A Reputational Model of Conflict: Why Die to Win?

Jihong Lee *

This paper considers a simple model of zero-sum conflict between two players (*e.g.*, territorial dispute) in which costly actions (*e.g.*, terrorism) are available to one side. We identify how reputation effects shape the outcome of such conflict. A small prior of fanatic commitment type induces the possibility of costly attack followed by withdrawal in equilibrium. The chance of withdrawal is proportional to the self-inflicted cost of attack.

Keywords: Conflict, Terrorism, Reputation

JEL Classification: C72, D72, D74

I. Introduction

Many conflicts are accompanied with violent acts. These acts, often referred to as acts of terrorism, are designed to coerce others into actions they would not otherwise undertake. While terrorist attacks generate significant loss to their targets, some tactics inflict large and vivid damages to the aggressors themselves. A case in point is suicide attack. In fact, last few decades have witnessed a rapid climb in the number of suicide attacks; it has grown from an average of fewer than five per year during the 1980s to 180 per year between 2000 and 2005, and from 81 suicide attacks in 2001 to 460 in 2005 (Atran 2006).

The purpose of this paper is to present a rational, game-theoretic an-

* Associate Professor, Department of Economics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-746, Korea, (Tel) +82-2-880-6365, (Fax) +82-2-886-4231, (E-mail) jihonglee@snu.ac.kr. The author is grateful to two anonymous referees for their helpful comments. This work was supported by the Je-Won Research Foundation Grant funded *via* the Institute of Economic Research at Seoul National University.

[**Seoul Journal of Economics** 2012, Vol. 25, No. 2]

alysis that links the level of self-inflicted costs of attacks to the outcome of the conflict. We aim to offer an explanation of why many terrorist campaigns resort to actions that impose large self-inflicted costs. The key ingredient of our analysis is indeed motivated by the popular perspective of terrorism, which views it as acts that are intended to create fear (terror). In our model, the potential victim of attack is uncertain about the nature of his opponent; in particular, he believes that his opponent has a small chance to be a fanatic commitment type who always attacks. With such incomplete information an attack may foster fear of further attacks and hence facilitate concession.

Our multi-period game begins with one player occupying the territory of another player. In each period, the occupier decides whether or not to continue occupation and the occupied has the option to undertake an attack on the occupier. The payoff of the occupied accrues directly as a function of the two parties' actions. The payoff of the occupier however depends on the outcome of voting that takes place at the end of the game. We model a median voter who prefers to vote out the occupier (*i.e.*, the government) if the attack occurs too frequently. The occupier earns a payoff if and only if he is kept in office.

This setup is motivated by Pape (2005). He finds that 95% of suicide attacks in recent times share the same specific strategic goal to cause an occupying state to withdraw forces from a disputed territory. Furthermore, the targeted countries are usually the ones where the government is democratic and public opinion plays a role in determining policy.

A perturbation of the complete information game, with a small initial prior of the occupied being a commitment type, creates the following equilibrium dynamics, which are unique if the players do not use weakly dominated strategies. In equilibrium the occupier initially opts to stay but attack is undertaken with a probability that builds the occupied's reputation (for being fanatic) to the level at which the occupier is indifferent between occupation and withdrawal in a later period. Indeed, to make this initial attack behavior optimal, the occupier responds by sometimes choosing to withdraw. The equilibrium frequency of attack depends on the prior and the nature of the commitment type (*i.e.*, how likely that he would attack). More importantly, the probability of withdrawal is proportional to the cost of attack that the occupied inflicts on himself.

This paper is related to previous works in economics that attempt to identify a variety of potential sources of costly conflicts *via* game-theoretic reasoning. Schelling (1963) and, more recently, Baliga and Sjöström (2004) propose a rational account of "spiral theory" of war. Other reasons

for a costly war to break out among rational players include bargaining frictions (Schelling 1966) and political bias (Jackson and Morelli 2007). Our paper suggests reputational concerns as another source of costly conflicts. In terms of modeling, this paper follows the adverse-selection approach of reputation that dates back to Kreps, Milgrom, Roberts, and Wilson (1982), Kreps and Wilson (1982), and Milgrom and Roberts (1982).

The paper is organized as follows. In Section 2, we lay out the model. In Section 3, we identify and analyze the equilibrium. Some concluding remarks are offered in Section 4.

II. The Model

We analyze the following game played by three players, A ("the occupied"), B ("occupier") and a median voter (of the "occupier"). The game begins with B occupying A's territory, and it lasts for three periods. In period 1, simultaneously, A chooses whether to "attack" or "not attack" and B chooses whether to "stay/occupy" or "withdraw." If B chooses to withdraw in period 1, he cannot return to occupy, *i.e.*, the game proceeds directly to period 3. If B chooses to stay, the game proceeds to period 2 and the two players face the same simultaneous action choices as in period 1. In period 3, the median voter chooses whether to vote B "in" or "out."

A is a either "fanatic" or "rational," which is A's privately known type, and $p \in (0, 1/2)$ is the commonly known prior of him being "fanatic" at the beginning of period 1. The fanatic is a *commitment type* who always attacks with probability $r \in (1/2, 1)$.[1]

In each period, rational A incurs cost a if occupied and 0 if left alone; attack costs c if B stays and 0 if B withdraws. B obtains payoff 1 if voted in; 0 otherwise. In period $t = 1, 2$, the median voter obtains b_t from occupation and loses d_t from an attack. If B withdraws, the median voter experiences neither gain nor loss. The median voter's payoff from voting out B is normalized to 0.

B's actions are accountable to its voters. We assume that the median voter votes *retrospectively* in the sense that, although his voting decision takes place after the actions of A and B over two periods, it reflects the payoff resulting from those actions. Specifically, if his total payoff at the end of period 2 is greater (less) than 0, the median voter votes B in

[1] We rule out the case of $r = 1$ purely to simplify the analysis. See footnote 3 below.

(out); if the total payoff is 0 then the median voter votes in B with probability 1/2.

Remark 1. There is a long-standing debate about the role of elections, whether they serve primarily as mechanisms of democratic accountability (*e.g.*, Key 1966) or as mechanisms of democratic selection (*e.g.*, Downs 1957). Using observational data to empirically distinguish between these two views is fraught with methodological difficulty.[2] In a recent laboratory experiment, Woon (2012) finds evidence of retrospective voting behavior and preferences for accountability.

We make the following assumptions:

***Assumption* 1:** $a > c$.
***Assumption* 2:** $b_1 = d_1$ and $b_2 < d_2 < b_1 + b_2$.

Assumption 1 ensures that a potential incentive exists on the part of rational A to instigate a costly attack. This can be motivated by interpreting A in our game as a collective such that the cost of occupation is the aggregation of the individual costs experienced by all members of the collective. The cost of attack, on the other hand, is often borne by an individual member.

Assumption 2 implies two things. The first part means that after attack in period 1 the accrued cost and benefit from occupation remains balanced for B. The second part says that, if B decides to stay over both periods, the median voter will vote him in as long as attacks do not occur in consecutive periods. Note that, if we were to have $b_1 > d_1$(or $b_1 + b_2 > d_1 + d_2$), B could guarantee re-election simply by staying for one period and then withdrawing in period 2 (or by staying for both periods). The other case where $b_1 < d_1$(and $b_1 + b_2 < d_1 + d_2$) also makes the problem uninteresting, as discussed in Remark 2 below. Overall, the main substance of Assumption 2 is that early withdrawal cannot induce sure election victory or failure for B. This is reasonable in a context where voters face uncertainty about the net benefits of occupation when it lasts less than its initially projected full term.

Next, let $\hat{p}$ represent A's *reputation*, or B's belief at the beginning of period 2 that player A is fanatic. We are interested in reputation-building behavior of rational A; hence, whenever we refer to A below, we shall

[2] See, for instance, Lewis-Beck, Nadeau, and Elias (2008).

mean the rational type.

The game effectively ends if B decides to withdraw in period 1. Thus, we can write A's behavioral strategy as a tuple (k_1, k_2^*, k_2^{**}), where $k_1 \in [0, 1]$ is the probability of attack in period 1 and $k_2^*, k_2^{**} \in [0, 1]$ are the probabilities of attack in period 2 given the period 1 history of attack and no attack, respectively. B's behavioral strategy is a tuple (s_1, s_2^*, s_2^{**}), where $s_1 \in [0, 1]$ is the probability of stay in period 1 and $s_2^*, s_2^{**} \in [0, 1]$ are the probabilities of stay in period 2 given the period 1 history of attack and no attack, respectively. Note that the median voter's voting decision is determined fully by the actions of A and B.

A strategy profile is a perfect Bayesian equilibrium (PBE) if the strategies are mutual best responses at every history and the posterior belief $\hat{p}$ satisfies Bayes' rule whenever possible.

III. Occupation and Terror

Our first Lemma establishes that a necessary equilibrium feature of the above game is that occupation and attack must occur in period 1.

***Lemma* 1:** In any PBE, A attacks with a positive probability and B stays for sure in period 1.

***Proof*:** We start by considering B's incentives. Note that, since $b_1 = d_1$, this player can guarantee an expected payoff of 1/2 whether he quits in period 1 or 2. If he stays in period 1, at least with probability $p(1-r) > 0$, attack will not occur, in which case Assumption 2 implies that staying on in period 2 will guarantee re-election. Thus, it is strictly dominant for B to stay for sure.[3]

Next, consider A. Suppose to the contrary of the claim; so, there is a PBE in which A attacks with zero probability in period 1. We have already show that B stays for sure. Thus, attack reveals the fanatic type and B's expected payoff from staying on in period 2 at such a history is equal to $1-r$, which is less than 1/2, *i.e.*, the payoff from withdrawal. Thus, in this equilibrium, B withdraws after attack in period 1. Now, consider A deviating in period 1 to attack. This incurs a cost of c but a gain of a from the subsequent withdrawal. Since $a > c$ (Assumption 1), such a deviation is profitable, a contradiction. ■

[3] Note that stay is weakly dominant if $r = 1$. Thus, Propositions 1 and 2 below remain valid for this case.

Note that if there is no attack in period 1, by Assumption 2, B will get voted in for sure if he continues to stay. Thus, B will stay for sure at such a history. A's incentives in period 2 are as follows.

***Lemma* 2:** Consider any PBE, and suppose that B chooses to stay in period 1. Then, attack is weakly dominated in period 2.

***Proof*:** Suppose that B stays with probability s at this history. Then, A's expected cost of attack in the continuation game is given by $(a+c)s$. Similarly, the expected cost of not attacking is equal to as. Thus, the claim follows. ■

Next, suppose that in period 1 A attacks and B stays. Since there was an attack in the previous period, continued occupation results in B getting voted out if and only if another attack arrives. Let $\hat{p}$ denote the posterior at the beginning of period 2. Then, assuming that rational A abstains from attack, the weakly dominated action, B's continuation expected payoff from occupation in period 2 amounts to

$$\hat{p}(1-r)+(1-\hat{p}). \tag{1}$$

By Assumption 2, withdrawal in period 2 makes the median voter indifferent and, hence, vote B in with probability a half. Thus, B will continue occupation only if (1) is at least 1/2, or only if

$$\hat{p} \leq \frac{1}{2r}.$$

Using these arguments, we construct the following equilibrium.

***Proposition* 1.** There exists a PBE with the following properties:

- In period 1, B stays for sure; A attacks with probability $k^*=\{p/(1-p)\}r(2r-1)$ such that, after attack, his reputation, $\hat{p}$, becomes exactly $1/2r$.
- In period 2, B stays with probability $s^*=(a-c)/a$ after attack in period 1, while staying for sure after no attack; A never attacks.

***Proof*:** Given Lemmas 1 and 2, it suffices to check the two players' indifference conditions and corresponding mixing probabilities.

First, A's equilibrium attack probability in period 1, k^*, is such that B is indifferent between staying and leaving in period 2 following an at-

tack in period 1. It therefore follows from previous arguments that, by Bayes' rule,

$$\hat{p} = \frac{pr}{pr + (1-p)k^*} = \frac{1}{2r}.$$

This implies

$$k^* = \frac{p}{1-p} r(2r-1) > 0,$$

as in the claim (where the inequality holds since $r > 1/2$).

Next, we compute B's mixing probability in period 2, s^*. For this, consider A's indifference condition in period 1. Let C^* and C^{**} represent A's continuation cost at period 2 after attack and no attack in period 1, respectively. If A attacks in period 1, he expects to incur a total cost of $a+c+C^*$ in equilibrium; if he does not attack, the corresponding cost is $a+C^{**}$. To have indifference, we thus need $C^{**}=C^*+c$. Clearly, $C^{**}=a$. So, $C^*=a-c$. In period 2, if B stays, A incurs cost a; if B withdraws he incurs zero cost. Thus, we need to have

$$as^*=a-c,$$

yielding $s^*=(a-c)/a$, as in the claim. ■

***Corollary* 1:** Lower cost of occupation, a, or higher cost of attack, c, to A decreases the equilibrium probability of occupation, s^*.

In the above equilibrium, A randomizes in period 1 in order to *build* his reputation of being fanatic to a level sufficient to induce an incentive for B's withdrawal in period 2. Furthermore, in order for A to be indifferent, B must himself randomize between occupation and withdrawal in period 2.

Lower cost of occupation or higher cost of attack to the occupied improves the chance of withdrawal. The intuition for this result is as follows. Note first that lower a or higher c means less incentive on the part of A to attack and mimic the fanatic type. However, A's equilibrium attack probability, k^*, is independent of these costs and given entirely by B's belief on the likelihood of an attack from the fanatic type (*i.e.*, p and r).

Therefore, in order to compensate for A's extra loss from reputation building, B must withdraw in period 2 with a greater frequency.

The equilibrium constructed above is not a unique equilibrium of our game. In particular, there exists an equilibrium in which A attacks for sure in both periods and B stays in both periods only if there is no attack in period 1 (by the fanatic type). Since attack incurs no cost if the other player pulls out, it is indeed mutual best responses for A to attack and B to withdraw after attack in period 1 (independently of the posterior belief since $r>1/2$). Also, A does not want to deviate in period 1 since no attack prolongs occupation for sure.[4] Nonetheless, we know that attack in period 2 is weakly dominated, which then leads to the following.

Proposition 2. Suppose that $p<1/2r$. Then, the PBE of Proposition 1 is the unique PBE in which no player uses a weakly dominated strategy.

Proof. Fix any PBE in which weakly dominated strategies are rule out. Then, by Lemmas 1 and 2, B must stay for sure in period 1 and A must choose not to attack for sure in period 2. We proceed in the following steps.

Step 1: A must attack with an interior probability in period 1.

Given Lemma 1, it suffices to show that A cannot attack for sure. To show this, suppose otherwise. But then, since $r<1$, the posterior belief following attack is strictly less than the prior p. In period 2, therefore, B's corresponding continuation payoff from staying is at least $p(1-r)+(1-p)=1-pr>1/2$, where the latter inequality follows from $p<1/2r$. Thus, B must stay for sure in period 2. This then implies that it is profitable for A to deviate in period 1 by choosing not to attack.[5]

Step 2: A's attack probability in period 1 and B's occupation probability in period 2 are given by k^* and s^* in Proposition 1, respectively.

We have already argued in the proof of Proposition 1 that B must mix with the stated probability s^* in order for A to be indifferent in period 1, while k^* makes B indifferent in period 2. ■

Remark 2. We have so far assumed that $b_1=d_1$ (Assumption 2). Alter-

[4] This is also a subgame perfect equilibrium of the complete information game with $p=0$. Note that, in such a game, there also exists an equilibrium in which A never attacks and B always stays. It is worth noting that this latter equilibrium possibility disappears in the perturbed game (Lemma 1).

[5] If p is close to 1, B will decide to withdraw after attack in period 1. However, A would not want to deviate in period 1 (from attack to no attack) because it would then surely keep B.

natively, we could assume that $b_1 < d_1$ (together with $b_1 + b_2 < d_1 + d_2$ so that two attacks remain a source of sure election loss). In this case, however, we obtain a very different equilibrium picture. The unique PBE is such that A abstains from attack and B stays in both periods for sure. This is because large period 1 damage from attack actually endows the occupier with commitment value. Once he has stayed, attack in period 1 leads to sure election loss from withdrawal and, since there is always a positive chance of no attack in the second period ($r < 1$), it is then strictly dominant for B to continue occupation in period 2. This takes away any incentive on the part of A to attack.

IV. Conclusion

In this paper, we have set up and analyzed a simple model of conflict between two parties in which reputation effects generate costly outcome. The occupier believes that the occupied has a small chance to be a fanatic commitment type who always attacks (with a probably greater than half), and this allows the occupied to build reputation and force withdrawal by randomly attacking his opponent. We identify the determinants of the equilibrium probabilities of attack and withdrawal. In particular, a higher self-inflicted cost of attack increases the chance of withdrawal.

Our model includes a median voter whose preferences are directly influenced by actions of the two other players while the occupier's payoffs are determined by the median voter's decision. One potentially fruitful channel of future research is to elaborate on this link between voting and the occupation decisions. Another extension is to consider two-sided incomplete information. In such a game, additional signaling by the occupier may not only affect the attack decisions of the occupied but also the voting decisions of his own citizens.

(Received 13 July 2011; Revised 9 February 2012; Accepted 16 February 2012)

References

Atran, S. "The Moral Logic and Growth of Suicide Terrorism." *The Washington Quarterly* 29 (No. 2 2006): 127-47.

Baliga, S., and Sjöström, T. "Arms Races and Negotiations." *Review of Economic Studies* 71 (No. 2 2004): 351-69.

Downs, A. *An Economic Theory of Democracy*. New York: Harper Collins, 1957.

Jackson, M. O., and Morelli, M. "Political Bias and War." *American Economic Review* 97 (No. 4 2007): 1353-73.

Key, V. O. *The Responsible Electorate*. Cambridge, MA: Belknap Press, 1966.

Kreps, D., Milgrom, P., Roberts, J., and Wilson, R. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* 27 (No. 2 1982): 245-52.

Kreps, D., and Wilson, R. "Reputation and Imperfect Information." *Journal of Economic Theory* 27 (No. 2 1982): 253-79.

Lewis-Beck, M. S., Nadeau, R., and Elias, A. "Economics, Party, and the Vote: Causality Issues and Panel Data." *American Journal of Political Science* 52 (No. 1 2008): 84-95.

Milgrom, P., and Roberts, J. "Predation, Reputation and Entry Deterrence." *Journal of Economic Theory* 27 (No. 2 1982): 280-312.

Pape, R. A. *Dying to Win*. New York: Random House, 2005.

Schelling, T. C. *The Strategy of Conflict*. New York: Oxford University Press, 1963.

_____. *Arms and Influence*. New Haven: Yale University Press, 1966.

Woon, J. "Democratic Accountability and Retrospective Voting: A Laboratory Experiment." *American Journal of Political Science*, Forthcoming.