Appendix for:

When Behavioral Data Betrays Users: A Diagnostic and Protective Framework Against Social Interaction Leakages

Appendix Contents

Appendix A: Economic Basis for the Spear-Phishing Analysis in Section 3.2

This appendix presents the financial underpinnings of our spear-phishing simulation, using the Return on Attack (ROA) calculation from Section 3.1 and Cremonini and Martini (2005). ROA measures the profitability of an attack as

$$\text{ROA} = \frac{G_{\text{illicit}} - C_{\text{total}}}{C_{\text{total}}} = \frac{|\mathcal{M}_{\text{correct}}|\,\alpha\,p - \big(C_{\text{fixed}} + c|\mathcal{M}_{\text{attack}}|\big)}{C_{\text{fixed}} + c|\mathcal{M}_{\text{attack}}|}, \tag{A.1}$$

where an attack is considered profitable if ROA > 0. In this expression,

- $|\mathcal{M}_{\text{correct}}|$ denotes the number of *correctly* impersonated social ties,
- α is the conditional monetary payoff per successful complaint/case (given that an impersonated contact yields compliance),
- p is the probability that a correctly impersonated individual ultimately yields financial losses (i.e., the attacker's success probability),
- C_{fixed} captures overhead costs that do not depend on the attack size,
- c is the marginal per-message cost through texts or emails, and
- $|\mathcal{M}_{\text{attack}}|$ refers to the number of edge-level impersonation messages in the phishing campaign. Multiple messages may target the same victim through different impersonated contacts, so this is not necessarily the number of distinct users.

A central focus in our analysis is that behavioral data leakage can significantly increase $|\mathcal{M}_{\text{correct}}|$, even for a fixed attack size $|\mathcal{M}_{\text{attack}}|$. Formally, if the attacker's attack precision is given by

$$\pi_M = \frac{|\mathcal{M}_{\text{correct}}|}{|\mathcal{M}_{\text{attack}}|},$$

then better knowledge of the network boosts π_M by allowing impersonation of genuine social ties—thus improving the attacker's expected profit. When social ties are leaked via behavioral data, π_M increases, thereby boosting $|\mathcal{M}_{\text{correct}}|$ and making large-scale attacks substantially more profitable. These considerations guide our simulation parameters and underscore why mitigating social network inference is critical to reducing the overall spear-phishing threat. Consequently, understanding how behavioral data influences $|\mathcal{M}_{\text{correct}}|$ is essential for evaluating the economic feasibility of large-scale spear-phishing.

A.1. Parameter Estimates

Conditional Illicit Payoff α and Compliance Probability p. From Table 1, we use $\alpha = \$581.29$, the unweighted average of the six annual loss-per-case entries; conditional on a successful, monetized impersonation, this is the per-case payoff to the attacker. Rows 2018–2020 use victim counts from the 2020 IC3 report's three-year comparison for "Phishing/Vishing/Smishing/Pharming," whereas rows 2021–2023 use complaint counts from the 2023 IC3 report's three-year comparison for "Phishing/Spoofing." The corresponding count-weighted average for the same six rows is \$346.07; using that alternative calibration would proportionally lower the expected gross-gain term and lower ROA according to Eqn. (4). We further assume a phishing success probability $p = 0.1$, indicating that 10% of correctly targeted attack cases yield a financial return when the attacker's impersonation is accurate.

Fixed Costs C_{fixed}. Cybercriminals often outsource phishing operations to freelance markets, incurring costs in:

1. *Data Collection:* \$20–\$50/hour for ~ 40 hours (e.g., web scraping, data purchasing).
2. *Computational Labor:* \$10–\$100/hour for ~ 40 hours (e.g., training inference algorithms).
3. *Phishing Message Writing:* \$15–\$40/hour for ~ 16 hours (e.g., crafting emails to impersonate trusted connections).

Based on these rates, we consider three fixed-cost levels:

$$C_{\text{fixed}} \in \{\,\$1{,}440,\ \$4{,}040,\ \$6{,}640\,\}.$$

These figures capture a range of plausible expenses, reflecting typical offers on freelance platforms. Data collection costs were obtained from Upwork's reported rates for data analysts: \$20–\$50 per hour (`https://www.upwork.com/hire/data-analysts/cost/`). Computational labor costs were extracted from Upwork's software developer rates: \$10–\$100 per hour (`https://www.upwork.com/hire/software-developers/cost/`). Message writing costs were estimated from Upwork content writing rates: \$15–\$40 per hour (`https://www.upwork.com/hire/content-writers/cost/`).

Variable (Per-Message) Cost *c*. We treat delivery-related expenses as scenario assumptions: \$0.01–\$0.05 per SMS message and \$0.10–\$0.50 per email message. Though modest, these fees significantly affect ROA as $|\mathcal{M}_{\text{attack}}|$ grows large, particularly when attack precision is high.

A.2. Minimum Attack Precision for Profitability

From Eqn. (A.1), the break-even condition $\text{ROA} = 0$ implies

$$\pi_M \;\geq\; \frac{|\mathcal{M}_{\text{attack}}|\, c + C_{\text{fixed}}}{p\,\alpha\,|\mathcal{M}_{\text{attack}}|}. \tag{A.2}$$

In practical terms, a high attack precision implies that the attacker accurately identifies a large fraction of genuine social ties, enabling more convincing impersonations. Conversely, platforms can reduce attack precision by obscuring user–user connections (e.g., via privacy-protection mechanisms), thus lowering adversaries' chances of successfully matching real edges.

Figures 1a–1c show how this minimum attack precision for profitability varies with the per-message cost c and the conditional illicit payoff per successful case α, holding the attack size fixed at $|\mathcal{M}_{\text{attack}}| = 5{,}000$. When adversaries can accurately identify real social ties, attack precision rises; platforms can push it lower by obscuring user–user relationships through privacy-preserving mechanisms.

Figures 1d–1f illustrate how ROA changes across different cost structures and attack-precision levels, further underscoring that high-precision attacks are significantly more profitable and therefore warrant stronger privacy measures.

Appendix B: Robustness Checks for Financial Incentive of Spear-Phishing Attacks for Attackers

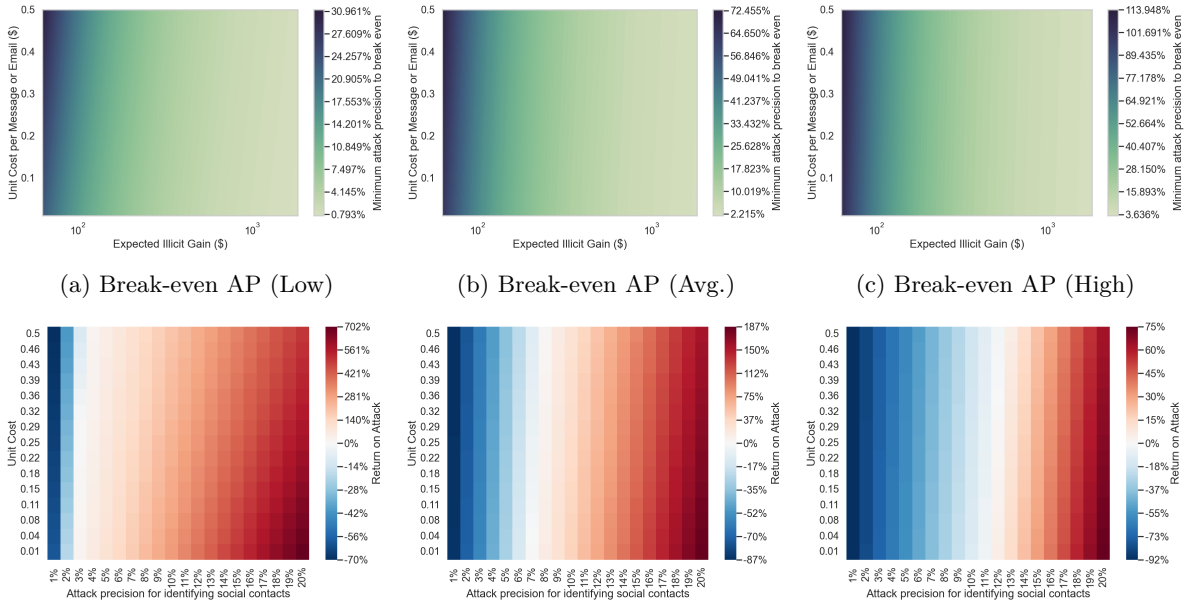
This appendix extends the analysis of spear-phishing economics to different campaign sizes beyond 5,000 impersonation messages, examining how changes in attack precision drive ROA (return on attack). The scenarios include 1,000, 10,000, and 50,000 messages, each evaluated under varying levels of unit cost per phishing attempt and the same illicit-gain calibration used in the main text.

For an attack of 1,000 impersonation messages, the upper panel of Figures B.1 shows that minimum attack-precision thresholds decline sharply as per-message costs fall or illicit gains rise, ranging from 0.8% to 31.0% for lower-end costs, 2.2% to 72.5% on average, and 3.6% to 113.9% at higher-end costs; values above 100% are infeasible for a precision and should be read as indicating that no feasible attack precision can make the attack profitable under that parameter combination. The lower panel demonstrates that once attack precision surpasses approximately 4% at lower-end costs, 8% at average costs, and 12% at higher costs, ROA becomes positive and continues growing with further increases in attack precision. At attack precision values near 20%, ROA can reach 702%, 187%, and 75% for low, average, and high labor costs, respectively.

When the attack size is increased to 10,000 impersonation messages, the upper panel of Figures B.2 confirms the same trend: attack-precision thresholds needed for breakeven reduce considerably as costs drop or gains climb, varying from 0.1% to 10.3% under lower-end costs, 0.2% to 14.4% at average levels, and 0.4% to 18.6% at higher-end costs. The lower panel reveals that ROA remains below zero if attack precision is too low or if per-message costs are high, but it grows rapidly once attack precision exceeds about 1% at lower-end costs and 2% at upper-end costs. At an attack precision of 20%, ROA can rise to 7,449%, 2,708%, and 1,625% across low, average, and high labor-cost conditions.

For large-scale attacks of 50,000 messages, the upper panel of Figures B.3 again shows a sharp decline in break-even attack-precision thresholds as gains increase or unit costs drop, with values ranging from 0.02% to 8.4% for lower-end costs, 0.05% to 9.3% on average, and 0.08% to 10.1% at higher-end costs. The lower panel indicates that ROA is already positive at an attack precision of around 1% if the cost per message and labor rates remain low, and it continues to expand at higher attack-precision values. When attack precision reaches 20%, ROA escalates to 29,863%, 12,704%, and 8,041% for low, average, and high labor costs, respectively.

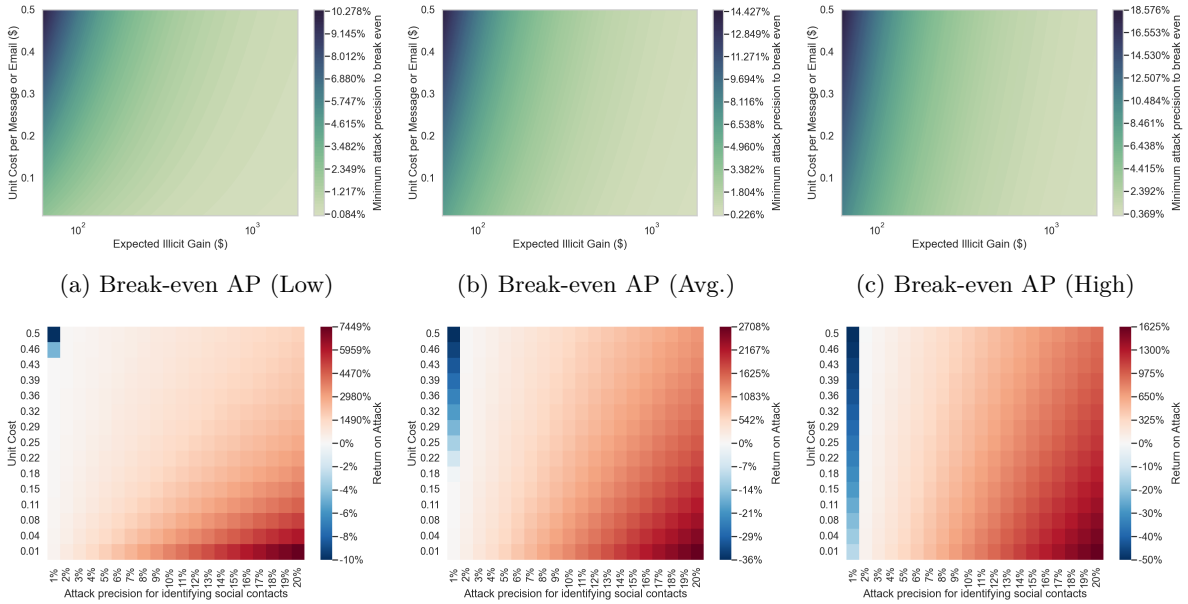
These findings consistently underscore how modest rises in attack precision can translate into steep gains in ROA, especially at larger attack scales where economies of scale increase overall profits. Even at smaller scales, ROA becomes positive at relatively low attack precision when labor costs are minimal or when per-case illicit gains are sufficiently high. The results confirm that mitigating adversarial attack precision—by limiting the ability to accurately identify social-network links—remains critical for reducing attackers' ROA and, by extension, deterring large-scale spear-phishing attempts.



(a) Break-even AP (Low) (b) Break-even AP (Avg.) (c) Break-even AP (High)

(d) ROA (Low) (e) ROA (Avg.) (f) ROA (High)

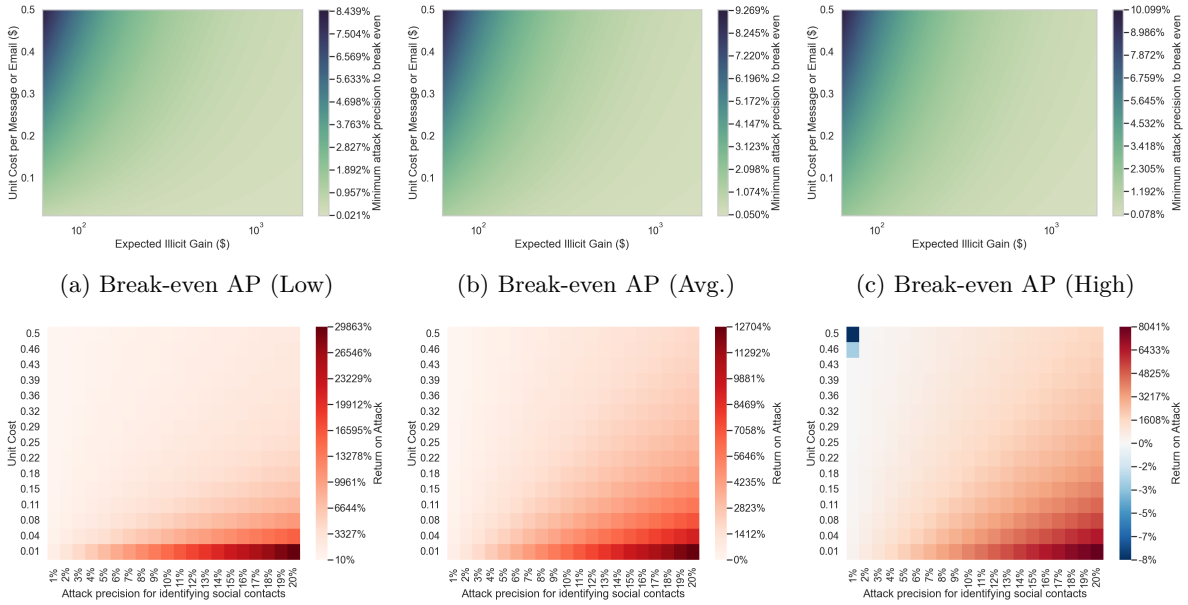
Figure B.1 Financial Incentives in Spear-Phishing Attacks (Attack Size = 1,000)



(a) Break-even AP (Low) (b) Break-even AP (Avg.) (c) Break-even AP (High)

(d) ROA (Low) (e) ROA (Avg.) (f) ROA (High)

Figure B.2 Financial Incentives in Spear-Phishing Attacks (Attack Size = 10,000)



(a) Break-even AP (Low) (b) Break-even AP (Avg.) (c) Break-even AP (High)

(d) ROA (Low) (e) ROA (Avg.) (f) ROA (High)

Figure B.3 Financial Incentives in Spear-Phishing Attacks (Attack Size = 50,000)

Appendix C: Proofs

C.1. Proofs for Network Identification

C.1.1. Proof of Theorem 1. The proof is conditional on a fixed nonzero hyperparameter β, so the coefficients $\beta a_i^{(k)}$ in the linear systems below are treated as fixed once the observed actions and the hyperparameter value are given. (Necessity) First suppose that it is possible to uniquely determine $\mathbf{G}$ from the K observed Nash equilibria. Without loss of generality, let $\mathcal{P}=\{1,2,\cdots,p\}$. We now proceed to prove the "only if " part of the statement.

Denote the first p rows of $\mathbf{G}$ by $\mathbf{G}^{\mathcal{P}}$ and the last $N-p$ rows of $\mathbf{G}$ by $\mathbf{G}^{\mathcal{V}-\mathcal{P}}$.

For the nodes $i\in\mathcal{P}$, we stack the $a_i^{(k)}$ and $b_i^{(k)}$ into vectors $\mathbf{a}_{\mathcal{P}}^{(k)}=[a_1^{(k)},\cdots,a_p^{(k)}]^\top$ and $\mathbf{b}_{\mathcal{P}}^{(k)}=[b_1^{(k)},\cdots,b_p^{(k)}]^\top$. Similarly, for the nodes $i\in\mathcal{V}-\mathcal{P}$, we stack the $a_i^{(k)}$ and $b_i^{(k)}$ into vectors $\mathbf{a}_{\mathcal{V}-\mathcal{P}}^{(k)}=[a_{(p+1)}^{(k)},\cdots,a_N^{(k)}]^\top$ and $\mathbf{b}_{\mathcal{V}-\mathcal{P}}^{(k)}=[b_{(p+1)}^{(k)},\cdots,b_N^{(k)}]^\top$. Furthermore, we stack the entries $\{G_{ij}|G_{ij}\in\mathbf{G}^{\mathcal{P}}\text{ and }i<j\}$ in rows on top of one another into $\mathbf{vec}(\mathbf{G}^{\mathcal{P}}):=[G_{12},G_{13},\cdots,G_{1N},G_{23},\cdots,G_{2N},\cdots,G_{pN}]^\top$. Similarly, we stack the entries $G_{ij}|G_{ij}\in\mathbf{G}^{\mathcal{V}-\mathcal{P}}$ and $i<j$ in rows on top of one another into $\mathbf{vec}(\mathbf{G}^{\mathcal{V}-\mathcal{P}}):=[G_{(p+1)(p+2)},G_{(p+1)(p+3)},\cdots,G_{(N-1)N}]^\top$.

Let $\mathbf{Q}^{\mathcal{P}(k)}$ be defined by $\mathbf{Q}^{\mathcal{P}(k)}=\{q_{ij}^{\mathcal{P}(k)}\}$ where

$$q_{ij}^{\mathcal{P}(k)}=\begin{cases}\beta a_t^{(k)}, & j=(t-1)N-\frac{t(t+1)}{2}+i \text{ and } \{t:t\in\mathbb{N},t<i\}\\ \beta a^{(k)}_{(i+j-\frac{(2N-i)(i-1)}{2})}, & \frac{(2N-i)(i-1)}{2}<j\le\frac{(2N-1-i)i}{2}\\ 0, & otherwise.\end{cases}\tag{C.1}$$

Since the full behavioral-inclination rows $\mathbf{B}_{i\cdot}$ for all $i\in\mathcal{P}$ are publicly known, the part of equations in Eqn. (8) related to the nodes in $i\in\mathcal{P}$ is represented by $\mathbf{A}^{\mathcal{P}}\mathbf{x}^{\mathcal{P}}=\mathbf{c}^{\mathcal{P}}$ where $\mathbf{x}^{\mathcal{P}}=\mathbf{vec}(\mathbf{G}^{\mathcal{P}})$, and

$$\mathbf{A}^{\mathcal{P}}=\begin{bmatrix}\mathbf{Q}^{\mathcal{P}(1)}\\ \mathbf{Q}^{\mathcal{P}(2)}\\ \vdots\\ \mathbf{Q}^{\mathcal{P}(K)}\end{bmatrix},\qquad \mathbf{c}^{\mathcal{P}}=\begin{bmatrix}-\mathbf{b}_{\mathcal{P}}^{(1)}+\mathbf{a}_{\mathcal{P}}^{(1)}\\ -\mathbf{b}_{\mathcal{P}}^{(2)}+\mathbf{a}_{\mathcal{P}}^{(2)}\\ \vdots\\ -\mathbf{b}_{\mathcal{P}}^{(K)}+\mathbf{a}_{\mathcal{P}}^{(K)}\end{bmatrix}.$$

Each $\mathbf{Q}^{\mathcal{P}(k)}$ has dimension $p\times\frac{(2N-p-1)p}{2}$, so $\mathbf{A}^{\mathcal{P}}$ has dimension $pK\times\frac{(2N-p-1)p}{2}$. Note that the elements in $\mathbf{A}^{\mathcal{P}}$ and in $\mathbf{c}^{\mathcal{P}}$ are known. Because the observed equilibria are assumed to be generated by a true graph satisfying Eqn. (8), there exists a true vector $\mathbf{vec}(\mathbf{G}^{\mathcal{P}})$ such that $\mathbf{A}^{\mathcal{P}}\mathbf{vec}(\mathbf{G}^{\mathcal{P}})=\mathbf{c}^{\mathcal{P}}$. Hence $\mathbf{c}^{\mathcal{P}}$ lies in the column space of $\mathbf{A}^{\mathcal{P}}$, so $\text{rank}(\mathbf{A}^{\mathcal{P}})=\text{rank}(\mathbf{A}^{\mathcal{P}}\mid\mathbf{c}^{\mathcal{P}})$. Therefore, $\mathbf{vec}(\mathbf{G}^{\mathcal{P}})$ can be uniquely solved when $\text{rank}(\mathbf{A}^{\mathcal{P}})=|\mathbf{vec}(\mathbf{G}^{\mathcal{P}})|$.

Let us now for the sake of the argument assume that $\mathbf{vec}(\mathbf{G}^{\mathcal{P}})$ is indeed uniquely recoverable, and for the sake of a contradiction argument assume that $|\mathcal{P}|<N-1$. We then focus on the solvability of the entries in $\mathbf{vec}(\mathbf{G}^{\mathcal{V}-\mathcal{P}})$. Define $\mathbf{Q}^{(\mathcal{V}-\mathcal{P})(k)}$ as $\mathbf{Q}^{(\mathcal{V}-\mathcal{P})(k)}=\{q_{ij}^{(\mathcal{V}-\mathcal{P})(k)}\}$ where

$$q_{ij}^{(\mathcal{V}-\mathcal{P})(k)}=\begin{cases}\beta a_{(t+p)}^{(k)}, & j=(t-1)(N-p)-\frac{t(t+1)}{2}+i \text{ and } \{t:t\in\mathbb{N},t<i\}\\ \beta a^{(k)}_{(i+p+j-\frac{(2N-2p-i)(i-1)}{2})}, & \frac{(2N-2p-i)(i-1)}{2}<j\le\frac{(2N-2p-1-i)i}{2}\\ 0, & \text{otherwise}\end{cases}\tag{C.2}$$

and $\mathbf{c}^{(\mathcal{V}-\mathcal{P})(k)}$ be defined by $\mathbf{c}^{(\mathcal{V}-\mathcal{P})(k)}=[a_{(p+1)}^{(k)}-\beta G_{(p+1)1}a_1^{(k)}-\cdots-\beta G_{(p+1)p}a_p^{(k)},a_{(p+2)}^{(k)}-\beta G_{(p+2)1}a_1^{(k)}-\cdots-\beta G_{(p+2)p}a_p^{(k)},\cdots,a_N^{(k)}-\beta G_{N1}a_1^{(k)}-\cdots-\beta G_{Np}a_p^{(k)}]^T$.

Provided that the intrinsic behavioral inclinations $\mathbf{b}_i$ for all $i \in \mathcal{V} \setminus \mathcal{P}$ are unknown and $\mathbf{vec}(\mathbf{G}^{\mathcal{P}})$ is uniquely solved, the remaining equations involve both the unknown edges inside $\mathcal{V} \setminus \mathcal{P}$ and the unknown behavioral-inclination rows for nodes in $\mathcal{V} \setminus \mathcal{P}$. The dimension count already indicates underdetermination, but non-identification of the graph entries can be shown directly. Because $p < N-1$, the set $U := \mathcal{V} \setminus \mathcal{P}$ contains at least two nodes. Pick two distinct nodes $r, s \in U$. For any sufficiently small nonzero scalar η such that $G_{rs} + \eta \geq 0$ and the model stability condition remains satisfied, define

$$\bar{\mathbf{G}} = \mathbf{G} + \eta(\mathbf{e}_r \mathbf{e}_s^\top + \mathbf{e}_s \mathbf{e}_r^\top).$$

Now define $\bar{\mathbf{B}}$ by leaving all rows unchanged except rows r and s:

$$\bar{\mathbf{B}}_{r\cdot} = \mathbf{B}_{r\cdot} - \beta\eta \mathbf{A}_{s\cdot}, \qquad \bar{\mathbf{B}}_{s\cdot} = \mathbf{B}_{s\cdot} - \beta\eta \mathbf{A}_{r\cdot}, \qquad \bar{\mathbf{B}}_{i\cdot} = \mathbf{B}_{i\cdot} \quad \text{for } i \notin \{r, s\}.$$

Then

$$(\mathbf{I} - \beta\bar{\mathbf{G}})\mathbf{A} = (\mathbf{I} - \beta\mathbf{G})\mathbf{A} - \beta\eta(\mathbf{e}_r \mathbf{A}_{s\cdot} + \mathbf{e}_s \mathbf{A}_{r\cdot}) = \bar{\mathbf{B}}.$$

Thus the same observed equilibria $\mathbf{A}$ are generated by two different graphs, $\mathbf{G}$ and $\bar{\mathbf{G}}$, while all revealed behavioral-inclination rows $\mathbf{B}_{i\cdot}$ for $i \in \mathcal{P}$ remain unchanged. Therefore $\mathbf{G}$ cannot be uniquely determined when $p < N-1$. This establishes the necessity statement of the theorem.

(Sufficiency) Next, we establish the "if " part of the statement. To this end, suppose that $|\mathcal{P}| \geq N-1$. In this case $\mathbf{vec}(\mathbf{G}^{\mathcal{P}}) = \mathbf{vec}(\mathbf{G})$. Because the observed data are generated by the model, the system is consistent; hence $\mathbf{G}$ is uniquely determined whenever $\text{rank}(\mathbf{A}^{\mathcal{P}}) = |\mathbf{vec}(\mathbf{G}^{\mathcal{P}})|$. ■

C.1.2. Proof of Theorem 2. As in Theorem 1, the proof is conditional on a fixed nonzero hyperparameter β, so the coefficients involving $\beta a_i^{(k)}$ are fixed by the observed actions and the selected hyperparameter value.

We introduce some useful notation in the subsequent proofs. Denote $\mathcal{V} \setminus \mathcal{P} = \{i : i \in \mathcal{V} \text{ and } i \notin \mathcal{P}\}$, $\mathcal{S} \setminus \mathcal{P} = \{i : i \in \mathcal{S} \text{ and } i \notin \mathcal{P}\}$, $\mathcal{V} \setminus (\mathcal{P} \cup \mathcal{S}) = \{i : i \in \mathcal{V} \text{ and } i \notin (\mathcal{P} \cup \mathcal{S})\}$, and $m = |\mathcal{P} \cap \mathcal{S}|$. (We use i as the node index throughout to avoid conflict with the behavior index $k \in \{1, \ldots, K\}$ used elsewhere.)

(Sufficiency) First suppose that $\mathcal{P} \cup \mathcal{S} = \mathcal{V}$. Without loss of generality, let $\mathcal{P} = \{1, 2, \cdots, p\}$ and $\mathcal{S} = \{t, \cdots, N\}$ where $t = p - m + 1$ and $0 \leq m \leq p$ since $m = |\mathcal{P} \cap \mathcal{S}|$. We proceed to

prove the "if" part of the statement. We show that each entry in $\mathbf{G}$ can be uniquely determined when $|\mathcal{P}|K \geq |\tilde{\mathbf{x}}^{\mathcal{P}}|$ and $\text{rank}(\tilde{\mathbf{A}}^{\mathcal{P}}) = |\tilde{\mathbf{x}}^{\mathcal{P}}|$, where $\tilde{\mathbf{A}}^{\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{P}}$ is the matrix form of the part of equations related to the nodes $i \in \mathcal{P}$ given that the subgraph $\mathbf{G}[\mathcal{S}]$ is publicly known. The row-count condition is necessary for full column rank, and model consistency gives existence of the true solution. The edges that have endpoints in $\{t, \cdots, N\}$ are publicly known. As a result, $[G_{t(t+1)}, G_{t(t+2)} \cdots, G_{tN}, G_{(t+1)(t+2)}, G_{(t+1)(t+3)} \cdots, G_{(t+1)N}, \cdots, G_{(N-1)N}]$ are uniquely determined. We then focus on the part of Eqn. (8) related to the nodes $i \in \mathcal{P}$. Since the edges $\{G_{t(t+1)}, G_{t(t+2)} \cdots, G_{tN}, \cdots, G_{p(p+1)}, G_{p(p+2)} \cdots, G_{pN}\}$ are publicly known, the unknowns of the part of Eqn. (8) related to the nodes $i \in \mathcal{P}$ become $\tilde{\mathbf{x}}^{\mathcal{P}} = [G_{12}, \cdots, G_{1N}, \cdots, G_{(t-1)t}, \cdots, G_{(t-1)N}]$ whose cardinality is $\frac{2N-1-|\mathcal{P}|+m}{2}(|\mathcal{P}| - m)$. Thus, the part of Eqn. (8) related to the nodes $i \in \mathcal{P}$ can be written as $\tilde{\mathbf{A}}^{\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{P}}$, where each entry of $\tilde{\mathbf{A}}^{\mathcal{P}}$ is either zero or $\beta a_i^{(k)}$, for $i \in \mathcal{V}, k = 1, 2, \cdots, K$. Equivalently, for each $i \in \mathcal{P}$ and $k = 1, \ldots, K$, the corresponding entry of $\tilde{\mathbf{c}}^{\mathcal{P}}$ is

$$\tilde{c}_i^{(k)} = \begin{cases} a_i^{(k)} - b_i^{(k)}, & i \in \mathcal{P} \setminus \mathcal{S}, \\ a_i^{(k)} - b_i^{(k)} - \beta \sum\limits_{\substack{j \in \mathcal{S} \\ j \neq i}} G_{ij} a_j^{(k)}, & i \in \mathcal{P} \cap \mathcal{S}. \end{cases}$$

Note that $\tilde{\mathbf{A}}^{\mathcal{P}}$ and $\tilde{\mathbf{c}}^{\mathcal{P}}$ depend on how $\mathbf{x}^{\mathcal{P}}$ is stacked. However, once $\mathbf{x}^{\mathcal{P}}$ is stacked, each entry of $\tilde{\mathbf{A}}^{\mathcal{P}}$ and $\tilde{\mathbf{c}}^{\mathcal{P}}$ is fully determined according to Eqn. (8). Because the observed equilibria and the revealed subgraph are generated by the model, $\tilde{\mathbf{c}}^{\mathcal{P}}$ lies in the column space of $\tilde{\mathbf{A}}^{\mathcal{P}}$. Therefore, if $\text{rank}(\tilde{\mathbf{A}}^{\mathcal{P}}) = |\tilde{\mathbf{x}}^{\mathcal{P}}|$, the reduced system has a unique solution for the remaining unknown edges, and all other edges are already known from the revealed induced subgraph.

(Necessity) Next, we proceed to the "only if" part of the statement. To this end, suppose that it is possible to determine $\mathbf{G}$ from the K observed Nash equilibria in the nontrivial case $p < N - 1$. Suppose that $(\mathcal{P} \cup \mathcal{S}) \neq \mathcal{V}$, i.e., the combination of $\mathcal{P}$ and $\mathcal{S}$ does not include all nodes in the network. Without loss of generality, let $\mathcal{P} = \{1, 2, \cdots, p\}$ and $\mathcal{S} = \{t, \cdots, s\}$ where $t = p - m + 1$ and $0 \leq m \leq p$. It is obvious to see $\mathcal{S} - \mathcal{P} = \{p+1, p+2, \cdots, s\}$ and $\mathcal{V} - (\mathcal{P} \cup \mathcal{S}) = \{s+1, s+2, \cdots, N\} \neq \emptyset$. Before we prove $\mathbf{G}$ cannot be uniquely determined, we first discuss the edges that are possible to be uniquely identified by either partially known graph structure or by uniquely solving $\tilde{\mathbf{A}}^{\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{P}}$. Since the induced subgraph $\mathbf{G}[\mathcal{S}]$ is revealed, $[G_{t(t+1)}, \cdots, G_{ts}, \cdots, G_{p(p+1)}, \cdots, G_{(t+1)s}, \cdots, G_{(s-1)s}]$ are publicly known. We then focus on $\tilde{\mathbf{A}}^{\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{P}}$ related to the nodes $i \in \mathcal{P}$. The unknowns become

$[G_{12}, G_{13}, \cdots, G_{1N}, \cdots, G_{(t-1)N}, G_{t(s+1)}, G_{t(s+2)}, \cdots, G_{tN}, \cdots, G_{p(s+1)}, \cdots, G_{pN}]$, whose cardinality is $\frac{(2N-t)}{2}(t-1)+(p-t+1)(N-s)=\frac{(2N-t)}{2}(t-1)+m(N-s)$. Further, each entry of $\tilde{\mathbf{A}}^{\mathcal{P}}$ is either zero or $\beta a_i^{(k)}$, for $i \in \mathcal{V}, k=1,2,\cdots,K$. Equivalently, for each $i \in \mathcal{P}$ and $k=1,\ldots,K$, the corresponding entry of $\tilde{\mathbf{c}}^{\mathcal{P}}$ is

$$\tilde{c}_i^{(k)} = \begin{cases} a_i^{(k)} - b_i^{(k)}, & i \in \mathcal{P} \setminus \mathcal{S}, \\ a_i^{(k)} - b_i^{(k)} - \beta \sum\limits_{\substack{j \in \mathcal{S} \\ j \neq i}} G_{ij} a_j^{(k)}, & i \in \mathcal{P} \cap \mathcal{S}. \end{cases}$$

Therefore, unique recovery of the subsystem $\tilde{\mathbf{A}}^{\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{P}}$ for the nodes $i \in \mathcal{P}$ would require $\tilde{\mathbf{A}}^{\mathcal{P}}$ to have full column rank equal to this number of unknown columns.

We then show that the edge set

$$\mathcal{E}^{\text{out}} = \{\overbrace{G_{(p+1)(s+1)}, \cdots, G_{(p+1)N}}^{N-s}, \cdots, \overbrace{G_{s(s+1)}, G_{s(s+2)}, \cdots, G_{sN}}^{N-s}, \overbrace{G_{(s+1)(s+2)}, \cdots, G_{(s+1)N}}^{N-s-1}, \cdots, \overbrace{G_{(N-1)N}}^{1}\}$$

cannot be uniquely identified. The unknowns $\tilde{\mathbf{x}}^{\mathcal{V}-\mathcal{P}}$ of the part of Eqn. (8) related to $i \in \mathcal{V} \setminus \mathcal{P}$ consist of $\mathcal{E}^{\text{out}}$ and the unknown behavioural inclination $\mathbf{b}_{\mathcal{V}-\mathcal{P}}^{(k)}, k=1,2,\cdots,K$. Thus, the remaining part of Eqn. (8) related to $i \in \mathcal{V} \setminus \mathcal{P}$ can be written as $\tilde{\mathbf{A}}^{\mathcal{V}-\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{V}-\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{V}-\mathcal{P}}$, where each entry of $\tilde{\mathbf{A}}^{\mathcal{V}-\mathcal{P}}$ is either one or zero or $\beta a_i^{(k)}$, for $i \in \mathcal{V}, k=1,2,\cdots,K$, and each entry of $\tilde{\mathbf{c}}^{\mathcal{V}-\mathcal{P}}$ is either $a_i^{(k)} - \sum_{j=1}^{s} \beta G_{ij} a_j^{(k)}$ or $a_i^{(k)} - \sum_{j=1}^{p} \beta G_{ij} a_j^{(k)}$. The preceding dimension count suggests underdetermination, but the non-identification of the unrevealed graph entries can be shown directly. Let

$$R := \mathcal{V} \setminus (\mathcal{P} \cup \mathcal{S}).$$

Under the maintained necessity argument, $R \neq \emptyset$. Since this theorem concerns the nontrivial case $p < N-1$, the set $U := \mathcal{V} \setminus \mathcal{P}$ contains at least two nodes. Choose $r \in R$ and choose $q \in U$ with $q \neq r$. The edge (r,q) is not incident to any node in $\mathcal{P}$ and is not contained in the revealed induced subgraph $\mathbf{G}[\mathcal{S}]$, because $r \notin \mathcal{S}$. For any sufficiently small nonzero scalar η such that $G_{rq} + \eta \geq 0$ and the model stability condition remains satisfied, define

$$\bar{\mathbf{G}} = \mathbf{G} + \eta(\mathbf{e}_r \mathbf{e}_q^{\top} + \mathbf{e}_q \mathbf{e}_r^{\top}).$$

This perturbation leaves all publicly revealed edges in $\mathbf{G}[\mathcal{S}]$ unchanged. Now define $\bar{\mathbf{B}}$ by

$$\bar{\mathbf{B}}_{r\cdot} = \mathbf{B}_{r\cdot} - \beta\eta \mathbf{A}_{q\cdot}, \qquad \bar{\mathbf{B}}_{q\cdot} = \mathbf{B}_{q\cdot} - \beta\eta \mathbf{A}_{r\cdot}, \qquad \bar{\mathbf{B}}_{i\cdot} = \mathbf{B}_{i\cdot} \quad \text{for } i \notin \{r,q\}.$$

Because $r, q \notin \mathcal{P}$, all revealed behavioral-inclination rows $\mathbf{B}_{i\cdot}$ for $i \in \mathcal{P}$ remain unchanged. Moreover,

$$(\mathbf{I} - \beta\bar{\mathbf{G}})\mathbf{A} = (\mathbf{I} - \beta\mathbf{G})\mathbf{A} - \beta\eta(\mathbf{e}_r\mathbf{A}_{q\cdot} + \mathbf{e}_q\mathbf{A}_{r\cdot}) = \bar{\mathbf{B}}.$$

Thus the same observed equilibria $\mathbf{A}$, the same revealed behavioral-inclination rows on $\mathcal{P}$, and the same revealed subgraph $\mathbf{G}[\mathcal{S}]$ are consistent with two different graphs, $\mathbf{G}$ and $\bar{\mathbf{G}}$. Therefore, when $p < N-1$ and $\mathcal{P} \cup \mathcal{S} \neq \mathcal{V}$, the full graph $\mathbf{G}$ cannot be uniquely determined. This establishes the necessity of the coverage condition in the nontrivial case. ■

C.2. Privacy Protection Mechanism

C.2.1. Necessary Assumptions

Assumption 1. *For each player $i \in \mathcal{V}$ in each game $k \in \mathcal{K}$, the intrinsic behavioral inclination $b_i^{(k)}$ is uniformly upper bounded, i.e., $|b_i^{(k)}| \le b_{\max}$.*

Assumption 2. *The spectral radius of $\beta\mathbf{G}$ is upper bounded, i.e., $\rho(\beta\mathbf{G}) \le \rho_{\max} < 1$.*

Assumption 3. *The intensity of peer effects between users is limited, i.e., $0 \le G_{ij} \le G_{\max}$.*

Remark 3. Assumption 1 assumes that each player's intrinsic behavioral inclination is constrained in each game. Assumption 2 guarantees the existence and stability of the Nash equilibrium actions. Assumption 3 assumes that each player has a limited impact on other players.

C.2.2. Necessary Lemmas

We denote the i-th eigenvalue of a square matrix $\mathbf{Q}$ by $\mathrm{eig}_i(\mathbf{Q})$. We also denote the column stack of an arbitrary matrix $\mathbf{P} \in \mathbb{R}^{p_1 \times p_2}$ by $\mathbf{P}_{\mathrm{vec}} \in \mathbb{R}^{p_1 p_2}$ as the vector obtained by stacking each column of $\mathbf{P}$. We further denote $\kappa = \max(2b_{\max}, G_{\max})$.

Lemma 1. *Let Assumption 2 hold. Then the eigenvalues of $(\mathbf{I} - \beta\mathbf{G})^{-1}$ satisfy*

$$\frac{1}{1+\rho_{\max}} \le \mathrm{eig}_i\big((\mathbf{I} - \beta\mathbf{G})^{-1}\big) \le \frac{1}{1-\rho_{\max}}, \qquad \forall i \in \mathcal{V}.$$

Proof. Since $\mathbf{G}$ is symmetric, $\beta\mathbf{G}$ is also symmetric, so all eigenvalues of $\beta\mathbf{G}$ are real. By Assumption 2, $\rho(\beta\mathbf{G}) \le \rho_{\max} < 1$, hence every eigenvalue $\lambda_i(\beta\mathbf{G})$ satisfies

$$|\lambda_i(\beta\mathbf{G})| \le \rho_{\max}.$$

Therefore, the eigenvalues of $\mathbf{I} - \beta\mathbf{G}$ are

$$\lambda_i(\mathbf{I} - \beta\mathbf{G}) = 1 - \lambda_i(\beta\mathbf{G}) \in [1-\rho_{\max},\, 1+\rho_{\max}],$$

so $\mathbf{I}-\beta\mathbf{G}$ is invertible. Taking reciprocals gives

$$\lambda_i\big((\mathbf{I}-\beta\mathbf{G})^{-1}\big) \in \left[\frac{1}{1+\rho_{\max}}, \frac{1}{1-\rho_{\max}}\right].$$

The proof is now completed. □

LEMMA 2. *Let Assumptions 1 and 2 hold. The Frobenius norm of Nash equilibrium actions is upper bounded as* $\|\mathbf{A}\|_F \le \frac{\sqrt{NK}b_{\max}}{1-\rho_{\max}}$, *where* $\mathbf{A}$ *is the matrix of Nash equilibrium actions and* $\|\cdot\|_F$ *denotes the Frobenius norm.*

Proof. Since $\mathbf{A} = (\mathbf{I}-\beta\mathbf{G})^{-1}\mathbf{B}$, by submultiplicativity,

$$\|\mathbf{A}\|_F = \|(\mathbf{I}-\beta\mathbf{G})^{-1}\mathbf{B}\|_F \le \|(\mathbf{I}-\beta\mathbf{G})^{-1}\|_2 \, \|\mathbf{B}\|_F.$$

For any matrix $\mathbf{P}$, the matrix 2-norm is $\|\mathbf{P}\|_2 = \sqrt{\lambda_{\max}(\mathbf{P}^\top\mathbf{P})}$. By Assumption 1,

$$\|\mathbf{B}\|_F \le \sqrt{NK}\, b_{\max}.$$

By the previous lemma,

$$\|(\mathbf{I}-\beta\mathbf{G})^{-1}\|_2 = \lambda_{\max}\big((\mathbf{I}-\beta\mathbf{G})^{-1}\big) \le \frac{1}{1-\rho_{\max}}.$$

Therefore,

$$\|\mathbf{A}\|_F \le \frac{\sqrt{NK}\, b_{\max}}{1-\rho_{\max}}.$$

The proof is now completed. □

C.2.3. Proof of Theorem 3 We stack the columns of $\mathbf{A} = (\mathbf{I}-\beta\mathbf{G})^{-1}\mathbf{B}$ into $\mathbf{A}_{\text{vec}}$. For any two $\Delta_{d_{\max}}$-adjacent databases $D_1 = (\mathbf{G}_1, \mathbf{B}_1)$ and $D_2 = (\mathbf{G}_2, \mathbf{B}_2)$, let

$$\mathbf{A}_1 = (\mathbf{I}-\beta\mathbf{G}_1)^{-1}\mathbf{B}_1, \qquad \mathbf{A}_2 = (\mathbf{I}-\beta\mathbf{G}_2)^{-1}\mathbf{B}_2.$$

By Lemma 2,

$$\|\mathbf{A}_1\|_F \le \frac{\sqrt{NK}b_{\max}}{1-\rho_{\max}}, \qquad \|\mathbf{A}_2\|_F \le \frac{\sqrt{NK}b_{\max}}{1-\rho_{\max}}.$$

Therefore,

$$\|\mathbf{A}_1 - \mathbf{A}_2\|_F \le \|\mathbf{A}_1\|_F + \|\mathbf{A}_2\|_F \le \frac{2\sqrt{NK}b_{\max}}{1-\rho_{\max}} := \bar{\Delta}_2.$$

Hence the ℓ_2-sensitivity Δ_2 of the column-stacked Nash-equilibrium query satisfies

$$\Delta_2 \le \bar{\Delta}_2.$$

By the Gaussian mechanism (e.g., Han and Pappas 2018, Theorem 6), adding Gaussian noise with standard deviation

$$\sigma_G > \frac{\sqrt{2\ln(1.25/\delta)}}{\epsilon}\,\bar{\Delta}_2$$

to each coordinate of $\mathbf{A}_{\text{vec}}$ yields (ϵ, δ)-DP. Rearranging the resulting noise vector into an $N \times K$ matrix $\mathbf{W}$ gives the mechanism

$$\mathcal{M}_D : (\mathbf{G}, \mathbf{B}) \mapsto (\mathbf{I} - \beta\mathbf{G})^{-1}\mathbf{B} + \mathbf{W}.$$

Thus, $\mathcal{M}_D$ satisfies (ϵ, δ)-DP. □

C.2.4. Proof of Theorem 4 Let h be the vector obtained by stacking all entries of $\hat{\mathbf{B}}$ together with the upper-triangular (i.e., $i < j$) entries of $\hat{\mathbf{G}}$. Under Definition 2, adjacent learned databases differ in exactly one coordinate of h by at most $\Delta_{d_{\max}}$, so the ℓ_1-sensitivity of the identity query is $\Delta_{d_{\max}}$. In our implementation, $\Delta_{d_{\max}} = \kappa$ with $\kappa = \max\{2b_{\max}, G_{\max}\}$. By the Laplace mechanism (e.g., Han and Pappas 2018, Theorem 5), adding i.i.d. $\text{Lap}(0, \Delta_{d_{\max}}/\epsilon)$ noise to each coordinate yields ϵ-DP for the noisy h. Mapping this noisy vector back to $(\hat{\mathbf{G}} + \boldsymbol{\eta}_G, \hat{\mathbf{B}} + \boldsymbol{\eta}_B)$ by mirroring the $i < j$ entries to enforce symmetry is deterministic post-processing and therefore preserves ϵ-DP. □

C.2.5. Proof of Theorem 5 By Theorem 4, the mechanism $\mathcal{M}_{ID}$ is ϵ-DP. The map g in Eqn. (15) is deterministic and depends only on the output of $\mathcal{M}_{ID}$, not directly on the underlying database. Therefore, by the post-processing property of differential privacy (Dwork et al. 2014, p. 18), the composition $g \circ \mathcal{M}_{ID}$ is also ϵ-DP. □

Appendix D: Illustrative examples for Theorem 1 and 2

We present the following example to illustrate the insights behind Theorem 1.

EXAMPLE 1. We consider a platform consisting of three users indexed in $\mathcal{V}=\{1,2,3\}$, whose decisions arise from the proposed quadratic game over a social interaction network $\mathbf{G}$. The K observed Nash equilibria $\mathbf{a}^{(k)}, k=1,\ldots,K$ as actions of the users lead to the following linear equation in the form of (8):

$$\begin{cases} a_1^{(1)}-\beta G_{12}a_2^{(1)}-\beta G_{13}a_3^{(1)}=b_1^{(1)} \\ -\beta G_{21}a_1^{(1)}+a_2^{(1)}-\beta G_{23}a_3^{(1)}=b_2^{(1)} \\ -\beta G_{13}a_1^{(1)}-\beta G_{32}a_2^{(1)}+a_3^{(1)}=b_3^{(1)} \\ \vdots \\ a_1^{(K)}-\beta G_{12}a_2^{(K)}-\beta G_{13}a_3^{(K)}=b_1^{(K)} \\ -\beta G_{21}a_1^{(K)}+a_2^{(K)}-\beta G_{23}a_3^{(K)}=b_2^{(K)} \\ -\beta G_{13}a_1^{(K)}-\beta G_{32}a_2^{(K)}+a_3^{(K)}=b_3^{(K)} \end{cases} \tag{D.1}$$

In each case we identify the unknowns explicitly in the accompanying matrix expressions and investigate two cases.

(i) Suppose $\mathcal{P}=\{1\}$, e.g., the intrinsic behavioral inclinations $b_1^{(1)}, b_1^{(2)}, \ldots, b_1^{(K)}$ for user 1 are publicly known (or known to a malicious eavesdropper).

$$\begin{cases} a_1^{(1)}-\beta G_{12}a_2^{(1)}-\beta G_{13}a_3^{(1)}=b_1^{(1)} \\ a_1^{(2)}-\beta G_{12}a_2^{(2)}-\beta G_{13}a_3^{(2)}=b_1^{(2)} \\ \vdots \\ a_1^{(K)}-\beta G_{12}a_2^{(K)}-\beta G_{13}a_3^{(K)}=b_1^{(K)} \end{cases} \quad \text{and} \quad \begin{cases} -\beta G_{21}a_1^{(1)}+a_2^{(1)}-\beta G_{23}a_3^{(1)}=b_2^{(1)} \\ -\beta G_{13}a_1^{(1)}-\beta G_{32}a_2^{(1)}+a_3^{(1)}=b_3^{(1)} \\ \vdots \\ -\beta G_{21}a_1^{(K)}+a_2^{(K)}-\beta G_{23}a_3^{(K)}=b_2^{(K)} \\ -\beta G_{13}a_1^{(K)}-\beta G_{32}a_2^{(K)}+a_3^{(K)}=b_3^{(K)} \end{cases} \tag{D.2}$$

Denote $\mathbf{A}$ and $\mathbf{A}^{\text{aug}}$ by

$$\begin{bmatrix} -\beta a_2^{(1)} & -\beta a_3^{(1)} \\ -\beta a_2^{(2)} & -\beta a_3^{(2)} \\ \vdots & \\ -\beta a_2^{(K)} & -\beta a_3^{(K)} \end{bmatrix} \text{ and } \begin{bmatrix} -\beta a_2^{(1)} & -\beta a_3^{(1)} & b_1^{(1)}-a_1^{(1)} \\ -\beta a_2^{(2)} & -\beta a_3^{(2)} & b_1^{(2)}-a_1^{(2)} \\ \vdots & & \\ -\beta a_2^{(K)} & -\beta a_3^{(K)} & b_1^{(K)}-a_1^{(K)} \end{bmatrix} \tag{D.3}$$

There is a unique solution for G_{12} and G_{13} when $\text{rank}(\mathbf{A})=\text{rank}(\mathbf{A}^{\text{aug}})=2$. However, it is always the case that the number of unknowns is larger than the number of equations

for the second part of Eq. (D.2), which means there are infinitely many solutions for the second part. Under the information set considered in Theorem 1, which excludes the HNL volume normalization as a public side equation, it is impossible to uniquely determine $\mathbf{G}$ regardless of the number of independent games played. This illustrates the non-identifiability result even in the presence of a user with known intrinsic behavioral inclinations.

(ii) Suppose $\mathcal{P}=\{1,2\}$, so $b_1^{(1)}, b_1^{(2)}, b_1^{(3)}$ and $b_2^{(1)}, b_2^{(2)}, b_2^{(3)}$ are publicly known. Let $K=3$. Eq. (D.1) can be written as

$$\begin{cases} a_1^{(1)} - \beta G_{12} a_2^{(1)} - \beta G_{13} a_3^{(1)} = b_1^{(1)} \\ -\beta G_{21} a_1^{(1)} + a_2^{(1)} - \beta G_{23} a_3^{(1)} = b_2^{(1)} \\ \vdots \\ a_1^{(3)} - \beta G_{12} a_2^{(3)} - \beta G_{13} a_3^{(3)} = b_1^{(3)} \\ -\beta G_{21} a_1^{(3)} + a_2^{(3)} - \beta G_{23} a_3^{(3)} = b_2^{(3)} \end{cases} \quad \text{and} \quad \begin{cases} -\beta G_{13} a_1^{(1)} - \beta G_{32} a_2^{(1)} + a_3^{(1)} = b_3^{(1)} \\ -\beta G_{13} a_1^{(2)} - \beta G_{32} a_2^{(2)} + a_3^{(2)} = b_3^{(2)} \\ \vdots \\ -\beta G_{13} a_1^{(3)} - \beta G_{32} a_2^{(3)} + a_3^{(3)} = b_3^{(3)} \end{cases} \tag{D.4}$$

It is straightforward to see that when the first part of Eq. (D.4) is uniquely solvable, which is certainly possible for a particular set of $\mathbf{a}^{(k)}, k=1,2,3$, $\mathbf{G}$ is uniquely determined. As a result, the fundamental privacy of the network influence structure for the platform has been lost when facing two users whose intrinsic behavioral inclinations are known publicly or to malicious eavesdroppers.

The impossibility and possibility for uniquely reconstructing $\mathbf{G}$ in the above two cases, are certainly consistent with Theorem 1. □

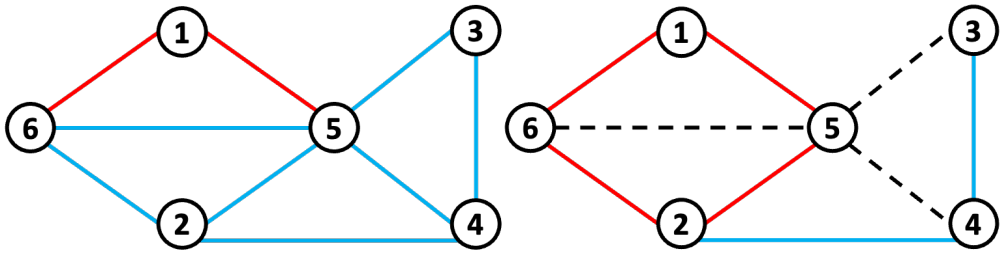
We present the following example to demonstrate the significance of Theorem 2.

Example 2. Consider a platform consisting of 6 users whose decisions form the proposed quadratic network game over a social interaction network $\mathbf{G}=(\mathcal{V},\mathcal{E})$ with $\mathcal{V}=\{1,2,3,4,5,6\}$.

We investigate two cases.

(i) Let $\mathcal{P}=\{1,2\}$ and $\mathcal{S}=\{2,3,4,5,6\}$. Then the social interaction links with endpoints in $\{2,3,4,5,6\}$ are publicly known , accounting for a set of links with cardinality 10. We then focus on all possible edges that are incident to nodes 1 and 2, for which the unknown social interaction links form the following vector

$$\mathbf{x}^{\mathcal{P}} = [G_{12}, G_{13}, G_{14}, G_{15}, G_{16}]^\top,$$



(a) $\mathcal{P}=\{1,2\}$ and $\mathcal{S}=\{2,3,4,5,6\}$ (b) $\mathcal{P}=\{1,2\}$ and $\mathcal{S}=\{2,3,4\}$

Figure D.1 **Learning the social interaction network $\mathbf{G}$ with prior knowledge of intrinsic behavioral inclinations of individuals in $\mathcal{P}$ and influence structure of individuals in $\mathcal{S}$. The edges in solid blue lines are publicly known, the edges in solid red lines can be uniquely reconstructed, and the edges in dashed black lines cannot be uniquely identified.**

whose dimension is 5. Define

$$\mathbf{A}^{\mathcal{P}}=\begin{bmatrix} \beta a_2^{(1)} & \beta a_3^{(1)} & \beta a_4^{(1)} & \beta a_5^{(1)} & \beta a_6^{(1)} \\ \beta a_1^{(1)} & 0 & 0 & 0 & 0 \\ & & \vdots & & \\ \beta a_2^{(K)} & \beta a_3^{(K)} & \beta a_4^{(K)} & \beta a_5^{(K)} & \beta a_6^{(K)} \\ \beta a_1^{(K)} & 0 & 0 & 0 & 0 \end{bmatrix}_{2K\times 5} \tag{D.5}$$

and

$$\mathbf{c}^{\mathcal{P}}=\begin{bmatrix} a_1^{(1)}-b_1^{(1)} \\ a_2^{(1)}-b_2^{(1)}-\beta a_3^{(1)}G_{23}-\cdots-\beta a_6^{(1)}G_{26} \\ \vdots \\ a_1^{(K)}-b_1^{(K)} \\ a_2^{(K)}-b_2^{(K)}-\beta a_3^{(K)}G_{23}-\cdots-\beta a_6^{(K)}G_{26} \end{bmatrix}_{2K\times 1}. \tag{D.6}$$

From the identity $\mathbf{A}^{\mathcal{P}}\mathbf{x}^{\mathcal{P}}=\mathbf{c}^{\mathcal{P}}$, $\mathbf{x}^{\mathcal{P}}$ can be uniquely solved when $\text{rank}(\mathbf{A}^{\mathcal{P}})=5$. It can also be deduced that $K\geq 3$ since the number of unknowns and equations are 5 and $2K$, respectively. Therefore, in the setting that $\mathcal{P}=\{1,2\}$ and $\mathcal{S}=\{2,3,4,5,6\}$, it is possible to uniquely determine $\mathbf{G}$ in Figure D.1 through the unique recovery of the full unknown vector

$$\mathbf{x}^{\mathcal{P}}=[G_{12},G_{13},G_{14},G_{15},G_{16}]^{\top}$$

when $\text{rank}(\mathbf{A}^{\mathcal{P}})=5$.

(ii) Let $\mathcal{P}=\{1,2\}$ and $\mathcal{S}=\{2,3,4\}$. In this setting, $\mathcal{P}\cup\mathcal{S}\subsetneq\mathcal{V}$. First, let us focus on the edges that are possible to be uniquely identified. Since the social interaction weights for

links in the induced subgraph $\mathbf{G}[\mathcal{S}]$ are revealed, $\tilde{\mathcal{E}}(\mathcal{S}) = \{G_{23}, G_{24}, G_{34}\}$ are publicly known. Among the links associated with nodes 1 and 2, the unknown edges are contained in the vector

$$\tilde{\mathbf{x}}^{\mathcal{P}} = [G_{12}, G_{13}, G_{14}, G_{15}, G_{16}, G_{25}, G_{26}]^{\top},$$

whose dimension is 7. Define

$$\tilde{\mathbf{A}}^{\mathcal{P}} = \begin{bmatrix} \beta a_2^{(1)} & \beta a_3^{(1)} & \beta a_4^{(1)} & \beta a_5^{(1)} & \beta a_6^{(1)} & 0 & 0 \\ \beta a_1^{(1)} & 0 & 0 & 0 & 0 & \beta a_5^{(1)} & \beta a_6^{(1)} \\ & & \vdots & & & & \\ \beta a_2^{(K)} & \beta a_3^{(K)} & \beta a_4^{(K)} & \beta a_5^{(K)} & \beta a_6^{(K)} & 0 & 0 \\ \beta a_1^{(K)} & 0 & 0 & 0 & 0 & \beta a_5^{(K)} & \beta a_6^{(K)} \end{bmatrix}_{2K\times 7} \tag{D.7}$$

and

$$\tilde{\mathbf{c}}^{\mathcal{P}} = \begin{bmatrix} a_1^{(1)} - b_1^{(1)} \\ a_2^{(1)} - b_2^{(1)} - \beta a_3^{(1)} G_{23} - \beta a_4^{(1)} G_{24} \\ \vdots \\ a_1^{(K)} - b_1^{(K)} \\ a_2^{(K)} - b_2^{(K)} - \beta a_3^{(K)} G_{23} - \beta a_4^{(K)} G_{24} \end{bmatrix}_{2K\times 1}. \tag{D.8}$$

The part of Eqn. (8) related to nodes 1 and 2 is in the form of $\tilde{\mathbf{A}}^{\mathcal{P}}\tilde{\mathbf{x}}^{\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{P}}$, which can be uniquely solved when $\text{rank}(\tilde{\mathbf{A}}^{\mathcal{P}}) = 7$. It can also be deduced in this case that $K \geq 4$ since the number of unknowns and equations are 7 and $2K$, respectively.

We now show that the links in $\mathcal{E}^{\text{not}} = \{G_{35}, G_{36}, G_{45}, G_{46}, G_{56}\}$ cannot be uniquely identified. Let $\tilde{\mathbf{x}}^{\mathcal{V}-\mathcal{P}} = [G_{35}, G_{36}, G_{45}, G_{46}, G_{56}, b_3^{(1)}, \cdots, b_6^{(1)}, \cdots, b_3^{(K)}, \cdots, b_6^{(K)}]^{\top}$ collect the corresponding unknown edges and intrinsic behavioral inclinations. Then

$$\tilde{\mathbf{A}}^{\mathcal{V}-\mathcal{P}} = \begin{bmatrix} \beta a_5^{(1)} & \beta a_6^{(1)} & 0 & 0 & 0 & \\ 0 & 0 & \beta a_5^{(1)} & \beta a_6^{(1)} & 0 & \\ \beta a_3^{(1)} & 0 & \beta a_4^{(1)} & 0 & \beta a_6^{(1)} & \\ 0 & \beta a_3^{(1)} & 0 & \beta a_4^{(1)} & \beta a_5^{(1)} & \\ & & & \vdots & & \mathbf{I}_{4K} \\ \beta a_5^{(K)} & \beta a_6^{(K)} & 0 & 0 & 0 & \\ 0 & 0 & \beta a_5^{(K)} & \beta a_6^{(K)} & 0 & \\ \beta a_3^{(K)} & 0 & \beta a_4^{(K)} & 0 & \beta a_6^{(K)} & \\ 0 & \beta a_3^{(K)} & 0 & \beta a_4^{(K)} & \beta a_5^{(K)} & \end{bmatrix}_{4K\times(4K+5)} \tag{D.9}$$

and

$$\tilde{\mathbf{c}}^{\mathcal{V}-\mathcal{P}} = \begin{bmatrix} a_3^{(1)} - \beta G_{13} a_1^{(1)} - \beta G_{23} a_2^{(1)} - \beta G_{34} a_4^{(1)} \\ a_4^{(1)} - \beta G_{14} a_1^{(1)} - \beta G_{24} a_2^{(1)} - \beta G_{34} a_3^{(1)} \\ a_5^{(1)} - \beta G_{15} a_1^{(1)} - \beta G_{25} a_2^{(1)} \\ a_6^{(1)} - \beta G_{16} a_1^{(1)} - \beta G_{26} a_2^{(1)} \\ \vdots \\ a_3^{(K)} - \beta G_{13} a_1^{(K)} - \beta G_{23} a_2^{(K)} - \beta G_{34} a_4^{(K)} \\ a_4^{(K)} - \beta G_{14} a_1^{(K)} - \beta G_{24} a_2^{(K)} - \beta G_{34} a_3^{(K)} \\ a_5^{(K)} - \beta G_{15} a_1^{(K)} - \beta G_{25} a_2^{(K)} \\ a_6^{(K)} - \beta G_{16} a_1^{(K)} - \beta G_{26} a_2^{(K)} \end{bmatrix}_{4K \times 1} . \tag{D.10}$$

The remaining part of Eqn. (8) related to nodes $i \in \{3, 4, 5, 6\}$ is in the form of

$$\tilde{\mathbf{A}}^{\mathcal{V}-\mathcal{P}} \tilde{\mathbf{x}}^{\mathcal{V}-\mathcal{P}} = \tilde{\mathbf{c}}^{\mathcal{V}-\mathcal{P}}.$$

It is obvious that the number of equations ($4K$) is less than the number of unknowns ($4K + 5$). Thus, it is impossible to uniquely determine the edge set $\mathcal{E}^{\text{not}}$. Therefore, in the setting that $\mathcal{P} \cup \mathcal{S} \subsetneq \mathcal{V}$, it is impossible to uniquely determine $\mathbf{G}$ from the K observed Nash equilibria.

The impossibility and possibility for uniquely determining $\mathbf{G}$ in these two cases are certainly consistent with Theorem 2.

Appendix E: Data Statistics

In this appendix, we present the summary statistics of the review length data ($\mathbf{A}$) and social network ($\mathbf{G}$) for Pennsylvania (PA) and Louisiana (LA).

We first describe the distribution of review lengths among users who have contributed reviews on the platform (Table E.1 and Figure E.1). We observe that a significant portion of entries in $\mathbf{A}$ are zero, indicating a user-business pair without a review. The percentage of nonzero entries is 0.3% in PA and 0.4% in LA.

The following statistics are computed only for the subset of nonzero entries, capturing the characteristics of review lengths among active users. The mean review length is 135.1 words in PA and 131.2 words in LA, with median values of 108 words and 104 words, respectively. The standard deviation is 103.8 in PA and 102.0 in LA, indicating considerable variation in review lengths. The minimum observed review length is 4 words in PA and 1 word in LA, while the longest reviews extend to 1,021 words in PA and 1,005 words in LA. The interquartile range (IQR) reveals that 50% of reviews fall between 64 and 174.5 words in PA and between 60 and 171 words in LA. To further assess variability, we report the 95% central interval for review length, with lower bounds of 22 words in PA and 21 words in LA, and upper bounds of 407 words in PA and 399 words in LA.

Table E.1 Descriptive Statistics of Nonzero Review Length Data ($\mathbf{A}$) for PA and LA

	PA	LA
Percentage Nonzero (%)	0.28	0.39
Mean	135.05	131.20
Median	108.00	104.00
Std Dev	103.80	101.97
Min	4.00	1.00
Max	1021.00	1005.00
Count	98921.00	48889.00
25% Quantile	64.00	60.00
75% Quantile	174.50	171.00
95% Central Interval Lower	22.00	21.00
95% Central Interval Upper	407.00	399.00

Table E.2 summarizes the structural properties of the social networks, based on standard network metrics. These metrics provide insights into the networks' connectivity, clustering, and overall efficiency. The PA network contains 2,234 nodes and 20,332 edges, whereas the LA network is slightly smaller, with 2,065 nodes and 14,257 edges. The larger edge count in PA suggests denser social interactions, which is further supported by a higher density

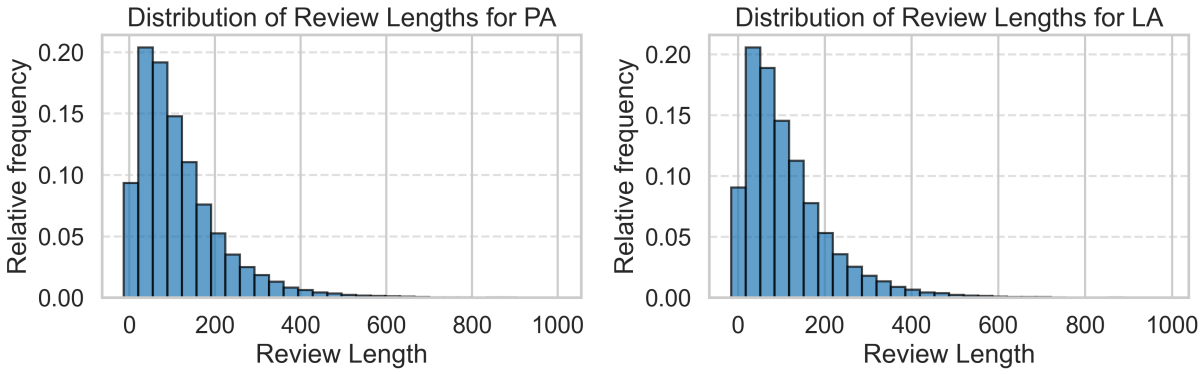


Figure E.1 Distribution of the Nonzero Review Length for PA and LA

value (0.008 in PA vs. 0.007 in LA). This indicates that a greater fraction of potential connections is realized in the PA network. In terms of node connectivity, the maximum degree—representing the highest number of connections held by any individual—is notably larger in PA (1,288) than in LA (860). This suggests the presence of more pronounced hubs or central influencers within PA's social structure. The network diameter, defined as the longest of all shortest paths in the graph, is smaller in PA (7) than in LA (8), implying that information can propagate more efficiently in PA.

Overall, while both networks exhibit similar properties, the PA network is more interconnected, with higher clustering, greater density, and a stronger core of highly connected nodes than the LA network. The fact that our method performs effectively across both networks—despite their structural differences—demonstrates its robustness with respect to network structures.

Table E.2 Summary Statistics of the Yelp Social Networks in PA and LA

	LA	PA
Num Nodes	2,065	2,234
Num Edges	14,257	20,332
Largest Degree	860	1,288
Diameter	8	7
Density	0.0067	0.0082
Average Clustering Coefficient	0.2970	0.3835

E.1. Implementation Details for the Yelp Benchmark

HNL configuration. We set the target peer-effect parameter $\beta_0 = 0.9$ and the regularization parameters $\theta_1 = \theta_2 = 10^{-7}$. The realized coefficient used by HNL for each Yelp state is

$\beta = -\beta_0/\rho(\mathbf{G}_{\text{ref}})$, where $\mathbf{G}_{\text{ref}}$ denotes the friendship reference graph for that state; the same specification is reused in the indirect-mechanism re-synthesis. Algorithm 1 is run with $T = 20$ outer iterations and $K_{\text{inner}} = 50$ inner mirror-descent steps, and HNL and the correlation-based baselines are fit to $\log(\mathbf{A} + 10^{-3})$.

Graphical-Lasso baseline. GL is computed from the empirical covariance after Ledoit–Wolf-style shrinkage (parameter 0.9) with graphical-lasso penalty $\alpha = 0.1$.

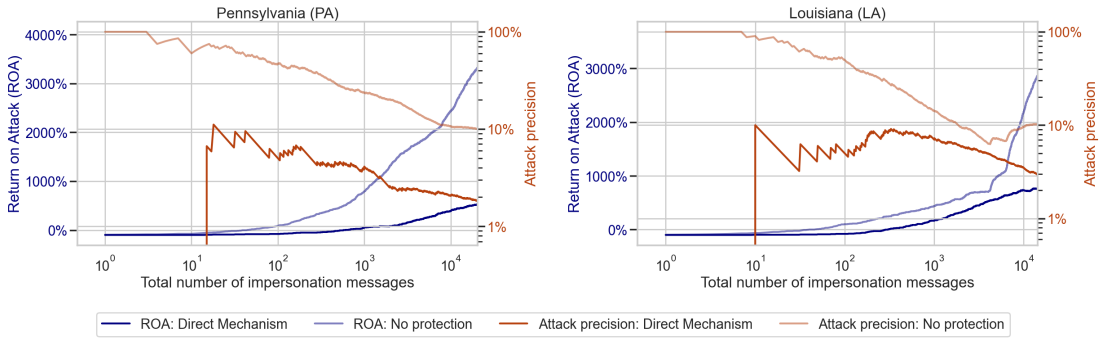
Mapping empirical τ to raw noise scales. The DM adds i.i.d. Gaussian noise with standard deviation $s_A = (\tau_G/2)\, a_{\max}$, where $a_{\max} := \max_{i,k} A_{ik}$. The IDM adds Laplace noise to $\hat{\mathbf{G}}$ and $\hat{\mathbf{B}}$ with scales $b_G = (\tau_L/3) \max_{i,j} |\hat{G}_{ij}|$ and $b_B = (\tau_L/3) \frac{1}{NK} \sum_{i,k} |\hat{B}_{ik}|$.

Appendix F: Sensitivity Analysis of ROA and Attack Precision under Varying Labor Costs and Privacy Mechanisms on the Yelp Platform

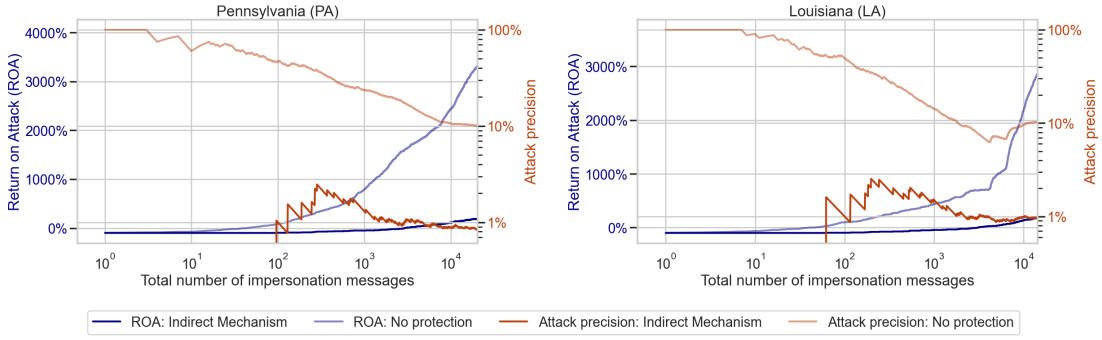
Figure F.1 presents a detailed sensitivity analysis evaluating how variations in labor costs (minimum versus maximum estimates obtained from Upwork) impact the ROA and attack precision for spear-phishing attacks across two datasets (PA and LA). We compare two privacy-protection mechanisms—direct mechanism and indirect mechanism—against a baseline scenario with no privacy protection. As in Figure 6, these appendix curves are constructed from single saved ranked-edge outputs rather than seed averages; only the labor-cost assumptions vary across panels.

Under the scenario with no protection, attackers experience the highest experiment-specific baseline profitability because attack precision remains high while fixed costs are spread over a growing campaign. The large ROA values reported in Appendix B, such as 702% for 1,000 messages and 7,449% for 10,000 messages under the low-labor-cost, 20%-precision calibration, are generic calibration benchmarks. Figure F.1 reports the corresponding Yelp-specific no-protection baselines under the PA/LA attack-precision and labor-cost assumptions.

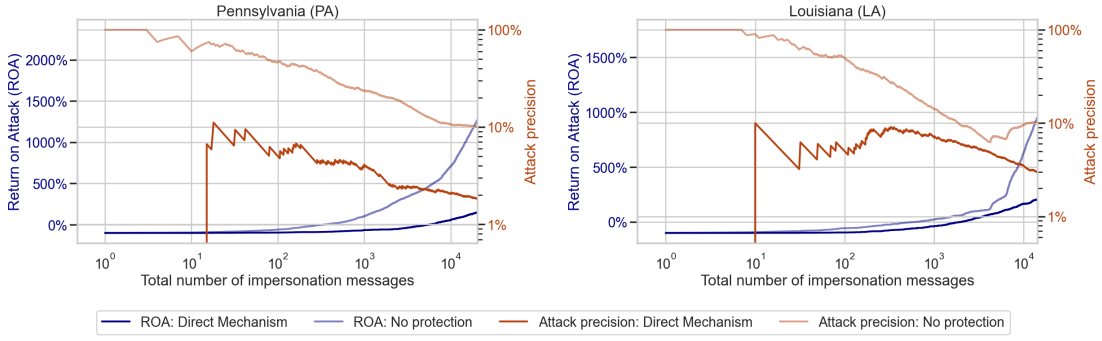
In the Yelp experiments, both the DM and the IDM reduce attack precision by disrupting the attacker's ability to infer social connections, thereby reducing realized ROA relative to the no-protection baseline. Under maximum labor costs, privacy protections become even more effective: the IDM yields negative ROA at smaller scales, making attacks economically infeasible, while the DM also provides substantial protection. The superior performance of the IDM over the DM across these scenarios indicates that, in our experiments, perturbing the learned interaction structure more strongly disrupts the attacker's inference pipeline. Hence, for platforms prioritizing stronger empirical privacy protection at comparable utility levels, the IDM is more effective in this setting.



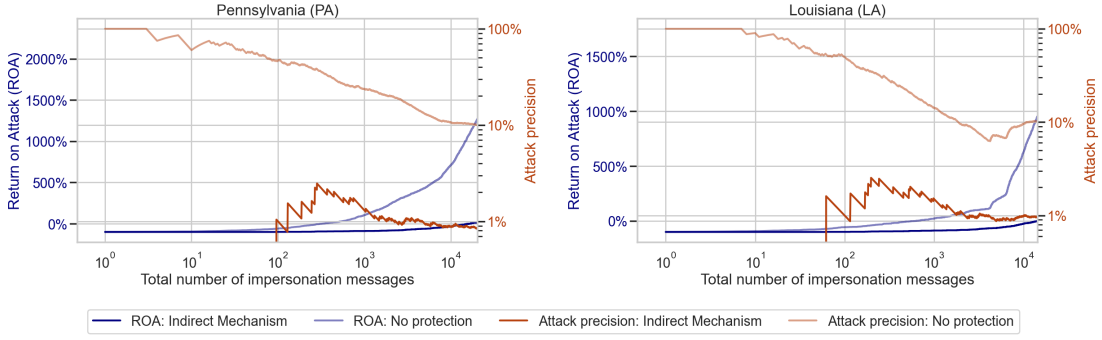
(a) Direct Mechanism (Minimum Labor Cost)



(b) Indirect Mechanism (Minimum Labor Cost)



(c) Direct Mechanism (Maximum Labor Cost)



(d) Indirect Mechanism (Maximum Labor Cost)

Figure F.1 **Sensitivity analysis of ROA and attack precision under minimum and maximum labor cost estimates for PA and LA. Both the direct and indirect mechanisms effectively mitigate attackers' financial incentives, with the indirect mechanism providing stronger protection across all scenarios.**