Supporting Information

"Kompromat Can Align Incentives But Ruin Reputations"

March 9, 2021

FOR ONLINE PUBLICATION

Table of Contents

A Semi-separating equilibria in the cheap talk stage

In the main text we derive conditions for a fully separating equilibrium where the agent sends a message $m = s$ with probability 1. As in a typical cheap talk games, there are lots of other fully separating equilibria which are essentially equivalent, e.g., sending the message $m = 1-s$, since in equilibrium the principal knows how to interpret each message and hence infers the agent's signal. There is also always a babbling equilibrium where both types send the same message (or mix on multiple messages with the same probability).

To fully characterize the equilibria, we also need to characterize equilibria where the types *sometimes* but not always send the same message. In any non-babbling equilibrium, there must be two messages which induce a different policy choice. And any optimal policy choice for the principal must be between $\tilde{\theta}_0$ and $\tilde{\theta}_1$. So, the $s = 1$ type always has a strict preference for whichever message induces a higher policy. As a result, we only need to look for semi-separating equilibria where the $s = 1$ type always sends the same message (again call this $m = 1$) and the $s = 0$ type mixes between this and another message.

Let q_1 be the probability that the $s = 0$ type sends $m = 1$, and so they send $m = 0$ with probability $1 - q_1$. Upon observing $m = 0$ the principal knows the signal observed was $s = 0$ and so chooses $x = \tilde{\theta}_0$. Upon observing $s = 1$ the posterior belief the optimal policy is:

$$\begin{aligned} x^*(1) &= \text{Pr}(s = 0|m = 1)\tilde{\theta}_0 + \text{Pr}(s = 1|m = 1)\tilde{\theta}_1 \\ &= \frac{(1-\pi)q_1}{(1-\pi)q_1 + \pi}\tilde{\theta}_0 + \frac{\pi}{(1-\pi)q_1 + \pi}\tilde{\theta}_1 > \tilde{\theta}_0 \end{aligned}$$

where π is the prior probability that $s = 1$. For the $s = 0$ type to be indifferent, this choice must be above $\tilde{\theta}_0 + b$; if not $\tilde{\theta}_0 < x^*(1) < \tilde{\theta}_0 + b$, and since $\tilde{\theta}_0 + b$ is the ideal policy for the $s = 0$ type they would strictly prefer to send $m = 1$ and get policy $x^*(1)$ rather then sending $m = 0$ and getting $\tilde{\theta}_0$.

The condition to make the $s = 0$ type indifferent is then:

$$x^*(1) - (\tilde{\theta}_0 + b) = (\tilde{\theta}_0 + b) - \tilde{\theta}_0$$
$$x^*(1) = \tilde{\theta}_0 + 2b$$

Solving for q_1 gives:

$$q_1 = \frac{\pi(2b + \tilde{\theta}_1 - \tilde{\theta}_0)}{2b(1 - \pi)}$$

For this to be a valid mixed strategy, this q_1 must lie between 0 and 1, which is true when:

$$b \in \left(\pi \frac{\tilde{\theta}_1 - \tilde{\theta}_0}{2}, \frac{\tilde{\theta}_1 - \tilde{\theta}_0}{2} \right)$$

Note the upper bound is the same upper bound for a truthful equilibrium. So, whenever this equilibrium exists, a truthful equilibrium exists as well. Further, it is easy to verify that, compared to the semi-separating equilibrium the truthful equilibrium gives a strictly higher payoff to the principal (who gets more information in the truthful equilibrium and hence makes better decisions) and the $s = 1$ type (since a higher policy is chosen upon observing $m = 1$ in the truthful equilibrium), while it gives the same payoff to the $s = 0$ type (as he gets the same payoff when sending $m = 0$). Since the semi-separating equilibrium holds less often and makes both players worse in expectation, we are comfortable ignoring it.

B Scope of participation in the accidental leaks model

We briefly discuss how the exogenous parameters of the accidental leaks model affect the scope for a productive relationship between the principal and the agent.

Comparative statics Several comparative statics for the equilibrium are natural. First, the truth-telling constraint is harder to meet when the agent is more biased (b increases). Second, the participation constraint is easier to meet when D is large, meaning there is more excess benefit from entering a relationship relative to the outside option. Third, as ν decreases, both conditions are easier to meet. This is because when leaks are likely, the threat to *purposefully* leak becomes less salient, and the cost of employment increases. As $\nu \to 0$, the participation constraint is always met, and so there is always some degree of kompromat (κ sufficiently high) which will allow for truthful employment for an agent of any bias, b.

Less obvious, as the organization's capacity C increases, this has an ambiguous effect on truthfulness if $b < 1$. In this case, as C increases, it is harder to satisfy both the constraints if $C < b/2$ but easier to satisfy them if $C > b/2$. This nonmonotonicity reflects two competing forces. As discussed in the text, as C increases, it is easier to sustain truth-telling even without kompromat. However, increasing C also increases the importance of making a good policy choice relative to the cost of kompromat. Combining, this result suggests that kompromat may be most effective in organizations with very low or very high capacity to collect information. However, when $b \geq 1$, kompromat is only effective at inducing truth-telling for low levels of capacity.

Trade-off between wages and kompromat As long as $\frac{D}{\nu} > \frac{2(b-C/2)C}{1-\nu}$, there are multiple truthful levels of kompromat (κ) and wages (w) which allow for truthful employment. We conclude this section by briefly analyzing the relationship between these two quantities. Recall from above that when kompromat makes it possible for the agent to be appointed to the organization, the range of acceptable wages is:

$$\overline{w}_A + \nu\kappa \leq w \leq \overline{w}_P$$

Suppose that there exists a range of κ that meets the conditions for a truthful equilibrium where the agent is appointed to the organization. Then when there is more kompromat (higher κ), then this "shifts up" the lower end of this range, meaning only higher wages will be acceptable for the agent. This is almost certainly bad for the principal, since it means she will need to offer a higher wage to offset the negative effects of kompromat. It is also likely bad for the agent, since unless the wage increases at a one-to-one rate with $\nu\kappa$, his total utility from working for the organization decreases. This suggests that if we were to explicitly model the wage bargaining process (holding fixed a level of bias $b > C/2$), the principal would like a level of kompromat κ such that the truth-telling constraint just binds.

C The imperfect monitoring model

Consider a modified version of the accidental leaking model with the following order of moves:

1. P chooses whether to offer an appointment ($a_P = 1$) or not ($a_P = 0$), and if $a_P = 1$ the agent can accept ($a_A = 1$) or not ($a_A = 0$). $a_A \in \{0, 1\}$, respectively. If $a_P = 0$ or $a_A = 0$, the game ends with reservation utilities $(\overline{u}_A, \overline{u}_P)$. If $a_A = a_P = 1$, then:
2. A privately observes signal $s \in \{0, 1\}$, and gives advice in the form of a message $m \in \{0, 1\}$
3. P observes m and chooses a policy $x \in \mathbb{R}$
4. P observes a validation signal $s_v \in \{0, 1\}$, where $s_v = s$ with probability $1 - \varepsilon \in (1/2, 1)$ and $s_v \neq s$ with probability ε, and chooses how much of whether to leak the kompromat $l \in \{0, 1\}$.
5. Payoffs realized.

So, compared with the model in the main text, we remove the possibility of accidental leaks, but now the principal is never certain if the agent has lied.

In this variant it will also be useful to allow for mixed strategies in leaking kompromat. Let $\lambda(m, s_v)$ be the probability of leaking kompromat, as a function of the agent's advice m and the validation signal s_v. Equivalently, one can think of releasing kompromat as continuous choice, and λ refers to the proportion released. Such a formulation leads to the same expected utilities, and hence the analysis is identical.

There are several classes of equilibria we could study. First, the amount of kompromat the principal leaks could be independent of whether she suspects the agent of lying (i.e., independent of her validation signal). In this case, the agent's incentives to be truthful are no different than in the baseline model above since the same amount of kompromat is leaked regardless of whether she is truthful. However, the fact that there is now kompromat that can be leaked will make it more costly for the agent to accept the principal's appointment to the organization.

In a more interesting class of equilibria, kompromat can provide additional incentives for truthfulness. In these equilibria, the principal leaks more or less kompromat depending on whether she suspects that the agent lied. A natural equilibrium of this form would be to choose to leak no kompromat when the principal suspects the agent told the truth (i.e., $\lambda = 0$ when $m = v$) and to leak some amount of kompromat when the principal suspects the agent lied (i.e., $\lambda > 0$ when $m \neq v$). This kind of strategy provides a stronger incentive for the agent to provide truthful advice.

Recall that the only situation in which the agent has an incentive to lie is when he observes $s = 0$. While threatening to leak kompromat when $m = 0$ and $s_v = 1$ would give even stronger incentives for an agent observing $s = 1$ to truthfully report $m = 1$, he doesn't have an incentive to lie in the first place. So, to start, we will focus on equilibria in which $\lambda(m = 1, s_v = 0) > 0$ and $\lambda(m, s_v) = 0$ otherwise. With some abuse of notation, we will denote $\lambda(m = 1, s_v = 0)$

by λ. After deriving our main result for this class of equilibria, we present an analysis of other potential leaking strategies, showing that any equilibrium with leaking in another contingency is Pareto dominated by an equilibrium where leaks only happen after $m = 1$ and $s_v = 0$.[17]

In the kind of equilibrium we consider, an agent observing $s = 0$ faces no chance of a leak if he tells the truth since the principal does not leak whenever $m = 0$. Note that this is true even if the validation signal disagrees with the message ($m = 0$ and $s_v = 1$). If he lies and reports $m_0 = 1$, then the agent's lie will be revealed by the principal's validation signal with probability $1 - \varepsilon$. So, the expected cost of the leak to the agent is $(1 - \varepsilon)\lambda\kappa$. The policy payoffs are the same as in the baseline, so the truth-telling constraint when $s = 0$ is now:

$$-(\tilde{\theta}_0 - (\tilde{\theta}_0 + b))^2 \geq -(\tilde{\theta}_1 - (\tilde{\theta}_0 + b))^2 - (1 - \varepsilon)\lambda\kappa$$

This reduces to

$$b \leq \frac{C}{2} + \left(\frac{1 - \varepsilon}{2C}\right)\lambda\kappa \tag{6}$$

The analysis of the participation constraint for the principal is the same as above: he will hire the agent if $w \leq \overline{w}_P$. For the agent, the difference is that kompromat will leak even if he is truthful not with an exogenous probability, but with probability λ when (1) he gets a signal of $s = 1$, (2) the validation signal is incorrect. So, the expected cost from kompromat is $\pi\varepsilon\lambda\kappa$, and the agent participation constraint is:

$$w \geq \overline{w}_A + \pi\varepsilon\lambda\kappa \tag{7}$$

17. Alternatively, if the utility to the principal is affected by the leak, our analysis is analogous to what we would get if assuming she can commit to a kompromat leaking strategy, and chooses the leaking strategy which most effectively provides incentives for truth-telling.

Combining, there is a wage both parties find acceptable if and only if:

$$\pi\varepsilon\lambda\kappa \leq \overline{w}_P - \overline{w}_A \tag{8}$$

Whether both constraints are satisfied depends on whether there is an amount of leaking $\lambda \in [0,1]$ where both conditions hold. Let $B \equiv b - \frac{C}{2}$ be the "excess bias" which kompromat may be able to solve, and let $D \equiv \overline{w}_P - \overline{w}_A$ as in the main text. Then:

> **Proposition A3.** There exists an equilibrium with truthful employment in the imperfect monitoring model if and only if
>
> $$\kappa \geq \frac{2BC}{1-\varepsilon} \tag{9}$$
>
> and
>
> $$D \geq \frac{2B\varepsilon\pi C}{1-2\varepsilon}. \tag{10}$$

Proof of Proposition A3. It is sequentially rational to tell the truth in equilibrium if and only if (TC1) is satisfied. Substituting and rearranging gives

$$\lambda\kappa \geq \frac{2B(\tilde{\theta}_1 - \tilde{\theta}_0)}{1-\varepsilon} \tag{11}$$

Conditional on truth-telling in equilibrium, it is sequentially rational for employment to occur if (PC1) is satisfied. Substituting and rearranging gives

$$\lambda\kappa \leq \frac{D}{\pi\varepsilon} \tag{12}$$

Both constraints can be satisfied if each condition holds for some $\lambda \in [0,1]$ and the range of $\lambda\kappa$ implied by the two inequalities exists.

First, note that since the right hand side of (12) is strictly positive, there always exists a λ such that (12) holds. Second, note that (11) only holds for λ sufficiently high since the right hand side is strictly positive. Then an equilibrium with truth-telling requires that

$$\kappa \geq \frac{2B(\tilde{\theta}_1 - \tilde{\theta}_0)}{1-\varepsilon}$$

Third, combining these two inequalities gives

$$\frac{2B(\tilde{\theta}_1 - \tilde{\theta}_0)}{1-\varepsilon} \leq \lambda\kappa \leq \frac{D}{\pi\varepsilon}$$

For there to be $\lambda\kappa$ where this holds, the lower bound must be higher than the upper bound, which becomes:

$$D \geq \frac{2B\varepsilon\pi(C)}{1-2\varepsilon}.$$

□

The comparative statics on this equilibrium are similar to the accidental leaking model. It is easier to sustain this equilibrium when ε is low, as this makes it easier to threaten agents from lying and lowers the probability that kompromat is leaked.

Next, we show that the restriction to the particular leaking strategy only rules out Pareto dominated equilibria:

> **Proposition A4.** Suppose there is no truthful equilibrium with no leaking, $b > C/2$, and there exist truthful equilibria with a leaking strategy of the form:
>
> $$\lambda(m, s_v) = \begin{cases} \lambda^* & m = 1 \text{ and } s_v = 0 \\ 0 & \text{otherwise} \end{cases}.$$
>
> Then the truthful equilibrium of this form with the least leaking possible, where

$\lambda^* = \lambda^{\min}$ is characterized by:

$$b \leq \frac{\tilde{\theta}_1 - \tilde{\theta}_0}{2} + \left(\frac{1-\varepsilon}{2(\tilde{\theta}_1 - \tilde{\theta}_0)}\right)\lambda^{\min}\kappa$$

Pareto dominates any truthful equilibrium with a different leaking strategy.

Proof of Proposition A4. Recall the principal can use any proportion of kompromat to leak as a function of the message and validation: $\lambda(m, s_v)$. So, there are four components to the leaking strategy, based on the message ($m = 0$ or $m = 1$) and validation ($s_v = 0$ or $s_v = 1$).

Assuming there is a truthful equilibrium with employment, (which we will check for subsequently), the expected value utility to the agent is:

$$w - y_A^T - L\kappa$$

where

$$L \equiv \pi(\varepsilon\lambda(1,0) + (1-\varepsilon)\lambda(1,1)) - (1-\pi)(\varepsilon\lambda(0,1) + (1-\varepsilon)(\lambda(0,0))$$

L is increasing in all of the $\lambda(m, s_v)$ terms. Since the agent utility is strictly decreasing in L, and the principal utility is not (directly) a function of L, if there are two truthful equilibria with expected leaking levels L^1 and $L^2 > L^1$, the equilibrium with L^1 Pareto dominates the equilibrium with L^2.

Given a leaking strategy, the condition for reporting truthfully when $s = 1$ is:

$$\begin{aligned} &-(\tilde{\theta}_1 - (\tilde{\theta}_1 + b))^2 - (\varepsilon\lambda(1,0) + (1-\varepsilon)\lambda(1,1))\kappa \\ &\quad \geq -(\tilde{\theta}_0 - (\tilde{\theta}_1 + b))^2 - (\varepsilon\lambda(0,0) + (1-\varepsilon)\lambda(0,1))\kappa \end{aligned}$$

and when $s = 0$:

$$\begin{aligned} &-(\tilde{\theta}_0 - (\tilde{\theta}_0 + b))^2 - (\varepsilon\lambda(0,1) + (1-\varepsilon)\lambda(0,0))\kappa \\ &\quad \geq -(\tilde{\theta}_1 - (\tilde{\theta}_0 + b))^2 - (\varepsilon\lambda(1,1) + (1-\varepsilon)\lambda(1,0))\kappa \end{aligned}$$

The $s = 1$ in equality is met with no leaking at all. However, since $b > C/2$, the $s = 0$ inequality does not hold without leaking. To make this equation hold while minimizing the loss from leaking, it is immediate that, as long as there is a truthful equilibrium where $\lambda(0, s_v) = 0$, any leaking strategy where the $s = 1$ constraint is met with $\lambda(0, s_v) > 0$ is Pareto dominated by a truthful equilibrium where $\lambda(0, s_v) = 0$.

Next, recall that in the statement of the proof, we assume the existence of a truthful equilibrium where leaking only happens in the $m = 1$, $s_v = 0$ information set, and at level $\lambda^{\min}$. Call this leaking strategy A:

$$\lambda^A(m, s_v) = \lambda^{\min}\mathbf{1}_{m=1,s_v=0}.$$

Since this is the lowest amount of leaking (in this information set) possible, it must meet:

$$(\tilde{\theta}_1 - (\tilde{\theta}_0 + b))^2 - b^2 = (1-\varepsilon)\lambda^A(1,0)\kappa$$

and any leaking strategy (B) where $\lambda^B(1,1) > 0$ (but does not lead to more of a collective level of leaking than necessary) must meet:

$$(\tilde{\theta}_1 - (\tilde{\theta}_0 + b))^2 - b^2 = ((1-\varepsilon)\lambda^B(1,0) + \varepsilon\lambda^B(1,1))\kappa$$

Combining gives:

$$(1-\varepsilon)\lambda^A(1,0) = ((1-\varepsilon)\lambda^B(1,0) + \varepsilon\lambda^B(1,1))$$
$$\frac{\varepsilon}{1-\varepsilon}\lambda^B(1,1) = (\lambda^A(1,0) - \lambda^B(1,0))$$

The difference in the losses associated with leaking strategy A versus B is:

$$\begin{aligned} L^A - L^B &= \pi\varepsilon\lambda^A(1,0) - (\pi\varepsilon\lambda^B(1,0) + \pi(1-\varepsilon)\lambda^B(1,1)) \\ &= \pi(\varepsilon(\lambda^A(1,0) - \lambda^B(1,0)) - (1-\varepsilon)\lambda^B(1,1)) \\ &= \pi\left(\varepsilon\frac{\varepsilon}{1-\varepsilon}\lambda^B(1,1) - (1-\varepsilon)\lambda^B(1,1)\right) \\ &= \left(\left(\varepsilon\frac{\varepsilon}{1-\varepsilon} - (1-\varepsilon)\right)\pi\lambda^B(1,1)\right) < 0 \end{aligned}$$

where the final inequality follows from:

$$\varepsilon\frac{\varepsilon}{1-\varepsilon} < \varepsilon < 1-\varepsilon$$

since $\varepsilon < 1/2$. So, any leaking strategy where $\lambda(1,1) > 0$ is Pareto dominated by a leaking strategy where λ is only strictly positive for $m = 1$ and $s_v = 0$. And, by the argument above, this implies that any other truthful equilibrium (i.e., those with $\lambda(0, s_v) > 0$ or $\lambda(1,1) > 0$) is Pareto dominated by the truthful equilibrium with leaking strategy $\lambda^A(m, s_v)$. □

D The bilateral kompromat model

Here is alternative model with a symmetric relationship between two actors, where kompromat may improve their ability to cooperate.

Two agents simultaneously make an effort choice e_i, observe the other agent's effort choice, and then choose whether to leak kompromat on the other, where $l_i = 1$ means i leaks on j. If i does not choose to leak on j, the kompromat will still leak "by accident" with probability $\nu \geq 0$. The utilities are:

$$u_i(e_i, e_{-i}) = e_i + e_{-i} - \frac{c}{2}e_i^2 - (l_{-i} + (1 - l_{-i})\nu)\kappa$$

With no leaking (which is always sequentially rational), the optimal effort choices solve the first order condition:

$$1 - ce_i = 0$$

so the optimal effort choice is $e_i^* = 1/c$.

The effort choice which would maximize the players collective utility (while not sequentially rational) is the e which maximizes:

$$u_i(e, e) = 2e - \frac{c}{2}e^2 - (l_{-i} + (1 - l_{-i})\nu)\kappa$$

which is $e^{\text{opt}} = 2/c$.

A natural equilibrium to search for is one where both players threaten to leak kompromat on each other if their effort choice does not meet (or exceed) some threshold $\hat{e} \in (1/c, 2/c]$. That is, both use a common leaking strategy $l_i = \mathbf{1}_{e_{-i}<\hat{e}}$. This leaking choice is always sequentially rational, so all that remains to check what levels of effort the actors will pick given the behavior at

this stage. The utility for making effort choice e_i is now:

$$u_i(e_i, e_{-i}; \hat{e}) = \begin{cases} e_i + e_{-i} - \frac{c}{2}e_i^2 - \kappa & e_i < \hat{e} \\ e_i + e_{-i} - \frac{c}{2}e_i^2 - \nu\kappa & e_i \geq \hat{e} \end{cases}$$

Since $\hat{e} \in (1/c, 2/c]$, the optimal choice on the first segment is $e^* = 1/c$, and the optimal choice on the second segment is $\hat{e}$. So they will both exert enough effort to avoid purposeful leaking if and only if:

$$\begin{aligned} 2\hat{e} - \frac{c}{2}\hat{e}^2 - \nu\kappa &\geq \frac{1}{c} + \hat{e} - \frac{c}{2}\left(\frac{1}{c}\right)^2 - \kappa \\ (1-\nu)\kappa &\geq \frac{1}{c} - \hat{e} - \frac{c}{2}\left(\left(\frac{1}{c}\right)^2 - \hat{e}^2\right) \end{aligned} \tag{13}$$

As $\hat{e} \to 1/c$, the right hand side goes to zero and this condition always holds. So, it is always possible to get a bit more effort with the promise of kompromat, which increases both agents' utilities. The optimizing value of $2/c$ is possible if:

$$(1-\nu)\kappa \geq \frac{1}{2c} \tag{14}$$

If this inequality is not met, then there is a maximal sustainable level of effort $e^{\max} \in (1/c, 2/c)$ which meets (13). By a bit of rearranging, this is given by:

$$e^{\max} = \frac{1}{c} + \sqrt{\frac{2(1-\nu)\kappa}{c}} \tag{15}$$

"Optimal kompromat" Is kompromat good for the agents in this model? More precisely, as κ increases, does this improve their equilibrium utility? There is a clear trade-off here, where higher values of κ make it possible to use the threat of leaks to attain a mutually better effort level, but even

when exerting this superior effort level the kompromat might still leak. An immediate consequence of (15) is that, at $\kappa = 0$, $\frac{\partial e^{\max}}{\partial \kappa} = \infty$. That is, the marginal impact of a bit of kompromat on effort is very large. (This is not particular to the functional forms used, as getting the agents to pick an effort level slightly higher than $1/c$ entails little loss on the margin, since the first derivative of the utility at this point is zero.) Since the derivative of the (symmetric) equilibrium utility with respect to $\hat{e}$ is strictly positive at $\hat{e} = 1/c$, both agents are always made better of with a little kompromat. However, there are diminishing returns to inducing more effort, and eventually there is enough kompromat to induce the optimal effort, at which point adding more kompromat is all cost and no benefit. Further, since there is little equilibrium loss to moving to an effort level slightly below $2/c$, the best level of kompromat is always less than that which leads to the optimal effort level (for a fixed level of kompromat).

Formally, rearranging equation (14), the level of kompromat required to induce the full effort level of $2/c$ is:

$$\kappa^{\text{full}} = \frac{1}{2c(1-\nu)} \tag{16}$$

Then:

> **Proposition A5.** If the agents pick the equilibrium which induces the optimal effort level, the amount of kompromat which maximizes their equilibrium utility (κ^*) is
> (i) between 0 and κ^{full},
> (ii) increasing in ν and decreasing in c.

Proof. The agent's equilibrium utility when putting in the maximal effort for a given level of kompromat equation (15) is:

$$\frac{3}{2c} - \kappa - \sqrt{2\kappa + \frac{1-\nu}{2c}} \tag{17}$$

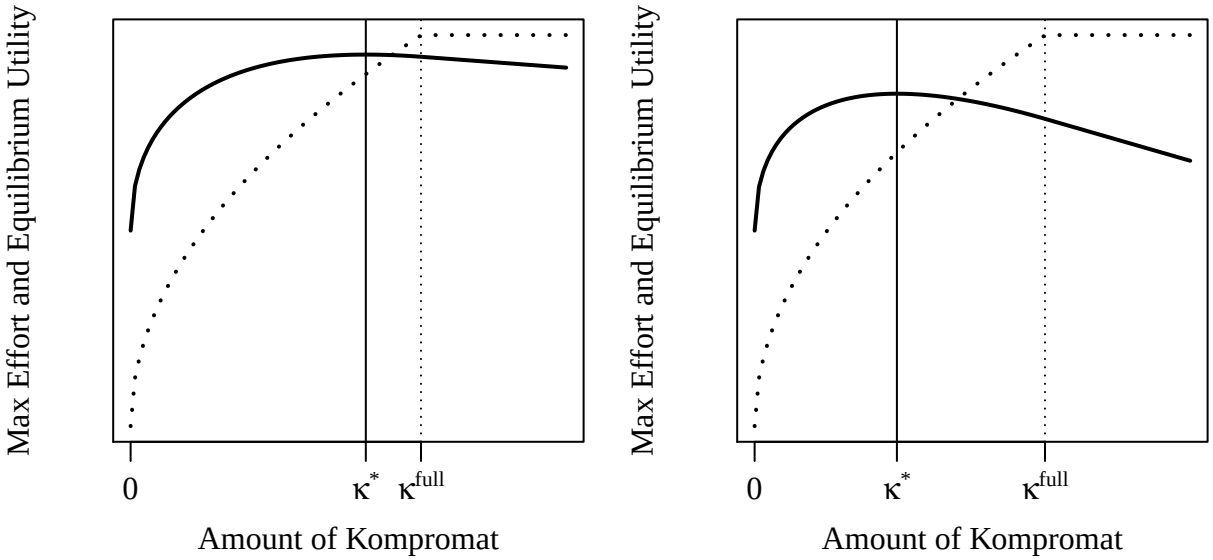

Figure D1: *Maximal effort and equilibrium utility as a function of the amount of kompromat* κ

The value of κ which maximizes this expression is:

$$\kappa^* = \frac{(1-\nu)}{2c} \tag{18}$$

Since $(1-\nu) < 1/(1-\nu)$ for $\nu \in (0,1)$, this is strictly less than κ^{full}, and it is decreasing in ν and c. □

Figure D1 illustrates. In each panel, the dotted line is the maximum amount of effort sustainable as a function of the amount of kompromat. The solid curve is the agents equilibrium utility when playing the strategies that lead to the optimal effort. In the left panel, leaks are rare ($\nu = 0.1$), and the best level of kompromat is just a bit below that which maximizes effort. In the right panel, leaks are more common ($\nu = 0.3$), and as a result the best level of kompromat involves only about half of the possible additional effort.

E Proof of Proposition 2

Proof. Part (i) immediately follows from the analysis in the text, as does the existence of the equilibria described in parts (ii)-(iii). What remains is to show that these are the equilibria (in the appropriate part of the parameter space) which maximize the probability of employment.

Write the probability of employment as:

$$P \equiv q \Pr(a = 1|\kappa = \kappa_H) + (1 - q) \Pr(a = 1|\kappa = \kappa_L)$$

Recall that there is no equilibrium where a low corruption type with $b > C/2$ is hired. And, in any equilibrium where those low corruption types with $b \leq C/2$ are not hired with positive probability, there is also an equilibrium where all of these types enter employment and report truthfully. This is because (1) reporting honestly is incentive compatible for these types, and (2) if more κ_L types enter employment, then this decreases $\tilde{q}_a$, and so if the participation constraints are met in the proposed equilibrium, they are also met for the modified equilibrium where all of the low-bias non-corrupt types are employed. So, no equilibrium without the proposed employment for low corruption types can lead to a higher probability of employment.

Given this employment strategy for the low corruption types, the final step is to show that no equilibrium where the corrupt type entry is nonmonotone can lead to a higher probability of employment than the identified monotone equilibrium.

Take any (potentially non-monotone) equilibrium where the non-corrupt types enter with probability one when $b \leq C/2$ and with probability zero otherwise. Let $b^{\max}$ be the highest-bias corrupt type who gets employed:

$$b^{\max} = \sup\{b|a^*(b, \kappa_H) = 1\}$$

and let $p_H = \Pr(a = 1|\kappa = \kappa_H)$ be the probability that a high-corruption type gets employed. In such an equilibrium:

$$\tilde{q}_a = \frac{qp_H}{qp_H + (1-q)F(C/2)}$$

and the truth-telling constraint is:

$$b^{\max} \leq \frac{C}{2} + \frac{\lambda r(1-\tilde{q}_a)}{2C}$$

However, there is always a monotone equilibrium where the high corruption types enter if and only if $b \leq \hat{b} \leq b^{\max}$, where $\hat{b}$ solves

$$\frac{qp_H}{qp_H + (1-q)F(C/2)} = \frac{qF(\hat{b})}{qF(\hat{b}) + (1-q)F(C/2)}$$

which generates the same $\tilde{q}_a$, and hence the participation constraint is met. Further, since $\hat{b} \leq b^{\max}$, the truth-telling constraint is met for all who get appointed. So, for any non-monotone equilibrium, we can find a monotone equilibrium with the same probability of employment. And the monotone equilibrium identified in the statement of the proposition is the one with the highest $\hat{b}$ (and hence probability of employment) where the participation constraint is not violated, and so it maximizes the probability of employment. □

F Accidental leaks in the reputation model

In our main analysis of the reputation model, we removed the possibility of accidental leaks. We do so in order to focus more directly on the reputation mechanism, which is our core focus. However, we now briefly consider what happens if we reintroduce the possibility of accidental

leaks. Assume that between steps 5 and 6 of the reputation model, there is an accidental leak of kompromat (if kompromat exists) with probability ν. Given this possibility, the truth-telling constraint must be revised as follows:

$$-(\tilde{\theta}_0 - (\tilde{\theta}_0 + b))^2 - r(\nu + (1-\nu)\tilde{q}_a) - V \geq -(\tilde{\theta}_1 - (\tilde{\theta}_0 + b))^2 - r(\lambda + (1-\lambda)\nu + (1-\lambda)(1-\nu)\tilde{q}_a) - V$$

This simplifies to:

$$b \leq \frac{C}{2} + \frac{\lambda(1-\nu)r(1-\tilde{q}_a)}{2C} \qquad \text{(TC2}')$$

The only difference between (TC2) in the main text and (TC2′) is the $(1-\nu)$ term. Moreover, given that accidental leaks are exogenous, the rest of the analysis can proceed as in the main text, substituting $\lambda(1-\nu)$ for λ. As before, the principal's *endogenous* leaks expand the scope for truth-telling by corrupt agents. However, the *exogenous* ("accidental") leaks shrink the scope for truth-telling. Accidental leaks make the principal's leaking strategy a less effective tool of control because it increases the probability of a leak after truth-telling *and* after lying.

G Unobserved bias at hiring

In our models of kompromat, we assume that the principal observes the agent's bias b, both at hiring and when working with the agent. We do so in order to focus our attention on the core agency problem, namely the potential conflict of interest between the principal and the agent.

The latter assumption that the principal observes the agent's bias when working with the agent is standard in the literature. We follow Crawford and Sobel (1982), as well as most of the papers

that build on it (as does ours), wherein the principal knows the bias of the agent, but not the information she possesses. This is reasonable: while we model a one-shot communication game (again, a standard simplifying assumption), these employment relationships last a long time, giving the principal time to learn about the agent preferences.

That the principal knows the bias at hiring is less standard, so we now briefly consider what happens if we relax it. Formally, suppose the principal a belief that b is distributed according to the cdf F described in the main text. After hiring, assume b is revealed and the analysis proceeds as in the main text. So, what changes with unknown b at hiring is the players' decision calculus around participation. We consider each player in turn.

Principal Let $\tilde{\beta}_\kappa$ be P's belief that b is such that TC holds (for a given κ). In this accidental leaks model this is $\Pr(b \leq \frac{C}{2} + \frac{1-\nu}{2C}\kappa)$ and in the reputation model this is $\Pr(b \leq \hat{b}_J)$ for $J \in \{H, L\}$. Given this uncertainty, the principal's participation constraint is:

$$-w - \tilde{\beta}_\kappa y_P^T - (1 - \tilde{\beta}_\kappa) y_P^B \geq -\overline{w}_P - y_P^T$$

$$w \leq \underbrace{\overline{w}_P - (1 - \tilde{\beta}_\kappa)(y_P^B - y_P^T)}_{<\overline{w}_P} \equiv \widehat{w}_P$$

Notice that $\widehat{w}_P < \overline{w}_P$, indicating that the principal's uncertainty over the agent's bias reduces her incentive to hire a given agent.

Agent We now consider the agent's participation constraint. Now that the principal is uncertain over b, it is possible that she hires an agent who would lie to her. This was not possible in our main models. So, we must consider the incentives of agents who would lie if hired and those who would be truthful if hired. For an agent for whom (TC1) or (TC2) is satisfied, the participation constraint is as in the main text.

For an agent for whom (TC1) does not hold in the model with accidental leaks, the participation constraint becomes

$$w - y_A^B - \kappa \geq \overline{w}_A - y_A^T \iff w \geq \underbrace{\overline{w}_A + \kappa + (y_A^B - y_A^T)}_{>\overline{w}_A + \nu\kappa} \equiv \widehat{w}_A^{\text{acc}}$$

And for an agent for whom (TC2) does not hold in the reputation model, the participation constraint becomes

$$\begin{aligned} w - y_A^B - (\lambda + (1-\lambda)\tilde{q}_a(\lambda))r \geq \overline{w}_A - y_A^T - r\overline{q} \\ \iff w \geq \underbrace{\overline{w}_A + (y_A^B - y_A^T) + r[(\lambda + (1-\lambda)\tilde{q}_a(\lambda)) - \overline{q}]}_{>\overline{w}_A + r(\tilde{q}_a(\lambda) - \overline{q})} \equiv \widehat{w}_A^{\text{rep}} \end{aligned}$$

Notice that $\widehat{w}_A^{\text{acc}} > \overline{w}_A + \nu\kappa$ and $\widehat{w}_A^{\text{rep}} > \overline{w}_A + r(\tilde{q}_a(\lambda) - \overline{q})$. This is because an agent who lies in equilibrium is more likely to have his kompromat revealed.

Taken together, the principal's uncertainty over the agent's bias shrinks the range of mutually agreeable wages, and makes its harder to sustain *any* relationship between the principal and agent. However, this observation does not alter the substantive take-away of our models. Moreover, since truth-telling agents and lying agents demand different wages, it is possible for the principal to screen for a truth-telling agent. In particular, if we were to explicitly model the wage bargaining process, the principal could offer a wage that would make a truth-telling agent accept the job but a lying agent reject the job. For example, in the accidental leaks model, as long as there is some possibility of a mutually acceptable wage, then a wage-setting principal could offer $w = \overline{w}_A + \nu\kappa < \widehat{w}_A^{\text{acc}}$ and induce *only* truth-telling agents to join the organization.