Can You Keep a Secret? Reputation and Secret Diplomacy in World Politics

Abstract

This paper explores how and under what conditions secret diplomacy allows leaders to cooperate. We present a novel game where private cooperation serves a screening function. We argue that private cooperation gives an adversary the opportunity to earn a reputation for trustworthiness by resisting the temptation to leak information from the adversary. Secret diplomacy, therefore, can be an effective tool of statecraft that allows adversaries to build mutual trust. We call this the screening equilibrium. At the same time, our model reveals that a state may also utilize secret diplomacy to induce an untrustworthy adversary to fake trustworthiness by initially keeping negotiations secret. We refer to this as the collusion equilibrium. We illustrate the logic of the model using three cases: secret negotiations between the United States and China leading to Nixon's visit to China; secret diplomacy between Kennedy and Khrushchev over the missiles in Turkey during the Cuban Missile Crisis; and secret dealings between Reagan and Khomeini during the Iran-Contra affair. This paper contributes to debates on information transmission, secrecy, and reputation in international politics.

manuscript length: 34.5 pages without cover page and appendix

1 Introduction

Secret diplomacy is a dangerous practice. Leaks of private communications and covert deals between leaders often result in significant public embarrassment, if not war. The disclosure of the Zimmermann Telegram, in which Germany secretly proposed a military alliance with Mexico by promising that Mexico would receive Texas, Arizona, and New Mexico, prompted the United States into entering World War I against Germany. The leaked private conversation between Wilhelm I and the French ambassador to Prussia, appearing to show insults between both parties regarding candidates for the Spanish throne, instigated intense public outrage in both France and Prussia, contributing to the outbreak of the Franco-Prussian War. At present, a special prosecutor is looking into reports of potentially unlawful collusion between Russia and the Trump transition team. Notwithstanding the risk of explosive leaks, however, leaders routinely rely on secret diplomacy to cooperate with foes. In the past 1,500 years, there have been almost 1,000 *known* secret treaties and agreements (Grosek 2007), and professional diplomats repeatedly emphasize the value of secrecy for successful diplomacy (Putnam 1988: 445). Yet few scholars have theorized the dynamics of secret diplomacy. Given the risks and benefits of secret diplomacy, when and why could secret diplomacy facilitate cooperation and foster trust between foes? Under what conditions would an adversary leak information about those secret negotiations?

Existing studies emphasize that secret diplomacy is valuable because it either serves a signaling function (Yarhi-Milo 2013) or permits concessions that would preserve peace (Kurizaki 2007). Our model highlights a novel mechanism by which secret diplomacy can effectively promote cooperation: leaders can test an adversary's ability to keep secrets. Adversaries who refrain from leaking information about secret overtures from their foe can acquire a reputation for secrecy-trustworthiness.[1] We define this as the beliefs of the leader initiating the secret cooperation about the perceived likelihood that the adversary will keep future assurances secret.[2] Whether an adversary can be trusted to keep

[1] *Nota bene*: To avoid cumbersome language, we use secrecy-trustworthiness and trustworthiness interchangeably.

[2] This is distinct from generalized trust (Rathbun 2011) or geopolitical trust (Kydd 2005). Nevertheless, as we note later, secrecy-trustworthiness could also affect larger beliefs by serving as a litmus

sensitive information secret is crucial for leaders seeking to cooperate. This is because, despite reasonable estimates and good judgment, leaders remain uncertain about their adversary's intentions—does the adversary seek long-term cooperation or simply an opportunity to betray and embarrass the initiator. Our model uncovers when and how secret diplomacy enables leaders to reduce this uncertainty, thereby allowing for more accurate assessment of the adversary's intentions and interests to cooperate.

Whether secret diplomacy reveals useful information about the adversary depends on the degree to which an untrustworthy adversary has incentives to fake trustworthiness. Two key factors determine the severity of this "adverse selection" problem: the adversary's costs from public cooperation, and the adversary's expected gain from betraying the initiator.

First, when an adversary leaks information, he[3] acquires a reputation for being untrustworthy—i.e., that he cannot be trusted to keep secrets —thereby making the initiator unlikely to approach him in private again, and leaving the adversary with only the option of public diplomacy in future rounds.[4] Conducting public diplomacy with a historical enemy, however, could carry significant reputational and political risks. Consequently, those adversaries who anticipate significant costs from engaging in public diplomacy would prefer the secret channel to remain open and thus will be less likely to initially reveal secret assurances. The trustworthy adversary, therefore, cannot credibly signal its trustworthiness by keeping secrets because an untrustworthy adversary would do the same. Secret diplomacy is least effective in screening for types when leaders face the strongest political pressures to go private.[5] Second, secret diplomacy is most effective in fostering future cooperation when the adversary has much to gain from betraying the initiator. When the adversary expects large political and strategic benefits from leaking, the decision not to reveal the initiator's secret assurance is a very costly signal that credibly communicates the adversary's type. Therefore, secret diplomacy is more

test of the adversary's intentions to cooperate.

[3] Though *Journal of Politics* usually uses gender-neutral terms whenever possible, this article will use male pronouns when discussing theoretical leaders to avoid cumbersome, awkward language.

[4] Our model assumes that the initiator deals with the same adversary in both rounds of the game. Reputation in our model, therefore, is leader-specific.

[5] On models in IR with screening, see, e.g., Powell 2004.

efficacious at fostering trust when the adversary faces strong temptations to leak.

When adverse selection is severe, the initiator may still approach an adversary in private even when such communication cannot reveal the adversary's type. This is because the initiator may benefit from temporary collusion before an untrustworthy adversary betrays him. Taken together, the incentives to fake trustworthiness determine whether we observe secret diplomacy that serves to separate types and facilitate long-term cooperation (i.e., the screening equilibrium), or one that merely facilitates temporary collusion without learning (i.e., the collusion equilibrium). The "screening" and "collusion" equilibria point to different rationales, logics, and consequences of secret diplomacy that existing literature has overlooked. We illustrate the empirical relevance of the screening equillibrium by examining the evolution of secret talks between the United States and China and revisiting several puzzling aspects of the Cuban Missile Crisis. We also briefly illustrate the plausibility of the collusion equilibrium in the context of the Iran-Contra arms trade.

This paper makes three contributions. First, existing studies often emphasize how adversaries conceal vital geo-strategic and military information for tactical and strategic advantages (Fearon 1995; Slantchev 2010). Our analysis, however, shows that leaders can strategically share sensitive private information in order to test the adversary's intentions. By explaining the conditions under which they choose to reveal private information, we refine our understanding of the origin and implications of information asymmetry in conflict.

Second, we offer the first formal analysis of the conditions under which a leader would use private assurances over public ones, and when an adversary would leak a secret assurance. This topic has received surprisingly little attention, although it is hardly conceivable that leaders would approach an adversary covertly without considering how that adversary might respond to secret diplomacy. Without studying the logic of leaks, we can only gain a partial understanding of the benefits and risks of secret diplomacy. Indeed, by analyzing leaders' calculus of discretion, we generate novel insights on when an adversary can effectively signal behind closed doors. Too much opposition from domestic or international audiences for public cooperation, we reveal, nullifies the communicative value of secret cooperation because it drives both trustworthy and untrustworthy adversaries to behave the same way and keep secret diplomacy quiet.

Third, our analysis advances the scholarship on reputation. We theorize the mechanisms and conditions under which reputation for secrecy-trustworthiness forms, a topic which existing studies of reputation in IR have neglected. Furthermore, while studies of reputation often focus on one type of reputation in isolation—whether they concern honesty (Sartori 2005), hostility (Crescenzi 2007), or resolve (Weisiger and Yarhi-Milo 2013)—we examine how different reputational considerations interact to shape international outcomes. Specifically, we focus on the interaction between the adversary's interest, on the one hand, to acquire a reputation for secret-keeping, and its interest, on the other hand, to avoid a reputation for weakness on national security in the eyes of domestic and international audiences.

2 The Limits of Existing Explanations

Why would leaders prefer secret diplomacy over public diplomacy? The first class of existing explanations highlights the signaling value of secret diplomacy. Classic works on crisis bargaining highlight how public diplomacy is informative because it generates domestic audience costs (Fearon 1994; Smith 1998; Schultz 2001). However, as Yarhi-Milo (2013) notes, private assurances are also costly, because of the risk of leaks from both the adversary and third parties. Consequently, private assurances could also serve as an important signal of the initiator's intentions. On the other hand, as Sartori (2005) shows, states pay reputational costs for dishonesty when they do not keep their promises, whether the negotiations leading to those promises are public or secret. Furthermore, Trager (2010) demonstrates that both public and private threats effectively convey resolve, because even a private threat may increase the risk of war by provoking an adversary to retaliate in response to the threat. Finally, private threats also communicate resolve in multi-dimensional bargaining spaces because states that care only about one particular issue are often reluctant to threaten to fight over an auxiliary issue, because such a "lie" can decrease their chances of getting a concession on the issue that they really care about (Trager 2011). Studies that emphasize the communicative value of secret diplomacy, despite their variations, all suggest that secret diplomacy is valuable because it credibly signals trustworthiness.

The second class of explanations suggests that secret diplomacy is valuable because it reduces the risk that negotiation would break down when an adversary faces strong domestic opposition for compromise. In Stasavage (2004), leaders face domestic reputation concerns regarding whether they share the same preference as their domestic constituents during an international dispute. Hence, leaders who wish to signal to their domestic audience that they are not "sell-outs" are predisposed to adopt hard-line bargaining positions when diplomacy is public. Therefore, leaders would sometimes prefer closed-door bargaining over public bargaining. Tarar and Leventoglu (2005), on the other hand, suggests that if diplomacy is public, the leader from each side of a dispute faces a strong incentive to make a public commitment to the leader's domestic audience in order to extract a bargaining concession from the opponent. Public diplomacy, therefore, generates a prisoner's dilemma dynamic where all leaders adopt intransigent positions that lead to bargaining deadlock; secret diplomacy, in contrast, is not associated with such problems. Finally, Kurizaki (2007), provides the first fully developed formal analysis of how crisis diplomacy unfolds in private. Private threats, according to Kurizaki (2007), are effective because they allow an adversary to capitulate to a challenge to avoid war without suffering domestic political consequences. In contrast with studies that emphasize the signaling value of secret diplomacy, this class of explanations suggests that leaders resort to secret diplomacy to expand the political space for them to maneuver, which makes it easier to strike a deal with an adversary.

Scholars have made much headway in explaining why secret diplomacy is valuable as a tool of statecraft. Nonetheless, there are still theoretical and empirical gaps. First, while Carson (2016) explores when leaders refrain from publicizing a foe's covert military action in order to manage escalation, no study to our knowledge has examined why and when leaders will keep diplomatic overtures or agreements secret, and under what conditions they will leak them.[6] Understanding these dynamics allows us not only to uncover when secret diplomacy is likely to result in trust-building, but also when it will likely end in a political disaster. Second, existing works focus on private versus public threats. Very little attention has been dedicated to studying the logic of private versus public assurance (with the exception of Yarhi-Milo 2013), and the conditions under which it is likely to succeed. Third, existing studies do not help answer why adversarial leaders sometimes engage in multiple rounds of

[6] Although both Kurizaki (2007, 550) and Yarhi-Milo (2013, 419) hint at this issue's importance.

secret cooperation, while, at other times, they only pursue one-off secret cooperation on a single issue. By filling these gaps, we are able to gain a better understanding of both the functions and limitations of secrecy in international politics.

3 Theory

Our theory seeks to answer three related questions: (1) When and why would a leader pursue secret rather than public diplomacy? (2) When would an adversary resist the temptation to leak? and (3) Under what conditions can secret diplomacy lead to effective screening of the adversary's types instead of one-off collusion? To address these questions, our model builds on two observations about diplomacy between adversaries.

First, political leaders may face costs for publicly cooperating with an adversary. All else equal, leaders who cooperate with a sworn enemy in public could face both domestic and international criticism for being "weak," if not treacherous, for reasons of ideology, historical acrimony, or strategy. Some Arab regimes, such as Saudi Arabia, have had to keep their dealings with Israel private to avoid raising the ire of fellow Arab countries and the Palestinian Authority." Similarly, one of the reasons the United States does not conduct public negotiations with terrorist organizations is so U.S. allies and adversaries do not perceive it as being "soft." We expect a leader who faces multiple adversaries to incur greater reputational costs, as well as a leader facing a hawkish domestic audience that could effectively impose political punishment. (Schultz 2005).

Furthermore, initiators of public diplomacy can be subject to other types of non-reputational costs for engaging with the adversary. For example, public diplomacy could induce domestic or international third parties to sabotage the negotiations (Kydd and Walter 2002), or demand strategic or political concessions from the initiator in return for their support (Putnam 1988: 451). Regardless of their nature or origins, these costs can prompt leaders to consider the option of secret rather than public diplomacy.

Second, when a leader secretly approaches an adversary, the adversary is often tempted to publicize the secret deal in order to humiliate the initiator. Domestic and international audiences who

would object to public diplomacy with the adversary are likely to be even more resentful when they find out through leaks that the leader offered secret concessions to the adversary. The revelation that secret diplomacy was pursued can make the initiator appear not just dishonest by concealing the his/her overtures (Stasavage 2004), but also incompetent given that he/she was played by the adversary (Smith 1998; Gelpi and Grieco 2015). Indeed, domestic or international audiences might believe that the initiating leader made himself vulnerable to exploitation by the adversary. In leaking the information, the adversary could divulge sensitive information, and embarrass the initiator in order to claim a diplomatic victory or a strategic advantage. Moreover, domestic actors may decry a lack of engagement with the public (especially in democracies), with government agencies, or with the ruling elite (in both democracies and non-democracies).

Allies affected by the secret deal could also be outraged and might argue that they should have been consulted. That the leader kept this secret from allies could be seen as an indication of his/her willingness to pursue a foreign policy inconsistent with the allies' shared interests. For example, when the United States secretly negotiated the details of the Iranian nuclear deal, Israel and Saudi Arabia complained that the United States was hiding damaging concessions that could affect their national security. Finally, revelation of secret dealings with an adversary could lead audiences to wonder what other secret dealings the leader has not yet disclosed, thereby undermining the leader's credibility. Thus, a strategic leak by the adversary could adversely affect the leader's trustworthiness and competence in the eyes of domestic and international audiences.

The above discussion implies, therefore, that all else being equal, adversaries who wish to humiliate their opponent can do so by leaking the adversary's secret overtures. The benefits from such betrayal are more substantial under some circumstances. By leaking, an adversary may signal to domestic audience and allies that he/she is honest and unwilling to keep secrets from these audiences. Moreover, the political "brownie points" that adversaries may score from leaking might be sizeable, especially the larger the ideological distance (Haas 2005) between an adversary and an initiator of secret diplomacy. Moreover, adversaries could see larger benefits from leaks when they believe those leaks are part of a broader plan. For instance, if the adversary wishes to topple the leader who initiated secret contact, s/he is more likely to leak when s/he believes the opposition in the initiator's country

could remove her/him from office. If the desired outcome by the adversary is to drive a wedge between the state that initiated secret diplomacy and that state's allies, then the adversary will have stronger incentives to leak when he believes such information could damage the alliance relationships.

The two observations above about the potential costs and benefits facing the initiator and the adversary when they engage in secret diplomacy are the key assumptions driving our model. While a state has the incentive to go private to avoid the reputational costs associated with appearing weak when publicly cooperating with a foe, its adversary has the incentive to take advantage of that strategy by leaking the fact of secret diplomacy out of its own self-interest. In the spirit of Yarhi-Milo (2013), we argue that it is precisely the temptation for the adversary to leak that makes secret diplomacy an effective tool of trust-building. Yet, we also show how and when the adverse selection problem significantly reduces the informative value of secret cooperation. Moreover, rather than emphasizing the ability of the initiator to signal intentions through secret diplomacy like Yarhi-Milo (2013), we focus on how the risky nature of a private deal can allow an initiator to screen out an untrustworthy opponent (and identify a trustworthy one).[7]

3.1 The secret diplomacy model

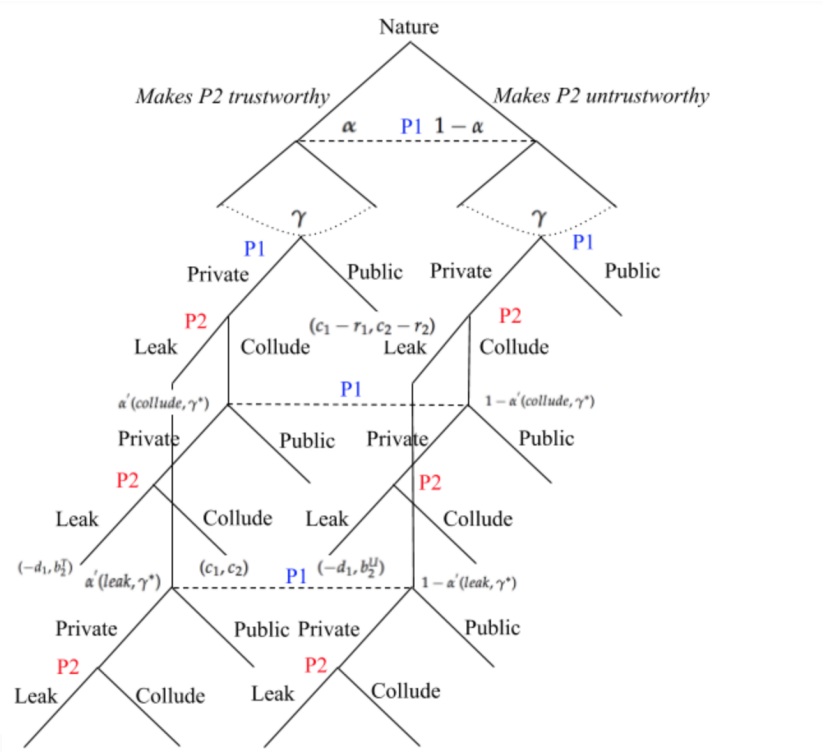
There are two players in this game: the initiator (P_1) and an adversary (P_2). The adversary has two types, trustworthy (P_2^T) and untrustworthy (P_2^U). When approached by P_1 in secret, the trustworthy adversary P_2^T receives a higher payoff for secret cooperation (reciprocating cooperation in secret) compared to betrayal (publicizing the initiator's attempt at "cutting a deal under the table"). In contrast, the untrustworthy adversary P_2^U receives a lower payoff for secret cooperation relative to betrayal. P_1 does not observe P_2's type. In other words, P_1 is uncertain about how the adversary would evaluate the benefits and costs of secret cooperation versus leaking, which are a function of the adversary's domestic political concerns and assessment of the strategic environment, as discussed earlier.

Nature starts the game by choosing the opponent's type, with probability α that the adversary is trustworthy. In this model, "trust" α refers only to the initiator's beliefs about the likelihood that

[7] Riley (2001) surveys screening models.

the adversary will keep the assurances secret. The initiator's level of trust in the adversary could be a result of a number of factors, including past interactions between the countries, the initiator's perceptions—whether accurate or not—about adversary leader's nature, as well as his/her personal assessment of the adversary's interests and intentions to leak. [8]

Figure 1: A Game of Secret Assurance and Reputation



Note: To avoid clutter, we did not specify all payoffs and the sequence of moves after public diplomacy.

Next, P_1 chooses the "stake" associated with the first round game γ relative to the second round $(1-\gamma)$, with $0 \leq \gamma \leq 1$. [9] The choice of γ may signify the importance of the issue driving P_1

[8] Put differently, trust is the initiator's belief in the probability the adversary will leak. Trust is thus leader and situation specific, and not an inherent trait of an actor.

[9] Leaders might not always have the choice on which issues they can offer assurances. There might be imminent issues that would take precedence over others as the Iran-Contra affair demonstrates. But even in such situations, leaders can still recognize the extent to which secret cooperation can reveal the trustworthiness of the adversary. See Watson (1999) and Kydd (2000; 2005) for models that endogenize the weights players put on different rounds of cooperation.

to initially approach P_2. There are a number of potential issues on which an adversarial dyad could choose to cooperate; part of the challenge for the initiator is to decide how to structure the cooperation regime. Specifically, the initiator must decide whether to approach the opponent in private with a small issue before a big issue, or vice versa. Because the number of possible issues on which any two countries could cooperate is limited, the choice over how to initially structure the cooperation regime has consequences for how the game proceeds. (Kydd 2000; Kydd 2005). [10]

Our game consists of two rounds, which allows P_2 to build a reputation in the first round that would affect P_1's decision to go private or public in the second round.[11] In each round, P_1 decides whether to cooperate with P_2 in public or in private. If P_1 approaches the opponent in public, P_2 can accept the offer of public cooperation, and the two players receive payoffs $(c_1 - r_1, c_2 - r_2)$.[12] The gain from cooperation is c, and r is the political cost associated with making a public deal with a foe. P_2 may also reject P_1's offer of public cooperation, and players will receive payoffs $(-r_1,0)$. The game ends in the first round if P_1 chooses public diplomacy.[13] Without loss of generality, P_1 has no option of retaining the status quo – which entails not reaching out to the opponent at all – because the focus is on examining P_1's incentive to cooperate privately versus publicly. We make this assumption in the spirit of Kurizaki (2007), who assumes that the initiator of coercive diplomacy does not have the option of retaining the status quo, in order to sharpen the comparison between public and private threats. Like Kurizaki (2007), the main results remain effectively unchanged when we allow P_1 to opt

[10] Our model inherits from Watson (1999) the assumption that both sides agree on the importance of an issue.

[11] We follow the canonical models of trust-building in international politics (Kydd 2000, 2005) and restrict our attention to studying the simplest dynamic problem: a two-round game. For a multi-round game where partners jointly craft a cooperation regime that allows for screening, see Watson (1999).

[12] We model secret assurance and reputation building, rather than the dynamic of *secret bargaining*, e.g., why the initiator will offer a particular secret deal c_2 to the adversary, because the latter has already been done (Stasavage 2004; Ramirez 2017).

[13] We have solved a version of the model where there is a second round after the public diplomacy. This model has an additional equilibrium where public diplomacy screens. This paper focuses on screening with secret diplomacy because our "simple" model already provides rich formal results and we wish to highlight the understudied phenomenon where reputation can be gained and lost behind closed doors.

for no diplomacy.

If P_1 attempts private cooperation, P_2 can reach a secret deal with P_1 that would give the players payoffs (c_1, c_2). Alternatively, P_2 can betray P_1 by leaking.[14]The reputational damage P_1 suffers due to a leak is d_1. For P_2^T, $c_2 > b_2^T$, while for P_2^U, $b_2^U > c_2$. The superscript i indexes the type of P_2,$i \in I \equiv T, U$; T indicates secrecy-trustworthiness while U indicates untrustworthiness. As discussed previously, P_2 benefits from a leak, because it is likely to tarnish P_1's reputation for resolve, honesty, and competence. *Ex ante*, P_1 does not know whether P_2 will value secret cooperation more than the potential benefits associated with betrayal.

The second round is identical to the first. P_1 decides whether to approach P_2 in private or in public. If P_1 goes public, P_2 either accepts or rejects the offer of public cooperation. If P_1 goes private, P_2 either accepts the offer of private cooperation or betrays P_1 by leaking. Importantly, in the second round, P_1 may form a new belief $\alpha'(s_2, \gamma)$ regarding the trustworthiness of P_2 given the history of the game in the first round and P_1's choice of γ. The $'$ superscript indicates that the belief is associated with the second round sub-game, the * signifies optimal choice, and s_2 denotes P_2's first round strategy with $s \in S \equiv cooperate, leak$. When P_1 updates his/her belief regarding P_2's type after observing P_2's past behavior, $\alpha'(s_2, \gamma)$ will be distinct from α, the prior belief that P_2 will not leak. When P_1 does not update his/her belief regarding P_2's type after observing P_2's strategy in the first round game, $\alpha'(s_2, \gamma)$ equals α. P_2^T can only acquire a reputation of secrecy-trustworthiness when P_1 considers past behavior as a good indicator of whether P_2 is going to leak in future secret cooperation.

Finally, adversaries often probe whether temporary cooperation on issues where state interests potentially overlap. For example, although Jordan and Israel had been embroiled in a bloody conflict for many years, they often discussed issues of mutual interest such as fighting against extreme Islamist

[14]To simplify the model, when P_2 rejects a secret offer, she also leaks. Crucially, allowing P_2 to reject the secret offer while not leaking will not affect the equilibrium strategic choice. For P_2^T, rejecting a secret deal without leaking gives her a pay-off of 0. Therefore P_2^T will always prefer secret cooperation over rejecting secret assurance without leaking and rejecting secret assurance with leaking. For P_2^U, rejecting a secret deal without leaking also gives her a pay-off of 0. Therefore P_2^U will always prefer rejecting secret assurance with leaking (which gives her a positive pay-off) over rejecting secret assurance without leaking. Importantly, since P_2^T will always accept the secret assurance, rejecting a secret assurance without leaking will not prevent P_2^U from acquiring a reputation of indiscretion.

terrorist groups or keeping the Jordanian monarchy in power. As such, we focus on the leaders' choice to conduct diplomacy secretly or publicly, rather than the decision whether to conduct diplomacy at all. We thus assume $c_1 > r_1$ and $c_2^i > r_2$ (see Appendix B for a model where we relax these constraints). The first inequality ensures that P_1 will always receive a positive payoff from public diplomacy. The second inequality ensures that P_2 will always cooperate when P_1 approaches him with a public deal. The two inequalities guarantee that P_1 will choose secret diplomacy not because public diplomacy is worse than the status quo, but rather because secret diplomacy could offer a larger payoff from cooperation relative to public diplomacy. Put differently, we assume that public cooperation allows leaders to improve their position over the status quo, even after paying a cost for engaging in public diplomacy. Figure 1 presents the extensive form of our secret diplomacy game.

Table 2: Player strategies and beliefs

First round	***Strategy***	***Belief(s)***
P_1	$\gamma \in [0,1]$; $\{Private, Public\}$	α
P_2^U	$\{collude, leak\}$(if $Private$); $\{cooperate, reject\}$ (if $Public$)	n/a
P_2^T	$\{collude, leak\}$(if $Private$); $\{cooperate, reject\}$ (if $Public$)	n/a
Second round		
P_1	$\{Private, Public\}$	$\alpha'(s_2^*, \gamma^*)$
P_2^U	$\{collude, leak\}$(if $Private$); $\{cooperate, reject\}$ (if $Public$)	n/a
P_2^T	$\{collude, leak\}$(if $Private$); $\{cooperate, reject\}$ (if $Public$)	n/a

3.2 Equilibrium analysis

The equilibrium concept employed here is Perfect Bayesian Equilibrium (PBE; Gibbons 1992: chapter 4; Morrow 1994: chapter 8), which requires that: (1) all players play strategies that correspond to their beliefs/information sets (*sequential rationality*); (2) beliefs – both on and off the equilibrium path – are determined by Bayes rule and the players' equilibrium strategies whenever possible (*consistency of belief*). There are three classes of pure strategy PBE for our model: private diplomacy equilibrium with screening (henceforth the "screening equilibrium"), private diplomacy equilibrium with collusion (henceforth the "collusion equilibrium"), and public cooperation equilibrium. For each

equilibrium, we specify the strategies of p_1, p_2^T and p_2^U in rounds 1 and 2, and player 1's beliefs α and $\alpha^{'}(s_2^*, \gamma^*)$. All proofs are in the appendix. Table 2 outlines the options and beliefs facing each actor in rounds 1 and 2.

The screening equilibrium

the screening equilibrium, the trustworthy adversary can acquire a reputation for secrecy-trustworthiness by not leaking the cooperation, because the untrustworthy adversary would be induced to leak, thereby revealing his/her true type in the first round. The screening equilibrium works because P_1 has set the importance of the first round secret diplomacy γ high enough that an untrustworthy adversary would leak and reveal his/her true colors in the initial round of the game. As γ becomes larger, an untrustworthy adversary would find it increasingly attractive to betray the initiator in the initial round. But the ability of the initiator to screen the adversary's intentions under such conditions comes with a significant risk. As γ becomes larger, the initiator of secret diplomacy could face a more explosive leak if the adversary betrays him.

The screening equilibrium emerges when two conditions are met: (1) the initiator is sufficiently trustful of the adversary to give secret diplomacy a chance; (2) the untrustworthy adversary has only a moderate incentive to mimic the trustworthy adversary by not leaking initially. On the requisite level of the initiator's trust in the adversary that would sustain screening, note that the initiator will reap a higher expected payoff from secret diplomacy with screening compared to public diplomacy whenever α – the the initiator's initial level of trust in the adversary's discretion - is above a critical threshold $\tilde{\alpha}_{screening}$:

$$\tilde{\alpha}_{screening} \equiv 1 - \frac{r_1}{\hat{\gamma}(d_1 - r_1) + r_1}$$

When $\alpha > \tilde{\alpha}_{screening}$, the initiator is sufficiently trustful of the adversary to employ secret diplomacy as a screen. When $\alpha < \tilde{\alpha}_{screening}$, the initiator is too fearful of a leak to use secret diplomacy. $\tilde{\alpha}_{screening}$ is a critical value that signifies the minimum level of trust that the initiator must have in the adversary for him to bear the risk of a leak and employ secret diplomacy for screening. $\tilde{\alpha}_{screening}$ is

increasing in $\hat{\gamma}$, the incentive for the untrustworthy adversary to keep cooperation secret in the first round:

$$\frac{c_2 - b_2^U - r_2}{2c_2 - 2b_2^U - r_2}$$

When $\hat{\gamma}$ is large, the untrustworthy adversary has a strong incentive to mimic the trustworthy type in the first round. When $\hat{\gamma}$ is small, the untrustworthy adversary has a weak incentive to mimic the trustworthy type in the first round. When the untrustworthy adversary is strongly inclined to not leak in the first round, the informational value of secret diplomacy diminishes, and the initiator becomes less willing to bear the risk of secret diplomacy. $\hat{\gamma}$ is a function of two factors: the adversary's anticipated cost from cooperating in public, and its potential gains from leaking. We discuss each in turn.

First, the untrustworthy adversary is compelled to fake trustworthiness early on when he anticipates facing high reputational costs if he cooperates with the initiator in public in the next round ($\hat{\gamma}$ is increasing in r_2). This is because when the initiator learns that the adversary is untrustworthy after the first round, he will only be willing to offer public cooperation in the subsequent round in order to prevent further opportunities for the untrustworthy adversary to leak. In other words, if the untrustworthy adversary cannot resist the temptation to leak in the first round, the initiator will only be willing to proceed with public cooperation, thereby forcing the untrustworthy adversary to pay the political cost associated with public diplomacy in the second round if he chooses to cooperate. Consequently, the untrustworthy adversary has a strong incentive to keep diplomacy behind closed doors in both rounds of the game when public cooperation is very undesirable because of its associated high political cost. Indeed, the net loss ("punishment") that the untrustworthy adversary suffers by acquiring a reputation of untrustworthiness if he leaks in the first round is $(1-\gamma)(d_2^{U'} - c_2' + r_2')$, which is the untrustworthy adversary's second round leaking payoff $(1-\gamma)d_2^{U'}$ minus his/her second round public cooperation payoff $(1-\gamma)(c_2' - r_2')$; $(1-\gamma)(d_2^{U'} - c_2' + r_2')$, and which increases as the cost for public cooperation r_2' rises.

Second, the gains the adversary could receive from leaking also affect an untrustworthy adver-

sary's initiatives to fake trustworthiness. Parameter $\hat{\gamma}$ is decreasing in b_2^U; therefore the untrustworthy adversary will less often fake trustworthiness when the payoff from leaking is high. By not leaking initially, the untrustworthy adversary forgoes the payoffs from immediate betrayal, γb_2^U, which is the opportunity cost of mimicking the trustworthy type. This opportunity cost rises as the payoff from leaking, b_2^U, increases. Hence as the gain from humiliating the initiator rises, the untrustworthy adversary's incentive not to leak early diminishes. This is good news for the trustworthy adversary, who would now find it easier to acquire a reputation of trustworthiness by not leaking, since the initiator knows that it would be difficult for the untrustworthy type to resist the temptation of a profitable early betrayal. Thus, secret diplomacy is an effective litmus test of an adversary's trustworthiness when the untrustworthy opponent does not have a strong incentive to cooperate in secret in round one.

Next, we explicate the initiator's incentives to test an adversary with secret diplomacy.[15] First, we find that when an untrustworthy adversary has a strong incentive to mimic a trustworthy type, screening is more difficult and therefore less likely. In such circumstances, initiators who wish to screen face a dilemma. On the one hand, the initiator could approach the adversary with a very important issue in the first round (set γ high) to make it worthwhile for the untrustworthy adversary to leak and thereby reveal its type. Structuring the secret cooperation regime in this way would allow the trustworthy type to credibly signal trustworthiness by not leaking. On the other hand, such a strategy comes with a significant risk for the initiator. Importantly, the initiator has to pay a "price" $\alpha\gamma d_1$—the expected damage from betrayal in the first round—to use secret diplomacy as an effective screen. Thus, the larger the stake, the more damaging a leak would be as the initiator raises γ in order to bait the untrustworthy adversary into leaking. In contrast, when the untrustworthy adversary has a weak incentive to mimic, the initiator could choose a moderately important issue to screen out the

[15] To be sure, we recognize that leaders practice secret diplomacy not just to test the adversary, but also because they seek cooperation with that adversary despite domestic and international opposition. Indeed, screening is valuable for the initiator because it allows him to decide whether to go attempt secret assurance again after initial secret contact. Importantly, the initiator is more likely to practice secret diplomacy instead of public diplomacy if he knows that secret diplomacy both facilitates cooperation and screens. Indeed, the adverse selection problem deters the initiator from going private because it eliminates the initiator's expected pay-off associated with learning the adversary's type (more discussion in the next sub-section).

untrustworthy adversary.

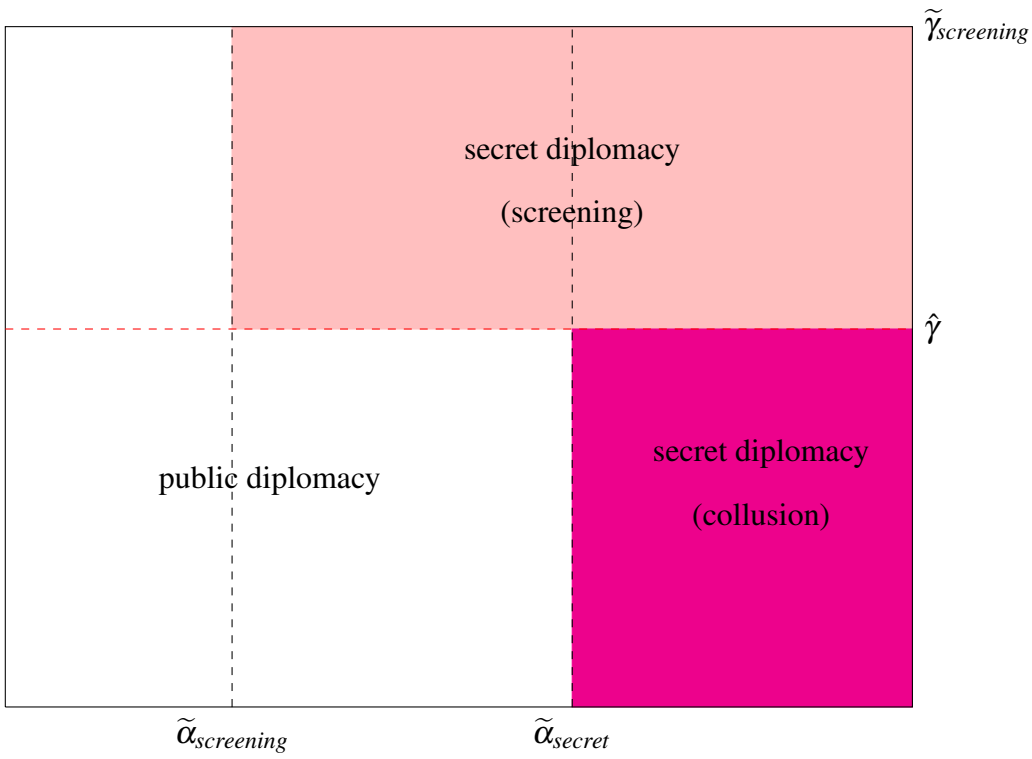
Given the costs and benefits of screening, the maximum stake that P_1 will be willing to put on the first round of the secret diplomacy game in order to screen is $\tilde{\gamma}_{screening}$:

$$\frac{c_1 - r_1 + d_1}{2c_1 - r_1 + 2d_1}$$

$\tilde{\gamma}_{screening}$ indicates the initiator's incentive to learn the adversary's type at the risk of exposing himself to a damaging leak. When $\tilde{\gamma}_{screening}$ is large, the initiator has a strong incentive to test the adversary with secret diplomacy. When $\tilde{\gamma}_{screening}$ is small, the initiator has a weak incentive to screen. Importantly, the initiator has less incentive to use secret diplomacy as a litmus test when the political cost r_1 associated with public cooperation is high ($\tilde{\gamma}_{screening}$ is decreasing in r_1). The intuition is as follows: After identifying the untrustworthy adversary in the first round, the initiator would go public, which not only eliminates the opportunity for the untrustworthy adversary to leak again, but also allows the initiator to cooperate with the untrustworthy adversary in public and receive payoff $(1-\gamma)(c_1' - r_1')$. Consequently, the initiator's incentive to identify the untrustworthy adversary is stronger when the public cooperation payoff is high, and lower when that payoff is small. Crucially, the public cooperation payoff diminishes as the political cost for public cooperation rises. Hence the initiator will have only a weak incentive to put up an effective screen in the first round when he anticipates facing high costs for cooperating with the adversary in public.

Taking the incentives of both sides into account, secret diplomacy is least likely to build trust when both the initiator and the adversary expect to pay high political costs for public cooperation. Specifically, when the adversary faces a sizable cost for a public deal, the untrustworthy type would have a strong incentive to resist the temptation to leak early on in order to keep diplomacy behind closed doors ($\hat{\gamma}$ is high); this makes it difficult for the trustworthy type to signal trustworthiness by not leaking. On the other hand, when the initiator faces a high cost for public cooperation, he has less incentive to gauge the trustworthiness of an adversary by approaching that adversary with an important issue ($\tilde{\gamma}_{screening}$ is low).

Figure 2: Secret and public diplomacy equilibria



Note: The x axis is α, the level of trust P_1 has in P_2. The y axis is $\tilde{\gamma}_{screening}$, the maximum stake that P_1 would be willing to put on the first round of the secret diplomacy game to screen. $\tilde{\alpha}_{secret}$ is the minimal level of trust that P_1 must have in P_2 for P_1 to go private. $\tilde{\alpha}_{screening}$ is the minimal level of trust that P_1 must have in P_2 for P_1 to go private and screen instead of going public. $\hat{\gamma}$ is the minimum weight associated with the first round issue that would induce P_2^U to leak early on.

Put differently, whether the screening equilibrium emerges depends critically on the initiator's incentive to screen the adversary through secret diplomacy (how large $\tilde{\gamma}_{screening}$ is), and the cost of employing secret diplomacy effectively, which is a function of the untrustworthy opponent's willingness to fake trustworthiness (how large $\hat{\gamma}$ is). Only when the gain of screening outweighs the cost ($\hat{\gamma} \leq \tilde{\gamma}_{screening}$) will the initiator choose a sufficiently important issue to screen the adversary.

Finally, the initiator's optimal stake γ^* is simply $\hat{\gamma} = \tilde{\gamma}_{screening}$, which is the lowest possible γ that would guarantee separation of the untrustworthy adversary and the trustworthy one in the first round of the game. Proposition 1 formally characterizes the screening equilibrium.

Proposition 1 (screening equilibrium): If $\tilde{\alpha}_{screening} \leq \alpha$ and $\hat{\gamma} \leq \tilde{\gamma}_{screening}$, there exists a unique perfect Bayesian equilibrium of the game with the following strategies. P_1 sets $\gamma = \tilde{\gamma}_{screening}$. In the first round of the game, P_1 goes private, while P_2^T colludes and P_2^U betrays. P_1 learns that P_2 is trustworthy if P_2 colluded in the first round ($\alpha^{'}(c_2^*, \hat{\gamma}) = 1$), and untrustworthy if P_2 leaked ($\alpha^{'}(l_2^*, \hat{\gamma}) = 0$). In the second round of the game, P_1 goes private if P_2 colluded in round 1 and goes public if P_2 leaked in round 1. P_2^T colludes again if P_1 goes private. Both P_2^T and P_2^U cooperates if P_1 goes public.

The collusion and the public diplomacy equilibria

There are two equilibria of the game where secret diplomacy does not solve a screening function: the collusion equilibrium and the public diplomacy equilibrium.

For the collusion equilibrium, the initiator always receives the private cooperation payoff c_1 (weighted by γ) in the first round. There is no risk of a leak, because the untrustworthy adversary will mimic the trustworthy type to avoid alarming the initiator until he is ready to betray him at a more opportune time (i.e., round two). Since both the trustworthy and the untrustworthy adversaries collude in the first round of this equilibrium, the initiator does not learn any useful information about the adversary's trustworthiness after the initial encounter with the adversary ($\alpha^{'}(s_2^*, \gamma_{collusion}^*) = \alpha$). In the second round, the initiator goes private again, the trustworthy adversary colludes, but the untrustworthy adversary now leaks. Under this equilibrium, temporary collusion paves the way for a possible leak later.

Some leaders will find this type of secret diplomacy appealing because it offers a possibility of short-term cooperation over an immediate issue of concern for the initiator. Importantly, the initiator is more inclined to adopt this approach if the untrustworthy adversary has a strong incentive to mimic the trustworthy type in the first round ($\hat{\gamma}$ is high). This is likely to be the case when the adversary faces sizable political cost from public cooperation and when the adversary has much to gain from an early betrayal, as discussed earlier. Both conditions will make screening riskier for the initiator, because the initiator needs to set a high γ to screen when adverse selection problem is severe. Furthermore, collusion will be more likely when the initiator has little incentive to screen, e.g., when he anticipates

high cost of public cooperation.

As Figure 2 shows, the collusion equilibrium requires the initiator to have a higher level of trust in the adversary's discretion compared to the screening equilibrium. This is the case because the initiator in this equilibrium must take the risk associated with private diplomacy in the second round, after learning nothing about the adversary's type in the first round ($\alpha' > \tilde{\alpha'}_{secret} = 1 - \frac{r'_1}{c'_1+d'_1}$, with $\alpha = \alpha'$, as the initial level of trust in the adversary's discretion is identical to the level trust in round two under the collusion equilibrium). Secret diplomacy in the second round, importantly, has no informational value for the initiator although it still facilitates cooperation between foes without reputational cost; the initiator does not benefit from learning the adversary's type in round two because the game ends in round two. The initiator's expected gain from going private in the second round, therefore, is smaller than its expected gain from going private in the first round (where secret diplomacy not only allows for the possibility of cooperation without reputational cost, but also serves a screening function that the initiator may decide to utilize). Consequently, the initiator must be quite confident in the adversary's trustworthiness for him to practice secret diplomacy with only temporary collusion in mind.

The discussion above highlights how a trustworthy adversary suffers from adverse selection. When the untrustworthy adversary has a strong incentive to mimic the trustworthy type, the initiator is more likely to go public. This is bad news for the trustworthy adversary, who always receives a higher pay-off from secret, rather than public, diplomacy. Thus, because an untrustworthy adversary enters the "secret diplomacy market" pretending to be trustworthy, the initiator goes public, and the trustworthy adversary is constrained to public diplomacy.

Proposition 2 summarizes our discussion of the collusion equilibrium.

> **Proposition 2 (collusion equilibrium)**: If $\tilde{\alpha'}_{secret} < \alpha'$ and $\hat{\gamma} > \tilde{\gamma}_{screening}$, there exists a unique perfect Bayesian equilibrium of the game with the following strategies. P_1 sets $\gamma = \hat{\gamma} - \varepsilon$. In the first round of the game, P_1 goes private, while both P_2^T and P_2^U collude. In the second round of the game, P_1 maintains his/her prior belief $\alpha' = \alpha$ and goes private, while P_2^T colludes and P_2^U betrays.

Lastly, under the public diplomacy equilibrium, the initiator goes public, and both the trustworthy and

untrustworthy adversaries accept a public deal. This equilibrium emerges under two conditions. First, the initiator will go public when he is fairly confident that the adversary will leak ($\tilde{\alpha}_{screening} > \alpha$). Second, the initiator will go public when screening is too costly and when he does not have sufficient trust in the adversary to collude. As discussed previously, the initiator can find it too risky to put up an effective screen to induce the untrustworthy adversary to reveal its true color early on, e.g., whenever $\hat{\gamma} < \tilde{\gamma}_{screening}$. On the other hand, if the initiator does not have a high level of trust in the adversary, he would not employ secret diplomacy again after the first round. This prevents the initiator from opting for a two-round secret cooperation regime that facilitates collusion. In brief, public diplomacy is the default option for the initiator when he has very little trust in the adversary or when he has medium level of trust in the adversary but finds it too difficult to screen.

The second condition highlights how the adverse selection problem deters the initiator from practicing secret diplomacy. We may decompose the initiator's expected pay-off from secret diplomacy into two components: (1) the pay-off from cooperating with an adversary without paying political cost for cooperation and; (2) the pay-off from learning the adversary's type.[16] When the adverse selection problem is severe, the initiator is unlikely to screen with secret diplomacy, thereby eliminating the second component of the initiator's expected pay-off from secret diplomacy and reduces the initiator's incentive to go private.

Proposition 3 formally characterizes the public diplomacy equilibrium.

> **Proposition 3 (public diplomacy)**: If (a) $\alpha < \tilde{\alpha}_{secret}$ or (b) $\tilde{\alpha}_{screening} < \alpha < \tilde{\alpha}_{secret}$ and $\hat{\gamma} < \tilde{\gamma}_{screening}$, there exists a unique perfect Bayesian equilibrium of the game with the following strategies. P_1 goes public and sets $\gamma = 1$. Both p_2^T and p_2^U accept the public deal.

Figure 2. summarizes propositions 1-3 graphically and underscores how a trustworthy adversary can *only* acquire a reputation for secrecy-trustworthiness when two conditions hold simultaneously: (1)

[16] We can further decompose the pay-off from learning the adversary's type into two components: (1) the pay-off from cooperating with a trustworthy opponent in the future after screening; (2) the pay-off from avoiding a leak in the future when the adversary turns out to be untrustworthy after screening.

the adverse selection problem is not too severe; (2) the leader who initiated secret diplomacy has at least a moderate level of trust in the adversary. If (1) holds but not (2), leaders would find it too risky to employ secret diplomacy (when $\alpha < \tilde{\alpha}_{secret}$) even if secret diplomacy can help them gauge the trustworthiness of an opponent. If (2) holds but not (1), leaders are unlikely to find it worthwhile to screen through secret diplomacy.

3.3 Applicability of the model

To sharpen the analytical focus of this paper, our model included a number of simplifying assumptions. In this section, we discuss how our model generates novel insights for a wider range of secret diplomacy scenarios.

Plausible deniability

Leaders sometimes rely on mediators to engage in private diplomacy because, if there is a leak, leaders may want to distance themselves from the secret communication by claiming that the mediators have acted without their consent (Pruitt 2007; Powell 2013). Our model has implications for understanding cases where leaders may deny their involvement in secret diplomacy. The ability of the initiator to deny involvement in private diplomacy affects values of two parameters in our model: (1) the reputational damage that the initiator suffers from a leak ($-d_1$) would decrease and (2) the untrustworthy adversary's payoff from betrayal (b_2^U) would decrease. Crucially, the initiator will become more tolerant of the risk of betrayal and employ secret diplomacy when the damage from a leak is small. Furthermore, the untrustworthy adversary faces a stronger incentive to mimic the trustworthy type when the adversary's pay-off from betrayal is small, because the opportunity cost of collusion in the first round has now decreased. In brief, our model suggests that plausible deniability would make it less risky for the initiator to go private, but it also reduces secret diplomacy's screening value. Consequently, an initiator who wishes to "test" an adversary's trustworthiness for secret-keeping, as opposed to a mere temporary collusion with the adversary, is likely to prefer direct secret communication over the use of secret mediators.

Autonomous leaks

In our model, the audiences do not find out about a secret deal if the adversary does not intentionally leak. We make this modeling choice not only because we wish to explicate the strategic logic behind the decision to leak, but also because leaders often take significant actions to minimize the risk of unwanted leaks by restricting knowledge of secret diplomacy to a very small number of people. Indeed, none of the other Politburo members had knowledge of the secret protocol attached to the 1939 Molotov-Ribbentrop Pact (Roberts 1992). Nixon kept his overture to China secret from most people, including William Rogers, his secretary of state and longtime confidant. By minimizing the involvement of unnecessary and potentially untrustworthy personnel, leaders significantly reduce the risk of autonomous leaks. Generally, the risk of autonomous leaks should be less severe when (1) the scope of the secret operation/agreement is relatively small, so fewer people have to be informed; (2) when the national security apparatus is more centralized so information is more tightly controlled within the bureaucracy; and (3) in dictatorships where the threat of severe punishment for leaking information that is not approved by the leader might act as a strong deterrent against potential leakers.

Nonetheless, our model also speaks to situations associated with substantial risk of "autonomous leaks" (Yarhi-Milo 2013) from third parties such as the media or disgruntled officials who are beyond the control of the negotiating leaders. Crucially, the threat of autonomous leaks would incentivize the untrustworthy adversary to fake trustworthiness in the first round, because the adversary may now benefit from a "windfall" from an autonomous leak even if it is not the one doing the leaking. To screen the untrustworthy adversary, the initiator must therefore increase the importance of the first round issue ($\tilde{\gamma}$), which would reduce the incentive for the initiator to employ secret diplomacy as a litmus test in the first place. Thus, we reveal that the risk of autonomous leaks "deters" secret diplomacy because it makes screening more costly.

Revelation and ex post leaks

Our model speaks most directly to secret agreements that remain secret for an appreciable period of time. There are abundant such examples in recent history. The secret protocol attached to the

1939 Molotov-Ribbentrop Pact remained secret until the fall of Nazi Germany, and the Soviet Union denied the existence of such a protocol until 1989. The 27 secret meetings between China and Taiwan from 1988 to 1992 were not revealed until 2001, 8 years after the two polities established quasi-diplomatic relations. Even when the implementation of a secret agreement is public, leaders are still able to keep the agreement itself secret. Indeed, leaders can often persuade domestic and international audiences that a foreign policy decision—included in a secret agreement— was taken purely based on the national interest. This was precisely Kennedy's strategy when he withdrew the Jupiter Missiles from Turkey while denying any secret deal between the United States and the Soviet Union (Criss 1997, 117).

As for agreements that require revealing the secret agreement itself, we expect leaders to pay a political cost for their secret cooperation. That cost, however, is smaller compared to the potential cost from public diplomacy. This is because leaders can often control the pace and timing of the revelation to reduce possible domestic and international opposition to the deal. For instance, leaders can embark on a propaganda campaign to improve the image of the adversary before such revelation (Goh 2004). They can also highlight the concessions the other side has made. Let $\hat{r}_2$ represent the reduced political cost associated with a revealed secret deal. Our results hold if $c_2 - \hat{r}_2 > b_2^T$, an inequality that ensures that the trustworthy adversary will still prefer secret cooperation over leaking even if the secret deal must be revealed at some point in the future.

Finally, what does our model say about situations where an untrustworthy adversary leaks right after striking a secret deal? The *ex post* leak scenario is a special case of our model where γ is exogenously fixed to be identical across two rounds of the game (since the *ex post* leak scenario involves the same issue, just in different time periods), with the adversary colluding in the first round and leaking in the second round after benefiting from the secret deal in the first round. *Ex post* leak, therefore, corresponds to our model's collusion equilibrium.

Retaining the status quo

We can complicate our "simple" model and allow the initiator to retain the status quo—e.g., not to reach out to an opponent at all, leaving all players with payoff 0—in addition to going private or

public. In this richer model, the initiator may penalize an untrustworthy opponent in the second round by denying him any cooperation when there is a leak. However, the initiator will never punish an untrustworthy opponent by refusing cooperation in round two if his/her expected payoff from public assurance is positive, e.g., when the political cost of a public deal is sufficiently low ($c_1 < r_1$) and when the adversary would accept a public deal. Crucially, whenever $c_2^i < r_2$, the adversary will always reciprocate public assurance with cooperation, and the initiator will never have to face the worst scenario when he goes public where he pays the reputational cost of public diplomacy without reaching any understanding with the adversary. In brief, when $c_1 < r_1$ and $c_2^i < r_2$ (end of 3.1 discusses why we assume these two inequalities), the initiator will always prefer public diplomacy over the status quo, and the equilibrium predictions of a model with the status quo option will be identical to a model without.

When $c_1 \leq r_1$ and/ or $c_2^i \leq r_2$, however, allowing the initiator to "not cooperate" and retain the status quo affects our results in two ways. First, the public diplomacy equilibrium disappears because the initiator now prefers "status quo" over public assurance, which gives the initiator zero or negative pay-offs either because a public deal is associated with a large reputational cost (if $c_1 \leq r_1$ and $c_2^i > r_2$) or because public diplomacy is guaranteed to fail (if $c_2^i \leq r_2$). Second, the comparative statics regarding the effect of the political costs of public cooperation on the incentives for the initiator to screen and the untrustworthy adversary to mimic changes, because the status quo is now the default option when the initiator refuses to go private. Nonetheless, the mechanisms discussed earlier that sustain the screening and collusion equillibria remain robust in this scenario.

4 Empirical evidence

We now turn to three historical case studies to shed light on the causal mechanisms underlying the leaders' calculations and decision-making (Lorentzen, Fravel, and Paine 2016). These illustrations also enable us to probe the relevance of our model to explain important cases of secret diplomacy. Leveraging qualitative analysis is appropriate not only because the nature of the topic prevents us from compiling a dataset that allows for large-N statistical analysis, but also because our model contains

core elements of strategic interaction that depend on the beliefs and preferences of leaders, which are relatively easier to establish in a case study (Goemans and Spaniel 2016). Nevertheless, we recognize that, by definition, secret diplomatic talks are intended to remain secret for some length of time; as such, the evidentiary record, even in these historical cases, is incomplete.

With these caveats in mind, we focus on several important observable aspects of our theory's causal mechanisms. First, leaders initiated secret diplomacy because the perceived political and reputational costs for engaging in public diplomacy were too high. Second, leaders understood that pursuing secret diplomacy was risky because the adversary might leak, but they were willing to take this calculated risk. Third, leaders used secret diplomacy either as a way to screen the intentions of the adversary (as in the case of the opening to China and the Cuban Missile Crisis) or temporarily collude with the adversary (as in the case of the Iran-Contra affair). Fourth, while engaging in secret diplomacy, leaders assessed whether the adversary was likely to pretend to cooperate in secret, only to leak this information later. Finally, in the case of the screening equilibrium, we probe the extent to which leaders' beliefs about the trustworthiness of the adversary changed as a result of the decision by the adversary not to reveal the secret. In the case of the collusion equilibrium, we look for evidence suggesting cooperation without updating of beliefs. Given space constraints, we have allocated our second case study illustrating the screening equilibrium (Cuban Missile Crisis) and the collusion equilibrium (Iran-Contra affair) to the Appendix.

The Secret Opening to China

The United States refused to recognize the new People's Republic of China for twenty years after the Chinese Communists took power in 1949. Upon coming to office in 1969, however, U.S. President Richard Nixon immediately began exploring the potential for a rapprochement with the People's Republic. This was an effort both to shore up the U.S. position in Asia, which was under strain as a result of the costs it had incurred in the Vietnam War, and to put pressure on the Soviet Union in the wake of the rupture in relations between the two Communist giants. The efforts stalled somewhat as the United States expanded the Vietnam War into Cambodia in 1970, but in October of that year

the United States reopened its Pakistani and Romanian back channels with China. Communication between the two countries was slow, but in April 1971, Chinese Premier Zhou Enlai invited the United States to send a high-level representative to Beijing. As a result, National Security Adviser Henry Kissinger visited in July—still in secret—setting the stage for Nixon's own public trip in February 1972.

As our theory assumes, reputational and political concerns led Nixon to seek secret diplomacy. Indeed, Nixon chose to approach China in secret because of the prospect of backlash (and potential sabotage) from domestic audiences, allies, and the Soviet Union. As Nixon himself explained to Mao, if his overtures regarding Taiwan, an old Cold War ally, were revealed, he would pay high political costs, noting: "Let me in complete candor tell the Prime Minister what my problem is, from a political standpoint...Our people, from both the right and the left, for different reasons, are watching this particular issue. The left wants this trip to fail, not because of Taiwan but because of the Soviet Union. And the right, for deeply principled ideological reasons, believes that no concessions at all should be made regarding Taiwan." (Memcon, February 22, 1972). As a result, Nixon wished to keep the assurances secret until these audiences could be presented with a fait accompli during his second term in office (Ibid.; Macmillan 2007: 161, 174, 257, 327).

Consistent with our model's screening logic, U.S. secret diplomacy started off with small stakes, both in terms of the U.S officials involved as well as the importance of the issues the United States was willing to raise with China. As a result of the Chinese willingness to keep the channels and assurances secret, U.S. secret diplomacy grew over time both in substance and the level of officials involved. Initially, U.S. overtures were made through intermediaries—most importantly Pakistan, but also Romania. In deciding on the timing for probes indicating U.S. interest in talks, American officials paid great attention to Chinese officials' rhetoric and to their willingness to reciprocate. In August-November 1969, Nixon put out tentative feelers through Pakistan and Romania that the United States was interested in accommodation with China. Then in December, the U.S. ambassador to Poland unofficially indicated to the Chinese Chargé d'Affairs that Nixon "would like to have serious concrete talks with the Chinese" (Memo from Stoessel to Rogers, "Contact with Communist Chinese," December 3, 1969).

As contacts were established through regular meetings in Warsaw in late 1969 and early 1970, U.S. officials began to consider ramping up their willingness to discuss sensitive issues and make assurances. These included issues of trade, travel restrictions from mainland China, and U.S. forces on Taiwan. Progress was somewhat undone by the U.S. intervention in Cambodia, but in late 1970 U.S. officials again began to probe Chinese willingness to resume talks and accept a U.S. envoy to China. In the years and months prior to Kissinger's July 1971 visit to Beijing, Nixon focused on explaining to the Chinese that the United States had no interest in colluding with the Soviets in order to encircle China, and that it would not simply use improved ties with Beijing as a bargaining chip vis-à-vis Moscow.

As the Chinese kept the form and content of the assurances secret, the United States was willing to upgrade the secret talks. In July 1971, National Security Adviser Henry Kissinger paid a secret visit to China, in which he expanded upon previous vague U.S. assurances. For one, he promised that the United States would keep China informed regarding Soviet-American agreements (Yarhi-Milo 2013: 422). Regarding Taiwan, he told the Chinese that: 1) the United States would "remove two-thirds of its armed force" from Taiwan as the Vietnam War drew to a close (Memcon July 9, 1971), with the remaining forces subject to withdrawal as Sino-American relations improved; 2) the United States would not support Taiwanese independence movements; and 3) the United States would oppose Japanese rearmament, as well as any Japanese military involvement in Taiwan. Finally, Kissinger conceded that the United States would move to a "One China" policy, in which Taiwan would be regarded as belonging to the mainland (Ibid). Then, during Nixon's visit, U.S. officials reiterated these assurances on Taiwan, Japan, and the Soviet Union. This time, however, Kissinger went a step further and shared top-secret American intelligence on the military assets of the USSR (Macmillan 2007: 241-242).

Importantly, Nixon and Kissinger interpreted China's decision to not leak as indicative of its desire for long-term secret cooperation. This is because the adverse selection problem Nixon and Kissinger faced was moderate; China was unlikely to have acted trustworthily in the initial rounds of secret diplomacy unless it was serious about continuing secret cooperation, as it could reap substantial rewards by leaking. Initially, as our model assumes, Nixon and Kissinger were uncertain about whether

the Chinese would fake trustworthiness only to leak the assurances at a more favorable time. As Nixon later wrote, while he was awaiting Beijing's reply to his proposal to send Kissinger to China, he understood the risk "for serious international embarrassment if the Chinese decided to reject my proposal and then publicize it" (Nixon 1978: 551). Specifically, Nixon and Kissinger were concerned about the possibility that the Chinese would cooperate initially but then "report what was said to the Soviets" in an effort to damage détente between the two superpowers (quoted in memo to the president's file, July 1, 1971). Yet, Nixon and Kissinger judged that the benefits from leaking, albeit significant, were only moderate relative to the benefits from cooperation. This is because, as Kissinger explained before his visit, the Chinese were worried about the Soviet threat along their border: "it would not help them to humiliate us if they want to use it in some way as a counterweight to the Soviets" (Scope paper, briefing book for Kissinger's July 1971 trip, p.2). Yet both Kissinger and Nixon were uncertain as to what the Chinese would do and thus took a calculated risk to test the Chinese with secret diplomacy. Similarly, just weeks before Nixon's trip in February 1972, Nixon again raised concerns about the possibility of the Chinese revealing "the secret record" of their meetings, but by that time Kissinger was confident enough in Chinese trustworthiness and asserted, "They won't make it come out." (Conversation between Nixon and Kissinger, February 14, 1972).

Consistent with our model, given that the adverse selection problem was only moderate in this case, the United States could have chosen small-stakes issues to initially screen China's type, and moved to more substantial assurances in the second round. Indeed, as noted above, U.S secret overtures began with low-level and indirect gestures to screen China and then turned into higher-level contacts and more explicit promises (Komine 2008: 152). By refusing to humiliate the United States and ultimately receiving a high-level envoy, the Chinese leaders demonstrated that they could resist their own hard-liners who opposed a rapprochement with the United States (Nixon 1978: 556). Kissinger, in particular, had been skeptical of Nixon's China initiatives, but once the Chinese had indicated their willingness to accept a high-level U.S. diplomatic visit, his faith in the potential for rapprochement increased enormously. He considered this a costly signal of Chinese intent, because if the visit did not bear fruit, it would diminish Chinese bargaining power vis-à-vis the Soviets by suggesting that China could not turn to the United States.

Following his trip, Kissinger wrote to Nixon, "We are building a solid record of keeping the Chinese informed on all significant subjects of concern to them, which gives them an additional stake in nurturing our new relationship." (Memo from Kissinger to Nixon, August 16, 1971). Kissinger also noted that "the Chinese were extremely suspicious of our desire for secrecy" (Kissinger 1979: 724), but that "[i]n time the Chinese came to understand our reasons; I have no doubt now that the secrecy of the first trip turned into a guarantee of a solid and well-managed improvement of relations" (Kissinger 1979: 725).

The secret talks between Kissinger and Chinese officials culminated in a historic public summit in which the president uttered unprecedented assurances regarding future relations with Taiwan. Even in this public meeting, the assurances Nixon conveyed to the Chinese were secret in nature, and the screening game continued with the stakes now higher than in previous rounds. In their meeting, Nixon stated that the way China had handled the secret negotiations allowed him to trust that the Chinese leadership would continue to keep those assurances secret, noting "I have never seen a government more meticulous in keeping confidences and more meticulous in keeping agreements than his (the Prime Minister's) government." (Memo of Conversation, February 22, 1972).

Evidence also suggests that the Chinese understood the screening logic of the game and maintained secrecy, in part, to build a reputation for trustworthiness in the eyes of Nixon and Kissinger. Before Nixon's visit, the Chinese acknowledged that maintaining secrecy might be difficult, but understood it was an important precondition to developing trust, noting, "If secrecy is still desired the Government of the People's Republic of China will on its part guarantee the strict maintenance of secrecy." (Letter from Zhou to Nixon, May 29, 1971). The United States, indeed, indicating to the Chinese that their willingness to maintain secrecy was a significant signal, stated, "President Nixon appreciates the fact that the Government of the People's Republic of China is prepared to maintain strict secrecy with respect to Dr. Kissinger's visit and considers this essential." (Message from the Government of the USA to the Government of China, June 4, 1971). Thus, by the time of Nixon's visit, Prime Minister Zhou reassured Nixon, "regarding to some things we have discussed secretly and in our secret meetings, that is not only regarding the questions of the Soviet Union, Japan and India but also things we have decided to do but not to say, we believe that we will maintain that secrecy and

that what happened after the two visits Dr. Kissinger paid to China can serve as proof to that. And we believe it can continue in that way." (Memo of conversation, February 28, 1972). Similarly, the Chinese Vice Chairman of the Military Commission told Kissinger that "the ability of our two sides to maintain secrecy has already been tested" with Kissinger affirming that "we have taken big steps toward establishing confidence." (Memcon, February 23, 1972).

Certainly the Chinese were aware that they had the option of leaking these assurances and imposing domestic costs on Nixon. In Zhou's words to Nixon, China was "not rushing to make use of the opponents of your present visit and attempt to solve all the questions and place you in an embarrassing position." But precisely because the Chinese decided to keep the assurances secret despite the payoffs they could have received from humiliating Nixon, the Chinese signaled to Nixon and Kissinger that they could be trusted. They hoped that by doing so, they could enjoy further cooperation with the United States on a variety of sensitive issues during Nixon's second term.

5 Conclusion

How, when, and why does secret diplomacy promote cooperation between foes? With a novel model, we argue that secret diplomacy is valuable because it could serve as a "litmus test" that allows a trustworthy opponent to acquire a reputation for secrecy-trustworthiness. Crucially, whether leaders employ secret diplomacy to test an adversary depends on the untrustworthy adversary's incentive to fake trustworthiness by not leaking initially, e.g. the adverse selection problem's severity. When the adverse selection problem is modest, leaders are more likely to employ secret diplomacy as a screen, which gives an adversary the opportunity to demonstrate his trustworthiness. In contrast, when the adverse selection problem is severe, leaders will often find it too difficult to realize secret diplomacy's screening value. In this scenario, secret diplomacy only leads to short-term collusion, and even a trustworthy adversary would find it difficult to acquire a reputation for trustworthiness. Importantly, we find that the adverse selection problem is more severe when the adversary anticipates high costs from public cooperation and less severe when the adversary expects large gain from betraying the initiator by leaking. Our model goes a long way in illuminating the logic behind secret diplomacy in

a number of high-profile historical cases.

Secrecy-trustworthiness is crucial for continued cooperation between adversaries behind closed doors. The extensive secret cooperation between Israel and Jordan in the pre-Oslo period, or, more recently, between Israel and Saudi Arabia in the absence of any formal relations between the two countries is a case in point. Nevertheless, secrecy-trustworthiness could induce or evolve into greater generalized trust over time. Adversaries who were able gain the trust of their counterparts by keeping secrets and refraining from humiliating the opponent, might also be subsequently trusted with public cooperation. Modeling those dynamics are beyond the scope of this paper. Nonetheless future studies could build on our insights to examine how secret diplomacy can engender trust for later public cooperation, or even institution-building, as well as the conditions under which a reputation for secrecy-trustworthiness could lead to a more generalized trust between adversaries.

Our theory and findings carry significant implications for our understanding of uncertainty and information transmission in international politics. While most studies have focused on leaders' incentives and capabilities to communicate their own intentions, we explicate how leaders devise strategies to learn their adversaries' intentions and the conditions under which those strategies will work. Indeed, our model is not only the first formal analysis of secret assurance, but also of the adverse selection problem in the realm of international security.[17] Importantly, screening and adverse selection are larger phenomenons that have relevance beyond secret diplomacy. Among U.S. foreign policy makers, one of the most powerful arguments for engagement instead of containment is the possibility for engagement to test whether an adversary (e.g., Iran) is willing to "play ball." When an adversary exploits the United States' goodwill, American leaders may garner domestic and international support for sanctions, if not for military deployment, against that adversary (Haass and O'Sullivan 2000: 4). Alternatively, when an adversary responds well to limited cooperation, American leaders may step up cooperation with that adversary, which could lay the foundation for transforming an adversarial relations into a burgeoning alliance. However, leaders may sometimes find it infeasible to test an adversary. In the case of secret diplomacy, we find that too much incentive for an adversary to go private—i.e., when the leader faces substantial hostility to public cooperation—nullifies the screening

[17] For a model in IPE that analyzes the adverse selection problem see, e.g., Bas and Stone 2013.

value of secret diplomacy by prompting all types of adversaries to behave the same way.

References

Abrahamian, Ervand. 1999. *Tortured confessions: Prisons and public recantations in modern Iran*. Univ of California Press.

Bas, Muhammet A and Randall W Stone. 2014. "Adverse selection and growth under IMF programs." *The Review of International Organizations* 9(1):1–28.

Byrne, Malcolm. 2014. *Iran-Contra: Reagan's Scandal and the Unchecked Abuse of Presidential Power*. University Press of Kansas.

Carson, Austin. 2016. "Facing off and saving face: covert intervention and escalation management in the Korean War." *International Organization* 70(01):103–131.

Carson, Austin and Keren Yarhi-Milo. 2017. "Covert Communication: The Intelligibility and Credibility of Signaling in Secret." *Security Studies* 26(1):124–156.

Crescenzi, Mark JC. 2007. "Reputation and interstate conflict." *American Journal of Political Science* 51(2):382–396.

Criss, Nur Bilge. 1997. "Strategic nuclear missiles in Turkey: The jupiter affair, 1959–1963." *The Journal of Strategic Studies* 20(3):97–122.

Dafoe, Allan and Devin Caughey. 2016. "Honor and War." *World politics* 68(02):341–381.

Dafoe, Allan, Jonathan Renshon and Paul Huth. 2014. "Reputation and status as motives for war." *Annual Review of Political Science* 17:371–393.

Gelpi, Christopher and Joseph Grieco. 2015. "Competency Costs in Foreign Affairs." *American Journal of Political Science* 59(2):440–456.

Gibbons, Robert. 1992. *Game theory for applied economists*. Princeton University Press.

Goemans, Hein and William Spaniel. 2016. "Multimethod Research: A Case for Formal Theory." *Security Studies* 25(1):25–33.

Goh, Evelyn. 2004. *Constructing the US Rapprochement with China, 1961–1974*. Cambridge University Press.

Grosek, Edward. 2007. *The secret treaties of history*. William S. Hein & Company.

Haas, Mark. 2005. *The ideological origins of great power politics, 1789-1989*. Cornell University Press.

Haass, Richard N and Meghan L O'Sullivan. 2000. "Terms of engagement: Alternatives to." *Survival* 42(2):113–35.

Hamilton, Lee and Daniel Inouye. 1987. *Report of the congressional committees investigating the Iran-Contra Affair*. Vol. 100 US House of Representatives Select Committee to Investigate Covert Arms Transactions with Iran.

Jervis, Robert. 1990. "Models and cases in the study of international conflict." *Journal of International Affairs* pp. 81–101.

Kissinger, Henry. 1979. "White House Years." *Boston: Little Brown and Co* .

Komine, Yukinori. 2008. *Secrecy in US foreign policy: Nixon, Kissinger and the rapprochement with China*. Aldershot, England: Ashgate.

Kornbluh, Peter and Malcolm Byrne. 1993. *The Iran-Contra scandal: the declassified history*. New Press.

Kurizaki, Shuhei. 2007. "Efficient secrecy: Public versus private threats in crisis diplomacy." *American Political Science Review* 101(03):543–558.

Kydd, Andrew. 2000. "Trust, reassurance, and cooperation." *International Organization* 54(02):325–357.

Kydd, Andrew. 2005. *Trust and mistrust in international relations*. Princeton University Press.

Kydd, Andrew and Barbara Walter. 2002. "Sabotaging the peace: The politics of extremist violence." *International Organization* 56(02):263–296.

Lorentzen, Peter, Taylor Fravel and Jack Paine. 2016. "Qualitative investigation of theoretical models: the value of process tracing." *Journal of Theoretical Politics* pp. 1–25.

MacMillan, Margaret. 2008. *Nixon and Mao: the week that changed the world*. Random House.

McFarlane, Robert and Zofia Smardz. 1994. *Special trust*. Cadell & Davies.

Morrow, James. 1994. *Game theory for political scientists*. Princeton University Press Princeton, NJ.

Nixon, Richard M. 1978. "The Memoirs of Richard Nixon." *(New York: Grosset and Dunlap, 1978)* .

Powell, Jonathan. 2015. *Terrorists at the table: Why negotiating is the only way to peace*. Macmillan.

Powell, Robert. 2004. "Bargaining and learning while fighting." *American Journal of Political Science* 48(2):344–361.

Putnam, Robert D. 1988. "Diplomacy and domestic politics: the logic of two-level games." *International organization* 42(03):427–460.

Rathbun, Brian. 2011. "Before hegemony: generalized trust and the creation and design of international security organizations." *International Organization* 65(02):243–273.

Riley, John. 2001. "Silver signals: Twenty-five years of screening and signaling." *Journal of Economic literature* 39(2):432–478.

Sartori, Anne. 2013. *Deterrence by diplomacy*. Princeton University Press.

Schultz, Kenneth. 2005. "The Politics of Risking Peace: Do Hawks or Doves Deliver the Olive Branch?" *International Organization* 59(01):1–38.

Smith, Alastair. 1998. "International crises and domestic politics." *American Political Science Review* 92(03):623–638.

Stasavage, David. 2004. "Open-door or closed-door? Transparency in domestic and international bargaining." *International Organization* 58(04):667–703.

Tarar, Ahmer and Bahar Leventoğlu. 2009. "Public commitment in crisis bargaining." *International Studies Quarterly* 53(3):817–839.

Trager, Robert. 2010. "Diplomatic calculus in anarchy." *American Political Science Review* 104(02):347–368.

Trager, Robert. 2011. "Multidimensional Diplomacy." *International Organization* 65(03):469–506.

Watson, Joel. 1999. "Starting small and renegotiation." *Journal of economic Theory* 85(1):52–90.

Weisiger, Alex and Keren Yarhi-Milo. 2015. "Revisiting reputation." *International Organization* 69(02):473–495.

Yarhi-Milo, Keren. 2013. "Tying hands behind closed doors: the logic and practice of secret reassurance." *Security Studies* 22(3):405–435.

Appendix

Proofs of Main Results

Proof for Proposition 1 (Screening Equilibrium)

We solve for this equilibrium by backward induction. If P_1 goes private in round 2, we know that P_2^T will always collude. P_2^U, in contrast, would always leak in round 2 to reap the payoff $b_2^{U'}$ because there is no political cost for doing so. On the other hand, if P_1 decides to go public in round 2, both P_2^T and P_2^U will settle for public cooperation. P_1 will opt for secret diplomacy in the second round if P_2 is trustworthy, and public diplomacy if P_2 is untrustworthy. Since P_2^U does not mimic P_2^T in this equilibrium, $\alpha'(collude, \gamma^*) = 1$ and $\alpha'(leak, \gamma^*) = 0$. In other words, in this equilibrium, P_1 learns about P_2's type *definitively* from the history of the first round sub-game.

In the first round, if P_1 goes private, P_2^T will always cooperate in private while P_2^U will cooperate in private when $\gamma^* > \hat{\gamma}$ and betray if $\gamma^* \leq \hat{\gamma}$. $\hat{\gamma}$ is the lowest possible γ that would make the first round important enough to discourage the untrustworthy opponent from imitating the trustworthy one. To derive $\hat{\gamma}$, note that P_2^U will mimic in the first round if and only if :

$$\gamma c_2 + (1-\gamma) b_2^{U'} \geq \gamma b_2^U + (1-\gamma)(c_2' - r_2') \tag{1}$$

(1) simply states that P_2^T must value the opportunity to betray in the second round enough (given P_1's choice of γ) to give up on the opportunity to betray the initiator in the first round. Rearranging the terms, we show:

$$\hat{\gamma} \equiv \frac{c_2 - b_2^U - r_2}{2(c_2 - b_2^U) - r_2} \tag{2}$$

Given $\hat{\gamma}$ (the incentive for P_2^U to act trustworthy in round 1), P_1 will go private in the first round if it has sufficient trust in P_2 , e.g. when $\tilde{\alpha}_{screening} < \alpha$. To solve for the threshold value of trust $\tilde{\alpha}_{screening}$ – the minimum level of trust that P_1 must have in P_2 for P_1 to attempt screening with private diplomacy

instead of going public – note that P_1 gains more from screening in this situation if and only if:

$$c_1 - r_1 \leq \hat{\gamma}[\alpha c_1 - (1-\alpha)d_1] + (1-\hat{\gamma})[\alpha'(s_2^*, \hat{\gamma})c_1' + (1-\alpha'(s_2^*, \hat{\gamma}))(c_1' - r_1')] \tag{3}$$

The right hand side (RHS) of (3) is the payoff associated with private cooperation for P_1 when secret diplomacy screens out P_2^U in the first round. The first round payoff for P_1 (the first component of RHS) is the expected payoff from private cooperation given the risk of betrayal, weighted by $\hat{\gamma}$, which is the optimal level of γ that P_1 would choose if he hopes to screen. The second round payoff for P_1 (the second component of RHS) is the *ex ante* expected payoff facing the initiator when she plays a complete information game of secret diplomacy, weighted by $1-\hat{\gamma}$. Note that $\alpha'(s_2^*, \hat{\gamma})$ *simplifies* to α for the RHS of the inequality, because P_1 expects in the "pre-diplomacy" stage where she sets γ that the opponent will turn out to be untrustworthy with probability α. Given the discussion above, we can rewrite (2) as (3):

$$c_1 - r_1 \leq \hat{\gamma}[\alpha c_1 - (1-\alpha)d_1] + (1-\hat{\gamma})[\alpha c_1' + (1-\alpha)(c_1' - r_1')] \tag{4}$$

Rearranging the terms, P_1 goes private in round 1 instead of relying on secret diplomacy to screen if and only if:

$$\alpha \geq \tilde{\alpha}_{screening} \equiv 1 - \frac{r_1}{\hat{\gamma}(d_1 - r_1) + r_1} \tag{5}$$

To finish characterizing the screening equilibrium, we solve for P_1's equilibrium choice of γ^*. Note that P_1 would screen if and only if the expected payoff from screening is larger than the expected payoff from collusion:

$$\gamma c_1 + (1-\gamma)(\alpha c_1' - (1-\alpha)d_1') \leq \gamma[\alpha c_1 - (1-\alpha)d_1] + (1-\gamma)[\alpha c_1' + (1-\alpha)(c_1' - r_1')] \tag{6}$$

The right hand side (LHS) of (6) is the payoff associated with private cooperation for P_1 when secret diplomacy facilitates temporary collusion but has no informational value. The first round payoff

for P_1 (the first component of LHS) is simply the payoff of collusion (since both P_2^T and P_2^U would collude now), weighted by γ. The second round payoff for P_1 (the second component of LHS) is the expected payoff he faces for a one round secret diplomacy game, weighted by $1-\gamma$. Recall that in the second round, P_2^U no longer has any incentive to fake trustworthiness and not leak, as the game now ends. Also note that the second round belief $\alpha'(s_2^*, \gamma)$ is simply α because no learning happens in round 1 under the collusion equilibrium.

Rearranging the terms, the maximum stake $\tilde{\gamma}_{screening}$ that P_1 is willing to put on round 1 of the game to induce P_2^U to leak early on is:

$$\frac{c_1 - r_1 + d_1}{2c_1 - r_1 + 2d_1} \tag{7}$$

For P_1 to opt for screening, $\hat{\gamma} \leq \tilde{\gamma}_{screening}$ must be true, which means that P_1 must have enough incentive to overcome the adverse selection problem by approaching the adversary with a sufficient important issue in the initial round of the secret diplomacy game; we can interpret $\tilde{\gamma}_{screening}$ as a parameter that captures the incentive for P_1 to screen. When $\hat{\gamma} \leq \tilde{\gamma}_{screening}$, P_1 will set γ as $\hat{\gamma}$. Any γ lower than $\hat{\gamma}$ will fail to induce P_2^U to reveal its true color. On the other hand, it is unnecessary for P_1 to set γ higher than $\hat{\gamma}$, which does not give P_1 any additional benefit while it will increase the cost that P_1 have to pay if the adversary turns out to be untrustworthy and leak.

Proof for Proposition 2 (Collusion Equilibrium)

Again, we start characterizing this equilibrium by looking at the three actors' second round choices. First, because P_2 faces no political cost of betrayal in round 2, P_2^U would always leak in the second round to reap the payoff $b_2^{U'}$ if P_1 goes private. P_2^T , in contrast, will still cooperate in round 2 if P_1 goes private or public. On the other hand, if P_1 decides to go public in round 2, both P_2^T and P_2^U will settle for public cooperation.

P_1 will opt for secret diplomacy in round 2 if and only if he has sufficient trust in P_2 after round 1. In other words, $\alpha'(s_2^*, \gamma^*) \geq \tilde{\alpha'}_{secret}$ must hold, with $\tilde{\alpha'}_{secret}$ as the threshold level of trust that would make p_1 indifferent between opting for public versus secret diplomacy in round 2. To solve for

$\tilde{\alpha'}_{secret}$, note that P_1 will opt for public diplomacy in round 2 if and only if:

$$c'_1 - r'_1 \leq \alpha'(s_2^*, \gamma^*)c'_1 - (1 - \alpha'(s_2^*, \gamma^*))d'_1 \tag{8}$$

The LHS of the inequality above is P_1's payoff for going public in round 2. The RHS is P_1's payoff for going private in round 2. The s_2^* denotes P_2's equilibrium strategy in round 1. Rearranging the terms, we show that P_1 would approach P_2 publicly in round 2 if and only if:

$$\alpha'(s_2^*, \gamma^*) \leq \tilde{\alpha}'_{secret} \equiv 1 - \frac{r'_1}{c'_1 + d'_1} \tag{9}$$

Crucially, $\tilde{\alpha'}_{secret}$ represents the level of trust necessary to sustain secret diplomacy when secret diplomacy serves neither screening nor collusion values. It is indeed the constraint that would guarantee that P_1 would go private in a one shot secret diplomacy game, which equivalent to the last (and second) round of the game.

In the first round, if P_1 goes private, P_2^T will always collude while P_2^U will cooperate in private when $\gamma^* > \hat{\gamma}$ and betray if $\gamma^* \leq \hat{\gamma}$. Whether P_1 would go public or collude depends on its trust in P_2. To solve for the threshold value of trust $\tilde{\alpha}_{collusion}$ necessary to sustain collusion, note that P_1 prefers to go public instead of colluding with P_2 if and only if:

$$c_1 - r_1 \geq (\hat{\gamma} - \varepsilon)c_1 + (1 - \hat{\gamma} + \varepsilon)(\alpha'(s_2^*, \hat{\gamma} - \varepsilon)c'_1 + (1 - \alpha'(s_2^*, \hat{\gamma} - \varepsilon))d'_1) \tag{10}$$

The right hand side (RHS) of (6) is the payoff associated with private cooperation for P_1 when secret diplomacy facilitates collusion. The first round payoff for P_1 (the first component of RHS) is simply the payoff of collusion (since both P_2^T and P_2^U would collude now), weighted by $\hat{\gamma} - \varepsilon$, which is the optimal level of γ that P_1 would choose if he hopes to collude. Any γ^* higher than $\hat{\gamma} - \varepsilon$ will induce P_2^U to leak, while γ^* lower than $\hat{\gamma} - \varepsilon$ will reduce the benefit of collusion for P_1. The second round payoff for P_1 (the second component of RHS) is the expected payoff it faces for a one round secret diplomacy game, weighted by $1 - \hat{\gamma} + \varepsilon$. Note that in the second round, P_2^U no longer has any incentive to fake trustworthiness and not leak, as the game now ends. Also note that $\alpha'(s_2^*, \hat{\gamma} - \varepsilon)$

simplifies to α because no learning happened in round 1.

Given the discussion above, we can rewrite (10) as (11):

$$c_1 - r_1 \geq (\hat{\gamma} - \varepsilon)c_1 + (1 - \hat{\gamma} + \varepsilon)(\alpha c_1' + (1 - \alpha)d_1') \tag{11}$$

Rearranging the terms, the level of trust that would allow P_1 to opt for collusion instead of going public in round 1 is:

$$\alpha \geq \tilde{\alpha}_{collusion} \equiv 1 - \frac{r_1}{(c_1 + d_1)(1 - \hat{\gamma} + \varepsilon)} \approx 1 - \frac{r_1}{(c_1 + d_1)(1 - \hat{\gamma})} \tag{12}$$

Note that for P_1 to opt for collusion, $\alpha \geq \tilde{\alpha}_{collusion}$ is a *necessary but not sufficient* condition. It is not sufficient because P_1 must also be trustful enough of P_2 to go private again in the second round (in other words, $\alpha \geq \tilde{\alpha'}_{secret}$). Importantly, the level of trust necessary for P_1 to go private instead of public in the first round $\tilde{\alpha}_{collusion}$ is strictly *smaller* compared to the level of trust necessary for P_1 to opt for temporary collusion instead of public diplomacy. This is unsurprising because in the first round, P_1 recognizes the value of secret diplomacy as a tool that has the potential to both facilitate cooperation without reputation cost of public diplomacy in round 1 and to induce P_2^U to collude in round 1. In contrast, in round 2, secret diplomacy is solely a useful tool to facilitate cooperation without political consequences. Because $\tilde{\alpha}_{secret} > \tilde{\alpha}_{collusion}$, $\alpha > \tilde{\alpha}_{secret}$ guarantees that $\alpha > \tilde{\alpha}_{collusion}$ must hold. In other words, we may focus on $\alpha \geq \tilde{\alpha}_{secret}$ as the *necessary and sufficient* constraint on P_1's level of trust that would sustain the collusion equilibrium.

Furthermore, note that $\tilde{\alpha}_{secret} > \tilde{\alpha}_{screening}$, which implies that the level of trust necessary to sustain the collusion equilibrium can also sustain the screening equilibrium. The key to distinguish the collusion and the screening equilibria therefore lies in identifying P_1's level incentive to screen $\tilde{\gamma}_{screening}$ relative to the severity of the adverse selection problem that P_1 faces $\hat{\gamma}$. When $\tilde{\gamma}_{screening} < \hat{\gamma}$ and $\alpha > \tilde{\alpha}_{secret}$, it would be too costly for P_1 to screen although it has sufficient trust in P_2 to bear the risk of betrayal necessary for screening. Consequently, P_1 would seek to induce P_2^U to collude in round 1 instead. In sum, both conditions $\tilde{\gamma}_{screening} \leq \hat{\gamma}$ and $\alpha \geq \tilde{\alpha}_{secret}$ are necessary to guarantee the existence of a collusion equilibrium.

Proof for Proposition 3 (Public Diplomacy Equilibrium)

To solve for the public diplomacy equilibrium, we specify the conditions under which P_1 would prefer to go public instead of going private to collude or screen. We start by examining round 2 of the game when P_1 goes private. We know that P_2^T will always collude, while P_2^U would always leak to reap the payoff $b_2^{U'}$ because there is no political cost for doing so. On the other hand, if P_1 decides to go public in round 2, both P_2^T and P_2^U will settle for public cooperation since $c_2 > r_2$. Given the two adversaries' optimal strategies, P_1 will opt for public diplomacy in round 2 if and only if:

$$c_1' - a_1' \geq \alpha'(s_2^*, \gamma^*)c_1' - (1 - \alpha'(s_2^*, \gamma^*))d_1' \tag{13}$$

Rearranging the terms, we show that P_1 would approach P_2 publicly in round 2 if and only if:

$$\alpha'(s_2^*, \gamma^*) \leq \tilde{\alpha}_{secret} \equiv 1 - \frac{r_1'}{c_1' + d_1'} \tag{14}$$

This condition guarantees that P_1 will prefer either going public or screening instead of collusion. This is because, as discussed earlier, $\alpha'(s_2^*, \gamma^*) \equiv \alpha \leq \tilde{\alpha}_{secret}$ must be true for P_2 to collude.

In round 1 of the secret diplomacy sub-game, we know that P_2^T will always cooperate in private, while P_2^U will leak if P_1 sets $\gamma \geq \hat{\gamma}$ and collude if P_1 sets $\gamma < \hat{\gamma}$. P_1 will go public instead of screening if the two-round expected payoff from screening is higher than the payoff from public cooperation, which implies P_1 would go public whenever $\alpha < \tilde{\alpha}_{screening}$ (see proof for proposition 1). Importantly, note that γ^* is 1, by construction, if P_1 opts for public diplomacy. This is condition (a) of proposition 3.

Note that $\alpha < \tilde{\alpha}_{screening}$ is a *sufficient but not necessary* condition for P_1 to go public. When $\tilde{\alpha}_{screening} < \alpha < \tilde{\alpha}_{secret}$ (bottom-middle section of Figure 2), P_1's level of trust can sustain screening, but P_1 may nonetheless go public if the adverse selection problem is too severe relative to P_1's incentive to screen ($\tilde{\gamma}_{screening} < \hat{\gamma}$). This corresponds to condition (b) of proposition 3. In sum, P_1 would go public either because he has very little trust in P_2, or because adverse selection problem makes screening undesirable when there is insufficient trust to sustain the collusion equilibrium.

Analysis of a model where public diplomacy fails

In the main model, we present a model where $c_1 > r$ and $c_2^i > r_2$ to ensure that secret diplomacy does not happen simply because public diplomacy is never desirable for either the initiator or the adversary. The second equality is particularly important, as it ensures that the adversary will receive a positive pay-off from public diplomacy, which guarantees public cooperation in equilibrium. What if we relax the second inequality and allow P_2 to reject P_1's public offer? Below we analyze a model where $c_2^i \leq r_2$. Note that for this "new" model, the results are identical whether $c_1 \leq r_1$ or $c_1 > r_1$ holds, as public diplomacy will always fail when $c_2^i \leq r_2$.

The results from our main model (propositions 1 - 3) remain unchanged for this version of the model. As in the case of our main model, there are three classes of pure strategy PBE (screening, collusion, and public diplomacy) associated with this model. The expressions for the cutting-off points, however, will be slightly different for this model.

SCREENING EQUILIBRIUM. As in the case of the main model, we solve for this equilibrium by backward induction. If P_1 goes private in round 2, we know that P_2^T will always collude. P_2^U, in contrast, would always leak in round 2 to reap the payoff $b_2^{U'}$ because there is no political cost for doing so. On the other hand, if P_1 decides to go public in round 2, both P_2^T and P_2^U will reject the public offer. P_1 will opt for secret diplomacy in the second round if P_2 did not leak in the first round, and public diplomacy if P_2 leaked in the first round. Since P_2^U does not mimic P_2^T in this equilibrium, $\alpha'(collude, \gamma^*) = 1$ and $\alpha'(leak, \gamma^*) = 0$.

In the first round, if P_1 goes private, we know that P_2^T will always cooperate in private while P_2^U will cooperate in private when $\gamma^* > \hat{\gamma}$ and betray if $\gamma^* \leq \hat{\gamma}$. To derive $\hat{\gamma}$, note that P_2^U will mimic in the first round if and only if :

$$\gamma c_2 + (1-\gamma) b_2^{U'} \geq \gamma b_2^U + (1-\gamma) \cdot 0 \tag{15}$$

The second period pay-off facing P_2^U when he betrays the initiator in the first round is simply 0, because he would reject the public offer. Rearranging the terms:

$$\hat{\gamma} \equiv \frac{b_2^U}{2b_2^U - c_2} \tag{16}$$

Given $\hat{\gamma}$ (the incentive for P_2^U to act trustworthy in round 1), P_1 will go private in the first round if he has sufficient trust in P_2 , e.g. when $\tilde{\alpha}_{screening} < \alpha$. To solve for the threshold value of trust $\tilde{\alpha}_{screening}$ – the minimum level of trust that P_1 must have in P_2 for P_1 to attempt screening with private diplomacy instead of going public – note that P_1 gains more from screening in this situation if and only if:

$$0 \leq \hat{\gamma}[\alpha c_1 - (1-\alpha)d_1] + (1-\hat{\gamma})[\alpha'(s_2^*, \hat{\gamma})c_1' + (1-\alpha'(s_2^*, \hat{\gamma})) \cdot 0] \tag{17}$$

This expression simplifies to:

$$0 \leq \hat{\gamma}[\alpha c_1 - (1-\alpha)d_1] + (1-\hat{\gamma})(\alpha c_1') \tag{18}$$

Rearranging the terms, P_1 would go private in round 1 instead of relying on secret diplomacy to screen if and only if:

$$\alpha \geq \tilde{\alpha}_{screening} \equiv 1 - \frac{c_1}{\gamma d_1 + c_1} \tag{19}$$

To finish characterizing the screening equilibrium, we solve for P_1's equilibrium choice of γ^*. Note that P_1 would screen if and only if the expected payoff from screening is larger than the expected payoff from collusion:

$$\gamma c_1 + (1-\gamma)(\alpha c_1' - (1-\alpha)d_1') \leq \gamma[\alpha c_1 - (1-\alpha)d_1] + (1-\gamma)[\alpha c_1' + (1-\alpha) \cdot 0] \tag{20}$$

Rearranging the terms, the maximum stake $\tilde{\gamma}_{screening}$ that P_1 is willing to put on round 1 of the game to induce P_2^U to leak early on is:

$$\frac{d_1(1-\alpha)}{c_1(1+\alpha)} \tag{21}$$

For P_1 to opt for screening, $\hat{\gamma} \leq \tilde{\gamma}_{screening}$ must hold.

Compared to the main model, the political cost of public cooperation r_1 disappears from the expressions defining $\hat{\gamma}$ and $\tilde{\gamma}_{screening}$. Because public cooperation is not a possible equilibrium outcome for this model, neither the initiator nor the adversary factors the political cost of public cooperation into their calculus of going public versus private or of betraying versus colluding.

COLLUSION EQUILIBRIUM. Again, we start characterizing this equilibrium by looking at the three actors' second round choices. First, because P_2 faces no political cost of betrayal in round 2, P_2^U would always leak in the second round to reap the payoff $b_2^{U'}$ if P_1 goes private. P_2^T , in contrast, will still cooperate in round 2 if P_1 goes private or public. On the other hand, if P_1 decides to go public in round 2, both P_2^T and P_2^U will settle for public cooperation.

P_1 will opt for secret diplomacy in round 2 if and only if he has sufficient trust in P_2 after round 1. In other words, $\alpha'(s_2^*, \gamma^*) \geq \tilde{\alpha'}_{secret}$ must hold, with $\tilde{\alpha'}_{secret}$ as the threshold level of trust that would make p_1 indifferent between opting for public versus secret diplomacy in round 2. To solve for $\tilde{\alpha'}_{secret}$, note that P_1 will opt for public diplomacy in round 2 if and only if:

$$0 \leq \alpha'(s_2^*, \gamma^*)c_1' - (1 - \alpha'(s_2^*, \gamma^*))d_1' \tag{22}$$

Rearranging the terms, we show that P_1 would approach P_2 publicly in round 2 if and only if:

$$\alpha'(s_2^*, \gamma^*) \leq \tilde{\alpha}_{secret}' \equiv \frac{d_1'}{c_1' + d_1'} \tag{23}$$

Crucially, $\tilde{\alpha'}_{secret}$ represents the level of trust necessary to sustain secret diplomacy when secret diplomacy serves neither screening nor collusion values (because the game ends in round 2). It is indeed the constraint that would guarantee that P_1 would go private in a one shot secret diplomacy game, which equivalent to the last (and second) round of the game.

In the first round, if P_1 goes private, P_2^T will always collude while P_2^U will cooperate in private when $\gamma^* > \hat{\gamma}$ and betray if $\gamma^* \leq \hat{\gamma}$. Whether P_1 would go public or collude depends on its trust in P_2. To solve for the threshold value of trust $\tilde{\alpha}_{collusion}$ necessary to sustain collusion, note that P_1 prefers

to go public instead of colluding with P_2 if and only if:

$$0 \geq (\hat{\gamma} - \varepsilon)c_1 + (1 - \hat{\gamma} + \varepsilon)(\alpha c_1' + (1 - \alpha)d_1') \tag{24}$$

Rearranging the terms, the level of trust that would allow P_1 to opt for collusion instead of going public in round 1 is:

$$\alpha \geq \tilde{\alpha}_{collusion} \equiv \frac{-d_1 + (\gamma + \varepsilon)[(d_1 - c_1)}{c_1 - d_1 - (\gamma + \varepsilon)(c_1 - d_1)} \tag{25}$$

Note that $\tilde{\alpha}_{collusion}$ is always strictly negative for this version of the model. In other words, $\tilde{\alpha}_{collusion}$ does not pose a constraint on the initiator. Therefore the only constraint that the initiator faces is $\tilde{\alpha}'_{secret} \equiv \frac{d_1'}{c_1' + d_1'}$. When $\alpha \geq \tilde{\alpha'}_{secret}$, the initiator goes private. When When $\alpha < \tilde{\alpha'}_{secret}$, the initiator goes public.

Furthermore, again note that $\tilde{\alpha'}_{secret} > \tilde{\alpha}_{screening}$ ($\frac{d_1'}{c_1' + d_1'} > 1 - \frac{c_1}{\gamma d_1 + c_1}$), which implies that the level of trust necessary to sustain the collusion equilibrium may also sustain the screening equilibrium. The key to distinguish the collusion and the screening equilibria therefore lies in identifying P_1's level incentive to screen $\tilde{\gamma}_{screening}$ relative to the severity of the adverse selection problem that P_1 faces $\hat{\gamma}$. When $\tilde{\gamma}_{screening} \leq \hat{\gamma}$ and $\alpha > \tilde{\alpha}_{secret}$, it would be too costly for P_1 to screen although it has sufficient trust in P_2 to bear the risk of betrayal necessary for screening. Consequently, P_1 would seek to induce P_2^U to collude in round 1 instead. In sum, both conditions $\tilde{\gamma}_{screening} \leq \hat{\gamma}$ and $\alpha > \tilde{\alpha'}_{secret}$ are necessary to guarantee the existence of a collusion equilibrium.

PUBLIC DIPLOMACY EQUILIBRIUM. Derivation of the public diplomacy equilibrium is identical to the derivation for proposition 3 of the main model. However, the values for $\tilde{\alpha'}_{secret}$, $\tilde{\alpha}_{screening}$, $\hat{\gamma}$, and $\tilde{\gamma}_{screening}$ are different, as we have shown in our discussion of the screening and the collusion equilibria earlier in this section.

Cuban Missile Crisis

The Cuban Missile Crisis of October 1962 is infamous for being perhaps the closest the world has ever come to nuclear war. In an effort to force the withdrawal of Soviet missiles from Cuba, U.S. President John F. Kennedy ordered a naval blockade of all incoming Soviet ships to Cuba, thus raising the risks of accidental confrontation. Ultimately, while it was largely the combination of the blockade and Soviet Premier Nikita Khrushchev's fear of a U.S. invasion of Cuba that created conditions favorable to a Soviet withdrawal from Cuba, there was an additional secret agreement that served to push the Soviets to agree to a withdrawal. In exchange for the removal of Soviet missiles, Kennedy agreed to remove all the Jupiter-class nuclear missiles the United States had placed in Turkey. Unlike the Soviets, however, the Americans insisted that their part of the bargain be kept secret, as the Kennedy Administration feared both domestic backlash and protests that the United States was selling out its NATO allies. Indeed, it was not until Robert F. Kennedy's posthumous memoirs that the discussion of removing of U.S. missiles from Turkey during the crisis even became public knowledge, and it would be even later still until the nature of the arrangement became clear.

One puzzling aspect of these events is why Khrushchev never revealed the deal, despite having incentives to do so, and why Kennedy trusted him to remain silent. Khrushchev could have humiliated Kennedy—both because the arrangement would have been unpalatable among U.S. domestic and allied audiences and because Kennedy had kept it hidden from them. Additionally, Khrushchev paid enormous domestic costs for the outcome of the crisis. Two years later, he lost power in no small part because his opponents saw his conduct during the crisis as weak and his decision to place missiles in Cuba as reckless. By making Kennedy's concessions public, Khrushchev might have spared himself this domestic punishment (Khrushchev 2000: 640-641; Jervis 2015: 31).

Our theory sheds light on the logic behind Kennedy's willingness to make a secret arrangement. The intuition behind the model is that leaders turn to secret diplomacy when they fear being perceived as "soft" or "weak" by domestic or international audiences. Indeed, scholars have identified this motivation as central to Kennedy's calculus. Kennedy was reluctant to make a trade with the Soviets in public, as "Kennedy and his colleagues did not want the American public or allies to know that he had

moved at least part of the way to meet Khrushchev's demands" (Jervis 2015: 25). Kennedy explicitly made secrecy a precondition of the deal. Robert Kennedy was instructed to warn Soviet Ambassador Anatoly Dobrynin that "the Jupiters would not be withdrawn if Moscow made any mention of the president's promise" (Lebow and Stein 1994: 122, 125, quote at p. 122; Jervis 2015: 21).

Second, our model indicates that the initiator understands the risks involved in reaching out to the adversary in secret, and thus views the adversary's unwillingness to leak as a way to screen his type. In the case of the Cuban Missiles Crisis, the historical record regarding Kennedy's willingness to accept the risk of trusting Khrushchev remains incomplete. But the evidence quite convincingly shows that he knew it was a risk, as knowledge of the deal would have caused considerable problems with allies, Congress, and the public. Secrecy only exacerbated the problem since U.S. allies and domestic audiences would likely have resented being deceived (Lebow and Stein 1994: 123, 127-129; Jervis 2015: 25). Furthermore, as Jervis (2015: 30) points out, "Kennedy knew that Khrushchev was impulsive," and the Soviet premier could easily have reconsidered his silence to score a political victory in the future. Nevertheless, by explicitly giving Khrushchev a weapon with which he could have humiliated the United States and curried domestic favor, Kennedy gained the ability to screen Khrushchev's trustworthiness while also signaling American willingness to cooperate. Averell Harriman, Assistant Secretary of State for Far Eastern Affairs, argued that making a deal on the Jupiters might "facilitate a 'swing' toward improved relations with the United States" on Khrushchev's part (Lebow and Stein 1994: 121).

Third, consistent with the screening equilibrium in our model, the U.S. chose a moderately important issue in the first round. Indeed, evidence suggests that the removal of the Jupiter missiles represented a moderately important issue—one sufficiently valuable to serve as a meaningful test of Soviet goodwill, but not so critical as to run major risks to U.S. national security if Khrushchev reneged. The missiles were of little-to-no military value, but nevertheless served an important political function. The administration considered the missiles "obsolete" (Trachtenberg 1985: 197, 199) in light of advances in U.S. submarine-launched ballistic missile technology (Allison and Zelikow 1999: 93, 114; Jervis 2015: 19-20, 26). However, Kennedy anticipated opposition from hawks in Congress, the media, the public, and even the Executive Committee of the National Security Council

(Lebow and Stein 1994: 128-129). Moreover, the administration expected to suffer reputational costs if other U.S. allies saw that it was willing to compromise on issues directly related to their own security for its own well-being. British Prime Minister Harold Macmillan argued that a deal involving the Jupiter missiles "would do great injury to NATO" (Jervis 2015: 26), a position with which U.S. officials—including both Kennedy and National Security Adviser McGeorge Bundy—agreed (Trachtenberg 1985: 199, 201; Jervis 2015: 21-23, 25). As Kennedy put it, knowledge of a deal involving the Jupiters "could break up the [NATO] Alliance by confirming European suspicions that we would sacrifice their security to protect our interests in an area of no concern to them" (Lebow and Stein 1994: 128).

Khrushchev's incentives to leak the Jupiter missiles deal, either before or after they had been removed, poses a more difficult question because we lack direct evidence about his calculations. It is fair to speculate, however, that Khrushchev had considerable short-term incentives to leak the deal, demonstrating Kennedy's dishonesty toward the United States' NATO allies, Congress, and the U.S. public. Moreover, Khrushchev could have scored points with his own domestic audiences by showing off his ability to extract concessions from the United States—and, perhaps just as importantly, by proving that he had not caved under U.S. pressure. Khrushchev not only faced pressure from Fidel Castro to take a hard line against the United States, but also faced a threat by China to Soviet leadership of the Communist bloc, and China was eager to portray the Soviet Union as weak (Lebow and Stein 1994: 115-116). Nevertheless, these benefits were outweighed by considerations of the longer-term benefits Khrushchev could have achieved from subsequent secret diplomacy with the United States over more important issues. Thus, the shadow of future cooperation with Kennedy (or, put differently, the cost of acquiring a reputation for untrustworthiness in the eyes of Kennedy) discouraged him from revealing the deal. Khrushchev used the agreement as a means to signal his willingness to Kennedy to continue to work together publicly or secretly in the future; in Jervis' (2015: 30) words, "if Khrushchev had revealed the secret he would have destroyed his relationship with Kennedy." Moreover, there is some evidence to suggest that Khrushchev was seeing potential benefit from keeping Kennedy engaged. As Khrushchev's son put it years later, "Father was now himself striving to achieve that 'special' relationship between two leaders that Kennedy had hoped to

establish when he left for the meeting in Vienna two years before" (Khrushchev 2000: 641).

The outcome of the crisis, as the model predicts, was greater levels of trust and cooperation between the two leaders. First, even after Kennedy withdrew the missiles, Khrushchev and consecutive Soviet leaders never revealed the secret agreement, thereby the secret diplomacy game continued. Second, first steps on the road to nuclear arms control were made possible. Khrushchev and Kennedy both began voicing support for a ban on nuclear testing, and in June 1963, Kennedy delivered his "A Strategy of Peace" speech at American University, in which he not only laid out his hopes for restraining the nuclear arms race but also discussed the need for superpower detente more generally (Jervis 2015: 30-31). This is not to argue that Khrushchev's willingness to keep his word and not reveal the secret agreement was the sole or even the primary reason for improved relations between the leaders; rather, it is one additional factor that scholars have not focused on sufficiently in their analysis. And yet, because Kennedy was assassinated a little over a year after the conclusion of the crisis, it is possible that Khrushchev's willingness to keep the arrangement secret would have contributed to greater secret cooperation between the two leaders had Kennedy remained in power.

Finally, we should be clear, the removal of the Jupiter missiles from Turkey was not purely a screening device. Both the Americans and the Soviets desperately sought a way to end the crisis before a nuclear war started. Indeed, the secret deal was not the only factor that ended the crisis, but it both facilitated the ending of the crisis (alongside the United States's signals and pledge not to invade Cuba), as well as served as a means to probe Soviet intentions. Kennedy's willingness to make the arrangement secret and his choice to approach the Soviets over a moderately important issue like the Jupiter missiles in particular reflects the logic of the model. If he had expected Khrushchev to reveal the deal, then making the Jupiter missiles concession public would have been preferable in order to avoid the humiliation of U.S. domestic audiences discovering that they had been deceived (Lebow and Stein 1994: 123-124). By making the arrangement secret, however, Kennedy both revealed a degree of trust in Khrushchev not to leak and gained the ability to utilize that trust in the future.

Iran-Contra Affair

The Iran-Contra affair is a case that illuminates the collusion equillibrium. In the mid-1980s, the United States began supplying Iran—considered to be a state sponsor of terrorism—with U.S. arms in exchange for the release of American hostages in Lebanon. Using Israel as a middleman, the United States made its first shipment to Iran in August 1985, and by early the following year the United States was instead directly negotiating with Iran. When the story was leaked to Al-Sharaa, a Lebanese publication, in November 1986, the Reagan administration attempted to cover up U.S. officials'—and in particular the President's—knowledge of what was being sent to Iran. Because the profits from those arms deals were then illegally used to sponsor the Contras in Nicaragua, this turned the Iran-Contra affair into a major scandal.

The extent to which President Ronald Reagan was aware of and involved in the arms-for-hostages deal remains unclear (Inouye and Hamilton 1987: 166-167). Nevertheless, there is no evidence to suggest that Reagan and other officials in his administration believed that secret cooperation with Iran would allow them to learn about Iran's type; on the contrary, evidence exists that Reagan had little faith in the United States' ability to cooperate with Ayatollah Ruhollah Khomeini's government in the long term. Rather, Reagan wanted first and foremost to save the hostages, and expected Iran to cooperate only as far as its self-interest demanded, as Iran was desperate for arms in its bloody war with Iraq. National Security Adviser Robert McFarlane put the matter bluntly: despite mistrust and ideological and rhetorical animosity, "today the force of events and self-interests has brought [the Iranians] to the point of realizing that we do have some common interests" (Byrne 2014: 199). McFarlane later recounted that upon being presented with a proposal to trade antiaircraft missiles for hostages, Reagan's response was to "cross your fingers or hope for the best, and keep me informed" (Kornbluh and Byrne 1993: 215).

For Reagan, the short-term incentive to free the hostages was paramount, and if it required acting illegally—and even if the attempts attained little success—it was better to do something than nothing. Indeed, even McFarlane, who feared that "we were being duped," whether by the Iranian government or by its middleman, arms dealer Manucher Ghorbanifar, resigned himself to continuing the effort

given that "matters seldom go the way one thinks they will in the Middle East" (McFarlane 1994: 40). According to Secretary of State George Schultz, Reagan claimed that "the American people will never forgive me if I fail to get these hostages out" (Inouye and Hamilton 1987: 198), and that "they can impeach me if they want" (Byrne 2014: 107). Defense Secretary Caspar Weinberger similarly recalled that Reagan could not tolerate the thought that "'big strong President Reagan passed up a chance to free hostages'" (Byrne 2014: 106).

Khomeini was desperate for armaments to gain advantage in Iran's conflict with Iraq—so much so that he was even willing to reach out to Israel—which gave him incentives not to leak the deal (Kornbluh and Byrne 193: 214-215, 243; Byrne 2014: 34, 91). Indeed, rather than signal goodwill, the Iranians repeatedly attempted to extort the United States by raising their demands and drawing out the hostage exchange via "sequencing," in which Iran would only give up a limited number of hostages at a time. (Kornbluh and Byrne 193: 217-218, 245, 248, 252). Funding for the Contras provided extra incentive for the Reagan administration to continue selling arms even without receiving hostages, though it is unclear to what extent Iran knew this and exploited it. What deterred the Iranians from leaking the deal was not only their desperation for arms, but also the political costs of any revelation that they were striking a deal with the United States, despite Khomeini's anti-American rhetoric. As Byrne (2014: 204) puts it, no Iranian official "was willing to strike a deal alone with the Great Satan without the political cover of involving the others."

Thus, the secret U.S.-Iranian arrangement to exchange arms for hostages followed the logic of our model's collusion equilibrium. U.S. and Iranian policymakers did not engage in secret negotiations as a means to reveal information about Iran's trustworthiness, and thereby allow for long-term cooperation. Rather, the U.S. was motivated by a desire to pursue short-term self-interest. A public deal was not feasible due to the U.S. arms embargo on Iran, Reagan's frequent references to Iran as a terrorist state, and the United States' status in Iran as the "Great Satan" (Byrne 2014: 67, 71, 75). Similarly, Iran faced high reputational costs for public cooperation with the United States, which created a severe adverse selection problem that in turn facilitated temporary collusion between the two sides. Each side's silence was ensured not by the needs of signaling, but by the dictates of short-term incentives.

Consistent with our predictions, after several shipments and the release of several hostages, the secret cooperation between the United States and Iran was ultimately leaked to the media by a member of the Iranian Revolutionary Guard who opposed any type of cooperation with the United States. Consequently, and as our theory explains, Reagan's secret and illegal dealings significantly undermined his presidency.

document