(t-2)$. When the defender is communicative, an international outcome in a specific conflict is determined by the challenger's and defender's issue values, i_t^c and i_t^d. For example, given $i_t^c < j$ and the challenger makes a threat, it does not attack if it hears the defender's counter threat and attacks if it hears no counter threat. Given $l < i_t^d < m$, the defender makes a counter threat and does not defend. With $i_t^c < j$ and $l < i_t^d < m$, the challenger will make a threat, the defender will make a counter threat, and an international outcome of this conflict will result in the challenger's back down. This case shows the defender's bluffing is successful in attaining its goals peacefully. The lower half of figure 2 describes the punishment stage under which the challenger never listens to the defender's message. In this stage, since the defender knows that the challenger never use its message to infer its type, all types of the defender send the same message, "will defend." Thus, a diplomatic talk does not affect conflict behavior in the punishment stage, under which the equilibrium strategies are the same as strategies in the non-informative equilibrium (technically, the bubbling equilibrium). All thresholds, j, l, m, o, and q, are the function of the exogenous variables, p and T.

Sartori (2005) assumed that the punishment stage continues over two stages and showed the existence of the informative equilibrium. In contrast, we assumed that the punishment stage continues till the defender fights a war and sought the informative equilibrium. We used the numerical methods to search any informative equilibrium because the system of equations is nonlinear and could not be solved in algebraic form. The detail of our procedures is explained in the appendix. We found no informative equilibrium given our assumption on the recovering reputation process and had only the bubbling equilibrium because we had only solutions for the system of equations with the thresholds l which is zero. When states believe that they have to fight in order to recover tarnished reputation, no diplomatic talk is informative and

[7] Sartori (2005) pp. 131-136.

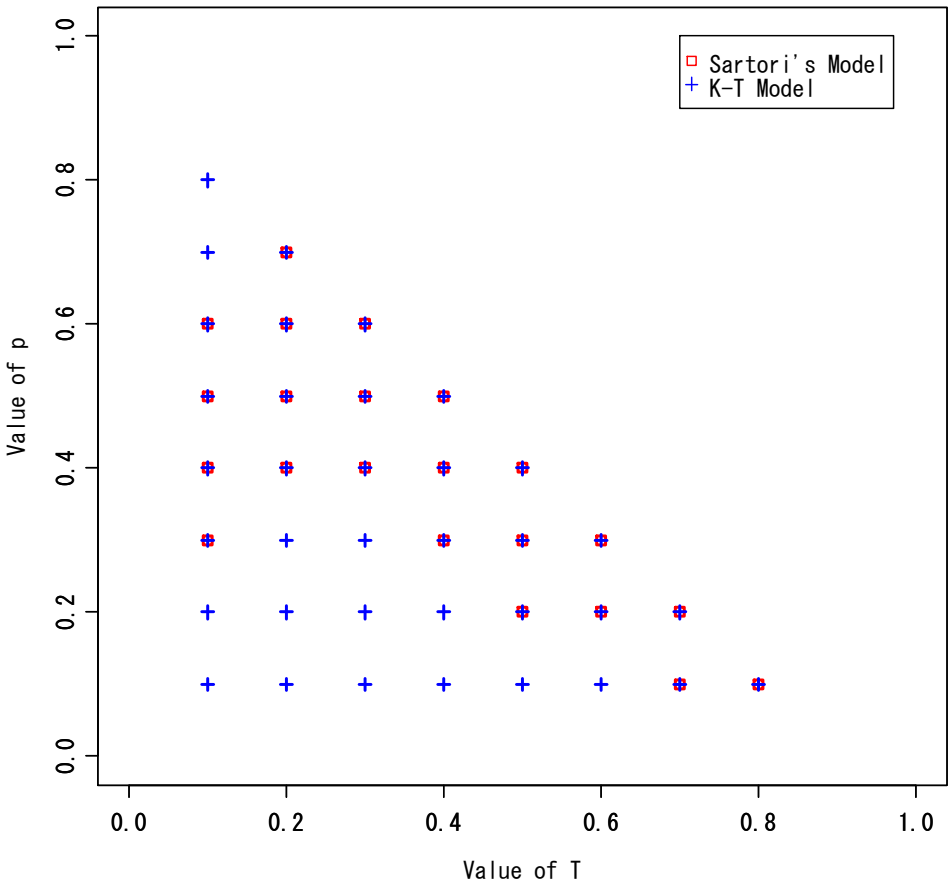
states do not use diplomatic messages to convey their intentions.

Figure 3 shows the equilibrium existence with sets of parameters, p and T. The informative equilibria of Sartori's model are represented by squares. The non-informative equilibria of Kagotani-Trager (KT model) are indicated by crosses. Given Sartori's setting, the model has both the informative and non-informative equilibria. However, if diplomacy serves as a value-added tool for information revelation, states will act like what the informative equilibrium tells. Given KT's setting, the model has only the non-informative equilibria. If both the informative and non-informative equilibria co-exist with the same parameter set of p and T, Sartori's and KT's reputation recovering processes generate different behavioral patterns in crises. These reputation recovering processes are norms among states. If states believe they can gradually recover their tarnished reputation over time, a tarnished reputation changes conflict behavior and diplomacy effectively function (Sartori's model). If states believe they can recover their tarnished reputation only by fighting, a tarnished reputation does not change conflict behavior and diplomacy does not work (KT model).

Why does an expectation about how reputations can be recovered once lost, an expectation about actors' behavior off the equilibrium path since reputations are never won or lost on the equilibrium path of the KT equilibrium, have such a dramatic effect on equilibrium path behavior? The answer is fairly straightforward. When reputation recovers naturally over time, the defender is more willing to fight in the honesty stage because it has a reputation to defend. In the KT equilibrium, by contrast, the defender has a reputation to defend in the honesty stage and a reputation *to regain* in the dishonesty stage. It's willingness to fight is therefore identical in both honesty and dishonesty stages. Challengers are therefore less willing to fight in the dishonesty stage than in the Sartori equilibrium. As a result, the expected utility to the defender of the dishonesty stage is greater (relative to the honesty stage) than in Sartori's informative equilibrium. In fact, the dishonesty stage in the KT equilibrium is not so unattractive as to justify not making a threat (conceding the issue for sure) in the honesty stage, even when the issue in question is of very little importance to the

defender. This, in turn, implies that there is no semi-separating equilibrium under the KT reputation recovery norm, and nothing can be learned from the defender's statements of intent.

Figure3: The Equilibrium Existence



Recovering Reputation and the Likelihood of War

The equilibrium analysis above showed that different norms on recovering reputation determine the existence of the informative equilibrium in the model. Unlike Sartori's work, we focus not on the relationship between tarnished reputation and the challenger's and the defender's behavior here, but on the relationship between tarnished reputation and the likelihood of war. In her work, she mentioned the

selection problem that the challenger targets the weak challenger and the challenger's decision not to challenge is not observable. When we consider the likelihood of war, we can avoid the selection problem based on strategic choice because the dependent variable is not a strategic choice but a social outcome.

Figure 4 and 5 shows the effect of p and T on the likelihood of war. As mentioned above, the challenger's and the defender's behavior in the equilibrium is a function of the exogenous parameters, p and T. Of course, since the likelihood of war is calculated by equilibrium behavior, it is also the function of p and T. In the graph, the square and cross marks represent the likelihood of war in the honesty stage and the punishment stage of the informative equilibrium in Sarotri's model. The circle mark indicates the likelihood of war in the non-informative equilibrium in KT model and it is the same as one in the punishment stage of Sartori's model. The effect of p on the likelihood depends on the size of T. With small T, as p increases, the likelihood of war increases and then decreases. With large T, as p increases, the likelihood of war only decreases. In contrast, the effect of T is monotonic. As T increases, the likelihood of war decreases.

Figure 4 and 5 also graphically illustrate an important point: war is more likely to happen in the honesty stage than in the punishment stage of Sartori's model. Of course, the likelihood of war in the non-informative equilibrium of the KT model is not affected by a reputation for honesty. We can now derive testable hypotheses relating reputations for honesty to the likelihood of war.

Hypothesis 1:

War is less likely happen when the defender has a tarnished reputation (consistent with a belief among states that reputations for keeping commitments gradually recover over time).

Hypothesis 2:

The likelihood of war is independent of the defender's reputation (consistent with a belief among states that reputations for keeping commitments are recovered only by

keeping costly commitments, that is, by fighting).

Figure 4: The Effect of p on the Likelihood of War

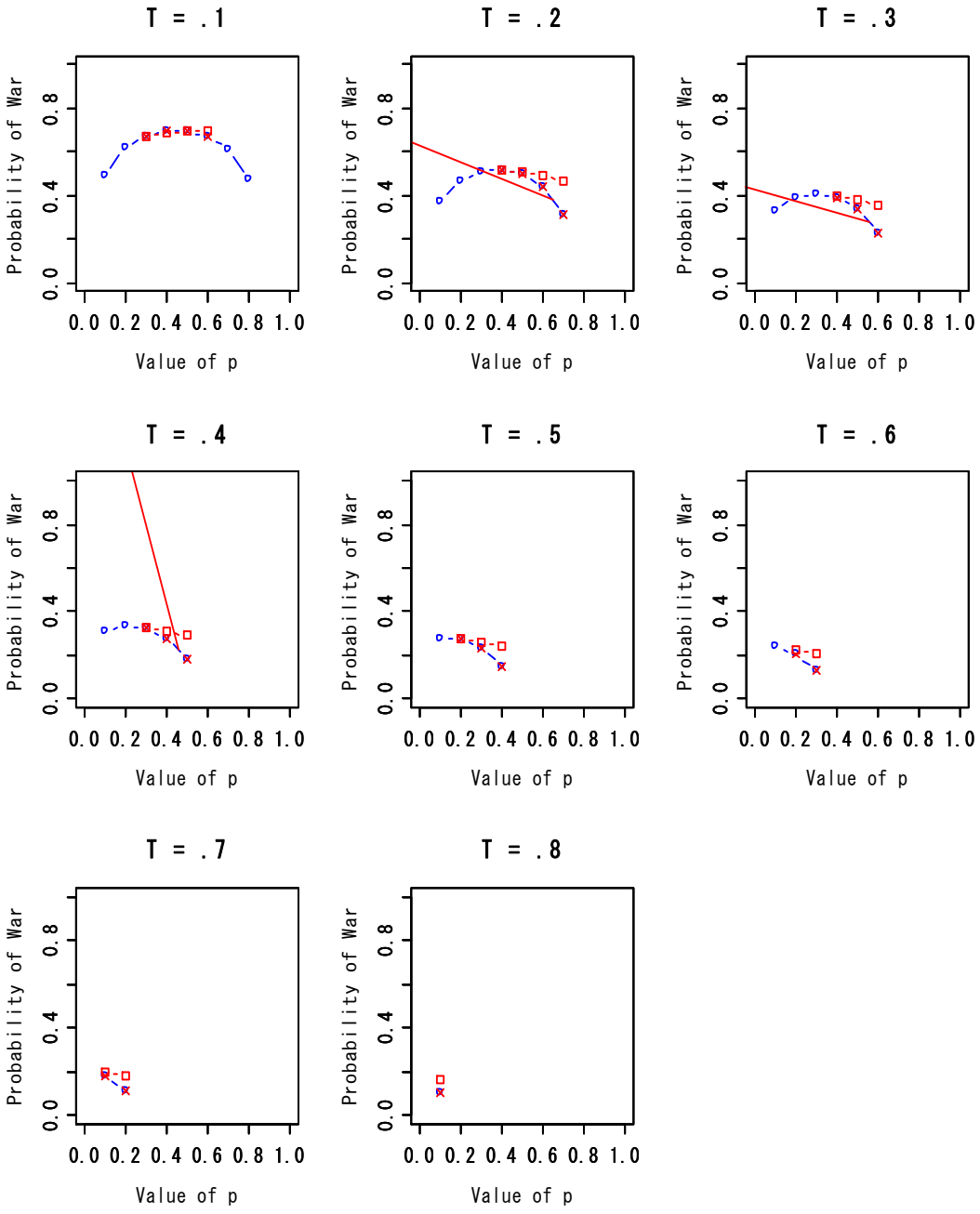
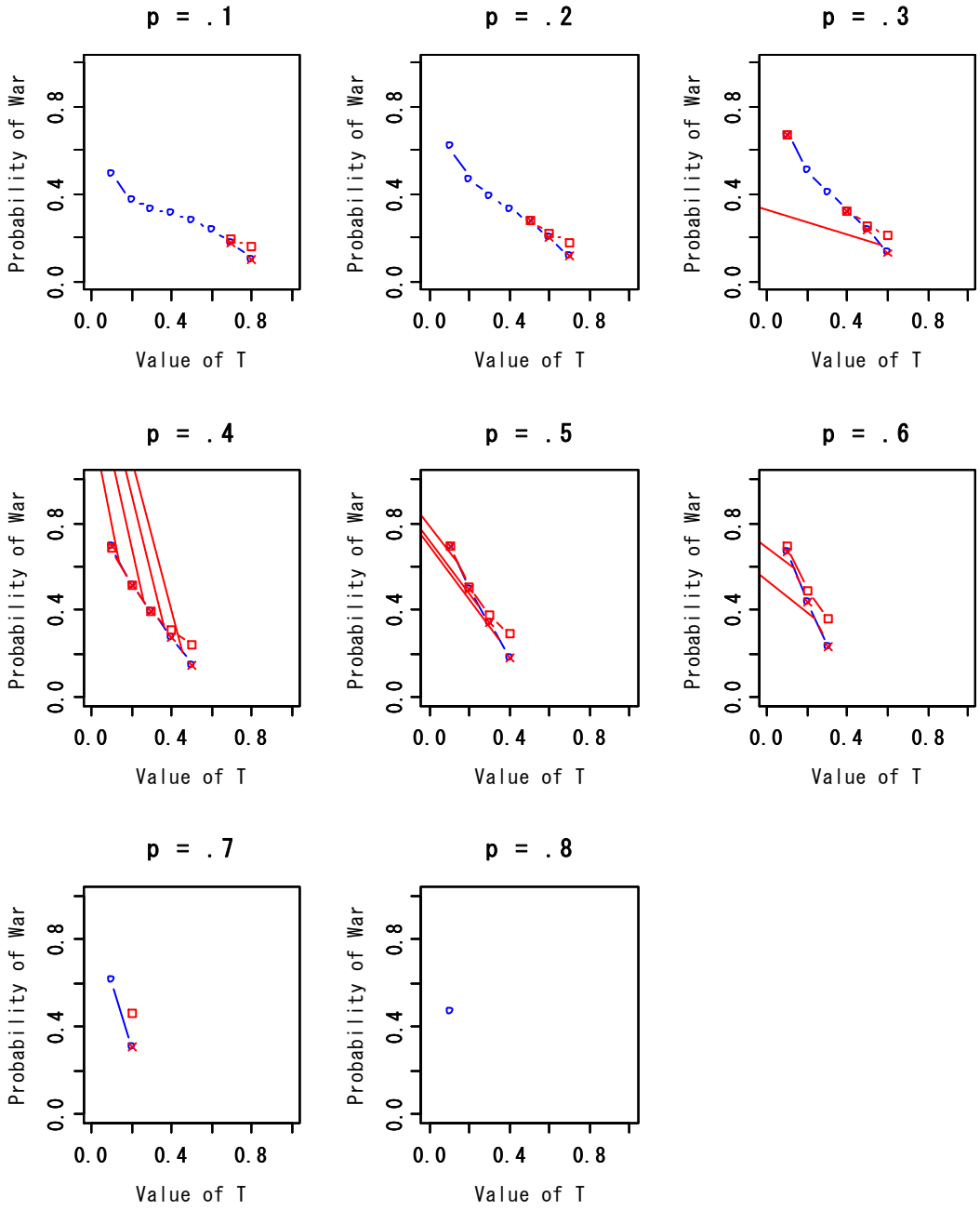


Figure 5: The Effect of T on the Likelihood of War



We followed Sartori's coding scheme and compiled the data about war and a reputation for bluffing. In her book, she used the directed-dyad data and investigated the relationship between the defender's tarnished reputation and its own choice or the challenger's choice. The directed dyad explains State A's decision against State B in a given year. Within a dyad of State A and State B, each state's decision is recorded as a case in a given year. She used information from the Correlates of War (COW) project and COW's Militarized Interstate Dispute data set to describe conflict behavior.[8] The hostility level indicates conflict behavior; 0 = No hostility, 1 = No militarized action, 2 = Threat to use force, 3 = Display of force, 4 = Use of Force, and 5 = War. According to her coding scheme, the hostility level less than 2 represents a case where a state did not make any threat. If the hostility level is 2 or 3, a state made a threat but did not fight the opponent. If the hostility level is more than 3, a state made a threat and fought the opponent.[9] We assumed that if both State A's and B's hostility levels are more than 3, a conflict happened within a dyad of State A and B. We coded a conflict as 1, and otherwise 0.

We replicate Sartori's reputation variable representing reputation for bluffing. We recognized a case as the challenger's choice if the variable of cwsideA1 is 1. Otherwise, we think a case as the defender's choice. Each case has the information about State A's and B's hostility level. If the challenger's hostility level is greater than 3 and the defender's hostility level is 2 or 3, we call a case as bluff. If the defender's bluffing is called in the period t, the state will have a tarnished reputation, $.85$, $.85^2$, …, $.85^{10}$ in the period $t+1, t+2, \ldots t+10$. That is, the defender's reputation for bluffing continues in the subsequent ten years.[10] We change one point from Sartori's setting and generated our own reputation variable. If a state with a tarnished reputation fights with the opponent in the period t, the value of the reputation immediately goes down to zero in the period $t+1$. We coded a case as honesty when the reputation variable is zero, and consider a case as bluff when the reputation

[8] On the MID data, See Ghosn, Faten, Glenn Palmer, and Stuart Bremer (2004). We used the computer program, EUGene version 3.201 (Bennett and Stam 2000), to manipulate the data.

[9] Sartori (2005), pp. 83-84.

[10] Sartori (2005), pp. 84-86.

variable is greater than zero. We used the directed dyad data but it is the same result as one with the dyad data because each dyad is counted twice and this scheme does not change the likelihood of war.

Table 1 and 2 show the results of a simple test on the effect of a tarnished reputation. The data covers the period 1816-2001. Table 1 tells that the war is more likely to occur when the defender has a reputation for bluffing and that the difference of the war likelihood in two stages is statistically significant at .05% level. In contrast, Table 2 tells that the likelihood of war is not affected by a reputation for honesty. Thus, the analysis of the raw data is against Sartori's model and consistent with KT model though we have to control other independent variables. With the aggregate data, states seemed to believe that they could recover their reputation by fighting. It is also possible that behavioral patterns changes over time. Investigating the disaggregated data of historically significant periods is for future research.

Table 1: Reputation and the Likelihood of War (Sartori's Model)

	No Conflict	Conflict
More honest (rep. for honesty)	99.76%	0.24%
Less honest (rep. for bluffing)	99.42%	0.58%
Total	99.68%	0.32%

Pearson chi2(1) = 852.9483 Pr = 0.0005

Table 2: Reputation and the Likelihood of War (KT Model)

	No Conflict	Conflict
More honest (rep. for honesty)	99.68%	0.32%
Less honest (rep. for bluffing)	99.69%	0.31%
Total	99.68%	0.32%

Pearson chi2(1) = 1.2978 Pr = 0.255

Conclusion

The results of the theoretical discussion demonstrate that, if states believe that tarnished reputations can be recovered by fighting, informative diplomacy will be impossible in the context of Sartori's (2005) model. In fact, states need not believe

that reputations *must* be recovered in that way, only that they *can* be for diplomacy to be ineffective. In such a world, the beliefs of states ("If a state develops a reputation for dishonesty, it can recover it by following through on a commitment to fight…") are entirely off the equilibrium path. States never do develop reputations for dishonesty. Nevertheless, expectations about how reputations would be recovered if they ever were lost ensure that diplomacy will be ineffective. Further, because these beliefs are off the equilibrium path, once such expectations are created, no event can occur that disabuses actors of them (for that matter, even if they are incorrect).

These results depend, of course, on the analytical context in which they were derived. We have studied only "emergent reputations", those that do not correspond to a disposition, preference or characteristic of agents, but are instead entirely socially constructed, emerging from state interactions in the context of particular sets of expectations about each other's behavior. It may be that if we studied dispositional reputations, or if uncertainty about the resolve of the other side were allowed to have a more general form (rather than the types being drawn from a uniform distribution in each period), then informative diplomacy would be possible in KT equilibrium. We will explore these, and other possibilities in future work.

Appendix

We do not show the proof of Sartori's informative equilibrium again (Sartori 2005, pp. 131-145). We explain our search for the informative equilibrium in our model. Suppose that we have the same informative equilibrium in our model as Sartori's. Let w_1 and w_2 be the defender's expected payoffs for the honesty and the punishment stages. When the defender's issue value i_t^d is less than l, it says "will not defend" and does not defend. All types of the challenger attacks after hearing the defender's "will not defend." The result of these choices is the defender's concession and the defender's payoff is i_t^d. In the next time period, the defender will have the expected payoff δw_1. δ is the discount factor which translate the future value into the current value and takes the value between 0 and 1. With the probability l, the defender will get $(-l/2+\delta w_1)$. In the same way, the defender's expected payoffs are calculated. One different point from Sartori's model is that the defender can recover its tarnished reputation only by fighting a war.

$$w_1 = l\left(\frac{-l}{2}+\delta w_1\right)+(m-l)\left(j\delta w_1+(1-j)\left(\frac{-m-l}{2}+\delta w_2\right)\right)$$
$$+(1-m)\left(\delta w_1+(1-j)\left(p\frac{-m-1}{2}-T\right)\right)$$

$$w_2 = o\delta w_2+(1-o)\left(q\left(\frac{-q}{2}+\delta w_2\right)+(1-q)\left(p\frac{-q-1}{2}-T+\delta w_1\right)\right)$$

The thresholds for each state are calculated as follows. The threshold j makes the challenger indifferent between attacking and not attacking after making a threat in the honesty stage.

$$0=(m-l)j+(1-m)(pj-T)$$
$$\Leftrightarrow j=\frac{(1-m)T}{m-l+p-mp}$$

The threshold l makes the defender indifferent between saying "will not defend" and "will defend" before choosing not to defend in the honesty stage.

$$-l+\delta w_1 = j(0+\delta w_1)+(1-j)(-l+\delta w_2)$$
$$\Leftrightarrow l = \frac{(1-j)(w_1-w_2)\delta}{j}$$

In the same way, other thresholds are calculated as follows.

$$-m+\delta w_2 = -pm-T+\delta w_1 \quad \Leftrightarrow m = \frac{\delta w_2 + T - \delta w_1}{1-p}$$

$$0 = qo+(1-q)(po-T) \quad \Leftrightarrow o = \frac{(1-q)T}{p+q+pq}$$

$$\delta w_2 - q = -pq-T+\delta w_1 \quad \Leftrightarrow q = \frac{\delta w_2 - \delta w_1 + T}{1-p}$$

We could not solve the system of seven equations in the general form and tried to solve these equations by the numerical methods. Mathematica version 3.0 showed that the system of equations has a real solution with l equal to zero. First, we reduce the system of seventh equations into two equations by substitution. The expected payoffs of w_1 and w_2 is represented by the functions of w_1, w_2, δ, p, and T. Let the two functions be f_1 and f_2. The defender's expected payoffs are expressed as flows.

$$w_1 = f_1(w_1, w_2, \delta, p, T)$$
$$w_2 = f_2(w_1, w_2, \delta, p, T)$$

Second, we fix exogenous variables δ, p, and T, and then solve the following objective function. We set δ equal to .9 and choose a set of parameters p, and T here.

$$v = (w_1 - f_1(w_1, w_2))^2 + (w_2 - f_2(w_1, w_2))^2$$

Given a set of exogenous variables, when we can minimize the objective function equal to zero, we have solutions for w_1 and w_2, and then can calculate all thresholds by these solutions. We wrote a code with the non-linear minimization function in the statistics software R and looked for the solutions for the objective function which generates all thresholds consistent with the assumption. However, we could not find any solution of the objective function to generate the informative equilibrium and had only solutions to lead to the non-informative equilibrium. We made sure for the result by Mathematica.

References

Bennett, D. Scott, and Allan Stam. 2000. "*EUGene*: A Conceptual Manual." *International Interactions* 26:179-204. Website: http://eugenesoftware.org.

Evron, Yair. 1987. *War and Intervention in Lebanon*. Baltimore, M.D.: Johns Hopkins University Press.

Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs." *Journal of Conflict Resolution* 41: 68-90.

Gibbons, Robert. 1992. *Game Theory for Applied Economics*. N.J.: Princeton University Press.

Ghosn, Faten, Glenn Palmer, and Stuart Bremer. 2004. "The MID3 Data Set, 1993–2001: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 21:133-154.

Jervis, Robert. 1970. *The Logic of Images in International Relations.* New York, N.Y.: Columbia University Press.

Jervis, Robert. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, N.J.: Princeton University Press.

Kifner, John. 1981. "Syria Resisting Pressure to Remove Missiles from Lebanon." *New York Times,* May 5.

Kurizaki, Shuhei. 2007. "More Effective Diplomatic Communication: The Might of the Pen Revisited." Presented at the *American Political Science Association* meetings in Chicago, August, 2007.

Kydd, Andrew. 2003. "Which Side Are You On? Bias, Credibility and Mediation." *American Journal of Political Science* 47(4): 597-611.

Lewis, Jeffrey, and Kenneth Schultz. 2008. *Modified International Crisis Behavior Dataset.*

Mercer, Jonathan. 1996. *Reputation and International Politics.* New York, N.Y.: Cornell University Press.

Niou, Emerson M. S. and Peter C. Ordeshook. 1990. "Stability in Anarchic International Systems." *American Political Science Review* 84(4): 1207-1234.

Niou, Emerson M. S. and Peter C. Ordeshook. 1991. "Realism Versus Neoliberalism:

A Formulation." *American Journal of Political Science* 35(2): 481-511.

Press, Daryl G. 2004. "The Credibility of Power: Assessing Threats During the "Appeasement" Crises of the 1930s." *International Security* 29(3): 136-169.

Sartori, Anne E. 2002. "The Might of the Pen; A Reputational Theory of Communication in International Disputes." *International Organization*, 56(1): 121-49.

Sartori, Anne E. 2005. *Deterrence by Diplomacy*. N.J.: Princeton University Press.

Trager, Robert. 2010. "Diplomatic Calculus in Anarchy: How Communication Matters." *American Political Science Review* 104(2):347-368.

Trager, Robert F. 2011. "Multi-Dimensional Diplomacy." *International Organization* 65:469-506.

Trager, Robert F. 2012. "Long-Term Consequences of Aggressive Diplomacy: European Relations after Austrian Crimean War Threats." *Security Studies* 21(2):232-265.

Trager, Robert F. 2013. "How the scope of a demand conveys resolve." *International Theory* 5(03):414-445.

Trager, Robert F. 2015a. "Diplomacy: Communication and the Origins of International Order." Manuscript.

Trager, Robert F. 2015b. "Diplomatic Signaling among Multiple States." *The Journal of Politics* 77(3):635-647.

Trager, Robert F. and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." American Journal Political Science 55(3):526-545.

Schelling, Thomas. 1966. *Arms and Influence.* New Haven, C.T.: Yale University Press.

Slantchev, Branislav. 2003. "The Power to Hurt: Costly Conflict with Completely Informed States." *American Political Science Review* 47(1): 123-135.

Wendt, Alexander. 1999. *The Social Theory of International Politics.* Cambridge, U.K.: Cambridge University Press.