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

Keren Yarhi-Milo† and George Yin‡

Abstract

This paper explores how and under what conditions secret diplomacy preserves peace during international crises. 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 initially by keeping negotiations secret. We refer to this as the collusion equilibrium. We illustrate the logic of the model using three cases: secret negotiations between 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 secrecy, information transmission, and reputation in international politics.

*For questions, comments, and discussions, we thank Brian Blankenship, Allison Carnegie, Thomas Chadeauf, Aditya Dasgupta, Songying Fang, Mindy Haas, Iain Johnston, Melissa Lee, Helen Milner, Jim Morrow, Michael Poznansky, Alex Weisiger, and especially Anne Sartori.

†Assistant Professor at Princeton’s Politics Department and the Woodrow Wilson School of Public and International Affairs, kyarhi@princeton.edu.

‡Dickey Fellow in U.S. Foreign Policy and International Security, Dartmouth College, george.yin1@gmail.com. Corresponding author.
1 Introduction

Secret diplomacy is a double-edged sword. On the one hand, leaders who secretly collude with adversaries engage in a dangerous practice. Leaks of private communications and covert deals between leaders often result in grave public embarrassment, if not war. The revelation of the Zimmermann Telegram, in which Germany secretly promised that Mexico would receive previous Mexican territory from the United States in exchange for a military alliance, prodded the United States into entering World War I against Germany. The leaked private conversation between Wilhelm I and the French ambassador to Prussia over the Spanish throne instigated intense public outrage in both France and Prussia that contributed to the outbreak of the Franco-Prussian War. Notwithstanding the risk of explosive leaks, leaders still routinely rely on secret diplomacy to cooperate with their foes. American presidents have repeatedly employed back-channels and cooperated with adversaries behind closed doors. Most recently, there have been media reports of potentially unlawful collusion between members of the Trump transition team and Russia. The European Union has regularly entered into secret deals with terrorist organizations to rescue kidnapped citizens. China and Taiwan enjoyed a brief détente in the early 1990s thanks to secret communication between their top leaders, and the Oslo Accords between Israel and the PLO were possible largely due to the secretive nature of their negotiations. 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?

In this paper, we theorize the unique risks and benefits of secret diplomacy between adversaries that scholars have not yet sufficiently explored. 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). Here, we present a novel model that highlights a different reason why secret diplomacy can effectively promote cooperation between adversaries: private cooperation provides the initiator with an opportunity to screen for trustworthiness.²

Revealing whether the adversary is trustworthy is crucial for leaders seeking to cooperate with a hated enemy behind closed doors. This is because, despite reasonable estimates and good judgment, leaders remain uncertain about whether the adversary seeks long-term cooperation or simply an opportunity to betray and embarrass the initiator. Our model thus uncovers when and how secret diplomacy

² Reputation for trustworthiness in our paper, therefore, refers only to the initiator’s beliefs about the likelihood that the adversary will keep the assurances secret, as opposed to a generalized trust (Rathbun 2011).
enables leaders to reduce this uncertainty, and thereby allow for more accurate assessment of the intentions and interests of the adversary to cooperate.

Specifically, our model shows that whether secret diplomacy reveals useful information about the adversary depends on the degree to which an untrustworthy adversary has incentives to fake trustworthiness and resist the temptation to leak early on. Two key factors determine the severity of this “adverse selection” problem: the adversary’s anticipated costs from public cooperation with the initiator, and the adversary’s expected gain from betraying the initiator by revealing the secret assurances.

First, if an adversary leaks what is offered to him in secret, he will develop a reputation for being untrustworthy (i.e. that he cannot be trusted to keep secrets in the future), and thus the initiator will be unlikely to approach him in private again.\(^2\) Hence, after he leaks, the adversary is left with only the option of public diplomacy in future rounds. As we explain, conducting public diplomacy with a historical enemy could carry significant reputational and political risks for leaders. 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 secret the assurances he received behind closed doors because an untrustworthy adversary would do the same. Counter-intuitively, then, we find that secret diplomacy is least effective in fostering trust between adversaries precisely when leaders face strong political pressures to cooperate in secret rather than publicly. Second, secret diplomacy is most effective in building trust between foes when the adversary has a lot to gain from betraying the initiator. As we explain, when the adversary expects large political and strategic benefits from leaking, his decision to not reveal the initiator’s secret assurance is a very costly signal that credibly communicates his type. Secret diplomacy is thus a more efficacious trust-building exercise between foes when the adversary resists strong temptations to leak.

When the adverse selection problem is severe, secret diplomacy is less likely to have informational value. Nonetheless, even in such conditions, the initiator may still approach an adversary in private. This is because the initiator can sometimes expect to benefit from temporary collusion before an untrustworthy adversary betrays him in the future. Taken together, the incentives to fake trustworthiness

\(^2\) 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. When there is a leadership transition, the initiator may opt for secret diplomacy again given his assessment of the costs and benefits of approaching the new leader in private.
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 short-term collusion without learning (i.e., the collusion equilibrium). The “screening” and “collusion” equilibria we find point to very different rationales, logics, and consequences of secret diplomacy that have not been explored in the literature to date. We illustrate the empirical relevance of our theory by explicating the evolution of secret talks between the United States and China and revisiting several puzzling aspects of the Cuban Missile Crisis. We show that, in both of these cases, leaders employed secret diplomacy as a way to screen the intentions of the adversary. We also briefly illustrate the plausibility of the collusion equilibrium in the context of the Iran-Contra arms trade.

Our theory highlights unexplored dynamics in secret diplomacy and reputation building. We advance the literature on secret diplomacy in two ways. First, building on Yarhi-Milo’s insights on the risks of leaks in secret diplomacy, we explicitly theorize when and why an adversary would leak if secretly approached by a foe, a topic that existing studies have largely overlooked. It is hardly conceivable that a country would open up a secret channel of communication with an adversary without considering how that adversary might respond to private cooperation. We can only gain a partial understanding of the benefits and risks of private diplomacy if we do not consider this scenario. Second, existing game-theoretic models mostly focus on explicating when one would opt for private over public threats. This study, however, is the first formal exposition to our knowledge of private versus public cooperation in international politics.

Furthermore, this study contributes to the growing literature on reputation in security studies (Dafoe, Renshon, and Huth 2014) by exploring two important questions that have received little or overly discrete analysis in the literature. First, when, why, and how do observers draw inferences about reputation? Most rationalist works on reputation adopt the perspective of the actor who is seeking to acquire or defend a particular reputation (Sartori 2002; Sartori 2005). Scholars have dedicated less attention to the circumstances driving the initiator to create windows of opportunity for an adversary to credibly communicate its intentions. Our analysis reveals how leaders strategically use the "backstage" to set up opportunities for their adversaries to acquire a reputation for trustworthiness, and the extent to which they revise their beliefs based on the adversary’s past actions behind closed doors. Second, how do leaders balance different reputational concerns when they make foreign policy decisions? Existing studies of reputation often focus on one type of reputation in isolation — that is, whether it concerns honesty (Sartori 2005), hostility (Crescenzi 2007), or resolve (Weisiger and Yarhi-Milo 2013; Dafoe and Caughey 4
Instead, our analysis suggests that the key to understanding how secret diplomacy works hinges on explicating the interaction between the adversary’s desire to acquire a reputation for trustworthiness on the one hand, and its desire to avoid acquiring a public reputation for appearing weak or soft on national security in the eyes of either domestic or international audiences, on the other hand.

2 The Limits of Existing Explanations

Why would leaders rely on secret diplomacy over public diplomacy in international politics? The first class of existing explanations highlight how secret diplomacy is effective because it is actually not so “cheap” given the downstream consequences associated with approaching an adversary in private. Classic formal works on crisis bargaining highlight how public diplomacy is informative because it generates domestic audience cost (Fearon 1994; Smith 1998; Schultz 2001; Ramsay 2004). However, as Yarhi-Milo (2013) notes, private assurance is also costly, because there is the risk of leaks from both the opponent and third parties. Consequently, under some conditions, private assurance could serve as a costly signal that conveys important information for the adversary about the initiator’s intention. On the other hand, as Sartori (2002; 2005) shows, states pay reputational costs for dishonesty when they do not follow on 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 arm and launch a strike in response to the threat. Finally, private threats also communicate resolve in multi-dimensional bargaining spaces because states that care a lot about one particular issue are often reluctant to tell their opponents that they are also willing to fight over another 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 international compromise. Stasavage (2004) supplies the first formal argument explicating this logic. In his model, 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, under some conditions, leaders would prefer closed-door bargain-
ing 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 his
domestic audience as a means to extract a bargaining concession from the opponent. Public diplomacy,
therefore, easily 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. Fi-

ally, 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 (for a related
argument see Carson 2015). In contrast with studies that explicate the signaling value of secret diplo-
macy, 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 bargaining is valuable as a tool of
statecraft. Nonetheless, there are still several theoretical and empirical gaps in the literature on secret
diplomacy. First, no study so far – to our knowledge – has examined when and why an adversary will
collude with the initiator to hide a deal from the public, and under what conditions it will likely reveal
the information. 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 for the initiator.
Second, existing works focus almost exclusively 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
us shed light on several empirical puzzles concerning secret diplomacy: specifically, why and when
would adversarial leaders sometimes engage in multiple rounds of sustained secret cooperation, while
other times, only pursue one-off secret cooperation on a single issue. In this paper, we hope to fill these
gaps, enabling a better understanding of both the functions and limitations of secrecy in international
relations.

3 Although both Kurizaki (2007, 550) and Yarhi-Milo (2013, 419) have hinted at the importance of this
issue.
3 Theory

Our theory seeks to answer three important, and related, questions in international politics: (1) When and why would a leader pursue secret rather than public diplomacy? (2) When would an adversary resist the temptation of leaking? and (3) Under what conditions does secret diplomacy lead to effective screening of the adversary’s types, and when does it lead at best to temporary collusion with untrustworthy adversaries? To address these puzzles, our model builds on two observations about the conduct of diplomacy between adversaries.

First, political leaders may face costs for cooperating with an adversary in public. All else equal, we argue, leaders who choose to cooperate with a sworn enemy in public could face criticism from both domestic and international actors for being “weak” or “soft”. Scholars have focused on the salience of reputational costs that domestic or international actors might impose for reasons of ideology, historical acrimony or strategy, such as maintaining the status quo. Some Arab regimes, such as Saudi Arabia, have had to keep their dealings with Israel private because doing so in public would raise the ire of fellow Arab countries and the Palestinian Authority, who might view them not only as "weak" but also as "traitors." Similarly, one of the reasons the United States does not conduct public negotiations with terrorist organizations is because U.S. allies and adversaries could perceive it as being "soft". We expect reputational costs for cooperating with the adversary in public to be more severe under some circumstances: when a leader faces multiple adversaries; when audiences could effectively impose political punishment on leaders who publicly negotiate with an adversary; or when the leader is facing a more hawkish domestic audience (Schultz 2005).

Furthermore, initiators of public diplomacy can also be subject to other types of costs that are not reputational in nature for engaging with the adversary. For example, public diplomacy could induce third parties (domestic or international) to devise strategies 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 the nature or origins of these costs, they can lead leaders to consider the option of secret rather than public diplomacy.

Second, when a leader secretly approaches an adversary, the adversary might be tempted to publicize the secret deal in order to humiliate the leader who initiated secret diplomacy. Domestic and

---

4 To be clear, here we conceive of strategic leaks that are clearly attributed to the adversary. Sometimes, however, adversaries might decide to leak information while concealing the source of the leak, or attributing it to third parties. With uncertainty regarding attribution, the adversary would acquire a
international audiences who objected to public diplomacy with the adversary are likely to be even more resentful when they find out from leaks that the leader offered concessions to the adversary in secret. The revelation that secret diplomacy was pursued can make the initiator appear not just dishonest by concealing his overtures (Stasavage 2004), but also incompetent given that he 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. After all, 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 nondemocracies).

Allies affected by the secret deal could also be outraged and might argue that they should have been consulted, and that the fact that the leader kept this secret from his allies is an indication of his willingness to pursue a foreign policy that is inconsistent with the allies’ shared interests. For example, when the United States negotiated in secret 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 with the leader has not yet disclosed, thereby undermining the credibility of the leader. Thus, a strategic leak by the adversary can 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, however, can be greater in some circumstance. By leaking, an adversary may signal to its domestic audience and allies that he is honest and unwilling to keep secrets from these audiences. Moreover, the political "brownie points" that adversaries may score from leaking might be substantial, especially when the ideological distance (Haas 2005) between an adversary and the state that initiates secret diplomacy is larger. Moreover, adversaries could see larger benefits from leaks when they believe those leaks could lead to desired outcomes. For instance, if the adversary wishes to topple the leader who initiated secret contact, he is more likely to leak when he believes the opposition in the initiator’s country would have the willingness and ability to capitalize on the leak and remove 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 negative reputation for leaking with a delay (until the initiator became fairly certain who leaked).
allies, then the adversary will have stronger incentives to leak when he believes such information could be damaging to the alliance relationships.

We take the two stylized facts above about the potential costs and benefits facing the initiator and the adversary when they engage in diplomacy as the key assumptions that drive 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 below that it is precisely this temptation for the adversary to leak that makes secret diplomacy an effective tool to help foes build trust. Nonetheless, rather than emphasizing the ability of the initiator to signal his intentions through secret diplomacy (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).

3.1 The secret diplomacy model

There are two players in this game: the initiator of diplomacy ($P_1$) and an adversary ($P_2$). The adversary has two types, trustworthy ($P_{2T}$) and untrustworthy ($P_{2U}$). When approached by $P_1$ in secret, the trustworthy adversary $P_{2T}$ 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 $P_{2U}$ 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 cooperation in secret versus leaking, which are a function of the adversary’s domestic political concerns and assessment of their nation’s strategic environment, as discussed earlier.

Nature starts the game by choosing the opponent’s type, with probability $a$ that the adversary is trustworthy. Importantly, we conceive of "trust" $a$ in this model to refer only to the initiator’s beliefs about the likelihood that the adversary will keep the assurances secret, as opposed to an ideological belief about the general trustworthiness of others (Rathbun 2011). 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 the nature of the leader of the adversary, as well as his personal assessment of the adversary’s interests and intentions to leak.

5 See Riley (2001) for a survey of screening models in economics.

6 Put differently, we conceptualize trust as the initiator’s belief in the probability the adversary will leak
After nature assigns a type to the adversary, $P_1$ chooses the “stake” associated with the first round game $\gamma$ relative to the second round $(1 - \gamma)$, with $0 \leq \gamma \leq 1$. The choice of $\gamma$ may signify the importance of the issue driving $P_1$ 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 regarding how the game proceeds. (Kydd 2000; Kydd 2005).

Our game consists of two rounds, which allow $P_2$ to build a reputation that would affect $P_1$’s decision to go private or public in the second round. In each round, $P_1$ decides whether to cooperate with $P_2$ in public or in private. If $P_1$ approaches his opponent in public, $P_2$ can accept the offer of public cooperation, and the two players receive payoffs $(c_1 - r_1, c_2 - r_2)$. $c$ is the gain from cooperation, and $r$ is the political cost associated with making a deal with a foe in public. $P_2$ may also reject $P_1$’s offer of public cooperation, and both players receive payoffs $(0,0)$. The game ends in the first round if $P_1$ chooses public diplomacy. Without loss of generality, $P_1$ has no option of retaining the status quo –

---

in a particular situation. Trust in thus leader and situation specific, and not an inherent trait of an actor. Moreover, trust is domain specific in that it refers only to the initiator’s belief about the probability the adversary will be able to keep secrets (concerning any policy issues) in the future.

---

7 See Watson (1999) and Kydd (2000: 336) for models that endogenize the weights players put on different rounds of cooperation.
8 Our model inherits from Kydd (2000; 2005) the assumption that both sides agree on what is considered a small issue or a big issue.
9 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. Watson (1999) presents a multi-round game where partners jointly craft a cooperation regime that allows for screening.
10 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).
11 In our model, $P_1$ does not suffer political cost from public negotiations when they fail. This is because in such situations, $P_1$ can simply claim that he did not reach a public deal with $P_2$ because he is unwilling to concede to $P_2$’s demand. Thus a failed public negotiation can actually make $P_1$ look “tough” instead of “weak”. Furthermore, there is no need to make costly concessions to buy off domestic opposition when there is no deal.
12 We have also solved a version of the model where there is a second round after the public diplomacy. Such a game has an additional equilibrium where public diplomacy can screen as well. For this paper, we focus solely on screening with secret diplomacy not only because our “simple” model already provides rich formal results, but also because we wish to highlight the understudied phenomenon where reputation for trustworthiness can be gained and lost behind closed doors.
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) where he 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 from our model remain effectively unchanged when we allow \( P_1 \) to opt for the status quo of no diplomacy.

If \( P_1 \) attempts private cooperation, \( P_2 \) can reach a secret deal with \( P_1 \) that would give the players payoffs \((c_1, c_2)\),\(^{13}\) Alternatively, \( P_2 \) can betray \( P_1 \) by leaking\(^{14}\), which would give the players payoffs \((-d_1, b_i^T)\)\(^{15}\). For \( P_2^T \), \( c_2 > b_i^T \), while for \( P_2^U \), \( b_i^U > c_2 \). The superscript \( i \) indexes the type of \( P_2, i \in I \equiv T, U; \) \( T \) indicates 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.

\(^{13}\)Some secret agreements require eventual public revelation if leaders want to implement them, as in the case of peace agreements that involve transfer of territories (Yarhi-Milo 2013: 419 - 420). In these cases, leaders will still pay a political cost for cooperating with a foe in secret, but that cost is smaller compared to public diplomacy. This is because leaders can control the pace and timing of the revelation to reduce possible domestic and international opposition to the deal. In a model where secret diplomacy is associated with some political cost because it must be revealed, our results below hold as long as \( c_2 - r^*_2 > b^T_2 \) holds, with \( r^*_2 \) as the reduced political cost associated with secret diplomacy with revelation. The inequality ensures that the trustworthy type will still prefer secret cooperation over leaking even if a secret deal is costly because it requires ultimate revelation.

\(^{14}\)We do not allow \( P_2 \) to reject a secret offer without leaking in our model, because giving \( P_2 \) this option will not affect his equilibrium strategic choice. For \( P_2^T \), rejecting a secret deal without leaking gives him a pay-off of 0. Therefore \( P_2^T \) will always prefer leaking (which gives him a positive pay-off) over rejecting secret cooperation without leaking. For \( P_2^U \), rejecting a secret deal without leaking also gives him a pay-off of 0. Therefore \( P_2^U \) will always prefer secret cooperation (which gives him a positive pay-off) over rejecting secret cooperation without leaking.

\(^{15}\)Some leaks are beyond the control of the adversary’s leader; we will discuss autonomous leaks later in the paper.
The second round of the game 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 $\gamma$. 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 $s \in S \equiv \text{cooperate, leak}$. When $P_1$ updates its belief regarding $P_2$’s type after observing $P_2$’s past behavior, $\alpha'(s_2, \gamma)$ will be distinct from $\alpha$, the prior belief that $P_2$ will not leak. When $P_1$ does not update its belief regarding $P_2$’s type after observing $P_2$’s strategy in the first round game, $\alpha'(s_2, \gamma)$ equals $\alpha$. $P_2^T$ can only acquire a reputation of trustworthiness when $P_1$ considers past behavior as a good indicator of whether...
\( P_2 \) is going to leak in future secret cooperation.

Finally, international politics often requires adversaries to probe whether temporary cooperation on issues where state interests potentially overlap is possible. For example, although Jordan and Israel had been embroiled in a bloody conflict for many years, they often discussed issues of mutual interest. As such, we zoom in on the leaders’ choice to conduct diplomacy secretly or publicly, rather than the decision whether to conduct diplomacy at all. In our main model, we thus assume \( c_1 > r_1 \) and \( c_2 > r_2 \) (see Appendix B for more information on a model where we relax these constraints). The first inequality ensures that \( P_1 \) will always receive a positive payoff from public diplomacy. Put differently, public cooperation allows leaders to improve their position over the status quo, even after paying a cost for engaging in 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. Figure 1 presents the extensive form of our secret diplomacy game.

### 3.2 Equilibrium analysis

The equilibrium concept employed here is Perfect Bayesian Equilibrium (PBE; see, e.g., 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 \( \alpha \) and \( \alpha' \left( s_2^*, \gamma^* \right) \). All proofs are in the appendix. Table 2 outlines the options and beliefs facing each actor in both rounds 1 and 2.
### Table 2: Strategies and beliefs facing the players

<table>
<thead>
<tr>
<th>First round</th>
<th>Strategy</th>
<th>Belief(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>$\gamma \in [0, 1]; {\text{Private, Public}}$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$P^U_2$</td>
<td>(collude, leak) (if Private); (cooperate, reject) (if Public)</td>
<td>n/a</td>
</tr>
<tr>
<td>$P^T_2$</td>
<td>(collude, leak) (if Private); (cooperate, reject) (if Public)</td>
<td>n/a</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second round</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>${\text{Private, Public}}$</td>
</tr>
<tr>
<td>$P^U_2$</td>
<td>(collude, leak) (if Private); (cooperate, reject) (if Public)</td>
</tr>
<tr>
<td>$P^T_2$</td>
<td>(collude, leak) (if Private); (cooperate, reject) (if Public)</td>
</tr>
</tbody>
</table>

#### The screening equilibrium

In the screening equilibrium, the initiator successfully induces the untrustworthy adversary to leak and reveal its true type in the first round, which in turn allows the trustworthy adversary to acquire a reputation of trustworthiness by not leaking early in the cooperation relationship. The screening equilibrium works because $P_1$ has set the importance of the first round secret diplomacy $\gamma$ high enough that an untrustworthy adversary would leak and reveal its true colors in the initial round of the game. As $\gamma$ 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 $\gamma$ becomes larger, the initiator of secret diplomacy could face a more explosive leak in the event the adversary decides to betray.

We observe the screening equilibrium under two main conditions: when the initiator is sufficiently trustful of the adversary to give secret diplomacy a chance, and when 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 $\alpha$ – the initial level of trust that the initiator has in the adversary - is above a critical threshold $\tilde{\alpha}_{\text{screening}}$:  

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

---

$\tilde{\gamma}$ is the minimum level of importance associated with the first round issue that would induce $P^U_2$ to reveal her true colors during the initial round of diplomacy, or what we call the minimum “strength” of an effective screen. We discuss it in more detail in the next section.
If $a > \tilde{a}_{\text{screening}}$, the initiator is sufficiently trustful of the adversary to employ secret diplomacy as a screen. If $a < \tilde{a}_{\text{screening}}$, the initiator is too fearful of a leak to use secret diplomacy. $\tilde{a}_{\text{screening}}$ 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. Note that $\tilde{a}_{\text{screening}}$ is increasing in $\hat{\gamma}$, a parameter that denotes the level of incentive for the untrustworthy adversary to resist the temptation of leaking in the first round. In other words, when the untrustworthy adversary is strongly inclined to fake trustworthiness by not leaking initially, the informational value of secret diplomacy diminishes, and the initiator becomes less willing to bear the risk of secret diplomacy.\footnote{When $\hat{\gamma}$ is 1, the requisite level of trust that sustains the screening equilibrium is actually equivalent to the requisite level of trust that sustains a collusion equilibrium.}

In addition to the initial level of trust, the second factor affecting whether the initiator can effectively screen using secret diplomacy concerns the incentive for an untrustworthy adversary to mimic a trustworthy one. We derive this parameter $\hat{\gamma}$ by comparing the expected two-round payoff facing the untrustworthy adversary when he leaks in the first round compared to when he does not leak:

$$\frac{c_2 - bU_2 - r_2}{2c_2 - 2bU_2 - r_2}$$

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 early. $\hat{\gamma}$ is a function of two factors: the adversary’s anticipated cost from cooperating with the initiator in public, and the adversary’s potential gains from leaking. We discuss each in turn.

First, the untrustworthy adversary is more compelled to fake trustworthiness and keep diplomacy behind closed doors 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 and humiliate him again. 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 its 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, in addition to political cost from public diplomacy, the gains the adversary could receive from leaking also affect incentives for an untrustworthy adversary to fake trustworthiness in secret. Formally, \( \hat{\gamma} \) is decreasing in \( b_{2}^{U} \); thereby the untrustworthy adversary is less compelled to fake trustworthiness when the payoff from leaking is high. By not leaking initially, the untrustworthy adversary forgoes the payoffs from imminent betrayal, \( \gamma b_{2}^{U} \), which is the opportunity cost of mimicking the trustworthy type. Clearly, 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 on 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 the first round.

Thus far, we have discussed the adversary’s incentives to mimic a trustworthy type. Now we explain the initiator’s incentives to use secret diplomacy to screen. Here, we find first that when an untrustworthy adversary has a strong incentive to mimic a trustworthy type, screening is more difficult but not impossible. In such circumstances initiators who wish to screen using secret diplomacy face a dilemma. On the one hand, the initiator could approach the adversary with a very important issue in the first round (set \( \gamma \) high) to make it worthwhile for the untrustworthy adversary to leak and thereby reveal his type. Structuring the secret cooperation regime in this way would also allow the trustworthy type to credibly signal its trustworthiness by not leaking. On the other hand, such a strategy comes with a significant risk for the initiator. Note that 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 \( \gamma \) in order to bait the untrustworthy adversary into leaking.

On the other hand, when the untrustworthy adversary has a weak incentive to mimic, the initiator
could choose a moderately important issue to screen out the untrustworthy adversary. The maximum stake that \( p_1 \) will be willing to put on the first round of the secret diplomacy game in order to screen – instead of collude, which we will discuss in the next section – is \( \tilde{g}_{\text{screening}} \):

\[
\frac{c_1 - r_1 + d_1}{2c_1 - r_1 + 2d_1}
\]

We can thus interpret \( \tilde{g}_{\text{screening}} \) as an indicator of the initiator’s incentive to learn the adversary’s type at the risk of exposing itself to a damaging leak. When \( \tilde{g}_{\text{screening}} \) is large, the initiator has a strong incentive to test the adversary with secret diplomacy. When \( \tilde{g}_{\text{screening}} \) is small, the initiator has a weak incentive to screen. Importantly, the initiator has less incentive to utilize secret diplomacy as a litmus test when the political cost \( r_1 \) associated with public cooperation is high (\( \tilde{g}_{\text{screening}} \) 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 the public cooperation payoff is small. We know that 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 (\( \tilde{g} \) is high); this makes it difficult for the trustworthy type to signal its 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 (\( \tilde{g}_{\text{screening}} \) is low).

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

Finally, note that the initiator’s optimal stake $\gamma^*$ is simply $\hat{\gamma} = \hat{\gamma}_{\text{screening}}$, which is the lowest possible $\gamma$ 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 $\tilde{\alpha}_{\text{screening}} \leq \alpha$ and $\hat{\gamma} \leq \hat{\gamma}_{\text{screening}}$, there exists a unique perfect Bayesian equilibrium of the game with the following strategies. $P_1$ sets $\gamma = \hat{\gamma}_{\text{screening}}$. 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 help adversaries learn about types and thus is not instrumental in building trust: the collusion equilibrium and the public diplomacy equilibrium.

For the collusion equilibrium, the initiator always receives the private cooperation payoff $c_1$ (weighted by $\gamma$) 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 2). 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 its initial encounter with the adversary ($\alpha'(s^*_2, \gamma^{\text{collusion}}) = \alpha$). In the second round, the initiator goes private again, the trustworthy adversary still colludes, but the untrustworthy adversary now leaks. Under this equilibrium, temporary collusion paves the way for a possible leak later.

When and why would leaders end up in this collusion equilibrium? Some leaders, we argue, will find this type of secret diplomacy appealing because it offers a possibility to achieve short term collusion over an immediate issue of concern for the initiator (albeit at the risk of betrayal in the second round). 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 \( \gamma \) 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.

Note that, as Figure 2 shows, the collusion equilibrium requires the initiator to have a higher level of trust in the adversary compared to the screening equilibrium.\(^{18}\) This has to be 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 (\( \alpha' > \tilde{\alpha}'_{\text{secret}} = 1 - \frac{\gamma}{c'_t + d'_t}, \) with \( \alpha = \alpha' \), as the initial level of trust is identical of the level trust in round 2 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 2 because the game ends in round 2. 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 has a screening function that the initiator may decide to utilize). Consequently, the initiator under the collusion equilibrium must be quite confident in the adversary’s trustworthiness. The screening equilibrium, in contrast, does not require the same level of trust.\(^{19}\)

The result above highlights how a trustworthy adversary suffers from adverse selection. As discussed above, 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 compared to 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 forced to practice public diplomacy.

Proposition 2 summarizes our discussion of the collusion equilibrium.

**Proposition 2 (collusion equilibrium):** If \( \tilde{\alpha}'_{\text{secret}} < \alpha' \) and \( \hat{\gamma} > \hat{\gamma}_{\text{screening}} \), there exists a unique perfect Bayesian equilibrium of the game with the following strategies. \( P_1 \) sets \( \gamma = \hat{\gamma} - \epsilon.\)

\(^{18}\) \( \tilde{\alpha}'_{\text{secret}} > \tilde{\alpha}_{\text{screening}} \) is always true unless \( \gamma = 1 \), e.g. when it is impossible to induce the untrustworthy adversary to reveal its true color in round 1.

\(^{19}\) *Nota bene:* the high level of trust necessary to sustain the collusion equilibrium can also sustain the screening equilibrium. The key distinction between the collusion and the screening equilibrium, therefore, lies in \( \hat{\gamma}_{\text{screening}} \).
In the first round of the game, $P_1$ goes private, while both $P_{2T}$ and $P_{2U}$ collude. In the second round of the game, $P_1$ maintains its prior belief $\alpha' = \alpha$ and goes private, while $P_{2T}$ colludes and $P_{2U}$ betrays.

We now briefly discuss the public diplomacy equilibrium. Here, 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 ($\tilde{\alpha}_{\text{screening}} > \alpha$). 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 $\hat{\gamma} < \tilde{\gamma}_{\text{screening}}$. 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: (1) it has very little trust in the adversary; (2) it is too risky to screen and not trustful enough to collude. Proposition 3 formally characterizes the public diplomacy equilibrium.

**Proposition 3 (public diplomacy):** If (a) $\alpha < \tilde{\alpha}_{\text{secret}}$ or (b) $\tilde{\alpha}_{\text{screening}} < \alpha < \tilde{\alpha}_{\text{secret}}$ and $\hat{\gamma} < \tilde{\gamma}_{\text{screening}}$, there exists a unique perfect Bayesian equilibrium of the game with the following strategies. $P_1$ goes public and sets $\gamma = 1$. Both $P_{2T}$ and $P_{2U}$ accept the public deal.

### 3.3 Summary and discussion

Our model reveals two reasons, overlooked in the existing literature, why leaders may employ secret diplomacy. First they may use secret diplomacy as a litmus test to determine the trustworthiness of an opponent, thereby providing an opportunity for a trustworthy adversary to acquire a positive reputation. Second, they may go private to secure temporary collusion with an untrustworthy adversary.

Secret diplomacy only serves informational purposes in the screening equilibrium, which corresponds to the pink area of Figure 2.\(^{20}\) Crucially, a trustworthy adversary can only acquire a reputation

---

\(^{20}\) The results of our model readily generalize to a game of more than two players. A third player $P_3$, if rational, will learn of $P_2$’s type after observing a betrayal, and treats $P_2$ as a sucker accordingly (refuse to approach $P_2$ in secret). We can simply multiply the payoff function facing $P_2$ in the second round by $t$ (the number of players facing $P_2$) to capture the added cost (benefit) of having a bad (good) reputation when $P_2$ faces multiple adversaries in the future.
for trustworthiness in this model when two conditions hold simultaneously: (1) 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 $a < \bar{a}_{secret}$) 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.

If leaders do not find it worthwhile to utilize secret diplomacy as a screen, they may still employ secret diplomacy to induce an adversary – including the untrustworthy type – to collude temporarily. This scenario corresponds to the collusion equilibrium, which is represented by the magenta area in Figure 2. Of course, leaders may also decide not to utilize secret diplomacy, and turn to public diplomacy instead. This public diplomacy equilibrium is represented by the white area in Figure 2.

Before proceeding to the case studies, we wish to discuss several features of our model. Thus far we have restricted our attention to a simple model to explicate the incentive for leaders to opt for secret instead of public diplomacy. Do our results remain robust if we allow the initiator to retain the status-quo (e.g., not to reach out to an opponent at all, leaving all players with the payoff 0)? In this richer model, the initiator can punish an untrustworthy opponent in the second round by denying the opponent any cooperation. We argue that our key results remain effectively unchanged in this extension of the model. Crucially, it is irrational for the initiator to punish an untrustworthy opponent by refusing cooperation in the second round if the payoff of a public deal is positive, e.g., when the political cost of a public deal is sufficiently low ($c_1$ is larger than $r_1$). Put differently, when the political cost of a public deal is not too high, “public cooperation” dominates “no cooperation” as a strategy. Consequently, the equilibrium predictions of the richer model are identical to the simple model that we presented in the main body of the text.
Figure 2: Secret and public diplomacy equilibria

Note: The x axis is $\alpha$, the level of trust $P_1$ has in $P_2$. The y axis is $\tilde{\gamma}_{\text{screening}}$, the maximum stake that $P_1$ would be willing to put on the first round of the secret diplomacy game to screen. $\tilde{\alpha}_{\text{secret}}$ is the minimal level of trust that $P_1$ must have in $P_2$ for $P_1$ to employ secret instead of public diplomacy. $\tilde{\alpha}_{\text{screening}}$ is the minimal level of trust that $P_1$ must have in $P_2$ for $P_1$ to employ secret diplomacy to screen instead of going public. $\hat{\gamma}$ is the minimum weight associated with the first round issue that would induce $P_2^{U1}$ to reveal its type by leaking early on.

Allowing the initiator to “not cooperate” and retain the status quo only affects our results when $c_1$ is smaller than $r_1$ and/or $c_2$ is smaller than $r_2$, e.g., the political cost of a public deal is so high that a public deal gives the initiator and the adversaries negative payoffs (See Appendix B for the case when $c_1 > r_1$ and $c_2 > r_2$). First, the public diplomacy equilibrium disappears because now “no cooperation” dominates “cooperation.” Second, the comparative statics regarding the effect of the political costs of public cooperation on incentives for the initiator to screen and the untrustworthy adversary to mimic will not hold, because the status quo is now the default option when the initiator refuses to go private. Nonetheless, the mechanisms that we identify earlier that sustain screening and collusion remain robust even in this richer model. In sum, the key insights from our model do not depend on its simplicity.
Second, leaders would 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 remains valuable for understanding cases where leaders may deny their involvement in secret diplomacy. Essentially, the ability of the initiator to deny involvement in private diplomacy affects the 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 ($b_U^2$) would decrease. From our earlier analysis, we know that the initiator will become more tolerant of the risk of betrayal and employ secret diplomacy when the damage from a leak is small. Furthermore, we also know that the untrustworthy adversary faces 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 sum, our model suggests that the ability of leaders to deny their participation in secret diplomacy would make it less risky for the initiator to employ secret diplomacy, although it also reduces secret diplomacy’s informational value.

Third, our model seeks to explicate the strategic logic behind an adversary’s decision to leak. Thus, to sharpen the analytical focus of our paper, the audiences in our model do not find out about a secret deal if the adversary does not intentionally leak. Nonetheless, our model has implications for understanding how the "autonomous risk of leaks" (Yarhi-Milo 2013) - leaks that originate from third parties such as the media or disgruntled officials and are beyond the control of the negotiating leaders - affects secret diplomacy. Crucially, the threat of autonomous leaks would incentivize the untrustworthy adversary to fake trustworthiness in the first round, because now the adversary may still benefit from a “windfall” from an autonomous leak even if it is not the one doing the leaking. Therefore the initiator must increase the importance of the first round issue ($\tilde{g}$) to screen the untrustworthy adversary, which would reduce the incentive for the initiator to employ secret diplomacy as a litmus test in the first place.

Finally, what does our model say about situations where an untrustworthy adversary leaks right after striking a secret deal? Can an untrustworthy adversary have it both ways? The ex post leak scenario is actually a special case of our model where $\gamma$ 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 he benefited from the secret

---

21 $\gamma$ is therefore 1/2.
deal in the first round. The situation of *ex post* leak, therefore, corresponds to the collusion equilibrium in our model.

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). In addition, these illustrations also enable us to probe the relevance of our model to explain important cases of secret diplomacy. Our choice to leverage 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 of secret diplomacy contains core elements of strategic interaction that depend on the beliefs and preferences of leaders. Those beliefs and preferences 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, and as such the evidentiary record, even in these historical cases, is still incomplete.

With these caveats in mind, when illustrating the empirical relevance of our model we focus on several important observable aspects of our theory’s causal mechanism: 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) to Appendix C.

---

22 We focus less on the comparative statics of our model in the case studies because of the difficulty involved in observing some parameters, such as expected reputation cost from public diplomacy or perceived expected gains from betrayal. To the extent that we find evidence on those perceptions, we report them in the cases.
The Secret Opening to China

For twenty years after the Chinese Communists took power in 1949, the United States refused to recognize the new People’s Republic of China, instead considering the Republic of China on Taiwan to be the legitimate government of all of China. 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 about publicly approaching China 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. \(^{23}\) 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.” 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 (Macmillan 2007: 161, 174, 257, 327).

More specifically, consistent with the screening logic of our model, 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 broach 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 Pak-

\(^{23}\) We do not argue that these were the only concerns shaping the desire for secrecy, but rather that there is evidence to suggest that they played an important role in the calculations.
istan, 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 (Komine 2008: 95-99, quote at p. 95). Then in
December, the U.S. ambassador to Poland unofficially indicated to the Chinese Chargé d’Affairs after a
fashion gala in Warsaw that Nixon “would like to have serious concrete talks with the Chinese” (Komine
2008: 117).

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 the 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 (Komine 2008:
123-124, 139).

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). Re-
garding 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 (Komine 2008: 167), 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 main-
land (Macmillan 2007: 259). Then, when Nixon visited the following year, U.S. officials reiterated these
assurances on Taiwan, Japan, and the Soviet Union. This time, however, Kissinger went a step further
and shared American intelligence on the Soviet Union, including its military assets (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 trustworthy 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. 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" (quoted in Komine, 162). 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.”

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 move 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.” He noted in his memoirs 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 his secret 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.”

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. Even 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.” 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.” 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.” 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.”

To be sure, 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.

**Iran-Contra Affair**

While the first two case studies illustrate the screening logic of our model, the Iran-Contra affair is a case of secret negotiations that illuminate the collusion equilibrium. 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 no longer using Israel and 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 incentive 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. What motivated both U.S. and Iranian policymakers to use secret diplomacy was not the expectation that this interaction would 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 the predictions of our theory, after several shipments and the release of several
hostages, the secret cooperation between the United States and Iran was ultimately leaked to the me-
dia by a member of the Iranian Revolutionary Guard who was against any type of cooperation with the
United States. Consequently, and as our theory explains, Reagan’s secret and illegal dealings signifi-
cantly undermined his presidency.

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 trustworthiness. Crucially, leaders are more inclined to employ
secret diplomacy to screen an adversary when the adverse selection problem is moderate, e.g. when the
adversary is unlikely to fake trustworthiness by not leaking in order to betray at a more opportune time
in the future. In addition to specifying clear conditions under which secret diplomacy leads to trust or
short-term collusion, our model goes a long way in illuminating the logic behind secret diplomacy in a
number of high-profile historical cases.

Our theory highlights an understudied aspect of interstate communication. While most studies have
focused on leaders’ incentives and capabilities to employ costly signals to communicate their own in-
tentions, we focus on how leaders can create conditions that would facilitate or hinder their adversaries’
ability to credibly signal their type. Put differently, the extent to which adversaries can credibly convey
intentions is shaped by whether their opponents give them the opportunity to do so. In the context of
this paper, we derived the conditions under which leaders will seek to use secrecy diplomacy to learn
whether the adversary is trustworthy. Future studies would benefit from exploring further when op-
ponents will actively set conditions that would allow them to credibly determine the intentions of their
adversaries, as well as additional tools and strategies that they can employ to effectively screen for types.

Our theory and findings also carry significant implications for the study of reputation in international
politics. We present clear mechanisms and conditions under which adversaries can acquire a reputation
for trustworthiness in the eyes of their counterparts. We show that it is possible to do so even in the
absence of public signals. As such we contribute to a growing literature on the informational value of se-
crecy (Yarhi-Milo 2013; Carson and Yarhi-Milo 2016). Moreover, we show that leaders who initiate secret
diplomacy recognize the effectiveness of the secrecy platform precisely because it allows their adver-
sary to gain and lose a reputation for keeping secrets. However, a reputation for trustworthiness that is
acquired through secret diplomacy has limitations: because learning about types is taking place "back-stage" between the leaders, this reputation might not translate into greater trust between the leaders’ publics, and thus might not allow leaders to move the cooperation into the "front stage." The extensive secret cooperation between the leaders of Israel and Saudi Arabia in the absence of any formal relations between the two countries is a case in point. Our paper explores for the first time the opportunities and limits of reputation building between leaders in the secret realm. We focus on multiple rounds of secret cooperation because leaders often have strong political incentives to keep cooperation with an adversary secret when they have the option. However, future studies could helpfully investigate how secret diplomacy might also be valuable for subsequent public cooperation.

We want to conclude by highlighting that all theoretical models are understandably simplified depictions of reality. For example, our model gives the initiator the opportunity to structure the cooperation regime in a way that would test the adversary’s trustworthiness. In reality, however, leaders might not always have the choice over the issues with which to approach an adversary. There might be imminent issues that would take precedence over others. The Iran-Contra case, for example, highlights this constraint. Nonetheless, we show that even in such situations, leaders can still recognize the extent to which cooperation in secret can reveal the trustworthiness of the adversary. As such, our model is valuable even in the absence of control over the issues. Another potential complication of our model is that for leaders to be able to effectively screen or collude requires accurate assessments about the adversary’s incentives to fake trustworthiness. In reality, however, leaders could misconstrue those incentives. Thus, they can believe, for example, that secret diplomacy in the initial round allowed them to uncover the adversary’s type, when in fact it has not. While we do not model those misperceptions, the value of our theory lies in explicating their potential consequences. Our paper sheds light on the conduct of leaders behind closed doors. Yet, much remains to be studied about secrecy in world affairs.
References


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


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 $P^T_2$ will always collude. $P^U_2$, in contrast, would always leak in round 2 to reap the payoff $b^U_2$ because there is no political cost for doing so. On the other hand, if $P_1$ decides to go public in round 2, both $P^T_2$ and $P^U_2$ 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 $P^U_2$ does not mimic $P^T_2$ in this equilibrium, $\alpha'(\text{collude}, \gamma^*) = 1$ and $\alpha'(\text{leak}, \gamma^*) = 0$. 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, we know that $P^T_2$ will always cooperate in private while $P^U_2$ will cooperate in private when $\gamma^* > \hat{\gamma}$ and betray if $\gamma^* \leq \hat{\gamma}$. $\hat{\gamma}$ is the lowest possible $\gamma$ that would make the first round important enough to discourage the untrustworthy opponent from imitating the trustworthy one. To derive $\hat{\gamma}$, note that $P^U_2$ will mimic in the first round if and only if:

\[ \gamma c_2 + (1 - \gamma)b^U_2 > \gamma b^U_2 + (1 - \gamma)(c_2' - r_2') \]  

(1) simply states that $P^T_2$ must value the opportunity to betray in the second round enough (given $P_1$’s choice of $\gamma$) to give up on the opportunity to betray the initiator in the first round. Rearranging the terms, we show:

\[ \hat{\gamma} = \frac{c_2 - b^U_2 - r_2}{2(c_2 - b^U_2) - r_2} \]  

(2)

Given $\hat{\gamma}$ (the incentive for $P^U_2$ 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 $\bar{\alpha}_{\text{screening}} < a$. To solve for the threshold value of trust $\bar{\alpha}_{\text{screening}}$ – 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:

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

(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 $\gamma$ 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 $a'(s_2^*, \hat{\gamma})$ simplifies to $a$ for the RHS of the inequality, because $P_1$ expects in the “pre-diplomacy” stage where she sets $\gamma$ that the opponent will turn out to be untrustworthy with probability $a$. Given the discussion above, we can rewrite (2) as (3):

$$c_1 - r_1 < \hat{\gamma}[ac_1 - (1 - a)d_1] + (1 - \hat{\gamma})[ac'_1 + (1 - a)(c'_1 - r'_1)]$$  (4)

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

$$a \geq \bar{a}_{screening} \equiv 1 - \frac{r_1}{\hat{\gamma}(d_1 - r_1) + r_1}$$  (5)

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

$$\gamma c_1 + (1 - \gamma)(ac'_1 - (1 - a)d'_1) < \gamma[ac_1 - (1 - a)d_1] + (1 - \gamma)[ac'_1 + (1 - a)(c'_1 - r'_1)]$$  (6)

The right hand side (LHS) of (6) is the payoff associated with private cooperation for $P_1$ when secret diplomacy facilitates temporary collusion in round 2 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 $\gamma$. 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 $a'(s_2^*, \gamma)$ is simply $a$ because no learning happens in round 1 under the collusion equilibrium.

Rearranging the terms, the maximum stake $\hat{\gamma}_{screening}$ that $P_1$ is willing to put on round 1 of the game to induce $P_2^U$ to leak early on is:
\[ \frac{c_1 - r_1 + d_1}{2c_1 - r_1 + 2d_1} \]  

(7)

For \( P_1 \) to opt for screening, \( \hat{\gamma} < \hat{\gamma}_{\text{screening}} \) must be true. In other words, \( P_1 \) must have enough 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. When \( \hat{\gamma} < \hat{\gamma}_{\text{screening}} \), \( P_1 \) will set \( \gamma \) as \( \hat{\gamma} \). Any \( \gamma \) 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 \( \gamma \) 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 \( b_{2U} \) 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 \( P_2^U \) and \( P_2^T \) 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, \( a'(s^*_2, \gamma^*) \geq a'_{\text{secret}} \) must hold, with \( a'_{\text{secret}} \) 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 \( a'_{\text{secret}} \), note that \( P_1 \) will opt for public diplomacy in round 2 if and only if:

\[ c'_1 - r'_1 < a'(s^*_2, \gamma^*)c'_1 - (1 - a'(s^*_2, \gamma^*))d'_1 \]  

(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:

\[ a'(s^*_2, \gamma^*) < a'_{\text{secret}} \equiv 1 - \frac{r'_1}{c'_1 + d'_1} \]  

(9)

Crucially, \( a'_{\text{secret}} \) 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 \( \tilde{a}_{\text{collusion}} \) necessary to sustain collusion, note that \( P_1 \) prefers to go public instead of colluding with \( P_2 \) if and only if:

\[
c_1 - r_1 > (\hat{\gamma} - \epsilon)c_1 + (1 - \hat{\gamma} + \epsilon)(\alpha'(s_2^*, \hat{\gamma} - \epsilon)c'_1 + (1 - \alpha'(s_2^*, \hat{\gamma} - \epsilon))d'_1)
\]  

(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 \( \hat{\gamma} - \epsilon \), which is the optimal level of \( \gamma \) that \( P_1 \) would choose if he hopes to collude. Any \( \gamma^* \) higher than \( \hat{\gamma} - \epsilon \) will induce \( P_2^U \) to leak, while \( \gamma^* \) lower than \( \hat{\gamma} - \epsilon \) 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 \( 1 - \hat{\gamma} + \epsilon \). 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 \( \alpha'(s_2^*, \hat{\gamma} - \epsilon) \) simplifies to \( \alpha \) because no learning happened in round 1.

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

\[
c_1 - r_1 > (\hat{\gamma} - \epsilon)c_1 + (1 - \hat{\gamma} + \epsilon)(\alpha c'_1 + (1 - \alpha)d'_1)
\]  

(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:

\[
\alpha > \tilde{a}_{\text{collusion}} \equiv 1 - \frac{r_1}{(c_1 + d_1)(1 - \hat{\gamma} + \epsilon)} \approx 1 - \frac{r_1}{(c_1 + d_1)(1 - \hat{\gamma})}
\]  

(12)

Note that for \( P_1 \) to opt for collusion, \( \alpha > \tilde{a}_{\text{collusion}} \) 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, \( \alpha > \tilde{a}'_{\text{secret}} \)). Importantly, the level of trust necessary for \( P_1 \) to go private instead of public in the first round \( \tilde{a}_{\text{collusion}} \) 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 both as a useful tool that has the potential to facilitate cooperation without
reputation cost of public diplomacy in round 1, but also as a tool that can induce even \( P^U_2 \) to collude in round 1. In contrast, in round 2, secret diplomacy is solely a useful tool to facilitate cooperation without political consequences. Because \( \tilde{a}_{\text{secret}} > \tilde{a}_{\text{collusion}} \), \( \alpha > \tilde{a}_{\text{secret}} \) guarantees that \( \alpha > \tilde{a}_{\text{collusion}} \) must hold. In other words, we may focus on \( \alpha > \tilde{a}_{\text{secret}} \) as the necessary and sufficient constraint on \( P_1 \)'s level of trust that would sustain the collusion equilibrium.

Furthermore, note that \( \tilde{a}_{\text{secret}} > \tilde{a}_{\text{screening}} \), 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 \( \tilde{g}_{\text{screening}} \) relative to the severity of the adverse selection problem that \( P_1 \) faces \( \tilde{g} \). When \( \tilde{g}_{\text{screening}} \leq \tilde{g} \) and \( \alpha > \tilde{a}_{\text{secret}} \), 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 \( P^U_2 \) to collude in round 1 instead. In sum, both conditions \( \tilde{g}_{\text{screening}} \leq \tilde{g} \) and \( \alpha > \tilde{a}_{\text{secret}} \) are necessary to guarantee the existence of a collusion equilibrium.

**Proof for Proposition 3 (Public Diplomacy Equilibrium)**

The key to solving for the public diplomacy equilibrium lies in specifying the conditions under which \( P_1 \) would prefer to go public instead of going private to collude or screen. We start by examining the round 2 game. In round 2 of the game, we know that \( P^T_2 \) will always collude. In contrast, \( P^U_2 \) would always leak to reap the payoff \( b^U_2 \) because there is no political cost for doing so. On the other hand, if \( P_1 \) decides to go public in round 2, both \( P^T_2 \) and \( P^U_2 \) will settle for public cooperation since \( c_2 > a_2 \). Given the two adversaries’ optimal strategies, \( P_1 \) will opt for public diplomacy in round 2 if and only if:

\[
c'_1 - a'_1 > a'(s^*_2, \gamma^*)c'_1 - (1 - a'(s^*_2, \gamma^*))d'_1
\]  

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

\[
a'(s^*_2, \gamma^*) < \tilde{a}_{\text{secret}} \equiv 1 - \frac{r'_1}{c'_1 + d'_1}
\]  

This condition guarantees that \( P_1 \) will prefer either going public or screening instead of collusion. This is because, as discussed earlier, \( a'(s^*_2, \gamma^*) \equiv \alpha < \tilde{a}_{\text{secret}} \) must be true for \( P_2 \) to collude.

\[24\] Conversely, the inequality implies that the level of trust necessary to sustain screening may not be sufficient to sustain collusion.
In round 1 of the secret diplomacy sub-game, we know that $P^T_2$ will always cooperate in private, while $P^U_2$ 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 $\alpha < \tilde{\alpha}_{\text{screening}}$ (see proof for proposition 1). Importantly, note that $\gamma^*$ is 1, by construction, if $P_1$ opts for public diplomacy. This is condition (a) of proposition 3.

Note that $\alpha < \tilde{\alpha}_{\text{screening}}$ is a sufficient but not necessary condition for $P_1$ to go public. When $\tilde{\alpha}_{\text{screening}} < \alpha < \tilde{\alpha}_{\text{secret}}$ (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 ($\hat{\gamma}_{\text{screening}} < \hat{\gamma}$). 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 assume $c_1 > r_1$ and $c_2^2 > r_2$. When the two inequalities hold simultaneously, both the initiator and the adversary are guaranteed to receive a positive pay-off form public diplomacy when the initiator goes private. Alternatively, if one of the inequalities fails to hold, both the initiator and the adversary will receive no pay-off when the initiator goes public. In this section, we will analyze a model where public diplomacy fails. The results from our main model (propositions 1 - 3) remain unchanged for this “new” model. As in the case of our main model, there are three classes of PBE (screening, collusion, and public diplomacy) associated with this model. The expressions for the cutting-off points, however, will be slightly different for this version of the 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 $P^T_2$ will always collude. $P^U_2$, in contrast, would always leak in round 2 to reap the payoff $b^U_2$ because there is no political cost for doing so. On the other hand, if $P_1$ decides to go public in round 2, both $P^T_2$ and $P^U_2$ 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 $P^U_2$ does not mimic $P^T_2$ in this equilibrium, $\alpha'(\text{collude}, \gamma^*) = 1$ and $\alpha'(\text{leak}, \gamma^*) = 0$.

In the first round, if $P_1$ goes private, we know that $P^T_2$ will always cooperate in private while $P^U_2$ will cooperate in private when $\gamma^* > \hat{\gamma}$ and betray if $\gamma^* \leq \hat{\gamma}$. To derive $\hat{\gamma}$, note that $P^U_2$ will mimic in the first

25 Note that $\tilde{\alpha}_{\text{screening}} < \tilde{\alpha}_{\text{secret}}$ is always true if $\gamma \geq 1$. 
round if and only if:

\[ \gamma c_2 + (1 - \gamma)b^{U}\ell_2 > \gamma b^{U}\ell_2 + (1 - \gamma) \cdot 0 \]  

(15)

The second period pay-off facing \( P_2^{U} \) when he betrays the initiator in the first round is simply 0, because there is no public cooperation. Rearranging the terms, we show:

\[ \hat{\gamma} \equiv \frac{b^{U}\ell_2}{2b^{U}\ell_2 - c_2} \]  

(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 \( \tilde{a}_{\text{screening}} < a \). To solve for the threshold value of trust \( \tilde{a}_{\text{screening}} \) – 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:

\[ 0 < \hat{\gamma}[ac_1 - (1 - a)d_1] + (1 - \hat{\gamma})[a'(s_2^*, \hat{\gamma})c_1' + (1 - a'(s_2^*, \hat{\gamma})) \cdot 0] \]  

(17)

This expression simplifies to:

\[ 0 < \hat{\gamma}[ac_1 - (1 - a)d_1] + (1 - \hat{\gamma})(ac_1') \]  

(18)

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

\[ a \geq \tilde{a}_{\text{screening}} \equiv 1 - \frac{c_1}{\gamma d_1 + c_1} \]  

(19)

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

\[ \gamma c_1 + (1 - \gamma)(ac_1' - (1 - a)d_1') < \gamma[ac_1 - (1 - a)d_1] + (1 - \gamma)[ac_1' + (1 - a) \cdot 0] \]  

(20)

Rearranging the terms, the maximum stake \( \tilde{\gamma}_{\text{screening}} \) that \( P_1 \) is willing to put on round 1 of the game to induce \( P_2^{U} \) to leak early on is:
For $P_1$ to opt for screening, $\hat{\gamma} < \hat{g}_{\text{screening}}$ must hold.

Compared to the main model, the political cost of public cooperation $r_1$ disappears from the expressions defining $\hat{\gamma}$ and $\hat{g}_{\text{screening}}$. 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. Consequently, the comparative statics regarding reputation cost in round 2 will be void.

**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, $P_2^{LU}$ would always leak in the second round to reap the payoff $b^{LU}_{2}$ 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 $P_2^{LU}$ 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, $\alpha'(s^*_2, \gamma^*) \geq \tilde{\alpha}'_{\text{secret}}$ must hold, with $\tilde{\alpha}'_{\text{secret}}$ 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 $\tilde{\alpha}'_{\text{secret}}$, note that $P_1$ will opt for public diplomacy in round 2 if and only if:

$$0 < \alpha'(s^*_2, \gamma^*)c_1' - (1 - \alpha'(s^*_2, \gamma^*))d_1'$$

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

$$\alpha'(s^*_2, \gamma^*) < \tilde{\alpha}'_{\text{secret}} \equiv \frac{d_1'}{c_1' + d_1'}$$

Crucially, $\tilde{\alpha}'_{\text{secret}}$ 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 $P_2^{LU}$ 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 $\tilde{\alpha}'_{\text{collusion}}$ necessary to sustain collusion, note that $P_1$ prefers to go public
instead of colluding with $P_2$ if and only if:

$$0 > (\hat{\gamma} - \epsilon)c_1 + (1 - \hat{\gamma} + \epsilon)(\alpha c'_1 + (1 - \alpha)d'_1)$$

(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:

$$\alpha > \tilde{\alpha}_{\text{collusion}} \equiv \frac{-d_1 + (\gamma + \epsilon)(d_1 - c_1)}{c_1 - d_1 - (\gamma + \epsilon)(c_1 - d_1)}$$

(25)

Note that $\tilde{\alpha}_{\text{collusion}}$ is always strictly negative for this version of the model. In other words, $\tilde{\alpha}_{\text{collusion}}$ does not pose a constraint on the initiator. Therefore the only constraint that the initiator faces is $\tilde{\alpha}'_{\text{secret}} \equiv \frac{d'_1}{c'_1 + d'_1}$. When $\alpha > \tilde{\alpha}'_{\text{secret}}$, the initiator goes private. When $\alpha \leq \tilde{\alpha}'_{\text{secret}}$, the initiator goes public.

Furthermore, note that $\tilde{\alpha}'_{\text{secret}} > \tilde{\alpha}_{\text{screening}} \left( \frac{d'_1}{c'_1 + d'_1} > 1 - \frac{c_1}{\gamma d_1 + c_1} \right)$, 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 $\hat{\gamma}_{\text{screening}}$ relative to the severity of the adverse selection problem that $P_1$ faces $\hat{\gamma}$. When $\hat{\gamma}_{\text{screening}} \leq \hat{\gamma}$ and $\alpha > \tilde{\alpha}_{\text{secret}}$, 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 $P_2^{LI}$ to collude in round 1 instead. In sum, both conditions $\hat{\gamma}_{\text{screening}} \leq \hat{\gamma}$ and $\alpha > \tilde{\alpha}'_{\text{secret}}$ 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 $\tilde{\alpha}'_{\text{secret}}, \tilde{\alpha}_{\text{screening}}, \hat{\gamma}$, and $\hat{\gamma}_{\text{screening}}$ 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

---

26 Conversely, the inequality implies that the level of trust necessary to sustain screening may not be sufficient to sustain collusion.
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 “sweeten the pot” (Jervis 2015: 18, 21-22; quote at p. 16). 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 Jupiters 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’ 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 some trust in Khrushchev
not to leak and gained the ability to utilize that trust in the future.