0.1. Proof of Lemma 1

Proof. For any covariate $a \neq a^*$, we have $\mathbb{P}[X_n^a = Y_n \mid \hat{X}_n^a = Y_n] = 1 - q$ and $\mathbb{P}[X_n^a = Y_n \mid \hat{X}_n^a \neq Y_n] = q$ for each observation $n$. Since $1 - q > q$, the distribution of the number of observations that match the outcome in the raw data for a covariate $\hat{a}$ with $\hat{X}_n^a = Y_n$ for every $n$ first-order stochastically dominates that of any other covariate. Reporting such a covariate thus maximizes the chance of passing the policymaker’s test, for any $\gamma$. Since the hacker will never find $a^*$, he will only maximize this passing chance. Similarly, among those covariates $\hat{a} \notin \{a^*, a'\}$, the maven’s expected payoff is maximized by covariates $\hat{a}$ with $\hat{X}_n^a = Y_n$ for every $n$.

It remains to show how the maven optimally chooses between $a^*$ and $a'$ based on his information. Suppose the maven learns the true cause and the red herring are in the set $\{a', a''\}$, where $\hat{X}^{a'}$ matches $Y$ in $k_1$ observations and $\hat{X}^{a''}$ matches $Y$ in $k_2$ observations with $k_1 < k_2$.

If $a^* = a'$, the maven’s data has likelihood $\tilde{p}^{k_1}(1 - \tilde{p})^{N - k_1} \cdot (1/2)^N$ where $\tilde{p} = \psi(1 - q) + (1 - \psi)q > 1/2$. This is because $\mathbb{P}[\hat{X}_n^{a*} = Y_n] = \tilde{p}$ for every observation $n$, while $\mathbb{P}[\hat{X}_n^{a'} = Y_n] = 1/2$ for every observation $n$. If $a^* = a''$, then the data likelihood is $\tilde{p}^{k_2}(1 - \tilde{p})^{N - k_2} \cdot (1/2)^N$ by the same reasoning. This second likelihood is larger, because $k_2 > k_1$ and $\tilde{p} > 1/2$. So the covariate that matches the outcome in more observations has a higher posterior probability of being the true cause.

Let $a \in \{a', a''\}$ and consider the probability that $a$ passes the policymaker’s test. Conditional on $a = a^*$, if $\hat{X}_n^a = Y_n$, then there is $\frac{\psi(1-q)}{\psi(1-q)+(1-\psi)q}$ chance that $X_n^a = Y_n$. If $\hat{X}_n^a \neq Y_n$, then there is $\frac{\psi q}{\psi q+(1-\psi)(1-q)}$ chance that $X_n^a = Y_n$. Conditional on $a = a'$, if $\hat{X}_n^a = Y_n$, then there is $1 - q$ chance that $X_n^a = Y_n$. If $\hat{X}_n^a \neq Y_n$, then there is $q$ chance that $X_n^a = Y_n$. By simple algebra, $\psi > 1/2$ implies that $\frac{\psi(1-q)}{\psi(1-q)+(1-\psi)q} > 1 - q$.
and \( \frac{\psi q}{\psi q + (1 - \psi)(1 - q)} > q \). So, the covariate \( a'' \) with a higher posterior probability of being the true cause also has a higher chance of passing the policymaker’s test. The maven gets strictly higher expected utility from proposing \( a'' \) than \( a' \) because the probability of being right, and of passing the test, are higher under \( a'' \) than under \( a' \).

\[ \square \]

0.2. Proof of Proposition

\[ \text{Proof.} \] Let such a \( q \) be fixed.

When the hacker proposes \( \hat{a} \) with \( \hat{X}_n^\hat{a} = Y_n \) for every \( 1 \leq n \leq N \), the number of observations \( n \) where \( X_n^a = Y_n \) has the distribution \( \text{Binom}(N, 1 - q) \). Let \( \ell_N \) be the probability that the hacker’s proposal is accepted. Since \( 1 - q < \gamma \), we have \( \ell_N \to 0 \) as \( N \to \infty \). So the principal’s utility conditional on facing a maven converges to 0 when \( N \to \infty \).

For the maven, consider the following three events.

\[ E_1 : |\{ n : \hat{X}_n^\hat{a} = Y_n \}| \leq |\{ n : X_n^{a^*} = Y_n \}|. \] The distribution of the LHS is \( \text{Binom}(N, \psi(1 - q) + (1 - \psi)q) \), where \( \psi(1 - q) + (1 - \psi)q > 1/2 \) since \( \psi > 1/2 \) and \( q < 1/2 \). Also, the distribution of the RHS is \( \text{Binom}(N, 1/2) \). By the law of large numbers, the probability of \( E_1 \) goes to 0 when \( N \to \infty \).

\[ E_2 : |\{ n : X_n^{a^*} = Y_n \}| < \lfloor \gamma \cdot N \rfloor. \] The LHS has the distribution \( \text{Binom}(N, \psi) \) where \( \psi > \gamma \). So by the law of large numbers, the probability of \( E_2 \) also approaches 0 when \( N \to \infty \).

Let \( E_3 \) be the event that the maven optimally proposes a covariate other than \( a^* \) or \( a'' \). Consider the feasible strategy for the maven of picking between the two covariates identified by his private information based on which one matches the outcome in more observations. Except on the event \( E_1 \cup E_2 \), this feasible strategy correctly identifies \( a^* \) and the proposal gets accepted by the policymaker. Therefore, the expected payoff of this feasible strategy for the maven is at least \( 1 - \mathbb{P}[E_1 \cup E_2] \) for any value of \( 0 \leq w_{\text{maven}} \leq 1 \). When the maven proposes some \( \hat{a} \notin \{a^*, a''\} \) with \( \hat{X}_n^\hat{a} = Y_n \) for every \( n \), he gets \( 1 - w_{\text{maven}} \) if the proposal is accepted and 0 otherwise, and acceptance happens with probability \( \ell_N \). In order for the maven’s optimal strategy to achieve at least an expected payoff of \( 1 - \mathbb{P}[E_1 \cup E_2] \), we therefore need

\[ \mathbb{P}[E_3] \cdot (1 - w_{\text{maven}}) \cdot \ell_N + (1 - \mathbb{P}[E_3]) \cdot 1 \geq 1 - \mathbb{P}[E_1 \cup E_2] \iff \mathbb{P}[E_3] \leq \frac{\mathbb{P}[E_1 \cup E_2]}{1 - (1 - w_{\text{maven}})\ell_N}. \]
Note that we have $\mathbb{P}[E_3] \to 0$ as $N \to \infty$, since we have both $\mathbb{P}[E_1 \cup E_2] \to 0$ and $\ell_N \to 0$ as $N \to \infty$.

Outside of the event $E_1 \cup E_2 \cup E_3$, the maven’s optimal strategy proposes $a^*$ and this proposal passes the policymaker’s test. So the principal’s utility conditional on facing a maven converges to 1 when $N \to \infty$.

\[ \square \]

0.3. **Proof of Proposition 2**

Proof. Consider the events $E_1$, $E_2$, and $E_3$ in the proof of Proposition 1.

First, we bound the probability of $E_1$. Using Hoeffding’s inequality, since $|\{n : \hat{X}_n^a = Y_n\}|$ has a binomial distribution with success rate $\psi(1 - q) + (1 - \psi)q$,

\[
\mathbb{P}[|\{n : \hat{X}_n^a = Y_n\}| \leq \frac{\psi(1 - q) + (1 - \psi)q + 0.5}{2}N \leq \exp(-2N[\psi(1 - q) + (1 - \psi)q - 0.5]^2).]
\]

Similarly, since $|\{n : \hat{X}_n^{a^*} = Y_n\}|$ has a binomial distribution with success rate $1/2$,

\[
\mathbb{P}[|\{n : \hat{X}_n^{a^*} = Y_n\}| > \frac{\psi(1 - q) + (1 - \psi)q + 0.5}{2}N \leq \exp(-2N[\psi(1 - q) + (1 - \psi)q - 0.5]^2).]
\]

Bounding the probability of each of these two events by $h/32$ requires $N \geq \frac{-2\ln(h/32)}{(\psi(1 - q) + (1 - \psi)q - 0.5)^2}$.

This ensures $\mathbb{P}[E_1] \leq h/16$.

Also by Hoeffding’s inequality, $\mathbb{P}[E_2] \leq \exp(-2N[\psi - \gamma]^2)$. So whenever $N \geq \frac{-\ln(h/16)}{2(\psi - \gamma)^2}$, we have $\mathbb{P}[E_2] \leq h/16$.

The probability that the hacker’s proposal gets accepted is bounded by $\exp(-2N \cdot [q - (1 - \gamma)]^2)$ by Hoeffding’s inequality. This quantity is less than $h/8$ whenever $N \geq \frac{-\ln(h/8)}{2[q - (1 - \gamma)]^2}$.

From the proof of Proposition 1, we have $\mathbb{P}[E_3] \leq \frac{\mathbb{P}[E_1 \cup E_2]}{1 - (1 - u_{\text{maven}})\ell_N} \leq \frac{\mathbb{P}[E_1 \cup E_2]}{1 - \ell_N}$. When $\mathbb{P}[E_1 \cup E_2] \leq h/8$ and $\ell_N \leq h/8$, we have $\mathbb{P}[E_3] \leq h/4$.

So whenever $N$ satisfies the bound in the statement of the proposition, we have $\mathbb{P}[E_1 \cup E_2 \cup E_3] \leq (3h)/8$. The principal’s expected payoff is at least $-1 \cdot (\frac{3h}{8}) + (1 - \frac{3h}{8})$ when the agent is a maven, and at least $-1 \cdot (h/8)$ when the agent is a hacker. So, the principal’s total expected payoff is larger than $(1 - h) \cdot [-1 \cdot (\frac{3h}{8}) + (1 - \frac{3h}{8})] + h \cdot (h/8) \geq 1 - \frac{7h}{4} - \frac{h^2}{8} \geq 1 - 2h$. By comparison, the principal’s payoff is smaller than $1 - 2h$ when $q = 0$. 
0.4. **Proof of Lemma 2.**

*Proof.* Any strategy of the hacker leads to zero probability of proposing the true cause, so the hacker finds it optimal to just maximize the probability of the proposal passing the test. If the hacker proposes a covariate that matches $Y$ in $n_1$ observations and mismatches in $n_0$ observations, then the distribution of the number of matches in the raw dataset is $\text{Binom}(n_0, q) + \text{Binom}(n_1, 1-q)$. A covariate that matches the outcome variable in every observation in noisy dataset will have a distribution of $\text{Binom}(n_0 + n_1, 1-q) = \text{Binom}(n_0, 1-q) + \text{Binom}(n_1, 1-q)$ as its number of matches in the raw dataset, and $\text{Binom}(n_0, 1-q)$ strictly first-order stochastic dominates $\text{Binom}(n_0, q)$ if $n_0 \geq 1$ and $q < 1/2$. Therefore the hacker finds it optimal to propose any $a \in A$ that satisfies $\hat{X}_n^a = Y_n$ for every $1 \leq n \leq N$.

For the maven, since $w_maven > 1/2$, it is never optimal to propose covariates other than $a^*$ or $a^\star$ since these have zero chance of being the true cause. Out of the two candidate covariates that the maven narrows down to, the one that matches $Y$ in more observations in the noisy dataset has a higher posterior probability of being the true cause. Note that if the maven proposes $a^\star$, the policymaker always rejects the proposal since $X^{a^\star} = 1-Y$ in the raw dataset. Also, if the maven proposes $a^*$, it always passes the test since $X^{a^*} = Y$ in the raw dataset. $\square$

0.5. **Proof of Lemma 3.**

*Proof.* The hacker picks a covariate $\hat{a}$ where $\hat{X}_n^\hat{a} = Y_n$ for every $n$. Given that the policymaker is using the most stringent test, we get $V_{\text{hacker}}(q) = (1-q)^N$. For the maven, there are $2N$ bits of observations on the variables $X^{a^*}$ and $X^{a^\star}$. If strictly fewer than $N$ bits are flipped, then the maven proposes the correct policy (and therefore it passes the test). If exactly $N$ bits are flipped, then the maven recommends the correct policy $1/2$ of the time. So, $V_{\text{maven}}(q) = (p[\text{Binom}(2N, q) < N] + \frac{1}{2}p[\text{Binom}(2N, q) = N])$.
We have

\[ V'_\text{maven}(q) = \frac{d}{dq} \left( \mathbb{P}[\text{Binom}(2N, q) < N] + \frac{1}{2} \mathbb{P}[\text{Binom}(2N, q) = N] \right) \]

\[ = \frac{d}{dq} \left( \mathbb{P}[\text{Binom}(2N, q) \leq N] - \frac{1}{2} \mathbb{P}[\text{Binom}(2N, q) = N] \right) \]

\[ = -2N \cdot \mathbb{P}[\text{Binom}(2N - 1, q) = N] - \frac{1}{2} \frac{d}{dq} \left( q^N (1 - q)^N \binom{2N}{N} \right), \]

where the last step used the identity that \( \frac{d}{dq} \mathbb{P}[\text{Binom}(M, q) \leq N] = -M \cdot \mathbb{P}[\text{Binom}(M - 1, q) = N] \). Continuing,

\[ -2N \cdot q^N (1 - q)^{N-1} \binom{2N-1}{N} - \frac{1}{2} \binom{2N}{N} N(q^{N-1}(1-q)^N - q^N(1-q)^{N-1}) \]

\[ = \left( \binom{2N-1}{N} \right) Nq^{N-1}(1-q)^{N-1} \left( -2q - \frac{1}{2} \cdot 2 \cdot ((1 - q) - q) \right), \]

using the identity \( \binom{2N}{N} = 2 \cdot \binom{2N-1}{N} \). Rearranging shows the lemma. \( \square \)

**0.6. Proof of Proposition 3**

**Proof.** Using the Lemma \(^3\)

\[ \frac{d}{dq} [-hV_{\text{hacker}}(q) + (1-h)V_{\text{maven}}(q)] = hN(1-q)^{N-1} - (1-h) \binom{2N-1}{N} Nq^{N-1}(1-q)^{N-1}. \]

The FOC sets this to 0, so \( h - (1-h) \binom{2N-1}{N} q^{N-1} = 0 \). Rearranging gives \( q^* = \left( \frac{h}{1-h} \binom{2N-1}{N} \right)^{1/(N-1)} \). We know \( h \mapsto \frac{h}{1-h} \) is increasing, so \( \frac{\partial q^*}{\partial h} > 0 \). We know \( N \mapsto \binom{2N-1}{N} \) is increasing in \( N \), therefore both the base and the exponent in \( q^* \) decrease in \( N \), so \( \frac{\partial q^*}{\partial N} < 0 \).

\( \square \)

**0.7. Proof of Proposition 4**

**Proof.** The principal’s expected utility conditional on the agent being a maven is the same for every \( N \in \{1,...,N\} \), since the maven always proposes either \( a^* \) or \( a^r \) depending on which covariate matches \( Y \) in more observations, and the proposal passes the \( N \) threshold if and only if it is \( a^* \), since \( X^{a^r} = 1 - Y \) does not match the outcome in any observation in the raw dataset.

As shown in the proof of Lemma \(^2\), the distribution of the number of matches between \( X^a \) and \( Y \) in the raw dataset increases in the first-order stochastic sense with the
number of matches between $\hat{X}^a$ and $Y$ in the noisy dataset. So, for any test threshold $N$, the hacker finds it optimal to propose a covariate $a$ with $\hat{X}^a_n = Y_n$ for every $n$.

Therefore, the only effect of lowering $N$ from $N$ is to increase the probability of the hacker’s misguided policies passing the test. □