Who Pays for Payment Fraud?
Detection and Liability Rules under Strategic Fraudster Adaptation

Vimal Balasubramaniam Thomas Mosk Antoine Uettwiller*

August 2026

Abstract

Payment fraud per capita has doubled across major economies since 2017. The intuitive regulatory response, shifting fraud losses onto payment service providers (PSPs), overlooks two forces: fraudsters adapt, and liability can trigger debanking. Using 1.19 million underground forum posts, we show that criminal discussion shifts toward new methods after ChatGPT's release, and using the staggered rollout of Bitcoin ATMs, that rising local fraud is followed by rising account closures. We embed these findings in a model and show that without liability PSPs underinvest, but under full liability competing PSPs overinvest, since each ignores how its own detection accelerates market-wide adaptation, so the optimal rule assigns partial liability. Once liability crosses a threshold, PSPs debank fraud-prone customers, so regulators face a trade-off between providing detection incentives and preserving financial access. Information sharing on money-mule accounts becomes more valuable as fraudster adaptation intensifies, and sharing and liability are complements. When prevention also depends on digital platforms, whose screening complements PSP detection, splitting liability between the two invites free riding: welfare requires full liability for both, supplemented by a regulatory fee.

Keywords: Payment fraud, automated fraud detection, financial regulation, consumer protection, financial inclusion, ChatGPT

JEL Classification: G21, G23, G28, L51, O33, D62

*We thank Michele Fabi and Patrick Blonien (Discussants), Dirk Niepelt, Radoslawa Nikolowa, Maarten van Oordt, Tarun Ramadorai, Jason Sturgess, Konstantinos Zachariadis, and seminar participants at Queen Mary University of London, the 2nd Financial Fraud, Misconduct and Market Manipulation Conference, the CEPR Fintech and Digital Currencies RPN Annual Meeting 2025, the 2026 Fintech and Financial Institutions Research Conference at the Federal Reserve Bank of Philadelphia, the Cambridge Cybercrime Centre Annual Conference 2026, and the 8th Future of Financial Information Conference. Balasubramaniam: Queen Mary University of London, CEPR. Email: v.balasubramaniam@qmul.ac.uk. Mosk: Queen Mary University of London. Email: t.mosk@qmul.ac.uk. Uettwiller: Queen Mary University of London. Email: a.uettwiller@qmul.ac.uk.

I Introduction

Payment fraud has emerged as one of the fastest-growing economic crimes of the digital age. Fraud incidents per capita have roughly doubled in the US and UK since 2017, with steep rises in China and the EU as well, and fraud's share of total reported crime has risen across all four jurisdictions (Figure 1). The financial losses are large: in 2024 alone, US consumers lost $12.5 billion to payment fraud, UK consumers £341 million, and EEA consumers €4.3 billion. Yet these headline figures capture only the direct cost of fraud. Who ultimately bears the cost of fraud, and what it does to consumers' access to the payment system itself, is far less well understood.

Payment scams are difficult and costly for individual consumers to detect, whereas payment service providers (PSPs) can exploit scale economies and invest in advanced anti-fraud technologies, such as machine-learning transaction-risk scoring, network-based detection of mule accounts, and, increasingly, large language models used to flag AI-generated impersonation attempts. To align incentives, regulators have shifted liability onto PSPs, but coverage remains incomplete. In the US, Regulation Z requires credit card issuers to reimburse unauthorized fraud, and in practice all major card networks waive even the statutory $50 cap.[1] While most unauthorized fraud is reimbursed, authorized fraud, where the consumer is deceived into initiating the payment, is not covered by any federal rule, leaving a regulatory gap precisely for the fastest-growing fraud category.[2] The UK PSR's 2024 rules were the first to close this gap, requiring PSPs to reimburse victims of authorized push-payment fraud.

Yet even where liability exists, this reasoning overlooks two central issues: First, a central feature of fraud is that fraudsters adapt. As PSPs intensify detection of prevalent scams, fraudsters substitute toward harder-to-detect techniques such as AI-generated impersonations (Stop Scams UK, 2023) or image-based malware delivery via messaging apps.[3] This adaptation creates a negative detection externality: when one PSP improves its models, it pushes fraud toward variants that are harder for all PSPs to detect or displaces it to other markets entirely.

Second, higher PSP liability may lead to debanking. As PSPs bear a higher share of the fraud loss liability, they may seek to use their anti-fraud technologies to exclude customers

[1] Peer-to-peer systems (Zelle, Venmo, PayPal) are covered by Reg E for unauthorized transactions only, with the same tiered exposure; and cryptocurrency transfers have no federal consumer-protection framework at all.

[2] In one case documented by TIME Magazine, a Wells Fargo customer authorized a $1,000 transfer to a scammer posing as the bank; Wells Fargo initially denied reimbursement, arguing that federal rules only cover unauthorized transactions. `https://time.com/6952817/financial-scams-banks`.

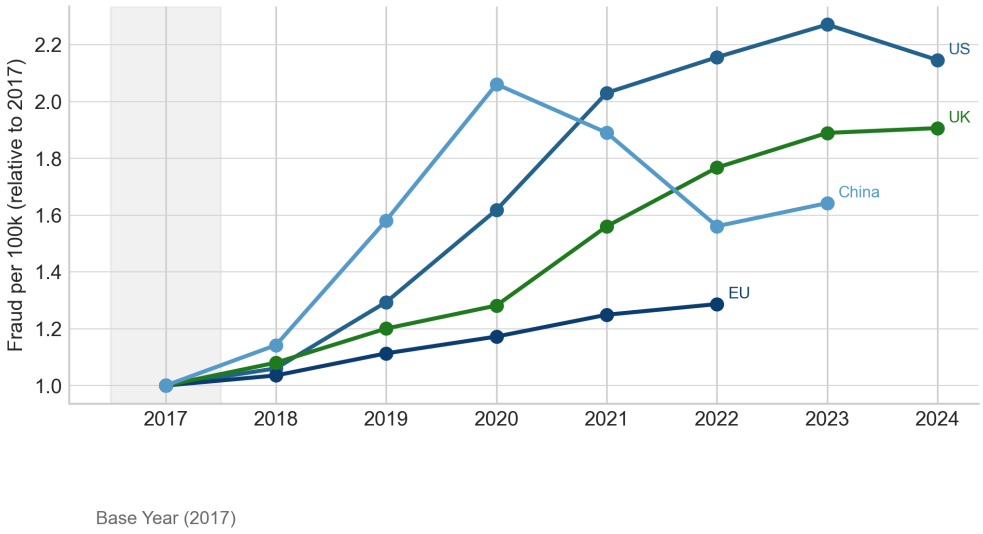
[3] See https://indianexpress.com/article/technology/he-downloaded-a-whatsapp-image-minutes-later-rs-2-lakh-was-gone-9951435/.

with a higher fraud risk. Consumer complaints to the CFPB about checking and savings account closures more than doubled between 2017 and 2024 (Sanchez-Adams, 2025). In one prominent case, the CFPB ordered Wells Fargo to pay $160 million to more than one million consumers whose deposit accounts were improperly frozen or closed between 2011 and 2016 based largely on an automated fraud detection system (Consumer Financial Protection Bureau, 2022). Debanking therefore creates a tradeoff for regulators between incentivizing fraud detection investment and preserving financial access.

This paper studies how fraudster adaptation shapes the optimal intensity and allocation of detection investment. We provide the first direct micro-level evidence that criminal communities respond to detection technologies by discussing and disseminating countermeasures, consistent with substitution toward harder-to-detect methods, and show that a rise in local fraud is followed by the exclusion of consumers from the payment system, through both account closures and denials of new accounts. We embed this evidence in a model of strategic adaptation and derive optimal liability rules, PSPs' incentives to debank fraud-prone consumers, and the welfare consequences of mandatory information sharing.

A critical assumption underlying both dynamics is that fraudsters adapt, developing and disseminating new circumvention techniques in response to detection technologies. The economics of crime literature shows that criminals respond to deterrence through spatial reallocation (Maheshri and Mastrobuoni, 2021), but not through method innovation. The closest evidence is Bian et al. (2024), who show that underground-forum discussion of consumer fraud contracts when a privacy change restricts fraudsters' access to personal data; whether criminals innovate over *methods* in response to detection technology itself remains undocumented. One reason is that official fraud statistics capture crime types but not the specific methods criminals use, making innovation difficult to observe in administrative data. We fill this gap using data from CrimeBB (Pastrana et al., 2018), one of the largest structured corpora of underground forum posts. From a panel of 1.19 million posts by 225,000 creators across five forums, we document two complementary patterns. First, the public release of ChatGPT in November 2022 serves as a natural experiment for the enabling-shock channel: within-creator AI–fraud co-discussion roughly doubled relative to its sample-average rate (+97%), with technically sophisticated creators driving early adoption and fraud-method discussion shifting relatively toward less skilled creators, consistent with a general-purpose shock that lowers skill barriers for fraud. Second, stacked event studies around the deployment dates of 27 anti-fraud technologies show that circumvention discussion builds as each defense matures and peaks around its deployment, while discussion of the targeted methods fades over the following years: the composition of criminal attention shifts from established methods toward countermeasures as defenses arrive, consistent with substitution toward

Panel A: Fraud per 100,000 inhabitants (Normalized to 2017)



Panel B: Fraud Share of Total Crime (Normalized to 2017)

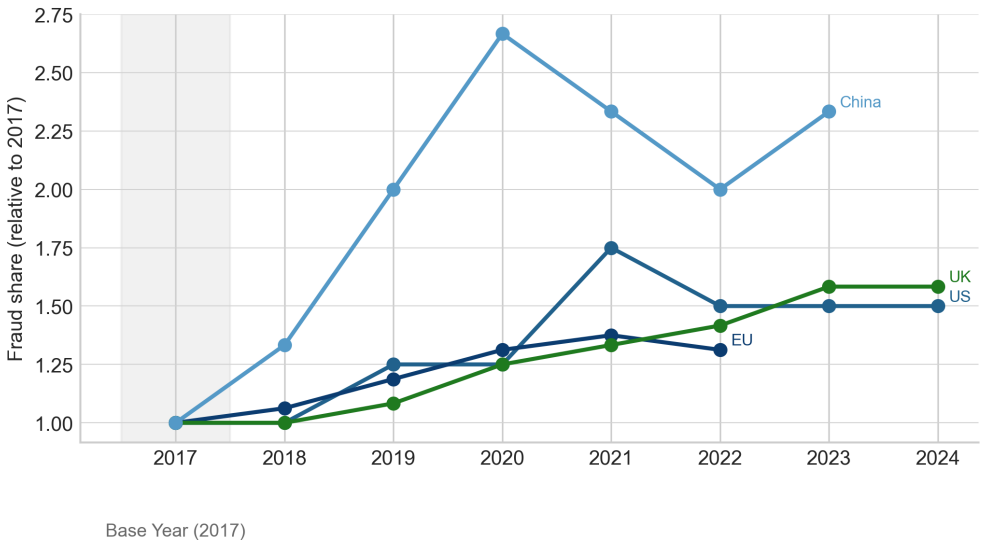


Figure 1
The Growth of Payment Fraud

Notes: Panel A shows fraud cases per 100,000 inhabitants in the EU, USA, UK, and China from 2017 to 2024, normalized relative to 2017. Panel B shows fraud as a share of total crime, also normalized relative to 2017. Sources: US: FBI UCR; EU: Eurostat; UK: ONS Crime in England and Wales; China: National Bureau of Statistics.

harder-to-detect methods.

We embed these empirical regularities in a model of payment fraud detection with strategic adaptation. The model yields four sets of results.

First, when PSPs do not bear fraud losses, detection is underprovided; but under full liability in competitive markets, PSPs overinvest because each ignores how its detection intensity accelerates fraudster adaptation market-wide. The socially optimal rule assigns partial liability, with the optimal share strictly below one and increasing in the number of PSPs.[4]

Second, liability interacts with financial inclusion. Using the staggered rollout of Bitcoin ATMs as a shock to local scam exposure, we show that a rise in local fraud is followed, with a lag, by a rise in account closures. In the model, a PSP debanks a flagged consumer once the liability it carries on that account exceeds the account's franchise value, so liability set purely to correct overinvestment can push PSPs to exclude the very consumers it was designed to protect. In practice, this margin is automated: the same machine-learning systems that score transactions for fraud also decide who keeps an account. The welfare-maximizing rule internalizes this margin directly: regulators must weigh the overinvestment correction against the risk of pricing vulnerable consumers out of the payment system.

Third, fraud detection also depends on information PSPs hold about each other's customers. Money-mule accounts, used to receive and launder fraud proceeds, are frequently held at multiple PSPs, so one PSP's identification has value to every institution exposed to the same mule, and can lower losses substantially once shared. Whether that translates into more detection is not automatic: sharing shrinks the pool an adapting fraudster targets, which can either free up detection capacity or blunt the incentive to invest in it, depending on how easily the fraudster's methods evade attribution. PSPs share voluntarily once liability is high enough to clear the fixed cost of running a registry, and because sharing weakens the same adaptation externality that justifies partial liability in the first place, the optimal liability share is higher than the one that would apply absent sharing.

Fourth, fraud prevention depends on the joint effort of multiple agents, a PSP and digital platforms, whose detection efforts are complementary. A fraudulent transaction requires both a platform to fail to screen the solicitation and a PSP to fail to flag the payment; placing too much responsibility on one weakens incentives for the other, creating a free-rider problem. A Reuters investigation based on internal Meta documents revealed that the company internally projected earning approximately $16 billion annually from fraudulent advertisements while its platforms were estimated to be involved in one-third of all successful US scams, yet platforms face no statutory reimbursement obligation (Horwitz, 2025). The

[4]We analyze consumer moral hazard in Internet Section VII.C.

welfare-maximizing rule sets full liability for all agents whose efforts are complementary, supplemented by a regulatory fee that internalizes the social cost of fraud; the UK PSR's 50/50 reimbursement split between sending and receiving PSPs falls short of this prescription, and the absence of platform liability represents the largest remaining regulatory gap. When PSPs and platforms instead detect entirely different fraud types, detection investment can displace fraud from one domain to the other rather than eliminating it, a cross-market Whac-A-Mole, and the logic of the baseline model suggests partial, rather than full, liability for both agents; we discuss this case informally rather than deriving it.

Our paper makes three principal contributions. First, our paper contributes to the theoretical literature on attack and defense games (Goyal and Vigier, 2014; Acemoglu et al., 2016). In both papers, strategic attackers and defenders allocate resources across a network, but the adversary's attack technology is fixed: it chooses *where* to strike, not *how*. We instead let the adversary respond to detection by shifting toward harder-to-detect attack methods, a mechanism we document directly in Section II.A and that neither fixed-technology model can generate. This differs from geographical displacement of crime (Maheshri and Mastrobuoni, 2021): there, technology is fixed and a constant threat reallocates across targets; here, the technology itself adapts, degrading every PSP's detection value through the aggregate level of investment. Since detection is degraded by the very adaptation it induces, we show that the standard security-economics prescription of full liability to correct underinvestment (Anderson and Moore, 2006) can reverse. Our results extend beyond payment fraud to the broader class of attack-and-defense situations in finance, including anti-money-laundering compliance, cyber risk in market infrastructure, and insider-trading surveillance, in which a regulated intermediary's defensive investment shapes an adaptive counterparty's behavior.

Second, while the literature on fraud in financial markets has studied perpetrators (Dimmock et al., 2018; Egan et al., 2019; Piskorski et al., 2015), we shift focus to the institutions that prevent and compensate fraud. This contrasts with existing theories of information acquisition in financial markets (Hauswald and Marquez, 2006; Vives and Ye, 2025), which treat detection technology as fixed. It also connects to work on big data and machine learning in household finance (Fuster et al., 2022; Bian et al., 2024), which shows that prediction technologies reshape who is served by financial institutions; the debanking margin extends this insight to fraud detection, where the same models that score transactions also decide who keeps an account. FinTech competition models such as Parlour et al. (2022) show that competition among payment providers promotes financial inclusion. We identify a channel that works in the opposite direction: liability can make debanking privately optimal for PSPs, since raising the cost of serving fraud-prone consumers can make it unprofitable to serve them at all, even when doing so is socially efficient.

Third, we make two empirical contributions to motivate our theory. The first is that fraudsters adapt: we provide direct micro-level evidence that fraudsters respond to detection investment by shifting toward harder-to-detect methods, the mechanism behind Proposition 1. Prior work documents the role of cryptocurrency in financing illegal activity (Foley et al., 2019; Griffin and Mei, 2024) and shows that profit pressure weakens bank incentives to report financial crime (Gao et al., 2020). Closest to us, Bian et al. (2024) use the same underground-forum corpus to show that fraud discussion contracts when a privacy change cuts off fraudsters' data inputs; we document the complementary response to defenses rather than inputs: detection technology does not suppress discussion of existing methods but redirects the community toward countermeasures, the first direct test of criminal method innovation in response to detection. The second links fraud to debanking: using the staggered rollout of crypto ATMs as a shock to the feasibility of local cash-based scams, following Cong et al. (2025) and Wei (2026), we show that ATM entry raises scam complaints and is followed, with a roughly one-year lag, by more account closures and more denials of new accounts; the timing and age profile of the responses point to fraud-driven de-risking rather than crypto exposure, motivating the debanking result in Section V.

The remainder of the paper proceeds as follows. Section II presents empirical evidence of strategic adaptation and the link between fraud and debanking. Section III develops the theoretical model of fraud detection with strategic adaptation. Sections IV–VII analyze optimal liability rules under various market structures, policies, and extensions. Section VIII concludes.

II Empirical Evidence: Adaptation and Debanking

Before turning to the model, we present empirical evidence motivating its two central mechanisms: fraudster adaptation and debanking. Administrative fraud statistics report outcomes but not the underlying behavioral responses of fraudsters and providers, so we draw instead on two settings that speak to them directly. Section II.A uses underground forum communications to show that fraudsters adapt their methods in response to detection technology, rather than persisting with a fixed set of techniques. Section II.B uses geographically granular consumer complaint data to show that fraud exposure is followed by the exclusion of consumers from the payment system, through both account closure and denial of new accounts.

II.A Fraudster Adaptation: Evidence from Underground Forums

Adversarial agents, whether fraudsters, spammers, or bot operators, adapt strategically to detection technologies, causing model performance to decay over time. While the economics of crime literature establishes that criminals respond to deterrence and price incentives through spatial reallocation (Maheshri and Mastrobuoni, 2021) and cross-type substitution (Draca et al., 2019), there is no direct evidence that criminals adapt their methods in response to detection technology. One empirical challenge is that official fraud statistics report crime types but do not contain information about the specific methods used by criminals, making adaptation difficult to observe in administrative data. We use CrimeBB (Pastrana et al., 2018), one of the largest publicly available underground forum datasets, to study the effect of technological change on how criminals discuss and share new fraud methods.

The empirical analysis exploits two complementary sources of variation, each with distinct advantages. The first is the public release of ChatGPT on November 30, 2022, which serves as a natural experiment for a general-purpose enabling shock. Its key advantage is precision: unlike most technological innovations, which diffuse gradually and lack a clearly defined adoption date, ChatGPT represents a single, sharp event that triggered an immediate and widely documented shift in public awareness and tool availability. The second source of variation consists of 27 targeted anti-fraud technology deployments spanning 2013–2023 (see Internet Appendix OA4). These events complement the ChatGPT analysis in two respects. First, they affect specific fraud types rather than the broad landscape of online activity, allowing us to test whether fraudster adaptation is targeted rather than general. Second, because they are distributed across different dates, fraud categories, and deploying institutions, they are not confounded by the large simultaneous shift in public discourse that accompanied ChatGPT's release.

II.A.1 Data and Measurement

CrimeBB Underground Forums Our primary micro-level data come from CrimeBB, a structured corpus of underground forum posts maintained by the Cambridge Cybercrime Centre (Pastrana et al., 2018).[5] CrimeBB contains full-text posts, timestamps, pseudonymous user identifiers, and board-level metadata from several of the largest English-language hacking and fraud forums. We extract posts from five forums: Hackforums, Nulled, Breached, BreachForums, and Cracked, spanning a range of specializations from general hacking discus-

[5]CrimeBB is available to vetted academic researchers through the Cambridge Cybercrime Centre (https://www.cambridgecybercrime.uk). A growing body of computational work has applied natural language processing to underground forum data to measure cybercriminal markets (Portnoff et al., 2017; Bhalerao et al., 2019).

sion and tool sharing to dedicated marketplaces for stolen credentials, carding, and account compromise.

ChatGPT analysis. We define the sample window around the release date of November 30, 2022. The pre-period covers November 2021 through November 2022 (13 monthly bins), and the post-period extends from December 2022 through September 2024 (22 months). The asymmetry reflects data availability: Hackforums, Nulled, and BreachForums snapshots extend to September 2024, while Cracked and Breached end in June 2023.[6] We restrict the sample to English-language, non-gibberish posts of at least five words.[7] Table 1 describes the sample construction. The final sample contains approximately 1.19 million posts by 225,534 unique creators. Hackforums is the largest forum by post volume (359,962 posts), followed by Breached (278,222) and Nulled (263,255). Each forum is organized into topical sub-boards such as Carding, Hacking Tutorials, and Marketplace, which we use as fixed effects. The median creator contributes 1 post over 1.5 months on average; 3.2% of creators post on more than one forum.

Table 1
CrimeBB Sample Construction and Descriptive Statistics

Forum	Raw Posts	Clean Posts	% Kept	Unique Creators	Boards
Breached	737,922	278,222	37.7	47,896	71
Breachforums	332,891	166,796	50.1	28,147	72
Cracked	898,326	122,619	13.6	52,528	104
Hackforums	392,578	359,962	91.7	43,977	142
Nulled	1,385,758	263,255	19.0	60,357	82
Total	3,747,475	1,190,854	31.8	225,534	471

Notes: Posts from CrimeBB PostgreSQL snapshots (Cambridge Cybercrime Centre). Sample window: 13 months pre to 22 months post ChatGPT launch (Nov 2021–Sep 2024). Clean posts: English-language, non-gibberish, $\geq$5 words. Hackforums and Nulled snapshots extend to Sep 2024; Cracked and Breached snapshots end Jun 2023.

Measures We construct five binary indicators at the post level that capture, with increasing specificity, the intersection of generative AI and fraud discussion. The measures are

[6] All regressions include forum-board and creator fixed effects, which absorb level differences arising from differential coverage.

[7] Some gibberish-classified posts may represent coded or obfuscated communication among experienced forum participants. Because our identification relies on dictionary-based keyword matching, such posts would not match any dependent variable terms regardless of inclusion, so their exclusion tightens estimation without removing informative observations. To the extent that sophisticated adaptation shifts toward coded communication channels, our estimates are a lower bound on total adaptation.

designed to distinguish genuine AI–fraud integration from a general increase in AI awareness following ChatGPT's release.

The broadest measure, $\mathbf{1}(\text{GenAI})$, equals one if a post mentions any generative AI term, including ChatGPT, GPT, LLM, deepfake, or Stable Diffusion, and captures general AI awareness without requiring fraud-related content. $\mathbf{1}(\text{Fraud})$ equals one if a post mentions any fraud-related term such as phishing, carding, ransomware, or credential stuffing, and serves as a placebo: if ChatGPT merely increased AI discussion without affecting fraud behavior, this measure should be unaffected.

The three remaining measures capture the intersection of AI and fraud discussion with increasing precision. $\mathbf{1}(\text{GenAI} \times \text{Fraud})$ equals one if a post contains both a GenAI and a fraud term. $\mathbf{1}(\text{Fraud Method})$ equals one if a post mentions a specific fraud method or tool, such as a phishing kit, fullz, CVV, or combo list, rather than fraud in general. $\mathbf{1}(\text{GenAI} \times \text{Method})$ equals one if a post contains both a GenAI term and a specific fraud method term, and is the most restrictive measure. A finding in which only $\mathbf{1}(\text{GenAI})$ increases after ChatGPT's release would indicate that the shock raised AI awareness without penetrating into fraud operations; a finding in which the intersection measures also increase would indicate integration of AI into concrete fraud activity. The complete keyword lists are reported in Internet Appendix Table OA5.

To test whether the ChatGPT shock differentially affected creators of varying skill levels, we construct a pre-period technical sophistication score for each creator. The score counts pre-period posts containing at least one term from a curated dictionary of 212 technical terms spanning 13 categories: exploitation and vulnerability research, malware development, reverse engineering, network attacks, OPSEC, and others.[8] Because the score is computed entirely from pre-period posts, it is a fixed creator attribute immune to contamination by the ChatGPT shock.

The distribution is heavily right-skewed: 97.1% of pre-period creators have zero sophistication posts, reflecting the concentration of technical expertise among a small core of underground forum participants. Among the 4,332 creators with at least one sophistication post, the median is 1 and the 90th percentile is 3. We use both the continuous measure $\log(1 + \text{sophistication posts})$ and a binary indicator $\text{TechSoph}^{\text{hi}}$ in the interaction specifications.

[8]The full dictionary is in Internet Appendix OA1. Categories covering deepfakes and AI-generated content are excluded to avoid overlap with the dependent variables.

II.A.2 Descriptive Statistics

Table 2 reports the prevalence of each dependent variable before and after ChatGPT's release. In the pre-period, only 0.11% of posts mention a GenAI term; post-ChatGPT, this rises to 0.79%, a seven-fold increase. More importantly, the intersection measures also surge: $\mathbf{1}(\text{GenAI} \times \text{Fraud})$ rises from 0.023% to 0.133%, a 5.8-fold increase, indicating that AI adoption on these forums penetrates into fraud-specific discussion. Meanwhile, $\mathbf{1}(\text{Fraud})$ remains roughly stable at approximately 6.3%, suggesting that the overall volume of fraud discussion did not change but its technological composition did. At the creator level, 1.1% of creators ever mention a GenAI term, while 9.9% mention a fraud term; 582 creators (0.3%) account for all posts in which AI and fraud co-occur.

That fraud terms appear in only 6.3% of posts reflects the nature of the forums rather than a shortage of criminal content. The five forums are general-purpose hacking and cracking communities: the bulk of posts concern hacking tutorials, account cracking, tool and configuration sharing, marketplace transactions, and community chatter, with fraud a specialized submarket concentrated on dedicated boards such as Carding. Our fraud dictionary is also deliberately narrow, requiring an explicit fraud term rather than inferring intent from context. Both features work against us: they compress the base rates of the dependent variables and make the estimates conservative. Forum-board fixed effects absorb the resulting cross-board heterogeneity in baseline fraud intensity.

Table 2
Fraud Discussion Before and After ChatGPT

	Pre-ChatGPT		Post-ChatGPT	
Variable	Count	%	Count	%
$\mathbf{1}$(GenAI)	766	0.11	4,050	0.79
$\mathbf{1}$(Fraud)	42,943	6.34	32,080	6.25
$\mathbf{1}$(GenAI × Fraud)	157	0.02	685	0.13
$\mathbf{1}$(Tech Sophistication)	9,167	1.35	9,270	1.81
Posts	677,451		513,403	
Unique creators	148,122		95,453	

Notes: Pre-ChatGPT: Nov 2021–Nov 2022. Post-ChatGPT: Dec 2022 onward. Each indicator equals 1 if the post contains at least one matching term from the respective dictionary. GenAI × Fraud requires co-occurrence of both AI and fraud terms in the same post.

II.A.3 Empirical Strategy

The release of ChatGPT on November 30, 2022 provides a single, precisely dated shock to the availability of generative AI tools. ChatGPT reached over 100 million users within weeks

of its release, making sophisticated language generation, code writing, and content creation accessible to anyone with an internet connection, including participants in underground fraud forums. Therefore, the ChatGPT shock is a general-purpose enabling shock that lowers the skill barrier for multiple fraud types simultaneously.

We estimate the effect of this shock on the five dependent variables using within-creator variation. The main specification is a linear probability model with creator fixed effects:

$$y_{it} = \alpha_i + \beta \cdot \text{After}_t + \gamma \cdot X_{it} + \varepsilon_{it}, \tag{1}$$

where $y_{it} \in \{0, 1\}$ is one of the five dependent variables for creator i in post t, α_i are creator fixed effects, $\text{After}_t = \mathbf{1}[t \geq \text{Dec 2022}]$, and X_{it} includes an indicator for thread-initiating posts and the log of the creator's pre-period posting count (the tables' "log reputation"); the latter is time-invariant and therefore absorbed wherever creator fixed effects are included. The dependent variable is standardized to zero mean and unit standard deviation, so $\hat{\beta}$ is interpretable in standard deviation units; multiplying by $\text{SD}(y)$ recovers the raw probability change. Standard errors are clustered at the creator level. Creator fixed effects ensure identification comes from within-creator changes in posting behavior, not from compositional shifts in who is active on the forums.[9]

The main threat to identification is confounding events coinciding with the ChatGPT release. We address this in three ways: event-study plots documenting flat pre-trends, creator fixed effects absorbing all time-invariant heterogeneity, and interaction specifications testing whether effects vary with pre-existing creator characteristics. These checks strengthen the attribution but cannot rule out an unobserved aggregate shock arriving at the same date; the estimates carry a causal interpretation under the assumption that ChatGPT was the dominant such shock to these outcomes.

II.A.4 Main Results

Table 3 reports the main results. The coefficient on **1**(GenAI) is large and highly significant, while the coefficients on **1**(Fraud) and **1**(Method) are small and statistically indistinguishable from zero. Crucially, both intersection measures are positive and significant. Column (1) shows a broad increase in AI-related discussion around the ChatGPT release, consistent with the shift in public discourse that accompanied the launch. However, ChatGPT did not increase general fraud discussion, but it did increase discussion that combines AI with fraud

[9]The identifying assumption is that, conditional on creator fixed effects, ChatGPT is the dominant shock to AI-related fraud discussion over this period. Earlier models such as GPT-3 (June 2020) and DALL-E 2 (April 2022) attracted attention but achieved nothing comparable in terms of mass adoption. Google Bard and Meta's LLaMA both post-date our treatment cutoff.

contexts and with specific fraud methods. The effect sizes are economically meaningful: AI–fraud co-discussion roughly doubles relative to its sample-average rate, while overall fraud discussion is unchanged. Wald tests (Internet Appendix Table OA3) confirm that the AI–fraud intersection grew significantly faster than overall fraud discussion, and the AI–method intersection faster than method discussion generally: the rise in co-discussion is not a mechanical reflection of trends in fraud discussion itself. General AI discussion grew faster still, as one would expect of a shock to AI itself; the intersections' growth indicates that part of that new AI attention landed in concrete fraud contexts.

Economic magnitudes. The base rates in Table 2 are small: AI–fraud co-discussion rises from 0.023% to 0.133% of posts, so the within-creator +97% (measured against the sample-average rate; the rise relative to the pre-period base rate is larger still) is off a low baseline. Three considerations govern how to read these magnitudes. First, the measurement is deliberately conservative. A post counts only if it contains both an explicit generative-AI term and an explicit fraud term, so paraphrased, coded, or private-channel discussion is excluded, and the measured response is a lower bound on adaptation. Second, the level of discussion is not the economically relevant margin; its concentration and reuse are. All AI–fraud co-occurrence originates with 582 creators, 0.3% of the panel, and the heterogeneity results below show technically sophisticated creators moving first with the effect subsequently broadening. Innovation on these forums works through exactly this structure: a small technical fringe develops and documents a technique, which then diffuses through tutorials, kits, and marketplace listings and can be reused at scale without generating further discussion. The economically relevant fact is that the frontier moved, not that the average post did. Third, the object the model requires from this evidence is a direction, not a level. The composition of criminal attention shifts toward the new technology—$\mathbf{1}(\text{GenAI} \times \text{Fraud})$ rises while total fraud discussion is flat—and it is this substitution, not its size, that the adaptation mechanism of Proposition 1 builds on; the aggregate, outcome-level counterpart of this shift appears in Section II.A.7.

Table 3
The Effect of ChatGPT on Underground Forum Discussion

	1(AI)	**1**(Fraud)	**1**(AI×Fr)	**1**(Method)	**1**(AI×M)
after_chatgpt	0.0960 * **	−0.0096	0.0254 * **	−0.0047	0.0138 * **
	(0.0088)	(0.0122)	(0.0049)	(0.0065)	(0.0044)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Creator	Creator	Creator	Creator	Creator
N	1,190,854	1,190,854	1,190,854	1,190,854	1,190,854
R^2	0.2043	0.5019	0.1897	0.3288	0.1896
Mean(y)	0.0040	0.0631	0.0007	0.0196	0.0003
SD(y)	0.0635	0.2431	0.0266	0.1387	0.0173

Notes: Sample: All boards. Y standardised to zero mean and unit standard deviation; X in natural units. $\hat{\beta}$ = effect of 0→1 change in SD(y) units. Creator FE + controls (initial post indicator, log reputation). Clustered standard errors (creator) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

II.A.5 Heterogeneity: Experience and Technical Sophistication

To test the skill-barrier-lowering mechanism, we interact After_t with two pre-determined creator characteristics. Table 4 reports the experience interaction. More experienced creators adopted AI terminology faster, but less experienced creators drove the increase in fraud method discussion. This pattern is consistent with a diffusion dynamic in which technically literate creators were early adopters, while ChatGPT provided the greatest marginal benefit to creators who previously lacked the expertise to generate convincing content or code.

Table 4
Early Adopters: The Role of Forum Experience

	1(AI)	**1**(Fraud)	**1**(AI×Fr)	**1**(Method)	**1**(AI×M)
after_chatgpt	0.0634 * **	0.0207	0.0186 * **	0.0279 * *	0.0202 * **
	(0.0105)	(0.0137)	(0.0071)	(0.0129)	(0.0074)
after × log_rep	0.0067 * **	−0.0063	0.0014	−0.0068 * **	−0.0013
	(0.0024)	(0.0043)	(0.0013)	(0.0024)	(0.0015)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Creator	Creator	Creator	Creator	Creator
N	1,190,854	1,190,854	1,190,854	1,190,854	1,190,854
R^2	0.2043	0.5019	0.1897	0.3288	0.1896
Mean(y)	0.0040	0.0631	0.0007	0.0196	0.0003
SD(y)	0.0635	0.2431	0.0266	0.1387	0.0173

Notes: Sample: All boards. after × log_rep = after_chatgpt × log(1+pre-period posts). Y standardised to zero mean and unit standard deviation; X in natural units. Creator FE + controls (initial post indicator, log reputation). Clustered standard errors (creator) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5 reports the interaction with pre-period technical sophistication. Technically so-

phisticated creators drove both AI adoption and AI–fraud co-discussion. For pure fraud and method discussion the sophistication gradient runs the other way: the interaction terms are negative, so the post-release movement in these outcomes, imprecisely estimated in levels, is relatively concentrated among less sophisticated creators. Together, the two interaction tests characterize the distributional incidence of the shock: sophisticated creators were first movers in integrating AI into fraud applications, while fraud-method discussion shifted relatively toward less skilled participants, for whom ChatGPT lowered entry barriers.

Table 5
Early Adopters: The Role of Technical Sophistication

	1(AI)	**1**(Fraud)	**1**(AI×Fr)	**1**(Method)	**1**(AI×M)
after_chatgpt	0.0767 * **	−0.0002	0.0163 * **	0.0099	0.0112 * **
	(0.0074)	(0.0112)	(0.0047)	(0.0067)	(0.0041)
after × log_tech_soph	0.0164 * **	−0.0080*	0.0077 * *	−0.0124*	0.0022
	(0.0060)	(0.0045)	(0.0038)	(0.0065)	(0.0046)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Creator	Creator	Creator	Creator	Creator
N	1,190,854	1,190,854	1,190,854	1,190,854	1,190,854
R^2	0.2044	0.5019	0.1897	0.3289	0.1896
Mean(y)	0.0040	0.0631	0.0007	0.0196	0.0003
SD(y)	0.0635	0.2431	0.0266	0.1387	0.0173

Notes: Sample: All boards. after × log_tech_soph = after_chatgpt × log(1 + pre-period tech-sophistication posts). Y standardised to zero mean and unit standard deviation; X in natural units. Creator FE + controls. Clustered standard errors (creator) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

II.A.6 Event Study

Figure 2 plots event-study coefficients from a variant of equation (1) that replaces After_t with month-by-month dummies relative to December 2022. Pre-period coefficients are close to zero and statistically indistinguishable from the reference period, indicating no pre-existing trend in fraud- and AI-related discussion prior to the ChatGPT release. Post-release coefficients show a sharp break at the treatment date, rising to a peak around months +4 to +7 before partially moderating; the coefficients remain positive over the remainder of the sample, though precision falls toward the endpoint. The results are consistent with rapid initial adoption followed by sustained diffusion through the underground community.

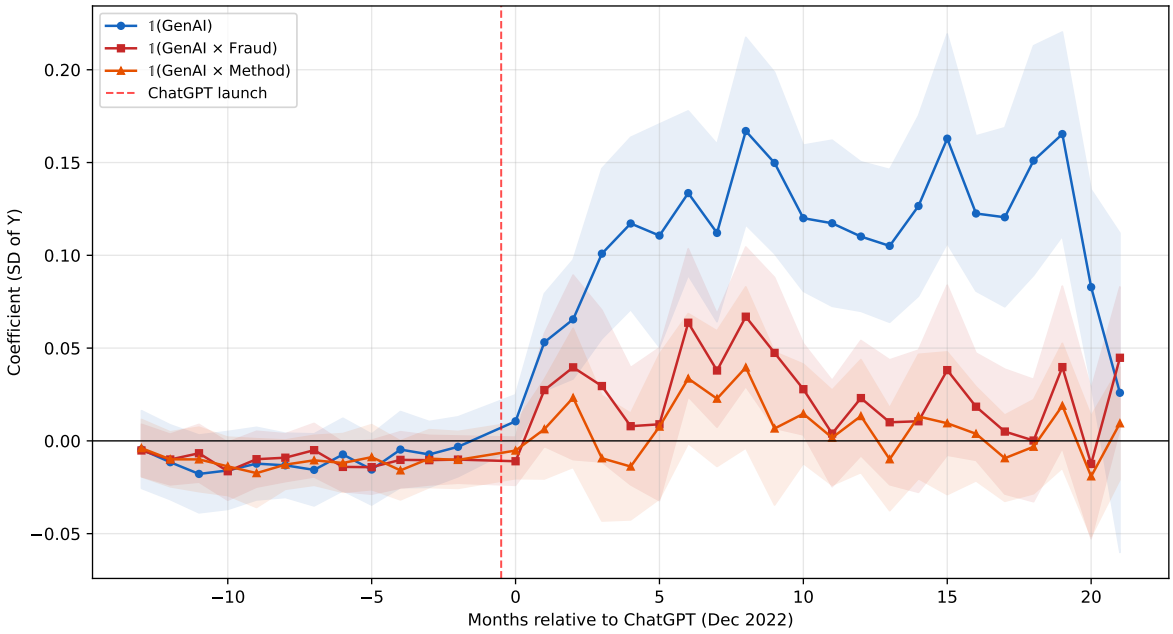


Figure 2
AI–Fraud Discussion Around the ChatGPT Release

Notes: Event-study coefficients from equation (1) with month-by-month dummies relative to December 2022 ($m = -1$ as reference). The dependent variable is standardized to zero mean and unit standard deviation. Shaded bands indicate 95% confidence intervals based on creator-clustered standard errors.

II.A.7 Corroborating Evidence

The ChatGPT analysis identifies a general-purpose mechanism, but it rests on a single shock. We corroborate this evidence with a stacked event-study design built around 27 targeted anti-fraud technology deployments spanning 2013–2023 and nine fraud categories in Online Appendix OA4, from device fingerprinting and behavioral biometrics to 3D Secure 2.0, SIM-swap defenses, and blockchain analytics (full technology list, deployment dates, and keyword dictionaries in Internet Appendix Tables OA4 and OA5). For each deployment, we score the full CrimeBB corpus against technology-specific vulnerable-method and countermeasure dictionaries and estimate a stacked event study pooling all 27 deployments in a single regression, with creator-by-technology fixed effects, seasonal or calendar-month time controls, and standard errors clustered by technology (methodology and full results in Internet Appendix OA4).

The estimates trace a slow reallocation of criminal attention over the deployment cycle. Countermeasure discussion builds over the three years *before* a defense's deployment date and peaks around the date itself, consistent with deployment dates marking the matura-

tion of defenses that diffuse gradually rather than surprise interventions; discussion of the targeted vulnerable methods fades over the years that follow. The composition of underground attention thus shifts from established methods toward circumvention over horizons of years—Whac-A-Mole at the level of criminal knowledge, and the targeted-technology counterpart of the general-purpose adaptation documented around ChatGPT—though we read these profiles as descriptive: deployment timing is not random, and the deployment-specific component, identified from the staggering of dates across technologies, is positive at long horizons but imprecisely estimated. The contrast with Bian et al. (2024) remains instructive. Using the same corpus, they find that restricting fraudsters' *inputs*—a privacy change cutting off access to consumer data—produces a net drop in fraud and identity-theft discussion; around detection deployments, criminal attention instead reallocates toward circumvention: starving fraud of inputs suppresses it, while defending against its methods redirects it.

Fraudster adaptation of this kind has two aggregate implications: the composition of fraud types shifts over time as fraudsters substitute toward harder-to-detect methods, and detection model performance decays as the attack distribution diverges from the training data. FBI IC3 and FTC Sentinel complaint data in Online Appendix OA2 show that the composition of fraud types shifts markedly over time: phishing and impersonation fraud grow steadily by complaint count, while investment fraud dominates financial losses and rises sharply after 2020. A prominent example of composition-driven substitution is EMV chip card adoption: Hayashi et al. (2015) documents that card-not-present fraud rose sharply in the United Kingdom, France, and Canada immediately following EMV deployment, as fraudsters substituted toward online channels where chip authentication offered no protection. These outcome-level shifts are the aggregate footprint of the within-forum displacement documented above: defenses that close one category of fraud move activity into the categories they do not cover.

II.B Fraud and Debanking: Evidence from Bitcoin ATMs

Fraud imposes welfare costs through two channels. The first is direct: consumers lose money to fraudsters. The second is indirect and less visible: payment providers respond to fraud exposure by cutting consumers off. That response has two margins. The sharpest is debanking, the unilateral closure of an existing account; its quieter counterpart is the refusal to open a new one. The first channel is straightforward to state, if hard to measure precisely. The second is the one our model speaks to and the one that is easy to miss, so we document both margins in detail. As in Section II.A, no single dataset observes the whole picture, so we piece it together from disparate and imperfect sources.

II.B.1 Direct Losses to Consumers

The direct cost of fraud is borne by the victim who loses the transferred funds. In the United States, reported consumer fraud losses run into the tens of billions of dollars annually and have risen sharply since 2020, with authorised push-payment scams, in which the victim is deceived into sending the money, the fastest-growing category (FTC Consumer Sentinel; FBI IC3). Because these payments are authorised by the consumer, they typically fall outside the reimbursement protections that cover unauthorised transactions, so the loss stays with the victim unless a liability rule reallocates it. This is the channel that liability regulation targets directly, and it is the starting point for the model in Section III.

II.B.2 Debanking: Fraud Exposure and Account Closure

The second channel operates through the provider's response. When fraud exposure rises, a provider that bears part of the loss has an incentive to shed the accounts it associates with that exposure[10]. The result is debanking. Unlike a fraud loss, a closure is rarely explained and almost never recorded as fraud-related, so it is hard to observe directly. We recover it from a pair of datasets, and read the two through a single chain from local scam exposure to account closure. This is the empirical counterpart to the debanking margin the model develops in Section V, where a provider excludes a consumer once the liability it bears on that account exceeds the franchise value of keeping it.

Two sources, one chain. The first source is the near universe of U.S. Bitcoin ATMs from Coin ATM Radar, roughly 19,850 machines, each with a street address and an install date. Bitcoin ATMs are of interest here not as a financial technology but as the cash-collection rail for a large class of consumer scams: victims of romance, "pig-butchering," tech-support, and government-impersonation scams are routinely instructed to feed cash into a nearby machine. The FTC singles the machines out as "a payment portal for scammers": reported scam losses paid through Bitcoin ATMs rose nearly tenfold between 2020 and 2023, topping $65 million in the first half of 2024 alone, with consumers over 60 accounting for the large majority of losses (Fletcher, 2024), and California now caps kiosk transactions at $1,000 per

[10]Little of this loss share is statutory in the United States: because the victim authorises the payment, formal reimbursement obligations rarely apply. Providers nonetheless bear a material *effective* share of every completed scam—dispute handling and investigation costs, Regulation E exposure when authorisation is contested, anti-money-laundering compliance and enforcement risk, recall and litigation costs, and the franchise value lost when defrauded customers leave—so the liability parameter γ of Section III may also be read as this effective share: low but positive in the United States, close to one under the UK's mandated reimbursement. This is why US banks invest in detection, and act on the closure margin documented here, despite bearing little formal liability.

person per day (California DFPI, 2026). A new machine is, in effect, a local shock to the feasibility of cash-based scams. Concurrent work exploits the same rollout at the county level and confirms its criminogenic bite: machine entry raises local cybercrime and financial fraud (Cong et al., 2025) and total crime by 6.9%, concentrated in offences that exploit anonymity and cash-to-crypto conversion (Wei, 2026). Neither studies the bank response—closures and denials—which is the margin here. The second source is the CFPB Consumer Complaint Database, from which we build five counts at the ZIP × quarter level over 2018–2026. Three concern existing accounts: *scam* complaints (the consumer reports being defrauded), *debanking* complaints (the institution closed the account, the involuntary-closure sub-issue), and *scam-cited debanking* (a closure whose narrative mentions a scam). Two concern the opening margin, taken up in Section II.B.3: *unable-to-open* complaints (the deposit sub-issue "Unable to open an account") and *application-denied* complaints (its credit-card analogue). Neither source is a census. CFPB records only complaints, so throughout we speak of the observable complaint margin, not population rates.

Design. The unit is a five-digit ZIP; the period is a quarter (2018Q1–2026Q1); the cohort is the quarter of a ZIP's first ATM install; the comparison group is not-yet-treated ZIPs. Machine placement is not random—operators favour cash-intensive, underbanked neighbourhoods—but identification uses only the timing of entry against not-yet-treated ZIPs, and the event studies below show flat pre-trends. Flat leads support, but cannot by themselves establish, the conditional parallel-trends assumption this reading requires; a host-venue instrument in the spirit of Wei (2026) is a natural robustness extension. We estimate group-time average treatment effects with the staggered difference-in-differences estimator of Callaway and Sant'Anna (2021), which is robust to the heterogeneous-timing bias that afflicts two-way fixed effects under staggered adoption. Standard errors are the analytic Callaway–Sant'Anna standard errors clustered on the ZIP.[11]

Scam exposure and account closures. Table 6 and Figure 3 report the results. A ZIP's first install raises scam complaints by +0.064 per ZIP-quarter ($p < 0.01$), a 137% increase over the pre-install mean, with flat pre-trends and a jump within one quarter of the install. The same install raises account closures by +0.053 per ZIP-quarter ($p < 0.01$), but with different dynamics: the closure response is flat for about a year and only then

[11]The results are not estimator-specific. A static two-way fixed-effects OLS of each outcome on a post-install indicator, with ZIP and quarter fixed effects and standard errors two-way clustered by ZIP and quarter, reproduces the headline effects almost exactly: scam +0.067 against a Callaway–Sant'Anna ATT of +0.065, debanking +0.054 against +0.053, card denials +0.015 against +0.014. With never-treated ZIPs two-thirds of the sample, the already-treated comparisons that bias two-way fixed effects under staggered adoption carry little weight.

Table 6

Bitcoin ATMs, Scam Complaints, and Debanking (Callaway–Sant'Anna ATT)

Treatment is the quarter of a zip code's first Bitcoin-ATM install; the control group is not-yet-treated zips. The coefficient is the Callaway–Sant'Anna overall ATT (complaints per zip-quarter). 'Implied scam effect' is the Wald/IV scam → outcome effect (the ATM instrumenting scam); brackets are a 95% CI from a block bootstrap over zip codes (FS/RF correlation included).

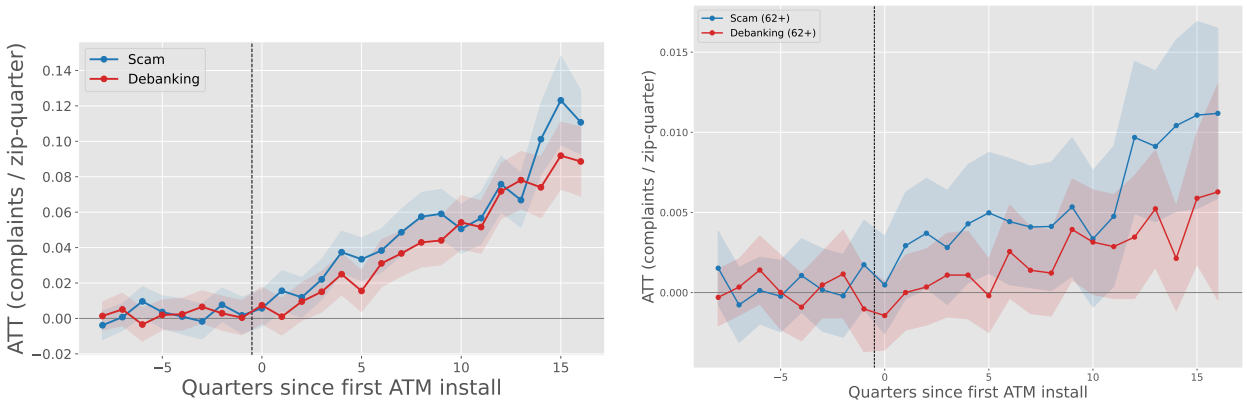
	Scam	Debanking	Debanking (scam victim)
Bitcoin ATM (ATT)	0.0645***	0.0534***	0.0026***
	(0.0039)	(0.0044)	(0.0004)
Baseline mean	0.0470	0.0510	0.0015
Implied scam effect		+0.828 [+0.68, +0.99]	+0.040 [+0.03, +0.05]
N	734976	734976	734976

Unit: zip code × quarter, 2018–2026. CS analytic SE (clustered by zip) in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

climbs, the signature of a downstream, screening-driven response rather than a mechanical contemporaneous correlation.

Falsification: scam victims, not crypto users. The leading threat to reading this chain causally is that ATMs proxy local crypto adoption and banks de-risk crypto users directly. We therefore split every outcome by the CFPB "Older American" tag and re-estimate. The debanking response is positive and significant for consumers aged 62+ (+0.0023, $p < 0.05$; Table 7 and Figure 3b), a group far more likely to be a scam victim than a crypto user. This addresses the leading alternative explanation on the one margin where it can be directly tested, though the aggregate closure response may still contain some direct crypto-related de-risking.

Reading the magnitude. If one is willing to treat the ATM rollout as an instrument for scam exposure, the just-identified Wald ratio (reduced form over first stage) implies that each additional ATM-induced scam complaint is accompanied by roughly 0.8 additional account closures, with a 95% confidence interval of $[0.68, 0.99]$ from a block bootstrap over ZIPs (de Chaisemartin and D'Haultfœuille, 2018). We report this magnitude as illustrative rather than as a clean local average treatment effect. The exclusion restriction is not proven: some closures among younger consumers may reflect direct de-risking or the closure of money-mule accounts, and the elderly tag is self-reported and undercounts, so the 62+ result is best read as a lower bound that rules out the dominant confound rather than as point identification. The robust content of the exercise is the reduced form: ATM entry is followed, with a lag, by more account closures.



(a) All consumers **(b)** Older Americans (62+)

Figure 3

Event-study ATTs of a ZIP's first Bitcoin-ATM install (Callaway–Sant'Anna; not-yet-treated controls). Panel (a): scam and debanking complaints, all consumers; the scam-cited-debanking response is reported as an overall ATT in Table 6. Panel (b): scam and debanking for Older Americans (62+) only. Scam responds within a quarter; debanking follows with a roughly one-year lag; pre-trends are flat, and the chain holds for a group that cannot plausibly be crypto users.

II.B.3 The Extensive Margin: Denial of New Accounts

Closure is one arm of the response; the other is refusing to bank a consumer in the first place. The distinction matters for welfare. A consumer whose account is closed but who can reopen across the street loses little; one who is also turned away at the door loses access to the payment system altogether, which is the financial-inclusion cost at the centre of Sections V and VII.D. The two margins are also institutionally linked: closures are reported to deposit-screening agencies such as ChexSystems and Early Warning Services—shared registries, the deposit-side counterpart of the mule registries of Section VI—so one institution's closure

Table 7

Age Heterogeneity: Bitcoin ATMs, Scams, and Debanking by Older-American Tag

Outcomes are complaint counts split by the CFPB 'Older American' tag (consumer reported as 62+); the coefficient is the Callaway–Sant'Anna overall ATT and the control group is not-yet-treated zips. A positive debanking effect for the 62+ supports the scam channel (the elderly are scam victims, not crypto users)—a test of the IV exclusion restriction.

	Scam (62+)	Scam (<62)	Debanking (62+)	Debanking (<62)
Bitcoin ATM (ATT)	0.0061***	0.0584***	0.0023**	0.0510***
	(0.0011)	(0.0037)	(0.0010)	(0.0043)
Baseline mean	0.0037	0.0433	0.0030	0.0479
N	734976	734976	734976	734976

Unit: zip code × quarter, 2018–2026. CS analytic SE (clustered by zip) in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8

Bitcoin ATMs and Account-Opening Denials (Callaway–Sant'Anna ATT)

Treatment is the quarter of a zip code's first Bitcoin-ATM install; the control group is not-yet-treated zips. Outcomes are the two account-opening denial measures—deposit 'unable to open an account' and card 'application denied'—for all consumers and for the 62+ (CFPB 'Older American' tag). The coefficient is the Callaway–Sant'Anna overall ATT (complaints per zip-quarter).

	Unable open	App. denied	Unable (62+)	Denied (62+)
Bitcoin ATM (ATT)	0.0203***	0.0141***	0.0004	-0.0006
	(0.0020)	(0.0027)	(0.0004)	(0.0008)
Baseline mean	0.0118	0.0174	0.0005	0.0012
N	734976	734976	734976	734976

Unit: zip code × quarter, 2018–2026. CS analytic SE (clustered by zip) in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

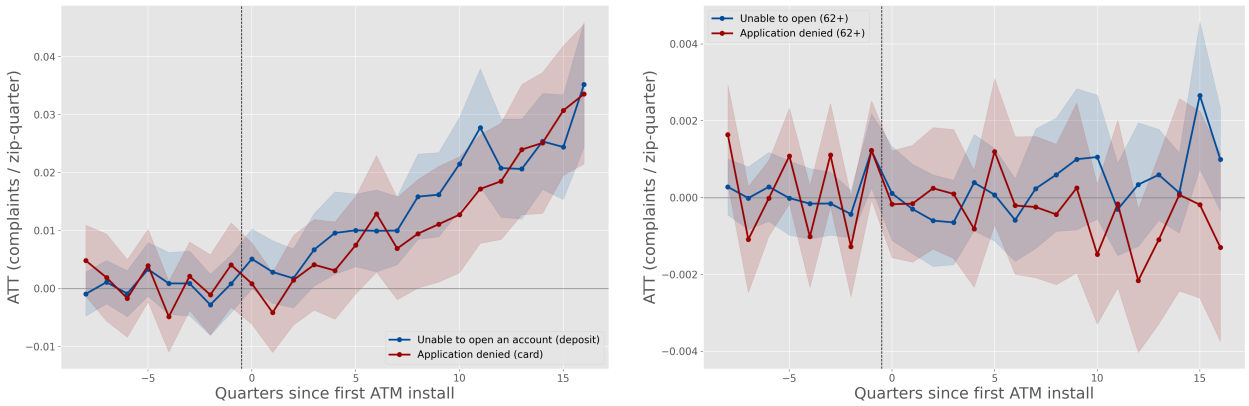
decision feeds every institution's opening decision.[12]

We therefore repeat the design on the two account-opening outcomes, estimated separately. Table 8 and Figure 4 report the results. A ZIP's first install raises deposit unable-to-open complaints by +0.0203 per ZIP-quarter ($p < 0.01$, a 173% increase over the pre-install mean) and card application-denied complaints by +0.0141 ($p < 0.01$, +81%), each with flat pre-trends and the same roughly one-year lag before the effect climbs—again the timing signature of a screening-driven response.

The age falsification of the closure margin has no power here: for the 62+ both responses are null (+0.0004 and −0.0006, n.s.) off near-zero baselines of 0.0005 and 0.0012 complaints per ZIP-quarter (Figure 4b). The elderly rarely apply for new accounts or cards, so the test is uninformative here rather than failed; the denial effects are concentrated among the young (+0.0199 and +0.0146, both $p < 0.01$), who do the applying. The causal claim thus rests on the closure margin above; we read the denial responses as suggestive rather than separately identified. What the extensive margin adds is breadth: providers respond to ATM entry both by closing existing accounts and by refusing new ones—exclusion from the payment system, not merely account churn. Of the two outcomes, deposit unable-to-open complaints are the closer analogue of exclusion from payment services; card application denials may partly reflect ordinary credit underwriting.

These findings motivate the theory that follows. In Section V a PSP closes an account once the fraud liability it carries exceeds the account's franchise value; in Section VII.D the same calculus, applied before an account is opened, prices fraud-prone consumers out of the system entirely.

[12]Consistent with this mechanism, CFPB complaints against the two screening agencies rise sharply after ATM entry in the same design (not reported).



(a) All consumers **(b)** Older Americans (62+)

Figure 4

Event-study ATTs of a ZIP's first Bitcoin-ATM install on the two account-opening denial measures, estimated separately: deposit "unable to open an account" and credit-card "application denied" (Callaway–Sant'Anna; not-yet-treated controls). Panel (a): all consumers—both margins show flat pre-trends and the same lagged post-install climb as debanking. Panel (b): the 62+ only—both flat off near-zero baselines; the elderly rarely apply for new accounts, so the age falsification is uninformative on the opening margin.

III The Model

Two types of agents interact over two periods: N payment service providers (PSPs) and a single strategic fraudster. All parties are risk neutral with a zero discount rate. [13] Consumers are not active players in the game. To isolate PSPs' investment incentives, we abstract from product market competition and consumer behavior in the baseline model.[14]

Transactions and fraud. Total transaction volume is normalised to one and split evenly across PSPs, so each PSP processes a mass $1/N$ of transactions per period. A share $\alpha \in (0,1)$ of transactions is targeted by the fraudster each period. Let $V > 0$ denote the loss per successful fraud attempt, which is also the fraudster's revenue.

Fraud technology. There are two fraud types. Type A succeeds at PSP i with probability $1 - I_{i0}$, where $I_{i0} \in [0,1]$ denotes detection intensity. Type B succeeds with probability $1 - \theta I_{i,0}$, where $\theta \in [0,1)$ measures the relative detectability of Type B. Lower θ means

[13] Risk neutrality and zero discounting are adopted for tractability, not because the qualitative mechanism requires them. Risk-averse PSPs would invest more aggressively in detection, which strengthens the overinvestment result under full liability; a positive discount rate attenuates but does not eliminate the model-decay externality

[14] We relax this in Section VII.C, where consumers and platforms, in addition to PSPs, can invest in fraud detection.

Type B is harder to detect; at $\theta = 0$ it is entirely undetectable. We treat θ as a property of the information environment rather than a margin any individual PSP controls: it indexes how observable the *adapted* fraud pattern is to the detection technology available to the industry at large, for instance whether it leaves a distinguishable transaction signature, and is shaped by industry-wide arrangements such as the information sharing analyzed in Section VI rather than by any single PSP's investment.

Detection cost Investment $I_{i,0}$ costs $c(I_{i,0})$, where $c : [0,1] \to \mathbb{R}_+$ is twice continuously differentiable, strictly increasing, and strictly convex, with $c(0) = 0$ and $c'(0) = 0$. The boundary condition $c'(0) = 0$ ensures PSPs always find it worthwhile to invest a strictly positive amount, so equilibria are interior. Unlike the adaptation-cost function k introduced below, c is never inverted in the analysis and enters the proofs only through c' and c''; C^2 regularity is therefore sufficient for c, whereas k requires one additional derivative because the fraudster's best response is characterized through $g \equiv (k')^{-1}$, and the uniqueness argument in Appendix C differentiates g itself.

We take $I_{i,0}$ as fixed for the entire two-period horizon, with no opportunity for PSPs to revise it after observing the fraudster's period-1 adaptation. It reflects the fact that fraud-detection systems are lumpy, model-based investments, involving staff training, vendor contracts, with redeployment lags that are long relative to the speed at which a fraudster can shift its script once detection is observed. Section VII.A relaxes this assumption directly: PSPs are allowed a second, costly reinvestment at $t = 1$ that fully restores detection against the fraudster's adapted attack.

Fraudster adaptation. In period 0 the fraudster employs only Type A fraud.

At the end of period 0, after observing average detection

$$\bar{I}_0 \;=\; \frac{1}{N}\sum_{i=1}^{N} I_{i,0},$$

it chooses $p_1 \in [0,1]$, the period-1 share of Type A fraud, shifting the remainder $1 - p_1$ to Type B. We interpret Type A as the fraud technique currently in circulation and observed by PSPs (e.g., a known social-engineering script), and Type B as an adapted variant that does not yet exist at $t = 0$. [15] Adapting away from the legacy script costs $k(1 - p_1)$, where $k : [0,1] \to \mathbb{R}_+$ is three times continuously differentiable, strictly increasing, and strictly convex, with $k(0) = 0$ and $k'(0) = 0$, and with $\overline{k'''} \equiv \max_{x\in[0,1]} k'''(x) < \infty$. As with c,

[15] In Internet Appendix OA6 we consider an extension in which the fraudster observes individual detection intensities $I_{i,0}$ and chooses, after period 0, whether to direct its attacks at customers of PSP 1 or PSP 2.

the condition $k'(0) = 0$ ensures the adaptation response is always interior; the additional third-derivative bound is used only in the proof of Proposition 1(i) to control a second-order term in the uniqueness argument, and is automatically satisfied (with $\overline{k'''} = 0$) under the quadratic specification of Corollary 1.

Liability. The parameter $\gamma \in [0, 1]$ denotes the PSP's liability share: for every unit of fraud loss the PSP reimburses fraction γ to the victim.

PSP objective. We assume PSP i earns exogenous surplus $u_{\text{psp}} > 0$ from serving its share of the market, the mass $1/N$ of consumers. PSP i's two-period objective function is

$$\Pi_i \;=\; u_{\text{psp}} - \frac{\gamma\alpha V}{N}\big(\phi_{i,0} + \phi_{i,1}\big) - c(I_{i,0}), \tag{2}$$

where $\phi_{i,t} \in [0, 1]$ is PSP i's period-t fraud incidence rate. In the baseline model every PSP always serves the same, homogeneous mass $1/N$ of transactions, so u_{psp} is a constant that does not vary with $I_{i,0}$: maximizing Π_i over $I_{i,0}$ is therefore equivalent to minimizing the two-period loss

$$L_i \;=\; \frac{\gamma\alpha V}{N}\big(\phi_{i,0} + \phi_{i,1}\big) + c(I_{i,0}), \qquad \phi_{i,0} = 1 - I_{i,0}, \tag{3}$$

Once consumers are heterogeneous, u_{psp} no longer drops out as a constant, and it reappears alongside consumer surplus u_{con} in the relevant objective. We relax the fixed-participation assumption along two distinct margins in extensions to the baseline model. Section VII.D lets each PSP unilaterally decide whether to debank a given customer; consumers remain entirely passive there, as in the baseline, so the margin is supply-side. The Online Appendix instead lets participation respond on the demand side. Section OA7 allows each PSP's market share to depend on its detection quality relative to rivals, reallocating volume across PSPs rather than excluding consumers from the payment system.

Timing. At $t = 0$: the regulator sets γ; PSPs simultaneously choose detection intensities $I_{i,0}$. A share α of transactions are targeted. We assume targeted consumers authorise the payment, consistent with the APP fraud setting in which victims are deceived into initiating transfers voluntarily. Only PSP detection can prevent fraud.[16] All flagged attempts are blocked; those that evade detection result in realised fraud. At the end of period 0 the fraudster observes $\bar{I}_0$ and chooses p_1^*. At $t = 1$: period-1 fraud occurs under portfolio $(p_1^*, 1 - p_1^*)$.

[16] We relax this assumption in Section OA8, where consumers choose whether to stay vigilant, which reduces their exposure.

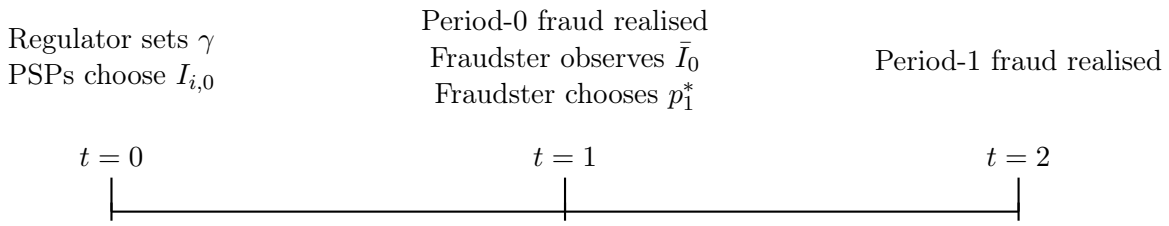


Figure 5
Timeline of Events

The game has Stackelberg structure: PSPs commit to detection intensities before observing fraud behaviour; the fraudster best-responds. Our model extends the standard single-defender framework (Korzhyk et al., 2011; Tambe, 2011) by introducing N defenders, continuous detection intensities, and an endogenous model decay externality arising from the fraudster's adaptive response.[17]

Proofs of all results are in Appendix C.

IV Equilibrium Detection and Overinvestment

Fraudster best response. The fraudster chooses p_1 to maximise

$$\Pi_{\text{fraudster}} \;=\; \alpha V\big[p_1(1-\bar{I}_0) + (1-p_1)(1-\theta\bar{I}_0)\big] - k(1-p_1). \tag{4}$$

The first-order condition gives

$$k'(1-p_1^*) \;=\; (1-\theta)\alpha V\bar{I}_0, \tag{5}$$

so adaptation $1-p_1^*$ is strictly increasing in $\bar{I}_0$ and strictly decreasing in θ: higher aggregate detection induces more substitution toward Type B, and more so when Type B is harder to detect. Let $g \equiv (k')^{-1}$, so $1-p_1^* = g\big((1-\theta)\alpha V\bar{I}_0\big)$; g is strictly increasing, C^1, and $g(0)=0$, since k is strictly increasing and strictly convex.

Model decay. Substituting p_1^* into the period-1 fraud success rate at PSP i,

$$\phi_{i,1} \;=\; (1-I_{i,0}) + \underbrace{(1-\theta)\,g\big((1-\theta)\alpha V\bar{I}_0\big)\,I_{i,0}}_{\text{model decay}}. \tag{6}$$

[17] An analogous mechanism operates in epidemiology, where pathogen populations evolve resistance under selective pressure from drugs or vaccines (Laxminarayan and Brown, 2001).

The second term is the key mechanism of the paper: as aggregate detection rises, the fraudster shifts toward Type B, and a PSP with higher own detection wastes a larger share of its detection capacity on a fraud type it can no longer catch. Model decay is therefore endogenous and self-reinforcing. The fraudster's optimal response to detection degrades the very technology PSPs invest in.

Private and social first-order conditions. Substituting (6) into (3) and using $\partial \bar{I}_0/\partial I_{i,0} = 1/N$ (PSP i takes rivals' investment as given), the symmetric private first-order condition is

$$c'(I) \;=\; \frac{\gamma \alpha V}{N}\Big[2-(1-\theta)g\big((1-\theta)\alpha V I\big)-(1-\theta)^2\alpha V I\, g'\big((1-\theta)\alpha V I\big)\cdot \tfrac{1}{N}\Big]. \qquad (7)$$

The planner sets every PSP to a common I directly, so the relevant derivative of $\bar{I}_0$ is 1 rather than $1/N$; minimizing $W(I) = \alpha V[\phi_0(I) + \phi_1(I, I)] + Nc(I)$ gives the social first-order condition

$$c'(I^W) \;=\; \frac{\alpha V}{N}\Big[2-(1-\theta)g\big((1-\theta)\alpha V I^W\big)-(1-\theta)^2\alpha V I^W\, g'\big((1-\theta)\alpha V I^W\big)\Big]. \qquad (8)$$

We impose two restrictions on the fraud prize αV:

Assumption 1 (Interiority)**.** $\alpha V < k'(1)$.

Assumption 1 ensures the fraudster's period-1 attack mix always retains strictly positive weight on Type-A fraud. Outside this region, when $\alpha V \geq k'(1)$, the fraudster corner-solves to pure Type-B fraud, $1 - p_1^* = 1$.

Let $\underline{c''} \equiv \min_{I\in[0,1]} c''(I)$, $\underline{k''} \equiv \min_{x\in[0,1]} k''(x)$, and $\overline{|k'''|} \equiv \max_{x\in[0,1]} |k'''(x)|$ denote the relevant curvature bounds on c and k, and define the composite curvature constant

$$\Gamma \;\equiv\; \frac{1}{\underline{k''}} \;+\; \frac{(1-\theta)\,\overline{|k'''|}}{\underline{k''}^3}.$$

Assumption 2 (Convexity)**.**

$$\alpha V \;<\; \frac{1}{1-\theta}\sqrt{\frac{N\,\underline{c''}}{\Gamma}}.$$

Assumption 2 bounds the fraud prize αV relative to the curvature of the cost functions c and k. The constant Γ has two roles: its first term, $1/\underline{k''}$, controls the curvature used in Section IV to bound the fraudster's adaptation elasticity, g', and is the only term needed when k is quadratic (as in Corollary 1, where $k''' \equiv 0$ and $\Gamma = 1/\underline{k''}$, so that the assumption is implied by the simpler sufficient bound $\alpha V < \sqrt{N\underline{c''}\,\underline{k''}}$ used informally in the main text,

since $1/(1-\theta) \geq 1$). Its second term additionally bounds how fast that elasticity itself can change, which the uniqueness argument in Appendix C requires when k is not quadratic.

The following proposition holds for *any* c and k satisfying the regularity conditions of Section III; it does not require quadratic costs (the strategic-substitutes claim in part (i) uses one additional mild curvature bound, stated as (42) in the proof and automatic when k is quadratic). Corollary 1 below specializes it to closed form.

Proposition 1 (Equilibrium detection and overinvestment)**.** *Suppose c and k satisfy the regularity conditions of Section III, Assumptions 1 and 2 hold, $\theta < 1$, and $N \geq 2$.*

(i) Existence, uniqueness, characterization, and strategic substitutes. *For any $\gamma \in (0, 1]$, there is a unique $I^*(\gamma) \in (0, 1)$ solving the private FOC* (7)*, and it constitutes the unique symmetric Nash equilibrium.*[18] *$I^*(\gamma)$ is strictly increasing in γ, and the game is one of strategic substitutes: best responses are strictly decreasing in rivals' investment.*

(ii) Underinvestment at $\gamma = 0$. *$I^*(0) = 0 < I^W$: without liability, PSPs provide no detection, where I^W is the unique solution to the social FOC* (8)*.*

(iii) Overinvestment at $\gamma = 1$. *$I^*(1) > I^W$. At any common intensity I, the private marginal benefit of detection exceeds the social marginal benefit by*

$$\frac{\alpha V}{N}(1-\theta)^2 \alpha V \, g'\big((1-\theta)\alpha V I\big)\, I \cdot \frac{N-1}{N} \;>\; 0. \tag{9}$$

(iv) Optimal liability. *There exists a unique $\gamma^* \in (0, 1)$ solving $I^*(\gamma^*) = I^W$, i.e. the unique γ^* for which* (7) *and* (8) *agree.*

Part (iii) is the central result. The private FOC internalizes the effect of $I_{i,0}$ on $\bar{I}_0$ at rate $1/N$; the planner internalizes the same feedback at rate 1. The gap (9) is the resulting overinvestment wedge, zero at $\theta = 1$, since a non-adaptive fraudster leaves nothing to internalize, and growing as fraudsters become more adaptive or the market more fragmented. The result does not rely on the quadratic specification; Corollary 1 merely specializes it to closed form.[19]

[18] $c'(0) = 0$ rules out the corner at $I = 0$ for $\gamma > 0$. Interiority at the upper corner requires in addition that the private marginal benefit of detection at $I_{i,0} = 1$ fall short of $c'(1)$, which we maintain for general c; under the quadratic specification it holds automatically, since $I^*(\gamma) < 2\gamma\alpha V/N < 1$ for every admissible $\alpha V < 1$ and $N \geq 2$.

[19] The mechanism is intertemporal rather than cross-sectional: unlike Acemoglu et al. (2016), our fraudster does not observe individual detection intensities and does not redirect attacks toward less-protected PSPs. The robustness of the two overinvestment results also differs: the strategic-attack overinvestment result in Acemoglu et al. (2016) requires the attacker's cost of concentrating an attack to lie in an intermediate range (their Theorem 6), whereas overinvestment here is unconditional on the strength of adaptation: it obtains

Corollary 1 (Closed form under quadratic costs)**.** *Under* $c(I) = \frac{1}{2}I^2$ *and* $k(x) = \frac{1}{2}x^2$*,* g *is the identity, the FOCs* (7)–(8) *are linear in* I*, and the objects in Proposition 1 have explicit closed forms:*

$$I^*(\gamma) \;=\; \frac{2\gamma\alpha V/N}{1+\gamma(\alpha V)^2(1-\theta)^2(N+1)/N^2}, \qquad \gamma^* \;=\; \frac{N^2}{N^2+(1-\theta)^2(\alpha V)^2(N-1)}. \quad (10)$$

γ^* *is strictly increasing in* θ*, and strictly increasing in* N *throughout* $N \geq 2$*: for any* $2 \leq N_1 < N_2$*,* $\gamma^*(N_1) < \gamma^*(N_2)$*, although* $\partial\gamma^*/\partial N$ *vanishes exactly at* $N = 2$*. Moreover,* $\gamma^* \to 1$ *as* $\theta \to 1$ *or* $N \to \infty$*.*

The comparative statics in Corollary 1 map directly onto policy: liability should be lower when Type-B fraud is hard to detect (θ low) or the market is fragmented (N small), since both raise the severity of the adaptation externality; it should be full when fraudsters cannot adapt ($\theta \to 1$) or the market is large ($N \to \infty$), where the externality disappears.

Figure 6 plots the optimal γ^* against θ for alternative (N), confirming that the welfare-maximizing γ^* is strictly below one whenever $\theta < 1$ and $N > 1$.

This reverses the standard security-economics prescription of full liability to correct underinvestment (Anderson and Moore, 2006). Without adaptation ($\theta = 1$) that prescription is exactly right and $\gamma^* = 1$; with adaptation, detection becomes a common-pool problem, each PSP's investment depletes the shared value of the technology, and the welfare-maximizing liability is strictly partial.

From detection to closure. Proposition 1 and Corollary 1 take the served population as fixed and ask only how hard a PSP should look for fraud. But a PSP facing liability γ has a second instrument besides detection: it can stop serving an account altogether. Section II.B documents this response directly, a scam-exposure shock is followed, with a roughly one-year lag, by a rise in account closures, and this lag is what a *closure margin* predicts: a decision made only after a PSP has observed something about a specific account, not a screening decision made once and for all at onboarding. The account-opening responses of Section II.B show a similar lag, so the timing is consistent with, rather than unique to, this margin; we model both. We model that margin here. A companion *denial margin*, in which the PSP instead screens consumers before an account exists, is analyzed in Section VII.D and mirrors the CFPB's "unable to open" and "application denied" outcomes, exactly as the

for every $\theta < 1$ and $N \geq 2$, because the wedge in (9) is driven purely by the divergence between private ($1/N$) and social (1) internalization of the model-decay externality, not by a threshold on attacker cost convexity. Appendix OA6 shows the result is robust to relaxing this: if the fraudster directs its attacks toward less-protected PSPs, a second, reinforcing overinvestment channel opens. The monopoly case $N = 1$, where the externality vanishes entirely, is in Appendix OA5.

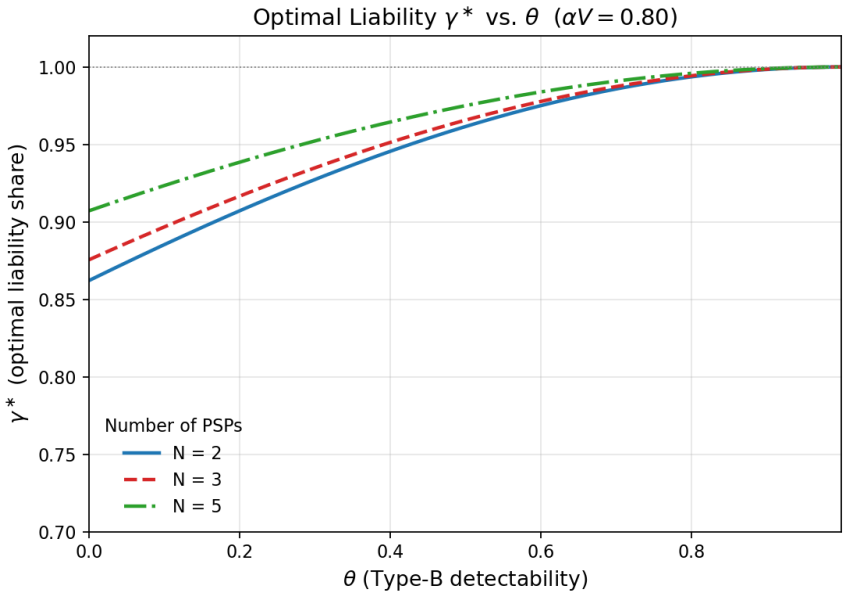


Figure 6
Optimal Liability is Partial

Notes: Optimal liability (γ^*) as a function of the the detectability of Type B fraud (θ) for alternative values of N. Baseline parameter values are reported in Table B2.

closure margin here mirrors "company closed your account."

V Fraud Detection and Debanking

A PSP facing fraud liability has a second lever besides detection: it can simply stop serving the customers most exposed to fraud. This section asks when that response is privately optimal, and shows that the same detection technology a PSP buys to fight fraud also produces the signal it uses to decide whom to exclude from the market. For the regulator, this margin is a constraint rather than an instrument: liability set to discipline detection can instead trigger exclusion, the trade-off Section V.C studies.

Consumers differ in their baseline susceptibility to fraud, which we link to financial sophistication.[20] A share $\lambda \in (0, 1)$ of consumers are type H, with per-period fraud exposure α_H, and the remaining share $1 - \lambda$ are type L, with exposure $\alpha_L < \alpha_H$; the α_τ play the role

[20]Financial sophistication is not the only primitive behind α_τ: a consumer's exposure also reflects how much exploitable information about them is in circulation, breached credentials, leaked phone numbers, account and identity details of the kind traded on the underground forums of Section II.A, and that information goes stale as passwords change, cards are reissued, and datasets age. Fraud risk is therefore a time-varying state rather than a fixed trait; the two types should be read as its snapshot over the model's two-period window.

α played in Section IV, now indexed by type. Write

$$\bar{\alpha} \equiv \lambda\alpha_H + (1-\lambda)\alpha_L, \qquad \beta \equiv \lambda\alpha_H^2 + (1-\lambda)\alpha_L^2, \tag{11}$$

for the pool's mean and second moment of exposure, so $\bar{\alpha}^2 < \beta < \bar{\alpha}$; the baseline model is the degenerate case $\lambda = 0$. Type is not observable. A PSP's only account-level information at $t = 1$ is whether the account's period-0 fraud attempt was blocked (a *flag*) or completed (a *complaint*); an account that was never targeted sends no signal. Because detection is applied uniformly to whichever account is targeted, ending up flagged rather than complained-about carries no information beyond the fact of having been targeted: both signals deliver the same posterior $\lambda\alpha_H/\bar{\alpha}$ of facing a type-H account. Closure is therefore all-or-nothing across signalled accounts, and what it removes is expected, not actual, risk: a type-τ account is signalled with probability α_τ, so the signalled mass is $\bar{\alpha}$ and carries expected period-1 exposure $\lambda\alpha_H \cdot \alpha_H + (1-\lambda)\alpha_L \cdot \alpha_L = \beta$, the second moment, precisely because a policy that cannot tell a truly fraud-prone customer from an unlucky one removes average risk rather than H-risk alone. What survives closure carries the balance

$$\delta \;\equiv\; \bar{\alpha} - \beta, \tag{12}$$

the exposure of the retained, non-signalled book. Losses throughout this section are stated per unit of the original, period-0 customer mass, so δ itself, not $\delta/(1-\bar{\alpha})$, is the object entering PSP losses and the fraudster's adaptation problem.

Let X denote the exposure of whichever pool the fraudster attacks in a given period; it both scales a PSP's fraud losses and calibrates the fraudster's adaptation. Generalising (6) to a calibrating pool that may differ from a PSP's own exposure,

$$\phi_1(I_{i,0}; \bar{I}_0; X) \;\equiv\; (1 - I_{i,0}) + (1-\theta)\, g\big((1-\theta)XV\bar{I}_0\big)\, I_{i,0}, \tag{13}$$

which coincides with (6) at $X = \alpha$. We maintain the quadratic specification $c(I) = \frac{1}{2}I^2$, $k(x) = \frac{1}{2}x^2$ of Corollary 1 for the remainder of this section, under which g is the identity and

$$\phi_1(I_{i,0}; \bar{I}_0; X) \;=\; (1 - I_{i,0}) + (1-\theta)^2 XV\bar{I}_0 I_{i,0}. \tag{14}$$

Below, X takes two values: $\bar{\alpha}$, while the fraudster faces the full pool, and δ, once signalled accounts have been closed.

V.A Timing, Losses, and the Closure Decision

The timing of Figure 5 gains one step between the two periods: after period-0 fraud is realised, but before the fraudster chooses its period-1 adaptation, each PSP observes which of its own accounts were signalled and decides whether to close them. Under no debanking the full pool persists into period 1, so PSP i's loss is exactly the baseline loss of Section IV at $\alpha = \bar{\alpha}$; under debanking, period 0 has already run its course on the full pool, while period 1 plays out on the retained exposure δ:

$$L_i^{\mathrm{nd}}(I_{i,0}; \bar{I}_0) = \frac{\gamma \bar{\alpha} V}{N}\Big[(1 - I_{i,0}) + \phi_1(I_{i,0}; \bar{I}_0; \bar{\alpha})\Big] + c(I_{i,0}), \tag{15}$$

$$L_i^{\mathrm{d}}(I_{i,0}; \bar{I}_0) = \frac{\gamma \bar{\alpha} V}{N}(1 - I_{i,0}) + \frac{\gamma \delta V}{N}\phi_1(I_{i,0}; \bar{I}_0; \delta) + c(I_{i,0}). \tag{16}$$

This mixed-exposure structure, full pool in period 0 and shrunken pool in period 1, is the defining feature of the closure margin: unlike a denial decided at onboarding (Section VII.D), a closure cannot protect a period that has already occurred.

Debanking also forfeits franchise value. Under no debanking, PSP i serves the full pool in both periods and earns the two-period surplus u_{psp} of Section III. Under debanking, the signalled mass $\bar{\alpha}$ is closed before period 1, so PSP i earns the period-0 half of u_{psp} on the full pool but the period-1 half only on the retained mass $1 - \bar{\alpha}$, for a total of $u_{\mathrm{psp}}(1 - \bar{\alpha}/2)$. Profits under the two regimes are

$$\Pi_i^{\mathrm{nd}}(I_{i,0}) = u_{\mathrm{psp}} - L_i^{\mathrm{nd}}(I_{i,0}; \bar{I}_0), \qquad \Pi_i^{\mathrm{d}}(I_{i,0}) = u_{\mathrm{psp}}\Big(1 - \frac{\bar{\alpha}}{2}\Big) - L_i^{\mathrm{d}}(I_{i,0}; \bar{I}_0). \tag{17}$$

At the closure stage, detection I is already sunk, so comparing the two profit streams at the common I gives the closure rule:

$$\text{PSP } i \text{ closes every signalled account} \iff \frac{\gamma V}{N}\Big[\bar{\alpha}\,\phi_1(I; I; \bar{\alpha}) - \delta\,\phi_1(I; I; \delta)\Big] > u_{\mathrm{psp}}\,\frac{\bar{\alpha}}{2}, \tag{18}$$

the period-1 liability saved by shedding the signalled book, whose members each carry expected exposure $\beta/\bar{\alpha}$, against the period-1 franchise value forfeited on the closed mass $\bar{\alpha}$. This is the mechanism behind the closure evidence of Section II.B: a rise in local scam exposure raises β, pushing more accounts across the break-even in (18), with the detect-review-close delay generating the observed lag. In equilibrium the closure decision is anticipated: a PSP that expects to debank chooses its detection for the debanking loss function, and conversely, so the operative comparison is between the two regimes at their own equilibrium detection levels, the object $\Delta(\gamma)$ of Lemma 2 below. Formally, the comparison is across the two *sym-*

metric regime equilibria, all PSPs anticipating retention against all anticipating closure, so the concept grants PSPs commitment to a closure policy at the investment stage and does not analyse unilateral closure deviations from the symmetric profile.[21]

V.B Debanking Threshold and Detection Substitution

Equilibrium detection under each anticipated regime follows directly from Corollary 1, applied at the relevant exposure. Throughout, write $s \equiv (N+1)\gamma(1-\theta)^2V^2$ for the strength of the adaptation feedback:

Lemma 1 (Optimal Detection Under Each Regime)**.** *Under quadratic costs,*

$$I_{\mathrm{nd}}(\gamma) = I^*(\gamma)\big|_{\alpha=\bar{\alpha}} = \frac{2\gamma\bar{\alpha}VN}{N^2 + s\bar{\alpha}^2}, \qquad I_{\mathrm{d}}(\gamma) = \frac{\gamma V(\bar{\alpha}+\delta)N}{N^2 + s\delta^2}. \tag{19}$$

The parallel between the two formulas is exact: across the two periods, detection guards exposure $\bar{\alpha}+\bar{\alpha}$ under retention but only $\bar{\alpha}+\delta$ under debanking, while the adaptation drag in the denominator is calibrated by the period-1 pool, $\bar{\alpha}^2$ against δ^2.

Lemma 2 (Debanking Threshold and Detection Substitution)**.** *Let* $\Delta(\gamma) \equiv \Pi_i^{\mathrm{nd}}\big(I_{\mathrm{nd}}(\gamma)\big) - \Pi_i^{\mathrm{d}}\big(I_{\mathrm{d}}(\gamma)\big)$ *denote the profit gain from retaining over debanking, each regime evaluated at its own equilibrium detection level.*

(i) Threshold. $\Delta(0) = u_{psp}\bar{\alpha}/2 > 0$*, and* Δ *is a rational function of* γ *with everywhere-positive denominator, so* $\tilde{\gamma} \equiv \inf\{\gamma \in (0,1] : \Delta(\gamma) < 0\}$*, with* $\tilde{\gamma} \equiv 1$ *when the set is empty, is strictly positive: PSP* i *retains all signalled accounts for* $\gamma < \tilde{\gamma}$ *and, whenever* $\tilde{\gamma} < 1$*, debanks on an interval of liability levels immediately above* $\tilde{\gamma}$*.*

(ii) Substitution. *Whenever the PSP debanks,*

$$I_{\mathrm{nd}} - I_{\mathrm{d}} = \frac{N\beta\gamma V\big[N^2 - s\,\bar{\alpha}(\bar{\alpha}+2\delta)\big]}{\big(N^2 + s\bar{\alpha}^2\big)\big(N^2 + s\delta^2\big)}. \tag{20}$$

Since the prefactor and both denominator factors are strictly positive, the sign of (20) *is the sign of the bracket; with* $s \propto (1-\theta)^2$*, the bracket rises in* θ*, so the comparison is a cutoff in detectability. Defining*

$$\theta^* \;\equiv\; 1 - \frac{N}{V\sqrt{(N+1)\gamma\,\bar{\alpha}(\bar{\alpha}+2\delta)}} \tag{21}$$

[21] Anticipation is time-consistent: at the calibration of Figure 7 we verify numerically that the sunk-I rule (18), evaluated at the anticipated regime's equilibrium detection level, reproduces the regime assignment of Lemma 2 at every point of the (γ,θ) grid, up to grid resolution at the boundary itself. A general equivalence proof is left open.

(interior whenever $(N+1)\gamma V^2\bar{\alpha}(\bar{\alpha}+2\delta) > N^2$*),*

$$I_{\mathrm{d}} < I_{\mathrm{nd}} \iff \theta > \theta^* \quad \textit{(substitutes)}, \qquad I_{\mathrm{d}} > I_{\mathrm{nd}} \iff \theta < \theta^* \quad \textit{(complements)}.$$

θ^* *is strictly decreasing in* N*.*

Part (i) states the threshold as the *first* crossing because Δ need not cross zero only once: in a thin corner of the admissible parameter region, $N = 2$ with θ high and $\bar{\alpha}V$ close to its admissible ceiling of $k'(1) = 1$, Δ returns above zero as $\gamma \to 1$, so PSPs debank at intermediate liability but retain again at the highest liability levels. The economics is that when Type B is easy enough to detect, the high detection that full liability induces under retention protects the book so effectively that shedding it stops being worthwhile. Numerically the re-entry region is negligible, and it is visible in Figure 7 only as the slight non-monotonicity of the boundary's top edge near $\theta \approx 0.6$; everywhere else the crossing is unique and PSPs debank exactly when $\gamma > \tilde{\gamma}$.

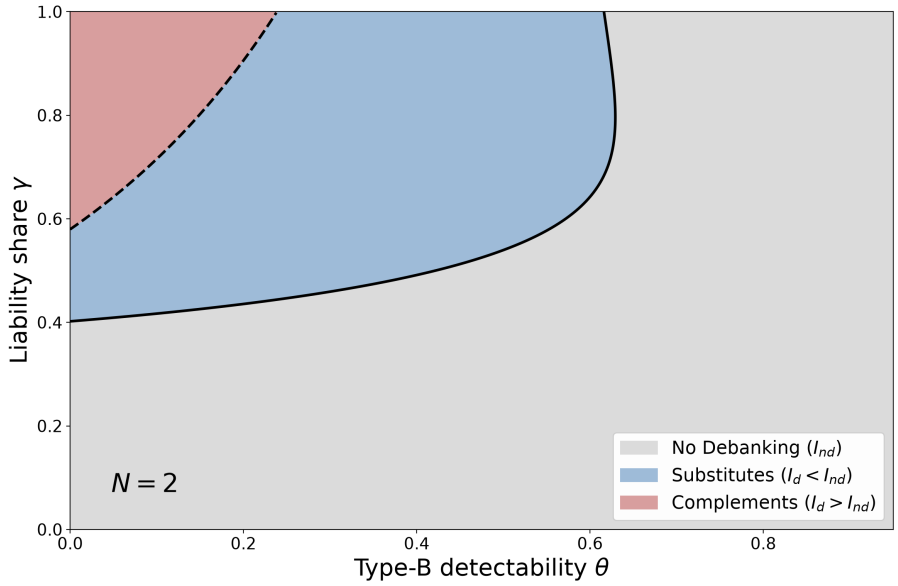


Figure 7
Debanking, Substitution, and Complementarity

Notes: The (γ, θ) plane, partitioned by Lemma 2 at $N = 2$. Grey: PSPs do not debank. Blue: PSPs debank and detection falls ($I_{\mathrm{d}} < I_{\mathrm{nd}}$). Red: PSPs debank and detection rises ($I_{\mathrm{d}} > I_{\mathrm{nd}}$). Solid line: the debanking boundary, the locus $\Delta(\gamma;\theta) = 0$. Dashed line: the substitution boundary $\theta^*(\gamma)$ of (21). Parameters: $N = 2$, $\lambda = 0.25$, $\alpha_H = 0.5$, $\alpha_L = 0.1$, $V = 5$, $u_{\mathrm{psp}} = 0.6$. V is set above the paper's baseline calibration (Table B2) so that all three regions are visible at this N; at the baseline V, the complement region is vanishingly small for $N \geq 2$ (see discussion below).

Figure 7 illustrates the lemma for a market of $N = 2$ PSPs. For low liability (grey), the

franchise value of an account exceeds the liability saved by closing it, and PSPs retain every account; detection follows the ordinary logic of Corollary 1. Once γ crosses the debanking boundary, PSPs close signalled accounts, and two offsetting forces determine what happens to detection. A smaller retained pool means less future exposure worth protecting, which lowers the return to detection (blue, the normal case: debanking substitutes for monitoring). But a smaller pool also gives the fraudster less to gain from migrating to the harder-to-detect Type B method, which weakens the model-decay channel of Section IV and makes remaining detection more effective (red: debanking complements monitoring). Both forces scale with the shed exposure β, which is why β multiplies but never signs (20): the size of the closure determines how much is at stake, detectability which way it cuts. The dividing line is $\theta^*(\gamma)$ in (21): complementarity requires Type B hard enough to detect and the protected prize $V^2\bar{\alpha}(\bar{\alpha}+2\delta)$ large relative to N^2. The red region is not a knife-edge artifact of this parameter draw: the substitution boundary θ^* enters the unit square exactly when $N < N^* \equiv [K+\sqrt{K^2+4K}]/2$, $K \equiv V^2\bar{\alpha}(\bar{\alpha}+2\delta)$, and the red region is the part of the debanking region below it, so it survives for any u_{psp} small enough that debanking occurs there, shrinks quickly in N, and vanishes once $N \geq N^*$. Complementarity between debanking and detection is thus a small-market, high-value-fraud phenomenon: in a large, competitive PSP market, or for low-value fraud, debanking substitutes for monitoring essentially everywhere in (γ, θ).

V.C Optimal Liability With and Without Debanking

Proposition 2 (Optimal Liability and the Cost of Preserving Full Service)**.** *Let $\gamma^{F,*} \equiv \gamma^*|_{\alpha=\bar{\alpha}}$, the liability of Corollary 1 evaluated at the pool's average exposure $\bar{\alpha}$, and let $\tilde{\gamma}$ be the debanking threshold of Lemma 2.*

(i) Absent the debanking margin, the regulator sets liability exactly as Corollary 1 prescribes for a homogeneous population with exposure $\bar{\alpha}$: $\gamma^{F,*}$ solves $I_{\text{nd}}(\gamma^{F,*}) = I^W(\bar{\alpha})$, where $I^W(\alpha) \equiv 2\alpha V/[N + 2(\alpha V)^2(1-\theta)^2]$ is the planner's efficient level of (8) under quadratic costs.

(ii) Subject to preserving full service (no debanking), optimal liability is $\gamma^*_{\text{d}} \equiv \min(\gamma^{F,*}, \tilde{\gamma})$: it implements $I^W(\bar{\alpha})$ when $\tilde{\gamma} \geq \gamma^{F,*}$ (Region I, the debanking margin does not bind), and sets $\gamma^*_{\text{d}} = \tilde{\gamma}$, accepting underinvestment $I_{\text{nd}}(\tilde{\gamma}) < I^W(\bar{\alpha})$ to preserve full service, when $\tilde{\gamma} < \gamma^{F,*}$ (Region II).[22]

[22] In the re-entry corner of Lemma 2(i) the no-debanking set can include a second interval near $\gamma = 1$; if $\gamma^{F,*}$ falls in that interval the regulator restores $I^W(\bar{\alpha})$ exactly and the constraint does not bind. Outside that corner, $\min(\gamma^{F,*}, \tilde{\gamma})$ is the exact constrained optimum.

(iii) $\gamma_{\mathrm{d}}^{*} \leq \gamma^{F,*}$*, strictly in Region II: the possibility of debanking weakly lowers optimal liability relative to Corollary 1's prescription for the same average exposure. The regulator now trades the overinvestment correction of the baseline model against a second, distinct cost, loss of accounts to closure, and the latter can bind first.*

Treating full service as a constraint is the natural reading of the financial-inclusion mandate: closure destroys the franchise surplus u_{psp} and, once consumers carry surplus of their own, consumer value as well. The proposition reprises the central tension of Section IV one level up: there, liability corrects underinvestment but can push PSPs to overinvest; here, correcting overinvestment can push PSPs to debank. A regulator who ignores the closure margin and sets $\gamma^{F,*}$ directly may therefore induce exactly the response documented in Section II.B: a rise in scam exposure followed, with a lag, by a rise in account closures.

There is an informational sting in the tail. The threshold $\tilde{\gamma}$ is not a market datum: it moves with the adaptation feedback s, so a regulator who wants to preserve full service must know how detectable the emerging fraud method is and how quickly detection decays—and near the boundary, errors are not smooth: overshooting $\tilde{\gamma}$ slightly does not buy slightly more detection but flips the regime, triggering closures and, in the substitute region, less monitoring of the surviving book. The re-entry corner of Lemma 2(i) sharpens the point: there the debanking window has two edges, so raising γ through $\tilde{\gamma}$ switches debanking on, while raising it through the window's upper edge switches it off again—near-full liability induces detection under retention strong enough that shedding the book stops paying. A regulator relying on the monotone rule of thumb "higher liability, more debanking" thus misreads both edges of the window; telling them apart requires knowing the decay technology itself.

V.D When Is Debanking Socially Optimal?

Proposition 2 treats full service as a constraint; whether the constraint is worth imposing is a welfare question, and answering it requires putting the excluded consumers into the objective. Let each served consumer enjoy participation utility u_{con} per period, the object Section VII.D endogenises, so closing the signalled mass $\bar{\alpha}$ costs $\bar{\alpha}u_{\mathrm{con}}$ of consumer value in period 1. Two-period market welfare under each regime, evaluated at that regime's private

equilibrium detection level, is

$$W_{\mathrm{nd}} = 2u_{\mathrm{con}} + Nu_{\mathrm{psp}} - \underbrace{\bar{\alpha}V\big[(1 - I_{\mathrm{nd}}) + \phi_1(I_{\mathrm{nd}}; I_{\mathrm{nd}}; \bar{\alpha})\big]}_{F_{\mathrm{nd}}} - \tfrac{N}{2} I_{\mathrm{nd}}^2, \tag{22}$$

$$W_{\mathrm{d}} = (2 - \bar{\alpha})u_{\mathrm{con}} + Nu_{\mathrm{psp}}\Big(1 - \tfrac{\bar{\alpha}}{2}\Big) - \underbrace{\big[\bar{\alpha}V(1 - I_{\mathrm{d}}) + \delta V\, \phi_1(I_{\mathrm{d}}; I_{\mathrm{d}}; \delta)\big]}_{F_{\mathrm{d}}} - \tfrac{N}{2} I_{\mathrm{d}}^2, \tag{23}$$

where fraud losses F_r enter at their full value V, since the planner counts the victim's share $(1 - \gamma)V$ that no PSP internalises,[23] detection costs are borne by all N PSPs, and the fraudster's payoff and adaptation cost are excluded, as in Section IV.[24]

Subtracting and using $NL^{\mathrm{nd}} = \gamma F_{\mathrm{nd}} + \frac{N}{2} I_{\mathrm{nd}}^2$ (and likewise for d) gives an exact decomposition of the private–social wedge:

$$W_{\mathrm{d}} - W_{\mathrm{nd}} \;=\; -N\,\Delta(\gamma) \;+\; (1 - \gamma)\big(F_{\mathrm{nd}} - F_{\mathrm{d}}\big) \;-\; \bar{\alpha}\, u_{\mathrm{con}}. \tag{24}$$

The first term is the PSPs' own comparison, scaled to the market. The two remaining terms are the two things their closure rule misprices, pulling in opposite directions: closure's effect on the fraud borne by *victims*, $(1 - \gamma)(F_{\mathrm{nd}} - F_{\mathrm{d}})$, which PSPs ignore because they value prevented fraud at only γV, and the excluded consumers' participation value $\bar{\alpha}u_{\mathrm{con}}$, which they ignore entirely. How a regulator best responds, with liability at its disposal, is a quantitative question.

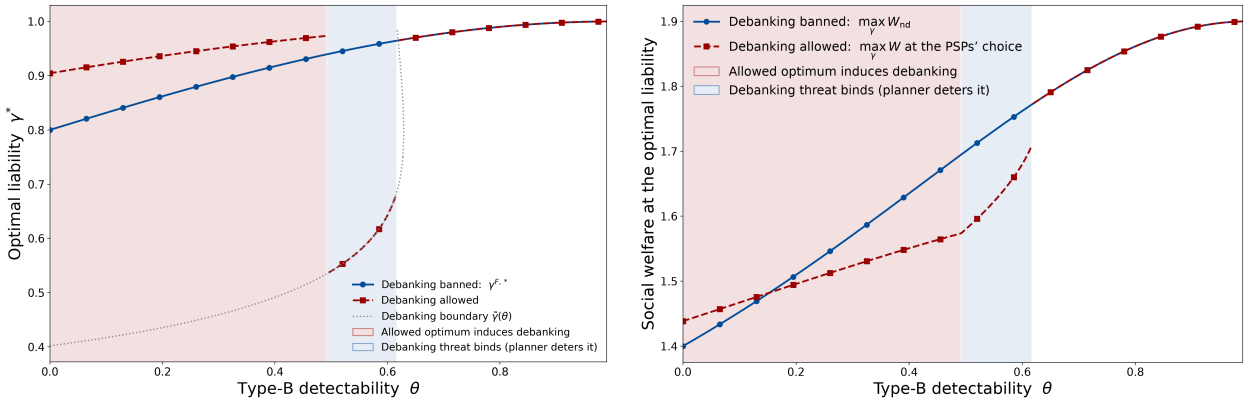
We answer it by letting the regulator choose γ to maximise welfare at each θ, under two policy regimes: closures *banned*, so PSPs always retain, and closures *allowed*, so PSPs debank wherever $\Delta(\gamma) < 0$ and the regulator anticipates that response. Under a ban the consumer and franchise terms in (22) do not vary with γ, so the optimum is exactly $\gamma^{F,*}$ of Proposition 2(i). Under the allowed regime the objective has two branches, each with a closed-form peak: on the retention branch again $\gamma^{F,*}$, feasible only below the debanking boundary $\tilde{\gamma}(\theta)$, and on the debanking branch

$$\gamma^{D,*} \;\equiv\; \gamma^*\big|_{\alpha=\delta} \;=\; \frac{N^2}{N^2 + (1 - \theta)^2(\delta V)^2(N - 1)} \;>\; \gamma^{F,*}, \tag{25}$$

Corollary 1's rule evaluated at the *retained* exposure. The ranking is the baseline model's

[23] Welfare weights are equal across agents, so the reimbursement γV is a pure transfer between PSP and victim and liability matters for welfare only through the detection it induces; in particular, we abstract from the fact that households are typically less able to bear a given loss than institutions.

[24] Nu_{psp} counts each consumer relationship exactly once: each consumer holds one account and the unit mass is split across PSPs, so the aggregate franchise is invariant to N; per-PSP franchise scales as $u_{\mathrm{psp}} = (\text{aggregate})/N$.



(a) Optimal liability **(b)** Welfare at the optimal liability

Figure 8
Optimal Liability With Debanking Banned versus Allowed

Notes: For each θ the regulator chooses γ to maximise welfare, (22)–(23), with closures banned (navy) or allowed (maroon, dashed). Panel (a): the chosen liability; the dotted curve traces the boundary of the PSPs' debanking region of Lemma 2, whose two edges pinch shut at the region's nose; maroon shading: the allowed optimum induces debanking, at $\gamma^{D,*}$ of (25); navy shading: the threat binds and the regulator deters it at $\tilde{\gamma} - \varepsilon$. Panel (b): welfare at the chosen liability; the shaded bands mark where allowing debanking gains (option value) and loses (cost of the closure margin) relative to a ban. Parameters as in Figure 7, with $u_{\text{con}} = 0.5$ per period.

logic run in reverse: partial liability exists to curb the overinvestment externality, whose strength scales with the prize the fraudster adapts over; debanking shrinks that prize from $\bar{\alpha}$ to δ, so once closure does part of the externality's work, liability can rise toward full. Closure and partial liability are substitute instruments for the same externality.

Figure 8 shows the resulting policy and its value; the allowed regulator's behaviour splits into three zones. While the emerging method is hard to detect, it *induces* debanking: it sets $\gamma^{D,*} > \gamma^{F,*}$, deliberately pushing PSPs over their boundary, and the signalled share $\bar{\alpha}$ of consumers loses period-1 access. On a narrow band of detectability above, the closure threat *binds without closure occurring*: $\gamma^{F,*}$ now exceeds $\tilde{\gamma}(\theta)$, debanking is no longer worth its exclusion costs, and the regulator deters it by retreating to $\tilde{\gamma} - \varepsilon$—Proposition 2's Region II made visible. The drop in γ^* at the switch is large, and its size is pinned by an indifference condition: the regulator abandons induced debanking only where deterrence requires a liability sacrifice just large enough to leave welfare unchanged, so a large policy discontinuity coexists with a continuous welfare path. At high detectability the threat dissolves and the two policies coincide at $\gamma^{F,*}$. In value terms (panel b), allowing debanking beats banning it while the method is hard to detect, the option value of exclusion; the ordering reverses well before PSPs themselves stop closing accounts, and on that intermediate range the regulator would prefer retention at first-best liability—exactly what the closure threat rules out.

Deterring there costs more in forgone detection than exclusion costs in lost access, so the regulator tolerates debanking it would rather avoid.

Behind the switch, the regulator chooses between two costs. Inducing debanking sacrifices the excluded consumers' access and the banks' forfeited franchise, in exchange for fraud savings that shrink as detection improves. Deterring it sacrifices detection instead: liability must stay below the banks' threshold, and cheap liability buys little monitoring. The regulator switches when deterrence becomes the cheaper of the two; at the switch both options are equally good, which is why the policy jumps while welfare does not. The two policies also treat banks very differently. A ban at first-best liability is loss-making for PSPs exactly where fraud is hardest to detect, so it would trigger exit rather than compliance; deterrence, by contrast, is a windfall for banks—low liability is cheap for them—so preserving full service doubles as a transfer to the industry it constrains.

VI Money Mules and Information Sharing

A PSP's second lever against fraud, closure, has so far applied to victims: customers who send money to accounts controlled by fraudsters (Section V). This section applies the same closure mechanic to the receiving side of the same crime, money mule accounts, used to receive and forward fraudulent proceeds. A mule account carries no franchise value for the PSP, so once identified it is always closed. But a mule typically holds accounts at more than one PSP, and a PSP that identifies a mule closes only the accounts it can see: sharing that identity with other PSPs lets them close the mule's remaining accounts too, before those accounts are drawn on again. This section studies when PSPs choose to share this information voluntarily, and how the availability of sharing affects the optimal liability rule.

Information-sharing regimes of exactly this kind already exist: in the United Kingdom, CIFAS operates a voluntary industry-funded fraud database that member firms pay to join; in the United States, Section 314(b) of the USA PATRIOT Act gives financial institutions a safe harbor to share suspected-fraud information voluntarily, with no obligation to do so and no central registry; and in the European Union, the proposed PSD3/PSR framework would instead mandate fraud-data sharing across payment providers.

Mule exposure and account overlap. A share $\lambda \in (0, 1)$ of accounts are money mules, indistinguishable from ordinary accounts at opening. Each mule holds accounts at $\rho \in [1, N]$ PSPs, an overlap we treat as common across mules,[25] on average, an observable feature of

[25]With heterogeneous overlap R, the retained pool under sharing is the account-weighted average $\lambda\,\mathbb{E}[R(1-\alpha_m)^R]/\mathbb{E}[R]$ rather than $\lambda(1-\alpha_m)^\rho$; the homogeneous case keeps the comparative statics trans-

the payment network.[26] In each period, a fraudster uses a mule account with probability $\alpha_m \in (0,1)$. As in Section V, we assume the PSP closes a mule account once it is identified, whether through the PSP's own detection flag or through a realized fraud case. We assume this as the baseline and treat the alternative in which only a *successful* fraud identifies a mule as an extension (Section VI.D and the Online Appendix).

Without sharing, a PSP closes only the mule accounts it identifies on its own book, leaving a mule's other $\rho - 1$ accounts open elsewhere. With sharing, identifying a mule at any one PSP closes all ρ of its accounts network-wide. Since detection outcomes are drawn independently across a mule's accounts, the pool retained into period 1 is

$$\lambda_1^{\mathrm{NS}} \;=\; \lambda(1-\alpha_m), \qquad \lambda_1^{\mathrm{S}} \;=\; \lambda(1-\alpha_m)^{\rho}, \tag{26}$$

with $\lambda_1^{\mathrm{S}} < \lambda_1^{\mathrm{NS}}$ strictly whenever $\rho > 1$; at $\rho = 1$ the two coincide and the registry has nothing to add. Both are exogenous constants, fixed by λ, α_m, ρ alone — unlike the retained exposure δ of Section V, they do not depend on the equilibrium detection intensity.

Account closure. A confirmed mule account carries no franchise value and no consumer utility: $u_{\mathrm{psp}} = u_{\mathrm{con}} = 0$ on a mule, since it is not a genuine customer relationship. Comparing Π_i^{nd} and Π_i^{d} of (17) at $u_{\mathrm{psp}} = 0$ makes closure strictly dominant for every $\gamma > 0$: there is no franchise value to weigh against the liability saved, so the closure rule (18) fires unconditionally. Unlike Section V, there is no debanking threshold $\tilde{\gamma}$ to characterize and no regime in which mules are knowingly retained; the only remaining choice is whether PSPs pool their identifications into a shared registry.

The fraudster's regime-specific adaptation. Because closure happens *before* the fraudster's period-1 adaptation choice (Section V.A), and λ_1^R, $R \in \{\mathrm{NS}, \mathrm{S}\}$, is common knowledge, the fraudster solves, regime by regime,

$$\max_{p_1 \in [0,1]} \;\; \lambda_1^R \alpha_m V \Big[p_1(1-\bar{I}_0) + (1-p_1)(1-\theta\bar{I}_0) \Big] - k(1-p_1), \tag{27}$$

the baseline problem (4) with the prize αV replaced by $\lambda_1^R \alpha_m V$: the fraudster calibrates its adaptation against the flow of fraud volume the *retained* mule pool generates each period, not against the gross stock λ it drew on in period 0. The first-order condition gives

$$1 - p_1^{*,R} \;=\; g\big((1-\theta)\lambda_1^R \alpha_m V \bar{I}_0\big), \tag{28}$$

parent.

[26] Nissen (2024) reports that one-third of Americans use three or more financial apps.

so p_1^* is regime-specific: since g is strictly increasing and $\lambda_1^{\mathrm{S}} < \lambda_1^{\mathrm{NS}}$, the fraudster adapts strictly less under sharing.

Losses enter the same way the retained exposure δ entered L_i^{d} in Section V.A, with λ in place of $\bar{\alpha}$, λ_1^R in place of δ, and $\alpha_m V$ in place of V wherever V multiplies a mule-related mass:

$$L_i^R(I_{i,0}; \bar{I}_0) \;=\; \frac{\gamma\lambda\alpha_m V}{N}(1 - I_{i,0}) + \frac{\gamma\lambda_1^R\alpha_m V}{N}\,\phi_1(I_{i,0}; \bar{I}_0; \lambda_1^R) + c(I_{i,0}), \tag{29}$$

where $\phi_1(I_{i,0}, \bar{I}_0; X) = (1-I_{i,0})+(1-\theta)g\big((1-\theta)X\alpha_m V\bar{I}_0\big)I_{i,0}$. Equation (28) is the fraudster's own optimization and should be read consistently with (29): substituting $X = \lambda_1^R$ into the general definition $\phi_1(I, \bar{I}_0; X) = (1-I) + (1-\theta)g((1-\theta)X\alpha_m V\bar{I}_0)I$ gives $1 - p_1^{*,R} = g((1-\theta)\lambda_1^R\alpha_m V\bar{I}_0)$, confirming the two equations agree. We maintain the quadratic specification $c(I) = \frac{1}{2}I^2$, $k(x) = \frac{1}{2}x^2$ throughout this section, under which g is the identity.

VI.A Equilibrium Detection

Lemma 3 (Equilibrium Detection With and Without Sharing)**.** *Under quadratic costs, writing* $\hat{s} \equiv (N+1)\gamma(1-\theta)^2\alpha_m^2 V^2$*, the symmetric equilibrium detection intensity under regime* $R \in \{\mathrm{NS}, \mathrm{S}\}$ *is*

$$I^R(\gamma) \;=\; \frac{\gamma\alpha_m V N(\lambda + \lambda_1^R)}{N^2 + \hat{s}(\lambda_1^R)^2}. \tag{30}$$

Since $\lambda_1^{\mathrm{NS}} > \lambda_1^{\mathrm{S}}$*, there is a cutoff*

$$\theta^{**} \;\equiv\; 1 - \frac{N}{\alpha_m V\sqrt{(N+1)\gamma\big[\lambda(\lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}}) + \lambda_1^{\mathrm{NS}}\lambda_1^{\mathrm{S}}\big]}}, \tag{31}$$

interior whenever $(N+1)\gamma\alpha_m^2 V^2\big[\lambda(\lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}}) + \lambda_1^{\mathrm{NS}}\lambda_1^{\mathrm{S}}\big] > N^2$ *and strictly decreasing in* N*, such that*

$$I^{\mathrm{S}} < I^{\mathrm{NS}} \iff \theta > \theta^{**} \quad \textit{(sharing substitutes for detection)},$$

$$I^{\mathrm{S}} > I^{\mathrm{NS}} \iff \theta < \theta^{**} \quad \textit{(sharing complements detection)}.$$

Lemma 3 traces a single primitive change, the fall in the retained mule pool from λ_1^{NS} to λ_1^{S}, through two channels in equilibrium detection (30) that pull in opposite directions. Directly, a smaller pool leaves less exposure for a PSP's own detection to protect, which by itself would lower I^{S} below I^{NS}. Indirectly, a smaller pool also gives the fraudster less to gain from migrating into the harder-to-detect Type-B method, the same channel formalized in (28): weaker adaptation means less of the model-decay externality of Section IV, which by itself would raise I^{S} above I^{NS}. Sharing is therefore not unambiguously good or bad for detection incentives; (31) identifies exactly which channel dominates.

The cutoff θ^{**} plays the same role here that θ^{*} plays in Lemma 2(ii): both compare a security-technology network effect (*fewer accounts exposed*, λ_1^{S} here and δ there) to the adaptation channel it defuses. Setting $\lambda = \bar{\alpha}$, $\lambda_1^{\mathrm{NS}} = \bar{\alpha}$, and $\lambda_1^{\mathrm{S}} = \delta$, and replacing $\alpha_m V$ by V, turns (31) into (21) exactly: the two comparisons are the same formula applied to two different closure technologies. The one substantive difference is the frequency correction α_m. Because $\alpha_m < 1$, the per-period prize a mule account carries, $\alpha_m V$, is smaller than V; a smaller prize means a weaker adaptation channel at any given θ, so using $\alpha_m V$ rather than V pushes θ^{**} down relative to a (misspecified) comparison that used V directly. Substitutes, the ordinary case in which sharing simply frees up detection capacity, therefore covers a wider range of θ here than the analogous comparison in Section V.

VI.B Information Sharing

Building and operating a shared mule registry costs $K \geq 0$ in total, split evenly, K/N per PSP, conditional on all N PSPs agreeing to participate, the same fixed-cost-sharing structure Pagano and Jappelli (1993) use for the credit bureaus their M lenders jointly fund and access.

Lemma 4 (Information Sharing Threshold)**.** *Suppose $\rho > 1$. Let $L^R(\gamma,\theta) \equiv L_i^R\big(I^R(\gamma,\theta); I^R(\gamma,\theta)\big)$ and $\Delta(\gamma,\theta) \equiv L^{\mathrm{NS}}(\gamma,\theta) - L^{\mathrm{S}}(\gamma,\theta)$ denote the loss avoided by sharing, gross of its cost. Then:*

(i) $\Delta(\gamma,\theta) > 0$ for every $\gamma \in (0,1]$ and every *$\theta \in [0,1]$, including $\theta = 1$; $\Delta(0,\theta) = 0$. Writing $A_R \equiv \gamma\alpha_m V(\lambda + \lambda_1^R)/N$, at $\theta = 1$,*

$$\Delta(\gamma,1) \;=\; \Big(A_{\mathrm{NS}} - \tfrac{1}{2}A_{\mathrm{NS}}^2\Big) - \Big(A_{\mathrm{S}} - \tfrac{1}{2}A_{\mathrm{S}}^2\Big) \;>\; 0. \tag{32}$$

(ii) Δ is continuous in (γ,θ) and strictly decreasing in θ in a neighbourhood of $\theta = 1$: $\partial\Delta/\partial\theta = -2(1-\theta)\Psi + o(1-\theta)$ as $\theta \to 1$, with

$$\Psi \;=\; \frac{\gamma^3(\alpha_m V)^4}{N^3}\big(\lambda_1^{\mathrm{NS}} - \lambda_1^{\mathrm{S}}\big)\big(\lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}} + \lambda\big)\big[(\lambda_1^{\mathrm{NS}})^2 + \lambda\lambda_1^{\mathrm{NS}} + (\lambda_1^{\mathrm{S}})^2 + \lambda\lambda_1^{\mathrm{S}}\big] \;>\; 0,$$

so the derivative itself vanishes at $\theta = 1$, where the dependence runs through $(1-\theta)^2$.

(iii) Suppose $K > 0$ and $\gamma\,\alpha_m V\big(2\lambda + \lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}}\big) \leq N$ on $(0,1]$, which holds for all admissible parameters once $N \geq 4$. Then Δ is strictly increasing in γ, and whenever $\Delta(1,\theta) > K/N$, a unique threshold $\tilde{\gamma}(K,\theta) \in (0,1)$ exists with $\Delta(\tilde{\gamma}(K,\theta),\theta) = K/N$; PSPs share exactly for $\gamma \geq \tilde{\gamma}(K,\theta)$, and $\tilde{\gamma}(K,\theta)$ is strictly increasing *in θ wherever Δ is decreasing in θ. (At $K = 0$, part (i) already gives sharing at every $\gamma > 0$.)*

Lemma 4 isolates the loss reduction sharing delivers, before weighing it against the registry's fixed cost K.

Part (i) establishes that sharing has positive gross value even in the complete absence of fraudster adaptation. At $\theta = 1$, closing a mule's account at one PSP does nothing to help a rival predict the fraudster's behavior, since there is no Type-B method to migrate into; the only thing sharing does at $\theta = 1$ is remove exposure that would otherwise sit on the books of $\rho - 1$ other PSPs until each of them independently identifies the same mule. Equation (32) quantifies exactly this: a pure network effect, present even in a world without model decay, and the reason $\Delta(\gamma, 1) > 0$ rather than $\Delta(\gamma, 1) = 0$.

Part (ii) shows this baseline value is not the whole story once adaptation is switched back on. Near $\theta = 1$, Δ is strictly decreasing in θ: gross of cost, sharing becomes *more* valuable as θ falls, i.e., as Type-B fraud becomes harder to detect and adaptation more costly for PSPs to live with. The reason is the same mechanism formalized in (28): sharing shrinks the pool the fraudster recalibrates its attack against from λ_1^{NS} to λ_1^{S}, and a smaller pool means less to gain from migrating into Type B. When θ is high this channel is nearly inert, since there is little migration to prevent; when θ is low, blunting that migration is worth increasingly more, on top of the pure exposure savings of part (i). Sharing, in other words, is doing double duty exactly where adaptation is most severe: it removes exposure *and* it defuses the model-decay externality of Section IV.

Part (iii) translates this into the liability space that determines whether sharing actually happens. Since the loss gap Δ must clear the per-PSP cost K/N before voluntary sharing occurs, and Δ is increasing in γ (Lemma 3 shows equilibrium detection, and hence the exposure sharing avoids, both scale with γ), higher liability makes sharing easier to justify. The threshold $\tilde{\gamma}(K, \theta)$ is strictly increasing in θ, the mirror image of part (ii): as θ falls and sharing's gross value rises, a smaller liability share is enough to clear the fixed cost K/N. PSPs facing severe model decay share sooner, at lower γ, than PSPs facing a fraudster who cannot adapt.

VI.C Optimal Liability With and Without Sharing

Proposition 3 (Optimal Liability With and Without Sharing)**.** *Under quadratic costs, the welfare-maximising liability share under regime $R \in \{\mathrm{NS}, \mathrm{S}\}$ is*

$$\gamma^{R,*}(\theta) \;=\; \frac{N^2}{N^2 + (1-\theta)^2(\lambda_1^R \alpha_m V)^2(N-1)}, \tag{33}$$

in which the direct exposure factor $\lambda + \lambda_1^R$ has cancelled, so the mule share λ enters only through the retained pool λ_1^R; smooth and strictly increasing in θ, with $\gamma^{R,*}(\theta) \to 1$ as $\theta \to 1$ and interior for every $\theta < 1$. Since $\lambda_1^{\mathrm{NS}} > \lambda_1^{\mathrm{S}}$,

$$\gamma^{\mathrm{NS},*}(\theta) \;<\; \gamma^{\mathrm{S},*}(\theta) \qquad \textit{for every } \theta < 1.$$

Since $\alpha_m < 1$, using the per-period prize $\alpha_m V$ rather than V pulls both $\gamma^{\mathrm{NS},*}$ and $\gamma^{\mathrm{S},*}$ closer to 1 than a specification that used V directly would imply: a smaller per-period prize means a weaker overinvestment externality at any $\theta < 1$, so less of a liability discount is needed to correct it. The ranking is unchanged, since it rests only on the sign of $\lambda_1^{\mathrm{NS}} - \lambda_1^{\mathrm{S}}$, which does not involve α_m.

Proposition 3 sets liability separately for each regime, sharing or not, to correct the overinvestment externality of Section IV. Sharing shrinks the retained pool from λ_1^{NS} to λ_1^{S}, so there is less for the fraudster to gain by migrating into the harder-to-detect Type-B method, and the overinvestment externality is correspondingly weaker. A weaker externality needs a smaller liability discount to correct it, so optimal liability is higher under sharing than without it, $\gamma^{\mathrm{S},*}(\theta) > \gamma^{\mathrm{NS},*}(\theta)$ for every $\theta < 1$. Both converge to full liability as $\theta \to 1$, since at $\theta = 1$ the fraudster cannot adapt and the externality vanishes regardless of regime.

Pagano and Jappelli (1993) value sharing for resolving static uncertainty about a fixed borrower type, with lending volume rising or falling depending on which types are revealed. Our fraudster instead reoptimizes against whatever is shared, migrating into the harder-to-detect Type-B method, so sharing's value depends on θ, a parameter with no counterpart under a fixed type distribution. And since an identified mule carries no surplus and is always closed, revelation here is never good news, so sharing works through removing exposure and defusing adaptation rather than through their volume margin.

VI.D Only Realized Fraud Generates a Signal

The identification technology assumed above (a mule is identified based on detection signals and realized fraud cases) keeps λ_1^R exogenous and lets detection and sharing operate as separable levers. Online Appendix OA9 relaxes this: if identification instead requires the routed payment to have actually cleared, a PSP's own detection mechanically destroys the evidence that would let it, or the network under sharing, identify and close the account, so detection and sharing are no longer cleanly separable instruments. Proposition OA.3 shows the resulting free-riding externality is real and unambiguously signed, amplified roughly ρ-fold under sharing, but does not overturn interior detection in the aggregate, because I remains a joint product protecting a PSP's ordinary customers as well as its mule book;

isolating the mule-specific return to detection shows that component can be driven to zero for realistic network breadth. The corner thresholds governing this result depend only on the ratio of the direct fraud-prevention benefit to the identification-destruction elasticity κ_R, and both of those are unaffected by whether the prize is V or $\alpha_m V$, since a common prefactor of $\alpha_m V$ cancels from the condition that determines the corner; writing the prize as $\alpha_m V$ instead of V therefore changes Lemmas 3–4 and Proposition 3 quantitatively, but leaves Proposition OA.3 exactly as it stood.

VII Extensions

VII.A Detection Reinvestment

PSPs can respond to model decay by reinvesting at $t = 1$: retraining on post-adaptation data restores the fraud success rate to $\phi_{i1}^{\text{new}} = 1 - I_{i1}$. Without reinvestment the old model yields $\phi_{i1}^{\text{old}} = (1 - I_{i0}) + (1 - p_1^*)(1 - \theta) I_{i0}$, which includes the decay term.

Corollary 2 (Reinvestment and the residual overreinvestment wedge)**.** *Let $I_{i,1}^*(\gamma)$ solve $c'(I_{i,1}^*) = \gamma\alpha V/N$, and define the private and social reinvestment gains, relative to retaining the decayed model,*

$$r_{PSP}(\gamma) \;\equiv\; \frac{\gamma\alpha V}{N}\Big(\phi_{i1}^{old} - \phi_{i1}^{new}\big(I_{i,1}^*(\gamma)\big)\Big) - c\big(I_{i,1}^*(\gamma)\big),$$

$$r_W \;\equiv\; \alpha V\Big(\phi_{i1}^{old} - \phi_{i1}^{new}\big(I_{i,1}^*(1)\big)\Big) - N\, c\big(I_{i,1}^*(1)\big) - \Delta,$$

where $\Delta > 0$ is a reduced-form cost representing the further round of fraudster adaptation that reinvestment triggers. At $\gamma = 1$,

$$r_W \;=\; N \cdot r_{PSP}(1) - \Delta \;<\; N \cdot r_{PSP}(1).$$

Consequently, whenever $0 < r_{PSP}(1) < \Delta/N$, PSPs reinvest ($r_{PSP}(1) > 0$) even though reinvestment is socially undesirable ($r_W < 0$): at full liability, PSPs reinvest too frequently.

The mechanism is the intertemporal analogue of Proposition 1: each PSP ignores how its reinvestment raises aggregate detection in period 1, triggering a further round of fraudster adaptation and increasing rivals' fraud rates. When $r_{\text{PSP}}(\gamma) \le 0$, no PSP reinvests, and the model reduces to the baseline of Section IV: the optimal rule is γ^* from Proposition 1. When $r_{\text{PSP}}(1) > 0$, setting $\gamma = 1$ leaves the *intensive* margin, the choice of $I_{i,1}^*$ conditional on reinvesting, undistorted, but not the *extensive*, whether-to-reinvest, margin. $\gamma = 1$ is therefore

a good approximation when reinvestment cycles are fast enough that the intensive margin is the first-order concern, as in cloud-native fintechs; the residual excess-triggering distortion is smaller in markets with slow retraining cycles, such as legacy bank infrastructure, and correcting it fully would require an instrument targeted at the reinvestment decision itself, beyond γ.

VII.B Deterrence and Fraudster Exit

The baseline has a single fraudster active in both periods. With F fraudsters and exit sensitivity $\lambda \geq 0$, a fraction $(1 - \lambda \bar{I}_0)$ remain active in period 1. The PSP's loss function becomes

$$L_i \;=\; \frac{\gamma \alpha V}{N}\big[\phi_{i0} + (1 - \lambda \bar{I}_0)\phi_{i1}\big] + c(I_{i0}). \tag{34}$$

Assumption 3 (Deterrence Interiority)**.** *The exit-sensitivity parameter λ is small enough that the deterrence term does not dominate the adaptation term in the private–social gap of equation* (47) *at $\gamma = 1$; equivalently, $I^*(1) \geq I^W$, where $I^*(1)$ solves the private first-order condition* (45) *at $\gamma = 1$ and I^W solves the social first-order condition* (46)*.*

Corollary 3 (Deterrence and optimal liability)**.** *Under the conditions of Proposition 1, Assumption 3, and $\lambda \geq 0$, there exists a unique $\gamma^* \in (0, 1]$ decentralising the social optimum. γ^* is strictly increasing in both θ and λ.*

Deterrence is the mirror image of adaptation. Own detection lowers rivals' period-1 fraud through fraudster exit (a positive externality) rather than raising it through adaptation (a negative externality). As λ rises, PSPs underinvest in deterrence relative to the social optimum, requiring a higher γ^* to close the gap. The two externalities pull in opposite directions: adaptation pushes γ^* below one, deterrence pushes it toward one, and Assumption 3 restricts attention to the region in which the two forces do not fully offset.

VII.C N-Agent Detection Game

The baseline model has a single class of detection-investing agent. Many applications instead involve two agents whose detection efforts are both necessary to stop a transaction: a sending and a receiving PSP, a PSP and the consumer it serves, or a PSP and the platform that originated the solicitation. This section generalizes Proposition 1 to that setting.

Setup. There are N_1 agents of type 1 and N_2 agents of type 2. Type-1 agent j chooses detection intensity $I_{1j} \in [0, 1]$ at cost $c(I_{1j})$; type-2 agent k chooses $I_{2k} \in [0, 1]$ at cost $c(I_{2k})$, with c satisfying the regularity conditions of Section III. Write $\bar{I}_1 = \frac{1}{N_1}\sum_j I_{1j}$ and

$\bar{I}_2 = \frac{1}{N_2}\sum_k I_{2k}$ for average investment within each type. A transaction is fraudulent only if it evades both the type-1 and the type-2 agent handling it: Type-A fraud succeeds with probability $(1-\bar{I}_1)(1-\bar{I}_2)$, Type-B fraud with probability $(1-\theta_1\bar{I}_1)(1-\theta_2\bar{I}_2)$, where $\theta_1, \theta_2 \in [0,1)$ are type-specific decay parameters. If type-2 agents are consumers, I_{2k} is naturally read as vigilance and γ_2 as the consumer's own residual liability; if they are platforms, I_{2k} is screening effort and γ_2 a platform reimbursement share. The results below apply to either reading.

Fraudster reallocation. Exactly as in the baseline model, the fraudster observes $(\bar{I}_1, \bar{I}_2)$ at the end of period 0 and chooses the period-1 share $p \in [0,1]$ of Type-A fraud to maximize $\alpha V[p(1-\bar{I}_1)(1-\bar{I}_2)+(1-p)(1-\theta_1\bar{I}_1)(1-\theta_2\bar{I}_2)]-k(1-p)$. Under the quadratic $k(x)=\frac{1}{2}x^2$, the first-order condition gives

$$p^*(\bar{I}_1,\bar{I}_2) \;=\; 1-\alpha V\,\Delta(\bar{I}_1,\bar{I}_2), \qquad \Delta(\bar{I}_1,\bar{I}_2) \equiv (1-\theta_1\bar{I}_1)(1-\theta_2\bar{I}_2)-(1-\bar{I}_1)(1-\bar{I}_2), \tag{35}$$

the two-type analogue of equation (5). Expanding Δ linearly in $\bar{I}_1$ at fixed $\bar{I}_2$,

$$\Delta(\bar{I}_1,\bar{I}_2) \;=\; \bar{I}_1\,a(\bar{I}_2)+b(\bar{I}_2), \qquad a(x)\equiv(1-\theta_1)+(\theta_1\theta_2-1)x, \quad b(x)\equiv(1-\theta_2)x. \tag{36}$$

Both coefficients vanish identically when $\theta_1 = \theta_2 = 1$: with no decay, Type-B fraud is indistinguishable from Type-A and the fraudster never reallocates.

Equilibrium detection. Each type-1 agent minimizes the two-period loss $L_{1j} = C_1[(1-I_{1j})(1-\bar{I}_2)+\varphi_{1j}]+\frac{1}{2}I_{1j}^2$, where $C_1 \equiv \gamma_1\alpha V/N_1$, $\gamma_1 \in [0,1]$ is the liability share borne by type-1 agents, and φ_{1j} is the period-1 fraud rate; type-2 losses are defined symmetrically. Every agent is a finite-population Cournot player, taking the *other* type's average as exogenous but internalizing its own $1/N_1$ contribution to its own type's average. Differentiating and imposing symmetry gives the best-response functions

$$I_1^* = f_1(I_2) = \frac{C_1\big[2(1-I_2)-K_1 b(I_2)\big]}{1+C_1K_1a(I_2)}, \qquad K_1 \equiv \alpha V\Big(1+\frac{1}{N_1}\Big)a(I_2), \tag{37}$$

with $I_2^* = f_2(I_1)$ defined symmetrically. The displayed f_1 is the unconstrained first-order-condition solution; the best response proper is its projection onto $[0,1]$, $\hat{f}_1 \equiv \min\{\max\{f_1,0\},1\}$, valid by strict convexity of each loss in own investment. Both projections are continuous single-valued self-maps of $[0,1]$, so a Nash equilibrium exists by Brouwer's fixed-point theorem; it is interior, where $\hat{f}_j = f_j$, and unique for V sufficiently small, since $\partial f_1/\partial I_2, \partial f_2/\partial I_1 \to 0$ as $V \to 0$.

Corollary 4 (Optimal Liability and the Regulatory Fee)**.** *Under the quadratic specification* $c(I) = \frac{1}{2}I^2$, $k(x) = \frac{1}{2}x^2$, *let* γ_1^* *decentralize the planner's efficient level* $I_1^W(I_2)$, *i.e. the unique root, in* γ_1, *of* $f_1(I_2) = I_1^W(I_2)$, *where* $I_1^W(I_2) = \dfrac{2\alpha V[(1-I_2) - \alpha V\, a(I_2)b(I_2)]}{N_1 + 2(\alpha V\, a(I_2))^2}$; γ_2^* *solves the symmetric equation.*

(i) Partial liability under adaptation. *When* $\theta_1, \theta_2 < 1$, $\gamma_1^* < 1$*: the model-decay externality of Proposition 1 reappears in the two-agent-type setting.*

(ii) Full liability under no adaptation. *When* $\theta_1 = \theta_2 = 1$, $\gamma_1^* = \gamma_2^* = 1$.

(iii) The regulatory fee. *Let* V *denote the loss from a realized fraud event and* $C \in [0, V]$ *any victim compensation target. At* $\theta_1 = \theta_2 = 1$, *paying both agents' liability* $\gamma_1^* V = \gamma_2^* V = V$ *directly to the victim would require compensation* $C = 2V > V$, *which is infeasible. Full liability is instead implementable by routing* C *to the victim and the residual* $\tau \equiv 2V - C \geq V > 0$ *to the regulator as a fee. By continuity of* γ_1^*, γ_2^* *in* (θ_1, θ_2), $\gamma_1^* + \gamma_2^* > 1$ *on an open neighborhood of* $\theta_1 = \theta_2 = 1$, *so a strictly positive fee remains necessary whenever fraudster adaptation is sufficiently mild.*

As θ_1, θ_2 fall and fraudster adaptation intensifies, the same overinvestment channel that pulls γ^* below one in the single-agent-type model of Corollary 1 eventually pulls $\gamma_1^* + \gamma_2^*$ below one as well, at which point ordinary bilateral reimbursement, with no fee, decentralizes the social optimum on its own. The regulatory fee is therefore a feature of environments with complementary detecting agents and limited fraudster adaptability, exactly the setting the UK PSR's sending-receiving PSP split and the absence of platform liability both fall into, rather than a universal prescription.

VII.D Heterogeneous Consumers, Fraud, and Financial Inclusion

This extension asks how the results of Section V change when a PSP can act on consumer type *before* an account exists rather than only after observing a signal. Consumers differ in baseline fraud susceptibility: a share $\lambda \in (0, 1)$ are type H, exposure α_H, the rest type L, exposure $\alpha_L < \alpha_H$, and type is observable at onboarding. A PSP's only participation margin is whether to also serve type H, $D = 1$, or exclude it, $D = 0$, retaining type L either way. Write

$$\alpha(D) \equiv D\bar{\alpha} + (1 - D)\tilde{\alpha}, \qquad \bar{\alpha} \equiv \lambda\alpha_H + (1 - \lambda)\alpha_L, \qquad \tilde{\alpha} \equiv (1 - \lambda)\alpha_L, \tag{38}$$

so $\tilde{\alpha} < \bar{\alpha}$ and $D = 1$ recovers the baseline model at $\lambda = 0$. Because the decision is made before period 0, the fraudster calibrates to $\alpha(D)$ in *both* periods, unlike the mixed exposure

of Section V.A: PSP i's loss under D is literally L_i from (3) at $\alpha = \alpha(D)$, so

$$I^D(\gamma) \equiv I^*(\gamma)\big|_{\alpha=\alpha(D)} \tag{39}$$

by Corollary 1 directly, no new derivation required. PSP i's profit and welfare follow (2) and its social counterpart at $\alpha = \alpha(D)$, scaled by $D + (1-D)(1-\lambda)$ to reflect the share of the market served; as in Section V, this scale factor is additively separable and irrelevant to the choice of $I_{i,0}$, mattering only for D.

Lemma 5 (Exclusion Threshold and Detection Substitution)**.** *Let* $\Delta(\gamma) \equiv \Pi_i(1, I^1(\gamma)) - \Pi_i(0, I^0(\gamma))$.

(i) Threshold. $\Delta(0) = \lambda u_{psp} > 0$*, and* Δ *is single-crossing from positive to negative on the admissible parameter region* $\{\bar{\alpha}V < 1,\ N \geq 2\}$ *(verified numerically; see the proof of Lemma 2), giving a unique* $\bar{\gamma} \in (0, 1]$*: PSPs serve both types for* $\gamma \leq \bar{\gamma}$*, exclude type H for* $\gamma > \bar{\gamma}$*.*

(ii) Substitution. *Define* $\alpha^\dagger \equiv N/[(1-\theta)V\sqrt{\gamma(N+1)}]$*. Then*

$$I^0(\gamma) < I^1(\gamma) \iff \bar{\alpha}\tilde{\alpha} < (\alpha^\dagger)^2, \qquad I^0(\gamma) > I^1(\gamma) \iff \bar{\alpha}\tilde{\alpha} > (\alpha^\dagger)^2.$$

On the maintained region $\{\bar{\alpha}V < 1, N \geq 2\}$*,* $\alpha^\dagger V \geq N/\sqrt{N+1} > 1 > \bar{\alpha}V$*, so* $\bar{\alpha}\tilde{\alpha} < (\alpha^\dagger)^2$ *holds throughout: exclusion strictly lowers equilibrium detection, and the complement case requires prizes above the maintained bound.*

$\alpha^\dagger$ is strictly increasing in N, exactly as in Lemma 2: the complement case is relatively more likely for small N under *either* margin, ex ante exclusion here or ex post closure in Section V. The two thresholds are not the same object, since exclusion shrinks both periods' exposure while closure shrinks only period 1's, but they move in the same direction in N, and the underlying single-peaked mechanism, a smaller pool lowers the direct return to detection but also lowers the fraudster's incentive to adapt, is identical across both margins.

Proposition 4 (Optimal Liability Subject to Full Service)**.** *Let* $\gamma^{F,*} \equiv \gamma^*|_{\alpha=\bar{\alpha}}$*, as in Proposition 2, and let* $\bar{\gamma}$ *be the threshold of Lemma 5. Then, subject to preserving service to type-H consumers,* $\gamma_D^* \equiv \min(\gamma^{F,*}, \bar{\gamma}) \leq \gamma^{F,*}$*, strictly whenever* $\bar{\gamma} < \gamma^{F,*}$*: the exclusion threat weakly lowers optimal liability below Corollary 1's prescription for the same average exposure, by the same argument as Proposition 2(ii)–(iii), with* $\bar{\gamma}$ *in place of* $\tilde{\gamma}$*. As in Section V.D, whether the regulator prefers this constrained optimum to permitting exclusion is a separate comparison of the two welfare branches, which we do not repeat here.*

VIII Conclusion

This paper studies payment fraud detection in a dynamic environment where fraudsters adapt strategically to detection technologies, causing model performance to decay over time. We provide the first direct micro-level evidence of criminal adaptation using CrimeBB, one of the largest publicly available underground forum datasets. The release of ChatGPT in November 2022 was followed by a near-doubling of within-creator AI–fraud co-discussion (+97%), with technically sophisticated creators driving early adoption and less skilled creators subsequently broadening participation, consistent with a general-purpose enabling shock that lowers skill barriers. Stacked event studies across 27 anti-fraud technology deployments show circumvention discussion building as each defense matures and peaking around deployment, while discussion of the targeted methods fades over the following years: criminal attention reallocates from established methods toward countermeasures over the deployment cycle, the Whac-A-Mole mechanism at the level of criminal knowledge.

Using the staggered rollout of Bitcoin ATMs as a shock to local scam exposure, we further document that a rise in local fraud complaints is followed, with a roughly one-year lag, by a rise in account closures, evidence consistent with fraud exposure driving debanking rather than the reverse. Together, these findings fill a gap in the economics of crime literature: while prior work establishes that criminals respond to incentives through spatial reallocation (Maheshri and Mastrobuoni, 2021) and cross-type substitution (Draca et al., 2019), direct evidence of criminal adaptation, and of its downstream consequences for who remains banked, has been absent because method development leaves no trace in administrative crime records.

We embed these empirical regularities in a model of payment fraud detection. This two-period model has two fraud types and endogenises PSP detection investment, yielding results along four margins.

First, when PSPs do not bear fraud losses, detection is underprovided relative to the social optimum. Second, under full liability in competitive markets, PSPs overinvest: each PSP ignores how its detection intensity accelerates fraudster adaptation market-wide, imposing a negative externality on rivals. The optimal liability share is therefore strictly below one and increasing in the number of PSPs. Third, when PSPs can reinvest in detection technology, competition generates a ratchet effect: reinvestment restores effectiveness but accelerates adaptation, raising the future return to yet further investment. An extension that prices the continuation cost of renewed adaptation in reduced form shows this ratcheting is individually rational for each PSP, not a mistake liability rules alone can prevent; under full liability PSPs then reinvest more frequently than is socially optimal, and the regulator must account for this when calibrating the liability share rather than treating detection investment as static.

Fourth, liability rules affect financial inclusion. In the model, a PSP debanks a flagged consumer once the liability it carries on that account exceeds the account's franchise value, so liability set purely to correct the overinvestment externality of the second result can push PSPs to exclude the very consumers it was designed to protect, exactly the pattern our Bitcoin ATM evidence documents in the data. The welfare-maximizing rule internalizes this margin directly: it weighs the overinvestment correction against the risk of pricing vulnerable consumers out of the payment system, and can require liability strictly below the level that a model of detection alone would prescribe.

Fraud detection also depends on information PSPs hold about one another's customers, particularly the money-mule accounts used to receive and launder fraud proceeds, which are frequently held at multiple institutions. We show that whether sharing this information helps or hurts detection incentives is itself ambiguous, and depends on how easily an adapting fraudster's methods evade attribution: sharing shrinks the pool of exploitable accounts, which can either free up detection capacity or blunt the incentive to invest in it. PSPs share voluntarily once liability clears the fixed cost of running a registry, and because sharing weakens the same adaptation externality that justifies partial liability in the first place, the liability level that supports sharing is higher than the one that would apply without it: sharing and liability are complements, not substitutes, and the gain from sharing rises as fraud becomes harder to detect.

The paper further shows that fraud prevention is a joint responsibility across multiple agents whose efforts are complementary. When platforms bear no reimbursement obligation despite originating the majority of fraud solicitations, they free-ride on PSP detection efforts, systematically undermining prevention incentives. Our N-agent framework shows that the welfare-maximizing liability rule sets full liability for all agents whose detection efforts are complementary, supplemented by a regulatory fee, regardless of whether the agents are sending and receiving PSPs, PSPs and platforms, or other combinations. The UK PSR's 50/50 reimbursement split between sending and receiving PSPs falls short of this prescription, and the absence of platform liability represents the largest remaining regulatory gap.

Taken together, the results imply that effective fraud regulation cannot be reduced to a single liability parameter. It requires liability rules calibrated to the model-decay externality rather than treating detection investment as static; explicit attention to the debanking margin, so that correcting overinvestment does not come at the cost of financial inclusion; mandatory information sharing paired with, not substituted for, higher liability; and liability extension to platforms, whose zero-liability status generates systematic underinvestment in fraud prevention. Future work could use proprietary transaction data to estimate adaptation rates in deployed fraud detection systems and calibrate optimal liability, debanking, and

information-sharing rules empirically.

References

Acemoglu, D., A. Malekian, and A. Ozdaglar (2016). Network security and contagion. *Journal of Economic Theory 166*, 536–585.

Aho, A. V. and M. J. Corasick (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM 18*(6), 333–340.

Anderson, R. and T. Moore (2006). The economics of information security. *science 314*(5799), 610–613.

Bhalerao, R., M. Katariya, B. Penber, S. Dashevskyi, J. Mugan, P. Neves, S. Le Blond, and J. van der Merwe (2019). Mapping the underground: Supervised discovery of cybercrime supply chains. In *Proceedings of the 2019 APWG Symposium on Electronic Crime Research*, pp. 1–16. IEEE.

Bian, B., M. Pagel, D. Raval, and H. Tang (2024). Consumer surveillance and financial fraud. NBER Working Paper 31692, National Bureau of Economic Research.

California DFPI (2026). Crypto ATM scams: Don't let fraudsters drain your wallet. California Department of Financial Protection and Innovation, consumer alert, updated May 20. `https://dfpi.ca.gov/consumers/crypto/crypto-atm-scams/`.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Cong, L. W., T. Feng, Y. Liu, and F. Lu (2025). Crypto atms: Material effects of virtual currencies. *Available at SSRN 5617630*.

Consumer Financial Protection Bureau (2022, December). In re Wells Fargo Bank, N.A. CFPB No. 2022-CFPB-0011.

de Chaisemartin, C. and X. D'Haultfœuille (2018). Fuzzy differences-in-differences. *The Review of Economic Studies 85*(2), 999–1028.

Dimmock, S. G., W. C. Gerken, and N. P. Graham (2018). Is fraud contagious? coworker influence on misconduct by financial advisors. *Journal of Finance 73*(3), 1417–1450.

Draca, M., T. Koutmeridis, and S. Machin (2019). The changing returns to crime: Do criminals respond to prices? *Review of Economic Studies 86*(3), 1228–1257.

Egan, M., G. Matvos, and A. Seru (2019). The market for financial adviser misconduct. *Journal of Political Economy 127*(1), 233–295.

Fletcher, E. (2024). Bitcoin ATMs: A payment portal for scammers. Federal Trade Commission, Consumer Protection Data Spotlight, September 3. https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2024/09/bitcoin-atms-payment-portal-scammers.

Foley, S., J. R. Karlsen, and T. J. Putniņš (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *Review of Financial Studies 32*(5), 1798–1853.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022). Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance 77*(1), 5–47.

Gao, J., J. Pacelli, J. Schneemeier, and Y. Wu (2020). Dirty money: How banks influence financial crime. Working paper. Available at https://ssrn.com/abstract=3722342.

Goyal, S. and A. Vigier (2014). Attack, defence, and contagion in networks. *Review of Economic Studies 81*(4), 1518–1542.

Griffin, J. M. and K. Mei (2024). How do crypto flows finance slavery? the economics of pig butchering. Working paper, University of Texas at Austin. Available at https://ssrn.com/abstract=4742235.

Hauswald, R. and R. Marquez (2006). Competition and strategic information acquisition in credit markets. *Review of Financial Studies 19*(3), 967–1000.

Hayashi, F., T. Moore, and R. J. Sullivan (2015). The economics of retail payments security. *Federal Reserve Bank of Kansas City Economic Review 100*(2), 21–66.

Horwitz, J. (2025, November). Meta is earning a fortune on a deluge of fraudulent ads, documents show. Reuters.

Korzhyk, D., V. Conitzer, and R. Parr (2011). Security games with multiple attacker resources. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Volume 22, pp. 273–279.

Laxminarayan, R. and G. M. Brown (2001). Economics of antibiotic resistance: A theory of optimal use. *Journal of Environmental Economics and Management 42*(2), 183–206.

Maheshri, V. and G. Mastrobuoni (2021). The race between deterrence and displacement: Theory and evidence from bank robberies. *Review of Economics and Statistics 103*(3), 547–562.

Nissen, K. (2024, August). One-third of americans use three or more financial apps. *S&P Global Market Intelligence*. Accessed: 2026-06-17.

Pagano, M. and T. Jappelli (1993). Information sharing in credit markets. *The journal of finance 48*(5), 1693–1718.

Parlour, C. A., U. Rajan, and H. Zhu (2022). When fintech competes for payment flows. *Review of Financial Studies 35*(11), 4985–5024.

Pastrana, S., D. R. Thomas, A. Hutchings, and R. Clayton (2018). CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1845–1854. International World Wide Web Conferences Steering Committee.

Piskorski, T., A. Seru, and J. Witkin (2015). Asset quality misrepresentation by financial intermediaries: Evidence from the RMBS market. *Journal of Finance 70*(6), 2635–2678.

Portnoff, R. S., S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson (2017). Tools for automated analysis of cybercriminal markets. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 657–666. International World Wide Web Conferences Steering Committee.

Sanchez-Adams, C. (2025, September). Fraud in Focus: Exposing Financial Threats to American Families. Testimony before the Oversight and Investigations Subcommittee of the U.S. House Financial Services Committee. Testimony of Carla Sanchez-Adams, National Consumer Law Center, on behalf of its low-income clients.

Stop Scams UK (2023). The impact of AI on fraud and scams. Technical report, Stop Scams UK.

Tambe, M. (2011). *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.

Vives, X. and Z. Ye (2025). Information technology and lender competition. *Journal of Financial Economics 163*, 103957.

Wei, J. (2026). The real effects of payment anonymity: Evidence from the crypto ecosystem. INSEAD Working Paper 2026/35/FIN. SSRN 6958685.

Who Pays for Payment Fraud? Detection and Liability Rules under Strategic Fraudster Adaptation
Internet Appendix

A Variable Definitions

Table A1
Variable Definitions

Variable	Definition
Panel A: Dependent Variables (Post Level)	
1(GenAI)	Binary indicator equal to one if the post contains at least one generative AI term (e.g., ChatGPT, GPT, LLM, deepfake, Stable Diffusion).
1(Fraud)	Binary indicator equal to one if the post contains at least one fraud-related term (e.g., phishing, carding, ransomware, credential stuffing).
1(GenAI × Fraud)	Binary indicator equal to one if the post contains both a generative AI term and a fraud-related term.
1(Fraud Method)	Binary indicator equal to one if the post mentions a specific fraud method or tool (e.g., phishing kit, fullz, CVV, combo list).
1(GenAI × Method)	Binary indicator equal to one if the post contains both a generative AI term and a specific fraud method term.
Panel B: Stacked Event-Study Outcomes (Creator-Month Level)	
vuln_{ijt}	Binary indicator equal to one if creator i posted at least one vulnerable-method term associated with technology j in month t. Constructed from a dictionary of 178 unique terms describing the attack vector targeted by each anti-fraud technology.
ctr_{ijt}	Binary indicator equal to one if creator i posted at least one countermeasure term associated with technology j in month t. Constructed from a dictionary of 270 unique terms describing circumvention techniques for each anti-fraud technology.
Panel C: Creator-Level Characteristics	
After_t	Binary indicator equal to one for months at or after December 2022 (the month of ChatGPT's public release).
Pre-period posts	Total number of posts by creator i in the pre-period (November 2021 through November 2022). Used in logarithmic form, $\log(1 +$ pre-period posts), as a measure of creator experience.
$\log(1 + \text{sophistication posts})$	Natural logarithm of one plus the number of pre-period posts by creator i containing at least one term from the 212-term technical sophistication dictionary. Measured entirely before December 2022.
$\text{TechSoph}^{\text{hi}}$	Binary indicator equal to one if the creator's pre-period sophistication post count is at or above the median among creators with at least one sophistication post.

Countermeasure dictionaries are reported in Internet Appendix Table OA5. The technical sophistication dictionary is reported in Internet Appendix OA1. The full list of anti-fraud technology events and their deployment dates is in Internet Appendix Table OA4.

B Baseline Parameters

Table B2
Baseline Parameters for Numerical Analysis

Parameter	Description	Range	Baseline Value
α	Share of transactions exposed to fraud	$[0,1]$	0.3
V	Loss from a successful fraud attempt	$\mathbb{R}_+$	3
N	Number of competing PSPs	$\mathbb{N}_+$	1, 2, 5
θ	Relative detectability of Type B fraud	$[0,1]$	0.3, 0.5
γ	PSP liability share (regulatory choice)	$[0,1]$	Endogenous
p_t	Share of Type A fraud in period t	$[0,1]$	Endogenous
I_{0i}	Detection intensity of PSP i at $t=0$	$[0,1]$	Endogenous
ϕ_{it}	Fraud success rate for PSP i in period t	$[0,1]$	Endogenous

Notes: Baseline parameter values used in the numerical analysis. "Endogenous" indicates that the variable is determined in equilibrium.

C Proofs

This appendix collects the proofs of all propositions in the order they appear in the main text. Each proof is self-contained and given in full; where a step relies on a standard technique (Berge's maximum theorem, Brouwer's fixed-point theorem, the implicit function theorem, the envelope theorem), we invoke it without reproducing its own proof. We use the quadratic specifications $c(I) = \frac{1}{2}I^2$ and $k(x) = \frac{1}{2}x^2$ throughout when closed forms are stated, noting where the qualitative result holds under the general regularity conditions of Section III and where the closed form requires the quadratic. All notation follows the main text.

C.1 Proof of Proposition 1

Proof. Throughout, fix $\theta < 1$ and $N \geq 2$. We first record two bounds implied by Assumptions 1–2 that are used repeatedly below. Let

$$\bar{g}' \;\equiv\; \max_{x\in[0,(1-\theta)\alpha V]} g'(x), \qquad \overline{|g''|} \;\equiv\; \max_{x\in[0,(1-\theta)\alpha V]} |g''(x)|,$$

both finite since $g = (k')^{-1}$ is C^2 on this compact interval (Section III). By Assumption 1, $(1-\theta)\alpha V \leq \alpha V < k'(1)$, so monotonicity of g gives

$$g\big((1-\theta)\alpha V\big) \;<\; 1. \tag{a}$$

Since $g'(x) = 1/k''(g(x))$ and $g''(x) = -k'''(g(x))/k''(g(x))^3$, we have $\bar{g}' \le 1/\underline{k''}$ and $\overline{|g''|} \le \overline{|k'''|}/\underline{k''}^3$. Assumption 2, $\alpha V < \frac{1}{1-\theta}\sqrt{N\underline{c''}/\Gamma}$, rearranges to $(\alpha V)^2(1-\theta)^2 < N\underline{c''}/\Gamma$, so

$$\underline{c''} \;>\; \frac{(\alpha V)^2(1-\theta)^2}{N}\,\Gamma \;\ge\; \frac{\alpha V}{N}(1-\theta)^2\alpha V\,\bar{g}' \;+\; \frac{(\alpha V)^2(1-\theta)^3}{N}\,\overline{|g''|},$$

and since $c''(I) \ge \underline{c''}$ for all I,

$$c''(I) \;>\; \frac{\alpha V}{N}(1-\theta)^2\alpha V\,\bar{g}' \;+\; \frac{(\alpha V)^2(1-\theta)^3}{N}\,\overline{|g''|} \qquad \text{for all } I \in [0,1]. \tag{b}$$

Part (i): existence, uniqueness, and strategic substitutes. By symmetry it suffices to find a fixed point of $\psi : [0,1] \to [0,1]$, the best response of PSP i when every rival plays a common I. Differentiating L_i twice with respect to $I_{i,0}$ and using $\partial \bar{I}_0/\partial I_{i,0} = 1/N$,

$$\frac{\partial^2 L_i}{\partial I_{i,0}^2} \;=\; \frac{\gamma\alpha V}{N}\left[\frac{2(1-\theta)^2\alpha V}{N}g'(x) + \frac{(1-\theta)^3(\alpha V)^2}{N^2}I_{i,0}\,g''(x)\right] + c''(I_{i,0}), \qquad x \equiv (1-\theta)\alpha V\bar{I}_0,$$

which exceeds $c''(I_{i,0})$ whenever $g' > 0$, as it always is since k' is strictly increasing. Bound (b) makes this strictly positive for any rival profile in $[0,1]^N$, so ψ is single-valued and, by Berge's maximum theorem, continuous; a continuous self-map of the compact interval $[0,1]$ has a fixed point by the intermediate value theorem, giving a symmetric equilibrium $I^*(\gamma)$ solving (7).

For uniqueness, ψ perturbs the *common* rival level I, so all $N-1$ rivals move together. Writing $F(I_i, I) \equiv \partial L_i/\partial I_{i,0}$ evaluated at $\bar{I}_0 = [(N-1)I + I_i]/N$, the implicit function theorem gives $\psi'(I) = -(\partial F/\partial I)/(\partial F/\partial I_i)$, with

$$\frac{\partial F}{\partial I} \;=\; (N-1)\left[\frac{\gamma(\alpha V)^2(1-\theta)^2}{N^2}g'(x) + \frac{\gamma(\alpha V)^3(1-\theta)^3}{N^3}I_i\,g''(x)\right], \tag{40}$$

i.e. $(N-1)$ times the single-rival cross-partial

$$\frac{\partial^2 L_i}{\partial I_{i,0}\partial I_{j,0}} \;=\; \frac{\gamma(\alpha V)^2(1-\theta)^2}{N^2}\Big[g'(x) + \frac{(1-\theta)\alpha V}{N}\,I_{i,0}\,g''(x)\Big] \tag{41}$$

evaluated at the symmetric point. The bracket is strictly positive: $g' = 1/(k''\circ g) > 0$, and since $|g''| = |k'''\circ g|/(k''\circ g)^3$, the second term is dominated by the first whenever $N\,(k''\circ g)^2 > (1-\theta)\alpha V\,|k'''\circ g|$, for which

$$N\,\underline{k''}^{\,2} \;\ge\; (1-\theta)\,\alpha V\,\overline{|k'''|} \tag{42}$$

is sufficient; (42) holds automatically for quadratic k ($k''' \equiv 0$, the case of Corollary 1) and we maintain it as part of the regularity requirements for the general-cost statement. Uniqueness below does not use it, relying only on the absolute bounds. A positive cross-partial of a loss function is the loss-minimization analogue of the profit-maximization substitutes condition: raising a rival's investment raises $\bar{I}_0$, hence the fraudster's adaptation, hence PSP i's marginal cost of its own detection — so best responses are strictly decreasing in rivals' investment and the game is one of strategic substitutes. Since $\gamma \leq 1$, $g'(x) \leq \bar{g}'$, and $|g''(x)| \leq \overline{|g''|}$, each term of (40) is bounded above by its counterpart in (b) scaled by $(N-1)/N^2 < 1/N$ for every $N \geq 2$, so $|\partial F/\partial I| < \partial F/\partial I_i$. Hence $|\psi'(I)| < 1$, ψ is a contraction, and the symmetric fixed point is unique.

For monotonicity in γ, write (7) as $F(I,\gamma) \equiv c'(I) - \gamma h(I) = 0$, where $h(I)$ is the bracketed term. Bound (a) ensures $h(I) > 0$ on $[0,1]$ (the direct benefit of detection is not overwhelmed by model decay), so $\partial F/\partial \gamma = -h(I) < 0$; bound (b) gives $\partial F/\partial I = c''(I) - \gamma h'(I) > 0$. By the implicit function theorem, $dI^*/d\gamma = -(\partial F/\partial \gamma)/(\partial F/\partial I) > 0$, and continuity of I^* on $(0,1]$ follows from the same argument. At $\gamma = 0$, $0 \leq I^*(\gamma) \leq (c')^{-1}(\gamma\, h(1)) \to (c')^{-1}(0) = 0 = I^*(0)$ as $\gamma \downarrow 0$, so I^* is continuous on all of $[0,1]$.

Part (ii): underinvestment at $\gamma = 0$. At $\gamma = 0$, $L_i = c(I_{i,0})$, minimized at $I^*(0) = 0$ since $c'(0) = 0$. The right-hand side of (8) equals $\frac{2\alpha V}{N} > 0$ at $I = 0$ (using $g(0) = 0$, since $k'(0) = 0$), while $c'(0) = 0$; since c' is strictly increasing, $I^W > 0 = I^*(0)$.

Part (iii): overinvestment at $\gamma = 1$. Subtracting the social FOC (8) from the private FOC (7) at $\gamma = 1$, evaluated at a common I: every term cancels except the $g'(\cdot)$ feedback term, weighted $1/N$ privately and 1 socially. The private right-hand side exceeds the social right-hand side by

$$(1-\theta)^2 \alpha V I\, g'\big((1-\theta)\alpha V I\big) \cdot \frac{\alpha V}{N}\Big(1 - \frac{1}{N}\Big) \;=\; \frac{\alpha V}{N}(1-\theta)^2 \alpha V\, g'\big((1-\theta)\alpha V I\big)\, I \cdot \frac{N-1}{N},$$

which is (9), strictly positive since $g' > 0$, $\theta < 1$, $N \geq 2$. Define $F_1(I) \equiv c'(I) -$ (private RHS)(I) and $F_0(I) \equiv c'(I) -$ (social RHS)(I) at $\gamma = 1$, so $F_1(I^*(1)) = 0 = F_0(I^W)$. Since private RHS exceeds social RHS at every I, $F_1(I) < F_0(I)$ for every I; in particular $F_1(I^W) < F_0(I^W) = 0 = F_1(I^*(1))$. By (b), F_1 is strictly increasing, so $F_1(I^W) < F_1(I^*(1))$ implies $I^W < I^*(1)$.

Part (iv): optimal liability. By parts (ii) and (iii), $I^*(0) < I^W < I^*(1)$. By part (i), $I^*(\gamma)$ is continuous on $[0,1]$ and strictly increasing on $(0,1]$ (hence on $[0,1]$, given $I^*(0) = 0 < I^*(\gamma)$ for $\gamma > 0$), so the intermediate value theorem delivers a unique $\gamma^* \in (0,1)$ with $I^*(\gamma^*) = I^W$. □

C.2 Proof of Corollary 1

Proof. Closed forms. Under $c(I) = \frac{1}{2}I^2$, $k(x) = \frac{1}{2}x^2$, $g(x) = x$ and $g' \equiv 1$. Substituting into (7),

$$I \;=\; \frac{\gamma\alpha V}{N}\Big[2 - (1-\theta)^2\alpha V I\Big(1 + \tfrac{1}{N}\Big)\Big],$$

linear in I; solving gives the first expression in (10). Substituting into (8) analogously gives $I^W = \dfrac{2\alpha V}{N + 2(\alpha V)^2(1-\theta)^2}$. Setting $I^*(\gamma^*) = I^W$ and cross-multiplying,

$$\gamma^* N\big[N + 2(\alpha V)^2(1-\theta)^2\big] \;=\; N^2 + \gamma^*(\alpha V)^2(1-\theta)^2(N+1),$$

which, using $2N - (N+1) = N - 1$, rearranges to $\gamma^*\big[N^2 + (\alpha V)^2(1-\theta)^2(N-1)\big] = N^2$, the second expression in (10).

Monotonicity in θ. Writing $D(\theta) \equiv N^2 + (\alpha V)^2(1-\theta)^2(N-1)$, so $\gamma^* = N^2/D$,

$$\frac{\partial\gamma^*}{\partial\theta} \;=\; -\frac{N^2 D'(\theta)}{D^2} \;=\; \frac{2(1-\theta)(\alpha V)^2(N-1)N^2}{D^2} \;>\; 0,$$

using $D'(\theta) = -2(1-\theta)(\alpha V)^2(N-1)$.

Monotonicity in N. Writing $c \equiv (\alpha V)^2(1-\theta)^2$ and $\gamma^*(N) = N^2/[N^2 + c(N-1)]$,

$$\frac{\partial\gamma^*}{\partial N} \;=\; \frac{2N\big[N^2 + cN - c\big] - N^2(2N + c)}{\big[N^2 + cN - c\big]^2} \;=\; \frac{cN(N-2)}{\big[N^2 + c(N-1)\big]^2},$$

strictly positive for $N > 2$ and exactly zero at $N = 2$. The vanishing derivative at $N = 2$ does not stop the function from increasing strictly across the whole range: for any $2 \le N_1 < N_2$, $\gamma^*(N_2) > \gamma^*(N_1)$ iff $(N_1 - 1)N_2^2 > (N_2 - 1)N_1^2$, and

$$(N_1{-}1)N_2^2 - (N_2{-}1)N_1^2 \;=\; (N_2{-}N_1)\big[N_1N_2 - N_1 - N_2\big] \;=\; (N_2{-}N_1)\big[(N_1{-}1)(N_2{-}1) - 1\big] \;>\; 0,$$

since $N_1 \ge 2$ and $N_2 > 2$ give $(N_1 - 1)(N_2 - 1) > 1$. Equivalently, $(N-1)/N^2$ attains its maximum on $[2, \infty)$ uniquely at $N = 2$, where $(N-1)/N^2 - 1/4 = -(N-2)^2/(4N^2)$, so γ^* attains its minimum there and rises strictly thereafter.

Limits. $\gamma^* \to 1$ as $\theta \to 1$ or $N \to \infty$ is immediate from the closed form in (10), since the correction term $(\alpha V)^2(1-\theta)^2(N-1)$ vanishes in the first limit and is dominated by N^2 in the second. □

C.3 Proof of Lemma 1

Proof. The no-debanking case. L_i^{nd} is literally L_i from (3) at $\alpha = \bar{\alpha}$, and u_{psp} is additively separable in Π_i^{nd}, so $I_{\mathrm{nd}}(\gamma)$ solves the same private FOC as Corollary 1, evaluated at $\alpha = \bar{\alpha}$: $I_{\mathrm{nd}}(\gamma) = I^*(\gamma)|_{\alpha=\bar{\alpha}}$.

The debanking case. Under quadratic costs,

$$L_i^{\mathrm{d}} = \frac{\gamma\bar{\alpha}V}{N}(1 - I_{i,0}) + \frac{\gamma\delta V}{N}\big[(1 - I_{i,0}) + (1-\theta)^2\delta V\bar{I}_0 I_{i,0}\big] + \tfrac{1}{2}I_{i,0}^2.$$

Differentiating with $\partial\bar{I}_0/\partial I_{i,0} = 1/N$ and imposing symmetry $I_{i,0} = \bar{I}_0 = I$:

$$I\Big[1 + \frac{\gamma(1-\theta)^2\delta^2V^2(N+1)}{N^2}\Big] = \frac{\gamma V(\bar{\alpha}+\delta)}{N} = \frac{\gamma V(2\bar{\alpha}-\beta)}{N},$$

using $\delta \equiv \bar{\alpha} - \beta$, which rearranges to the stated $I_{\mathrm{d}}(\gamma)$. Both losses are strictly convex in $I_{i,0}$ for any rival profile under Assumption 2, by the same argument as Proposition 1(i), so these first-order conditions characterise the unique symmetric equilibria. □

C.4 Proof of Lemma 2

Proof. Part (i): the threshold.

Substituting $I_{\mathrm{nd}}(\gamma), I_{\mathrm{d}}(\gamma)$ into $\Pi_i^{\mathrm{nd}}, \Pi_i^{\mathrm{d}}$ from (17) gives $\Delta(\gamma) = u_{\mathrm{psp}}\bar{\alpha}/2 - f(\gamma)$, where $f \equiv L^{\mathrm{nd}} - L^{\mathrm{d}}$ is the gap in equilibrium losses. At $\gamma = 0$ both detection levels, and hence both losses, vanish, so $\Delta(0) = u_{\mathrm{psp}}\bar{\alpha}/2 > 0$: at zero liability debanking only forfeits franchise value. The equilibrium levels (19) are rational in γ with everywhere-positive denominators, so Δ is continuous on $[0,1]$ with finitely many sign changes. Hence $\tilde{\gamma} = \inf\{\gamma : \Delta(\gamma) < 0\}$ is well defined and strictly positive; $\Delta > 0$ on $[0, \tilde{\gamma})$ gives retention below the threshold; and when $\tilde{\gamma} < 1$, continuity gives $\Delta(\tilde{\gamma}) = 0$ with $\Delta < 0$, hence debanking, on an open interval immediately to the right of $\tilde{\gamma}$. This is everything part (i) asserts.

Part (ii), Step 1: the difference formula.

Write $D_1 \equiv N^2 + s\bar{\alpha}^2$ and $D_2 \equiv N^2 + s\delta^2$ (reserving k for the adaptation cost), so $I_{\mathrm{nd}} = 2\gamma\bar{\alpha}VN/D_1$ and $I_{\mathrm{d}} = \gamma V(\bar{\alpha}+\delta)N/D_2$. Then

$$I_{\mathrm{nd}} - I_{\mathrm{d}} = \gamma VN\,\frac{2\bar{\alpha}D_2 - (\bar{\alpha}+\delta)D_1}{D_1D_2},$$

and using $2\bar{\alpha} - (\bar{\alpha}+\delta) = \beta$ and $2\delta^2 - \bar{\alpha}\delta - \bar{\alpha}^2 = (2\delta+\bar{\alpha})(\delta - \bar{\alpha}) = -\beta(\bar{\alpha}+2\delta)$,

$$2\bar{\alpha}D_2 - (\bar{\alpha}+\delta)D_1 = N^2\beta + s\,\bar{\alpha}\big(2\delta^2 - \bar{\alpha}\delta - \bar{\alpha}^2\big) = \beta\big[N^2 - s\,\bar{\alpha}(\bar{\alpha}+2\delta)\big],$$

which gives (20) after dividing by $D_1 D_2 > 0$.

Part (ii), Step 2: the cutoff.

The sign of $I_{\text{nd}} - I_{\text{d}}$ is the sign of $N^2 - s\,\bar{\alpha}(\bar{\alpha} + 2\delta)$, increasing in θ, giving θ^* at the zero crossing. Since $(1-\theta^*)^2 = N^2/[(N+1)\gamma V^2 \bar{\alpha}(\bar{\alpha}+2\delta)]$ and $N^2/(N+1)$ is strictly increasing in N, θ^* is strictly decreasing in N: as N grows, the complement region shrinks and debanking substitutes for detection over an increasingly wide range of θ. □

Remark: the shape of the debanking set. The lemma deliberately claims only the first crossing; the shape of the full debanking set, invoked in the discussion following the lemma in Section V, is established as follows. *Near $\theta = 1$, analytically:* as $\theta \to 1$ ($s \to 0$), both detection levels are linear in γ and each regime's equilibrium loss reduces to $L = I(1 - I/2)$ at its own detection level, so the loss gap factors as $f = (I_{\text{nd}} - I_{\text{d}})\big[1 - (I_{\text{nd}} + I_{\text{d}})/2\big] = (\gamma V \beta/N)\big[1 - \gamma V(3\bar{\alpha} + \delta)/(2N)\big]$, a concave parabola through the origin. Δ is then strictly convex with $\Delta(0) > 0$: it has at most two zeros, the debanking set is exactly the interval $\{f > u_{\text{psp}}\bar{\alpha}/2\}$, empty once $u_{\text{psp}} \geq \beta/[\bar{\alpha}(3\bar{\alpha} + \delta)]$, and by continuity this two-edged window persists for θ near one—the re-entry sliver discussed in the text. *For general θ, numerically:* over the admissible region $\{\bar{\alpha}V < 1,\ N \geq 2\}$ the loss gap is everywhere increasing or single-peaked in γ, so the debanking set is an interval with $\tilde{\gamma}$ as its lower edge. We state this as a numerical observation, not part of the lemma; none of the paper's results rely on it.

C.5 Proof of Proposition 2

Proof. (i) $I_{\text{nd}}(\gamma) = I^*(\gamma)|_{\alpha=\bar{\alpha}}$ by Lemma 1, so the equation defining $\gamma^{F,*}$ is identical to the one Corollary 1 already solves at $\alpha = \bar{\alpha}$; $I^W(\alpha)$ follows from (8) under quadratic costs by the same steps as (10), with $I^W[1 + 2(\alpha V)^2(1-\theta)^2/N] = 2\alpha V/N$.

(ii)–(iii) Immediate from $\gamma_{\text{d}}^* \equiv \min(\gamma^{F,*}, \tilde{\gamma})$ together with part (i): $\gamma \mapsto I^*(\gamma)|_{\alpha=\bar{\alpha}}$ is strictly increasing on $[0,1]$ (Proposition 1(i) at $\alpha = \bar{\alpha}$), so $\gamma_{\text{d}}^* = \tilde{\gamma}$ implements the highest detection level consistent with no debanking whenever Region II obtains, and $\gamma_{\text{d}}^* \leq \gamma^{F,*}$ holds by construction of the minimum, with strict inequality exactly when $\tilde{\gamma} < \gamma^{F,*}$. □

C.6 Proof of Lemma 3

Proof. Step 0: the fraudster's problem. Because closure precedes adaptation (Section V.A) and λ_1^R is common knowledge, the fraudster solves (27), the baseline problem (4) with the prize αV replaced by $\lambda_1^R \alpha_m V$, giving (28). Interiority requires $\lambda_1^R \alpha_m V < k'(1)$, the analogue of Assumption 1 with α replaced by λ_1^R and V replaced by $\alpha_m V$. Since g is strictly increasing and $\lambda_1^{\text{S}} < \lambda_1^{\text{NS}}$, the fraudster adapts strictly less under sharing at any common $\bar{I}_0 > 0$.

Step 1: closed form. Differentiating (29) with $\partial \bar{I}_0/\partial I_{i,0} = 1/N$ and imposing symmetry $I_{i,0} = \bar{I}_0 = I$ gives, under quadratic costs,

$$I\Big[1 + \frac{\hat{s}(\lambda_1^R)^2}{N^2}\Big] = \frac{\gamma \alpha_m V(\lambda + \lambda_1^R)}{N},$$

which rearranges to (30). Strict convexity of L_i^R in $I_{i,0}$ under the analogue of Assumption 2, evaluated with α replaced by λ_1^R and V replaced by $\alpha_m V$, makes this the unique symmetric equilibrium, by the same argument as Proposition 1(i): the period-0 term $\gamma\lambda\alpha_m V(1-I_{i,0})/N$ is affine in $I_{i,0}$ and contributes nothing to the curvature conditions that argument relies on.

Step 2: comparison. Write $D_r \equiv N^2 + \hat{s}X_r^2$ for a generic period-1 exposure X_r sharing a common period-0 multiplier Λ, so $I^r \propto (\Lambda + X_r)/D_r$ for whatever prize parameter enters $\hat{s}$. Then

$$\begin{aligned} I^a - I^b \;&\propto\; (\Lambda + X_a)D_b - (\Lambda + X_b)D_a \\ &=\; N^2(X_a - X_b) + \hat{s}\big[(\Lambda + X_a)X_b^2 - (\Lambda + X_b)X_a^2\big] \\ &=\; (X_a - X_b)\Big[N^2 - \hat{s}\big(\Lambda(X_a + X_b) + X_a X_b\big)\Big], \end{aligned}$$

using $(\Lambda + X_a)X_b^2 - (\Lambda + X_b)X_a^2 = -(X_a - X_b)[\Lambda(X_a + X_b) + X_a X_b]$. Setting $(\Lambda, X_a, X_b) = (\lambda, \lambda_1^{\mathrm{NS}}, \lambda_1^{\mathrm{S}})$ gives

$$I^{\mathrm{NS}} - I^{\mathrm{S}} \;=\; \gamma\alpha_m V N\big(\lambda_1^{\mathrm{NS}} - \lambda_1^{\mathrm{S}}\big) \cdot \frac{N^2 - \hat{s}\big(\lambda(\lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}}) + \lambda_1^{\mathrm{NS}}\lambda_1^{\mathrm{S}}\big)}{D_{\mathrm{NS}} D_{\mathrm{S}}},$$

while setting instead $(\Lambda, X_a, X_b) = (\bar{\alpha}, \bar{\alpha}, \delta)$ with the prize left at V (since $\bar{\alpha}, \delta$ need no frequency correction) recovers (20), confirming the two comparisons are instances of the same abstract formula applied with different prize parameters. The bracket $N^2 - \hat{s}(\lambda(\lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}}) + \lambda_1^{\mathrm{NS}}\lambda_1^{\mathrm{S}})$ is strictly decreasing in $\hat{s}$, hence strictly increasing in θ, zero exactly at (31), positive for $\theta > \theta^{**}$ and negative for $\theta < \theta^{**}$, signing $I^{\mathrm{NS}} - I^{\mathrm{S}}$ as claimed since $\lambda_1^{\mathrm{NS}} - \lambda_1^{\mathrm{S}} > 0$ and $D_{\mathrm{NS}}, D_{\mathrm{S}} > 0$. $(1 - \theta^{**})^2 \propto N^2/(N+1)$ is strictly increasing in N, so θ^{**} is strictly decreasing in N. □

C.7 Proof of Lemma 4

Proof. (i). Fix $\gamma > 0$, $\theta \in [0, 1]$, and view the symmetric-equilibrium loss as a function of the retained pool: for $X \in [\lambda_1^{\mathrm{S}}, \lambda_1^{\mathrm{NS}}]$ let $M(X) \equiv L_i\big(X; I^X, I^X\big)$, where $I^X = \gamma\alpha_m V N(\lambda + X)/D_X$ and $D_X \equiv N^2 + \hat{s}X^2$ is the equilibrium of (30) at pool X. Then $\Delta = M(\lambda_1^{\mathrm{NS}}) - M(\lambda_1^{\mathrm{S}})$, so it suffices to show $M'(X) > 0$. Because I^X is a Nash equilibrium, not the

minimiser of the diagonal map $I \mapsto L_i(X; I, I)$, there is no envelope theorem here; instead, using the equilibrium condition $\partial L_i/\partial I_{i,0} = -(1/N)\,\partial L_i/\partial \bar{I}_0$ to eliminate the own-investment derivative,

$$M'(X) \;=\; \frac{\partial L_i}{\partial X} \;+\; \frac{N-1}{N}\,\frac{\partial L_i}{\partial \bar{I}_0}\,\frac{dI^X}{dX},$$

the non-envelope correction being exactly the $(N-1)/N$ adaptation wedge of Proposition 1. Evaluating both derivatives at the symmetric point, with $I \equiv I^X$,

$$M'(X) \;=\; \frac{\gamma\alpha_m V}{N}\Big\{(1-I) + (1-\theta)^2\alpha_m V X\, I\,\Big[2I + \frac{N-1}{N}\,X\,\frac{dI^X}{dX}\Big]\Big\},$$

and substituting $dI^X/dX = \gamma\alpha_m V N\big[N^2 - \hat{s}X^2 - 2\hat{s}\lambda X\big]/D_X^2$ gives

$$2I + \frac{N-1}{N}\,X\,\frac{dI^X}{dX} \;=\; \frac{\gamma\alpha_m V}{D_X^2}\Big\{N^2(3N-1)X + 2N^3\lambda + \hat{s}X^2\big[(N+1)X + 2\lambda\big]\Big\} \;>\; 0,$$

every term in the brace being positive ($3N^3X$ dominates the lone negative cross term N^2X). Since $I < 1$, both terms in $M'(X)$ are positive, so $M' > 0$ and $\Delta > 0$ for every $\gamma > 0$ and every $\theta \in [0,1]$, including $\theta = 1$. At $\theta = 1$ ($\hat{s} = 0$), $I^R(\gamma, 1) = A_R$ and $L^R(\gamma, 1) = A_R - \frac{1}{2}A_R^2$, giving (32), strictly positive since $x \mapsto x - \frac{1}{2}x^2$ is strictly increasing on $[0,1)$ and $A_{\mathrm{NS}} > A_{\mathrm{S}}$. At $\gamma = 0$, $I^{\mathrm{NS}} = I^{\mathrm{S}} = 0$ and both losses are zero.

(ii). Collecting terms in (29) at symmetry and substituting $I^R = A_R N^2/D_R$, $D_R \equiv N^2 + \hat{s}(\lambda_1^R)^2$, gives

$$L^R(\gamma, \theta) \;=\; A_R - \frac{A_R^2 N^2}{D_R} + \frac{A_R^2 N^3 \hat{s}(\lambda_1^R)^2}{(N+1)D_R^2} + \frac{A_R^2 N^4}{2D_R^2}.$$

Expanding in $\hat{s}$ around $\hat{s} = 0$, using $D_R^{-1} = N^{-2}\big(1 - \hat{s}(\lambda_1^R)^2/N^2\big) + O(\hat{s}^2)$, the $O(\hat{s}^0)$ term is $A_R - \frac{1}{2}A_R^2$ and the three $O(\hat{s}^1)$ contributions sum to $A_R^2(\lambda_1^R)^2\hat{s}/[N(N+1)]$, since A_R does not depend on θ. Hence

$$\Delta(\gamma, \theta) \;=\; \Delta(\gamma, 1) + \frac{\hat{s}}{N(N+1)}\Big[A_{\mathrm{NS}}^2(\lambda_1^{\mathrm{NS}})^2 - A_{\mathrm{S}}^2(\lambda_1^{\mathrm{S}})^2\Big] + O(\hat{s}^2),$$

and since $A_{\mathrm{NS}} > A_{\mathrm{S}} > 0$ and $\lambda_1^{\mathrm{NS}} > \lambda_1^{\mathrm{S}} > 0$, the bracket is strictly positive. Factoring it, $A_{\mathrm{NS}}^2(\lambda_1^{\mathrm{NS}})^2 - A_{\mathrm{S}}^2(\lambda_1^{\mathrm{S}})^2 = \gamma^2(\alpha_m V)^2 N^{-2}(\lambda_1^{\mathrm{NS}} - \lambda_1^{\mathrm{S}})(\lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}} + \lambda)[(\lambda_1^{\mathrm{NS}})^2 + \lambda\lambda_1^{\mathrm{NS}} + (\lambda_1^{\mathrm{S}})^2 + \lambda\lambda_1^{\mathrm{S}}]$, and since $\hat{s} = (N+1)\gamma(1-\theta)^2(\alpha_m V)^2$ depends on θ through $(1-\theta)^2$, the chain rule gives $\partial\Delta/\partial\theta = -2(1-\theta)\Psi + O\big((1-\theta)^3\big)$ with Ψ as displayed in the statement: Δ is strictly decreasing in θ on a left-neighbourhood of $\theta = 1$, and the derivative vanishes exactly at $\theta = 1$.

(iii). At $\theta = 1$, $\Delta(\gamma, 1) = (A_{\mathrm{NS}} - A_{\mathrm{S}})\big[1 - \frac{1}{2}(A_{\mathrm{NS}} + A_{\mathrm{S}})\big]$ with A_R linear in γ, so $\partial\Delta/\partial\gamma|_{\theta=1} = (c_{\mathrm{NS}} - c_{\mathrm{S}})\big[1 - \gamma(c_{\mathrm{NS}} + c_{\mathrm{S}})\big]$ with $c_R \equiv \alpha_m V(\lambda + \lambda_1^R)/N$, which is strictly positive under the maintained condition $\gamma\alpha_m V(2\lambda + \lambda_1^{\mathrm{NS}} + \lambda_1^{\mathrm{S}}) \leq N$, i.e. $\gamma(c_{\mathrm{NS}} + c_{\mathrm{S}}) \leq 1$; this condition holds for all admissible parameters (each prize below $k'(1) = 1$) once $N \geq 4$. Away from $\theta = 1$ we verify $\partial\Delta/\partial\gamma > 0$ under the same condition numerically over the admissible region, in the same sense as the numerical step flagged in the remark following Lemma 2. Monotonicity in γ, continuity, $\Delta(0, \theta) = 0$, and $\Delta(1, \theta) > K/N$ then deliver a unique $\tilde{\gamma}(K, \theta) \in (0, 1)$ with sharing exactly on $[\tilde{\gamma}, 1]$; when $K = 0$, part (i) gives sharing at every $\gamma > 0$ and no interior threshold exists. The comparative static $d\tilde{\gamma}/d\theta = -(\partial\Delta/\partial\theta)/(\partial\Delta/\partial\gamma) > 0$ follows wherever $\partial\Delta/\partial\theta < 0$, which part (ii) establishes near $\theta = 1$ and which holds numerically throughout $[0, 1)$. □

C.8 Proof of Proposition 3

Proof. The planner sets every PSP to a common I directly and minimizes

$$W^R(I) \;=\; \lambda\alpha_m V(1 - I) + \lambda_1^R \alpha_m V\big[(1 - I) + (1 - \theta)^2 \lambda_1^R \alpha_m V I^2\big] + \frac{N}{2} I^2,$$

giving $I^{W,R} = \alpha_m V(\lambda + \lambda_1^R)/[N + 2(1 - \theta)^2(\lambda_1^R)^2(\alpha_m V)^2]$. Setting $I^R(\gamma^{R,*}) = I^{W,R}$ and cross-multiplying, using $2N - (N + 1) = N - 1$ as in the proof of Corollary 1, gives

$$\gamma^{R,*}\big[N^2 + (1 - \theta)^2(\alpha_m V)^2(\lambda_1^R)^2(N - 1)\big] \;=\; N^2,$$

i.e. (33); λ cancels since it multiplies $\alpha_m V(\lambda + \lambda_1^R) \neq 0$ on both sides before cross-multiplication. Monotonicity in θ, the limit at $\theta \to 1$, and interiority follow exactly as in Corollary 1 with α replaced by λ_1^R and V replaced by $\alpha_m V$; the ranking follows since $\gamma^{R,*}(\theta)$ is strictly decreasing in λ_1^R and $\lambda_1^{\mathrm{NS}} > \lambda_1^{\mathrm{S}}$. □

C.9 Proof of Corollary 2

Proof. Fix the symmetric period-0 outcome $(I_{i,0}, \bar{I}_0)$ with $I_{i,0} = \bar{I}_0$, inherited from Proposition 1. Absent reinvestment, PSP i's period-1 fraud rate is the decayed rate of equation (6), $\phi_{i1}^{\mathrm{old}} = (1 - I_{i,0}) + (1 - \theta)^2 \alpha V \bar{I}_0 I_{i,0}$. If PSP i reinvests, it pays $c(I_{i,1})$ at $t = 1$ to redeploy a technology of intensity $I_{i,1}$ retrained on post-adaptation data, achieving the decay-free rate $\phi_{i1}^{\mathrm{new}}(I_{i,1}) = 1 - I_{i,1}$.

Private reinvestment choice. PSP i chooses $I_{i,1}$ to minimise $\frac{\gamma\alpha V}{N}(1 - I_{i,1}) + c(I_{i,1})$, giving the first-order condition $c'(I_{i,1}^*(\gamma)) = \gamma\alpha V/N$; strict convexity of c makes $I_{i,1}^*(\gamma)$ the unique

minimiser. The private gain from reinvesting, relative to retaining the decayed model, is

$$r_{\mathrm{PSP}}(\gamma) \;\equiv\; \frac{\gamma\alpha V}{N}\Big[\phi_{i1}^{\mathrm{old}} - \phi_{i1}^{\mathrm{new}}\big(I_{i,1}^{*}(\gamma)\big)\Big] - c\big(I_{i,1}^{*}(\gamma)\big) \;=\; \frac{\gamma\alpha V}{N}\Big[(1-\theta)^2\alpha V\bar{I}_0 I_{i,0} + \big(I_{i,1}^{*}(\gamma) - I_{i,0}\big)\Big] - c\big(I_{i,1}^{*}(\gamma)\big), \tag{43}$$

using $\phi_{i1}^{\mathrm{old}} - (1 - I_{i,1}^{*}(\gamma)) = (1-\theta)^2\alpha V\bar{I}_0 I_{i,0} + I_{i,1}^{*}(\gamma) - I_{i,0}$.

The planner's problem and the source of the wedge. Taken literally, the two-period model of Section III ends after period 1, so a reinvested intensity $I_{i,1}$ faces no further fraudster response, and the planner's symmetric problem in I_1 is $\min_{I_1} \alpha V(1 - I_1) + Nc(I_1)$, with first-order condition $c'(I_1^W) = \alpha V/N$. At $\gamma = 1$ this coincides exactly with the private condition at $\gamma = 1$, so within the literal two-period model the reinvestment margin carries no distortion at $\gamma = 1$: $I_1^W = I_{i,1}^{*}(1)$. Generating overreinvestment therefore requires the additional channel already flagged in the main text: that reinvestment raises $\bar{I}_1$, triggering a further round of fraudster adaptation of the same qualitative form as Section IV, informally representing a continuation beyond the model's two periods.[27] We capture this in reduced form as an additional social cost of reinvestment,

$$\Delta \;\equiv\; \frac{N-1}{N}(1-\theta)^2(\alpha V)^2\bar{I}_0 I_{i,1}^{*}(1) \;>\; 0, \tag{44}$$

constructed by the same $(N-1)/N$ logic as the wedge in (9): an atomistic PSP internalises only $1/N$ of the effect of its own reinvestment on the pool of adaptation a further round would draw on, while the planner internalises the effect in full. With Δ so defined, the social reinvestment gain, evaluated at the privately chosen intensity $I_{i,1}^{*}(1)$ and scaled to the full N-PSP market, is

$$r_W \;\equiv\; \alpha V\Big[(1-\theta)^2\alpha V\bar{I}_0 I_{i,0} + \big(I_{i,1}^{*}(1) - I_{i,0}\big)\Big] - Nc\big(I_{i,1}^{*}(1)\big) - \Delta.$$

Comparing to N times (43) evaluated at $\gamma = 1$, the first two terms coincide exactly, since setting $\gamma = 1$ in (43) and multiplying by N gives $N \cdot r_{\mathrm{PSP}}(1) = \alpha V[(1-\theta)^2\alpha V\bar{I}_0 I_{i,0} + (I_{i,1}^{*}(1) - I_{i,0})] - Nc(I_{i,1}^{*}(1))$, so

$$r_W \;=\; N \cdot r_{\mathrm{PSP}}(1) - \Delta \;<\; N \cdot r_{\mathrm{PSP}}(1),$$

since $\Delta > 0$ for $N \geq 2$ and $\theta < 1$. Consequently, whenever $0 < r_{\mathrm{PSP}}(1) < \Delta/N$, we have $N \cdot r_{\mathrm{PSP}}(1) > 0$ but $r_W = N \cdot r_{\mathrm{PSP}}(1) - \Delta < 0$: PSPs strictly prefer to reinvest ($r_{\mathrm{PSP}}(1) > 0$) in a region where reinvestment is socially undesirable ($r_W < 0$). Hence, at full liability, PSPs

[27] Formalising this continuation as an explicit third period or an infinite-horizon repeated game is beyond the scope of this extension; we adopt the reduced-form representation below, by direct analogy with Proposition 1, and flag it as a maintained modelling assumption rather than a result derived from Section III as stated.

reinvest too frequently.

Optimal liability. When $r_{\mathrm{PSP}}(\gamma) \leq 0$, no PSP reinvests and the model reduces to the baseline of Section IV, so the optimal rule is γ^* from Proposition 1. When $r_{\mathrm{PSP}}(1) > 0$, PSPs reinvest; at $\gamma = 1$ the intensive margin, the choice of $I_{i,1}$ conditional on reinvesting, is efficient, since Δ operates only through the extensive, whether-to-reinvest, margin and the private and planner first-order conditions for $I_{i,1}$ itself coincide exactly at $\gamma = 1$. Setting $\gamma = 1$ is therefore a good approximation whenever reinvestment cycles are fast enough that the intensive margin is the first-order concern; correcting the residual excess-triggering distortion identified above would require an instrument targeted at the reinvestment decision itself, beyond γ. □

C.10 Proof of Corollary 3

Proof. Under quadratic costs and symmetry $I_{i,0} = \bar{I}_0 = I$, the private FOC associated with (34) is obtained by differentiating $L_i = \frac{\gamma\alpha V}{N}\big[(1 - I_{i,0}) + (1 - \lambda\bar{I}_0)\phi_{i1}\big] + \frac{1}{2}I_{i,0}^2$, with $\phi_{i1} = (1 - I_{i,0}) + (1-\theta)^2\alpha V\bar{I}_0 I_{i,0}$, and using $\partial\bar{I}_0/\partial I_{i,0} = 1/N$:

$$NI \;=\; \gamma\alpha V\left\{1 + \frac{\lambda}{N}\phi_1(I) + (1 - \lambda I)\left[1 - (1-\theta)^2\alpha V I\Big(1 + \frac{1}{N}\Big)\right]\right\}, \tag{45}$$

where $\phi_1(I) \equiv (1 - I) + (1-\theta)^2\alpha V I^2$ is the baseline decayed rate. The planner's symmetric problem sets $\partial\bar{I}_0/\partial I = 1$ throughout, giving

$$NI \;=\; \alpha V\Big\{1 + \lambda\phi_1(I) + (1 - \lambda I)\big[1 - 2(1-\theta)^2\alpha V I\big]\Big\}. \tag{46}$$

Subtracting the private bracket (at $\gamma = 1$) from the social bracket in (45)–(46),

$$\text{(social)} - \text{(private)}\big|_{\gamma=1} \;=\; \frac{N-1}{N}\Big[\lambda\phi_1(I) - (1-\theta)^2\alpha V I(1 - \lambda I)\Big], \tag{47}$$

which contains two terms of opposite sign: a deterrence term $\lambda\phi_1(I)$, by which the planner values PSP i's contribution to market-wide fraudster exit more than PSP i does itself, and an adaptation term $-(1-\theta)^2\alpha V I(1-\lambda I)$, the same overinvestment channel as Proposition 1(iii), scaled down by $(1 - \lambda I)$ to reflect that fewer fraudsters remain active to adapt.

Existence. At $\gamma = 0$, $I^*(0) = 0$ as in Proposition 1(ii), and $I^W > 0$ by the same argument applied to (46). Existence of $\gamma^* \in (0, 1]$ solving $I^*(\gamma^*) = I^W$ requires $I^*(1) \geq I^W$, i.e. that the deterrence term in (47) does not dominate the adaptation term at $\gamma = 1$: this is exactly Assumption 3. Under it, continuity and strict monotonicity of $I^*(\gamma)$, established as in Proposition 1(i), deliver a unique $\gamma^* \in (0, 1]$ by the intermediate value theorem.

Monotonicity in θ and λ. Apply the implicit function theorem to $I^*(\gamma^*(\theta,\lambda);\theta,\lambda) = I^W(\theta,\lambda)$. As θ rises, the adaptation term $(1-\theta)^2$ falls in both FOCs, but by a larger amount in the planner's, which weights it by 1 rather than $1+1/N$; the wedge in (47) therefore shrinks in θ, and γ^* must rise to keep $I^*(\gamma^*) = I^W$. As λ rises, the deterrence term $\lambda\phi_1(I)$ enters the planner's FOC at full weight and the private FOC at weight $1/N$; the planner's marginal benefit of detection therefore rises by more than the private one, so I^W rises relative to $I^*(\gamma)$ and γ^* must rise to compensate. □

C.11 Proof of Lemma 5

Proof. Both parts follow the proof of Lemma 2 verbatim, with δ replaced by $\tilde{\alpha}$ and $\bar{\alpha}(3\bar{\alpha}-2\beta)$ replaced by $\bar{\alpha}^2$ throughout: since exclusion uses the same exposure $\alpha(D)$ in both periods, the mixed-exposure algebra of that proof collapses to the single-pool case, and $I^X(\gamma) = AX/(N^2+CX^2)$ from (39) is single-peaked at $\alpha^\dagger = \sqrt{N^2/C}$ by the same argument used there. For part (ii), the exact comparison follows from the difference formula $I^1 - I^0 = A(\bar{\alpha}-\tilde{\alpha})\big[N^2 - C\bar{\alpha}\tilde{\alpha}\big]/\big[(N^2+C\bar{\alpha}^2)(N^2+C\tilde{\alpha}^2)\big]$, whose sign is that of $(\alpha^\dagger)^2 - \bar{\alpha}\tilde{\alpha}$; on the maintained region, $C = (N+1)\gamma(1-\theta)^2V^2$ gives $(\alpha^\dagger V)^2 = N^2/[(N+1)\gamma(1-\theta)^2] \geq N^2/(N+1) > 1$, while $\bar{\alpha}\tilde{\alpha}V^2 < 1$, so the substitution branch obtains throughout. As in that proof, the single-crossing property of $\Delta(\gamma)$ underlying part (i) is verified numerically over the admissible region $\{\bar{\alpha}V < 1,\ N \geq 2\}$ rather than established in closed form for arbitrary parameter values; we do not claim strict monotonicity of Δ beyond this region. $\Delta(0) = \lambda u_{\text{psp}}$ replaces $\bar{\alpha}u_{\text{psp}}/2$ since here the entire type-H share λ, not merely the signalled fraction $\bar{\alpha}$, is at stake. □

C.12 Proof of Proposition 4

Proof. Part (i). By (39) at $D=1$, $\alpha(1) = \bar{\alpha}$, so $I^1(\gamma) = I^*(\gamma)|_{\alpha=\bar{\alpha}}$ directly from Corollary 1. The equation defining $\gamma^{F,*}$, $I^1(\gamma^{F,*}) = I^W(\bar{\alpha})$, is therefore identical to the one Corollary 1 already solves at $\alpha = \bar{\alpha}$, so $\gamma^{F,*}$ coincides with the object of the same name defined in Proposition 2.

Parts (ii)–(iii). Immediate from $\gamma_D^* \equiv \min(\gamma^{F,*}, \bar{\gamma})$ together with part (i): $\gamma \mapsto I^1(\gamma) = I^*(\gamma)|_{\alpha=\bar{\alpha}}$ is strictly increasing on $[0,1]$ (Proposition 1(i) at $\alpha = \bar{\alpha}$), and by Lemma 5(i) PSPs serve both types, $D=1$, exactly on $\gamma \leq \bar{\gamma}$. On this region, raising γ strictly raises detection toward $I^W(\bar{\alpha})$, so $\gamma_D^* = \bar{\gamma}$ implements the highest detection level consistent with serving type H whenever $\bar{\gamma} < \gamma^{F,*}$, exactly as $\tilde{\gamma}$ does for the debanking margin in Proposition 2. Consequently $\gamma_D^* \leq \gamma^{F,*}$ holds by construction of the minimum, with strict inequality exactly when $\bar{\gamma} < \gamma^{F,*}$. □

Online Appendix:

Who Pays for Payment Fraud?
Detection and Liability Rules under Strategic Fraudster Adaptation

OA1 Technical Sophistication Dictionary

The technical sophistication measure described in Section II.A.1 is based on a curated dictionary of 212 terms spanning 13 categories of advanced technical knowledge. Categories 11 (Deepfake/Liveness Bypass) and 12 (AI-Generated Content for Fraud) are excluded from the sophistication measure because they overlap with the generative AI dependent variables. All terms are matched case-insensitively as substrings against post content.

Table OA1
Technical Sophistication Dictionary (212 terms, 13 categories)

Category Name	*N*	Terms
Exploitation & Vulnerability Research	34	zero day, 0-day, cve, proof of concept, poc, privilege escalation, remote code execution, rce, buffer overflow, heap spray, rop chain, return oriented, use after free, memory corruption, heap corruption, sql injection, cross site scripting, xss, path traversal, local file inclusion, lfi, command injection, server side request, ssrf, deserialization, dll inject, dll hijack, process injection, process hollow, shellcode, sandbox escape, type confusion, exploit chain, authentication bypass
Malware Development & Architecture	32	rootkit, bootkit, uefi, kernel driver, kernel module, ring-0, crypter, obfuscat*, polymorphic, metamorphic code, dropper, stager, reverse shell, bind shell, meterpreter, c2 server, command and control, cobalt strike, reflective inject, thread hijack, amsi bypass, etw patch, edr bypass, av bypass, fully undetectable, fileless malware, botnet stub, persistence mechanism, ransomware payload, remote access trojan, infostealer, keylogger
Reverse Engineering & Low-Level	18	reverse engineer*, decompil*, disassembl*, ida pro, ghidra, ollydbg, hex editor, memory dump, windows api, ntapi, syscall, windows kernel, lsass, mimikatz, dcsync, token impersonation, lolbins, dll sideload

Continued on next page

Category Name	N	Terms
Network Attacks & Infrastructure	31	port scan, nmap, packet sniff, man in the middle, mitm, arp spoof, arp poisoning, dns poison, ssl strip, lateral movement, credential dump, pass the hash, kerberoast, active directory, eternalblue, metasploit, msf, msfvenom, hashcat, rainbow table, session hijack, cookie hijack, packet injection, ja3, nat traversal, winrm, ss7, signaling attack, diameter attack, gtp attack, roaming exploit
Reverse Proxy Phishing & AiTM	13	evilginx, modlishka, muraena, necrobrowser, transparent proxy, real-time phish*, aitm, adversary in the middle, reverse proxy phish, relay phish, oauth abuse, consent phishing, token hijack
Stealer Logs & Credential Infrastructure	9	stealer log, redline log, raccoon log, fresh combo, private combo, combo list, log market, cookie stealer, token replay
Evasion & OPSEC	19	opsec, operational security, tails os, proxy chain, runtime detection, scantime fud, heuristic detection, antidetect, browser fingerprint, multilogin, gologin, puppeteer stealth, undetected chromedriver, cloudflare bypass, akamai bypass, imperva bypass, ja3 spoof, device associ*, behavioral link
CAPTCHA Solving & Bot Infrastructure	6	captcha solv*, 2captcha, anticaptcha, captcha bypass, recaptcha bypass, captcha farm
OSINT & Recon	10	osint framework, subdomain enumeration, google dorking, breach database, metadata extraction, certificate transparency, social media enumeration, dns enumeration, email reconnaissance, theharvester
Specific Tools & Frameworks	12	setoolkit, evilginx2, openbullet, hydra brute, sqlmap, aircrack, impacket, burp suite, wireshark, bluetoothctl, hcitool, blue hydra
Crypto Privacy & Obfuscation	10	monero, xmr, tornado cash, wasabi wallet, coinjoin, chain hop, crypto mixer, btc mixer, non custodial, self custody

Continued on next page

Category Name	N	Terms
Account & Device Manipulation	9	buy kyc account, sell verified, synthetic identity, manufactured identity, complex synthetic, esim hack, esim exploit, qr code sim, remote sim provision
Advanced Payment Evasion	9	velocity limit, velocity evasion, pattern evasion, drip charge, slow charge, warm card, mimic legit, avs match, low and slow

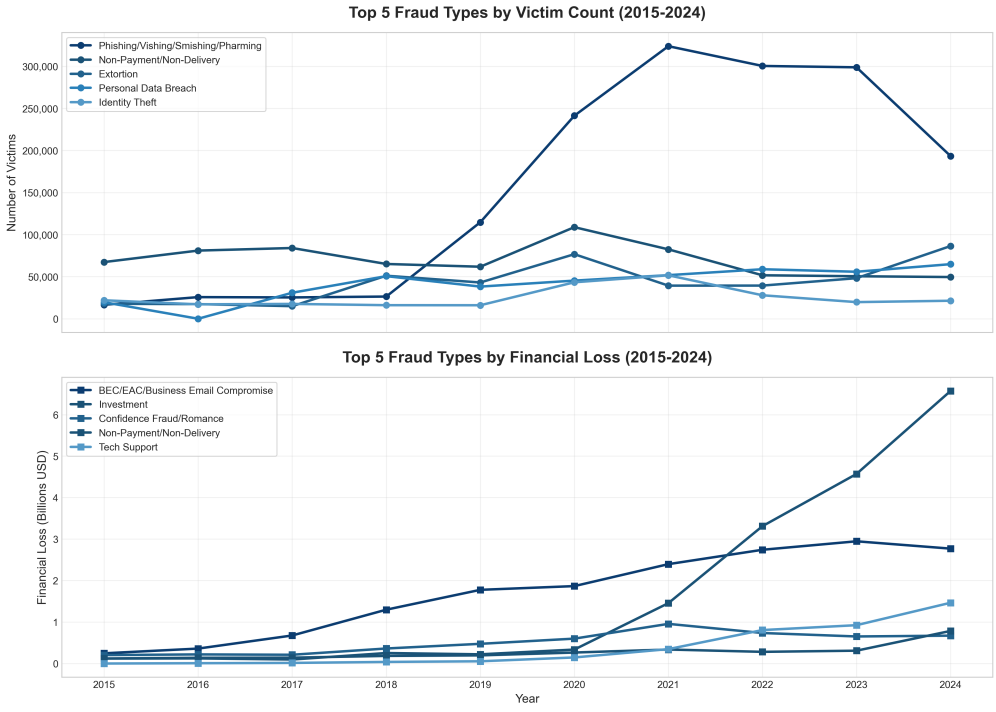
Notes: 212-term dictionary (13 categories) proxying technical sophistication (Section II.A.1), matched as case-insensitive substrings. Categories 11–12 (deepfake/liveness bypass, AI-generated content) excluded as they overlap the generative-AI dependent variables. Asterisk (*) denotes prefix match; `cve` and `xss` matched as whole words only.

OA2 Corroborating Evidence

OA2.1 Changes in Fraud Type Share over Time

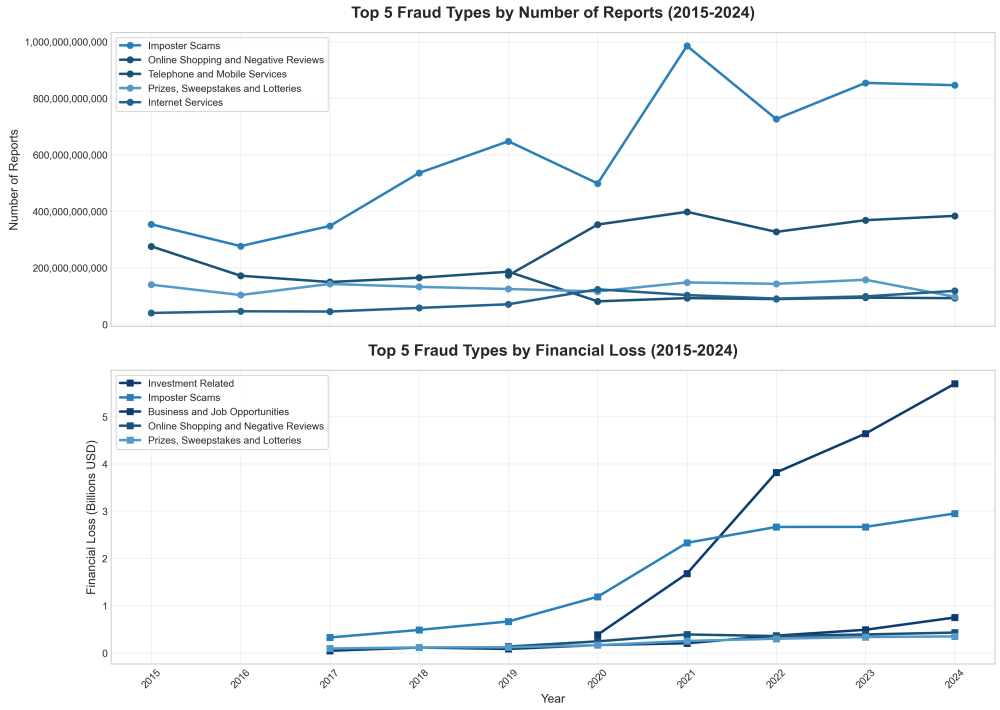
Figure OA1 shows the five most common fraud types in the United States by complaint volume (top panel) and financial losses (bottom panel) over 2015–2024. The composition shifts markedly: phishing and impersonation fraud grow steadily by complaint count, while investment fraud dominates financial losses and rises sharply after 2020. The types generating the highest losses are not the same types generating the highest complaint volume, reflecting heterogeneity in V across fraud types. Figure OA2 corroborates these patterns using FTC Sentinel data.

Figure OA1
US Fraud Trends (FBI IC3)



Notes: Evolution of the five most common internet fraud categories in the United States by complaint volume (top panel) and financial losses (bottom panel). Source: FBI Internet Crime Complaint Center (IC3).

Figure OA2
US Fraud Trends (FTC Sentinel)



Notes: Evolution of the five most common internet fraud categories in the United States by complaint volume (top panel) and financial losses (bottom panel). Source: FTC Sentinel.

These aggregate shifts have multiple potential sources, including the expansion of digital payment channels, changes in consumer behavior, and technological change. Most relevant for the present paper is the technology channel: anti-fraud technologies raise θ_k for targeted fraud types, inducing substitution toward harder-to-detect alternatives. Hayashi et al. (2015) provide direct evidence for this mechanism, documenting that card-not-present fraud rose sharply and became the dominant source of card fraud losses in the United Kingdom, France, and Canada in the years immediately following EMV chip card adoption, as fraudsters substituted away from counterfeit card-present fraud toward online channels where chip authentication offered no protection. Table OA2 reports aggregate fraud rates and crime shares for the four jurisdictions of Figure 1; the channel-level evidence around EMV adoption is in Hayashi et al. (2015).

Table OA2
Fraud Statistics by Jurisdiction

Panel A: Fraud per 100,000 inhabitants				
Year	US	EU	UK	China
2017	153.8	481.6	1087.5	24.6
2018	163.1	498.7	1175.1	28.1
2019	199.0	536.2	1305.7	38.9
2020	248.9	564.6	1393.8	50.7
2021	312.3	601.6	1697.0	46.5
2022	331.6	619.7	1922.0	38.4
2023	349.3	–	2054.8	40.4
2024	330.1	–	2073.2	–
Panel B: Fraud as Share of Total Crime				
Year	US	EU	UK	China
2017	0.04	0.16	0.12	0.03
2018	0.04	0.17	0.12	0.04
2019	0.05	0.19	0.13	0.06
2020	0.05	0.21	0.15	0.08
2021	0.07	0.22	0.16	0.07
2022	0.06	0.21	0.17	0.06
2023	0.06	–	0.19	0.07
2024	0.06	–	0.19	–

Notes: Fraud cases per 100,000 inhabitants and fraud as share of total reported crime in the US, EU, UK, and China from 2017 to 2024. Sources: US: FBI UCR; EU: Eurostat; UK: ONS Crime in England and Wales; China: National Bureau of Statistics.

OA3 Wald Tests

OA4 Adversarial Adaptation to Anti-Fraud Technologies

OA4.1 Empirical Strategy

We complement the ChatGPT analysis with a stacked event-study design that examines fraudster adaptation to 27 targeted anti-fraud technology deployments spanning 2013–2023. While the ChatGPT analysis identifies the effect of a general-purpose enabling shock, the stacked design tests whether anti-fraud technology deployments trigger the accumulation of countermeasure knowledge without suppressing discussion of the targeted attack methods.

Stacked event study. For each of the 27 anti-fraud technology deployments, we retain all creator-months in the window $[T_{0j} - 36, T_{0j} + 60]$ months around the deployment date T_{0j}. The full CrimeBB corpus of approximately 8 million posts is scored against all 443 dictionary

Table OA3
Pairwise $\hat{\beta}$ Comparison — Full Sample

DV_1	DV_2	$\Delta\hat{\beta}$	SE	z	p
1(AI)	**1**(Fraud)	0.1055	(0.0150)	7.04	0.0000 ∗ ∗∗
1(AI)	**1**(AI×Fr)	0.0706	(0.0100)	7.03	0.0000 ∗ ∗∗
1(AI)	**1**(AI×M)	0.0822	(0.0098)	8.37	0.0000 ∗ ∗∗
1(Fraud)	**1**(AI×Fr)	−0.0350	(0.0131)	−2.67	0.0076 ∗ ∗∗
1(Fraud)	**1**(Method)	−0.0049	(0.0138)	−0.36	0.7211
1(Method)	**1**(AI×M)	−0.0185	(0.0079)	−2.35	0.0189 ∗ ∗
1(AI×Fr)	**1**(AI×M)	0.0116	(0.0066)	1.76	0.0786∗

Notes: Full sample, Creator FE specification. Wald test under independence: $z = (\hat{\beta}_1 - \hat{\beta}_2)/\sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

terms simultaneously using the Aho–Corasick algorithm (Aho and Corasick, 1975), producing for each technology j a creator-month panel with binary indicators (vuln_{ijt}, ctr_{ijt}) capturing whether a creator posted about vulnerable methods or countermeasures in a given month. Each technology yields approximately 1.7 million creator-months across roughly 1 million unique creators.

Technology Events and Dictionaries We identify 27 anti-fraud technology deployments spanning 2013–2023 across nine fraud categories: carding, account takeover, banking, identity, payment systems, SIM and telecom, AI techniques, circumvention, and cryptocurrency fraud. Each deployment event T_0 is anchored to a precisely dated landmark, a product launch, regulatory milestone, or market-defining acquisition, to ensure the event date is not endogenous to forum activity. The full list with deployment rationales is in Internet Appendix Table OA4.

For each technology, we construct two curated keyword dictionaries. *Vulnerable-method terms* (178 unique terms) describe the attack vector targeted by the technology; *countermeasure terms* (270 unique terms) describe techniques for circumventing the deployed defense. The asymmetry reflects the greater diversity of circumvention strategies relative to the attack vectors they target. All dictionaries are in Internet Appendix Table OA5.

We stack the 27 technology sub-experiments into a single regression:

$$y_{ijt} = \alpha_{ij} + \gamma_t + \sum_{k \neq \text{ref}} \beta_k \cdot \mathbf{1}[\text{bin}(t - T_{0j}) = k] + \varepsilon_{ijt}, \tag{48}$$

where $y_{ijt} \in \{\text{vuln}_{ijt}, \text{ctr}_{ijt}\}$, α_{ij} are creator-by-technology fixed effects, and the β_k are event-time coefficients relative to the reference window $[-3, -1]$ months before T_{0j}. Event-time bins

are: < -24, $[-24, -13]$, $[-12, -7]$, $[-6, -4]$, $[-3, -1]$ (reference), $[0, 2]$, $[3, 5]$, $[6, 11]$, $[12, 23]$, $[24, 35]$, $[36, 47]$, $48+$. We report two variants of the time controls γ_t. The *month-of-year* variant absorbs seasonality and identifies the event-time profile from within-creator variation around each deployment; it describes how attention evolves around deployments but does not separate deployment-specific movements from market-wide trends. The *calendar-month* variant absorbs every common shock; because a full set of calendar-month effects is collinear with event time within any single technology, this variant is identified only by the staggering of deployment dates across technologies, and isolates the deployment-*specific* component of the profile. Standard errors are clustered by technology, the level at which the treatment varies.

OA4.1.1 Results: Adaptation Around Deployment

Figure OA3 presents the stacked event-study estimates under both time-control variants. We read the estimates as descriptive of the deployment cycle rather than as sharp causal effects of the deployment date, for a reason the estimates themselves reveal: attention moves *before* the dates as well as after them.

Two patterns organize the figure. First, countermeasure discussion builds *into* deployment: under month-of-year controls, every earlier pre-deployment bin lies significantly below the $[-3, -1]$ reference window, so circumvention attention rises steadily over the three years preceding each deployment date and peaks around the date itself. This is what one expects when deployment dates mark the maturation of defenses that diffused gradually, an observation consistent with how several of the events in Table OA4 are dated, and it cautions against reading the dates as surprise interventions. Second, discussion of the *targeted vulnerable methods* fades over the years after deployment: the long-run (36+ month) coefficients are negative in both variants, indicating that attention drifts away from the attack surface a defense covers. The deployment-specific component of the countermeasure profile, the calendar-month variant identified from staggered timing, is positive at long horizons but imprecisely estimated with technology-clustered errors.

Together the two patterns describe a slow reallocation of criminal attention over the deployment cycle: circumvention knowledge accumulates as a defense arrives, and the methods the defense targets lose ground afterwards. The composition of underground attention shifts from established methods toward countermeasures over horizons of years, the corpus-level counterpart of the substitution mechanism in Section III, though we emphasize that these profiles are descriptive: deployment timing is not random, and the market-wide component of the drift is absorbed only in the calendar-month variant.

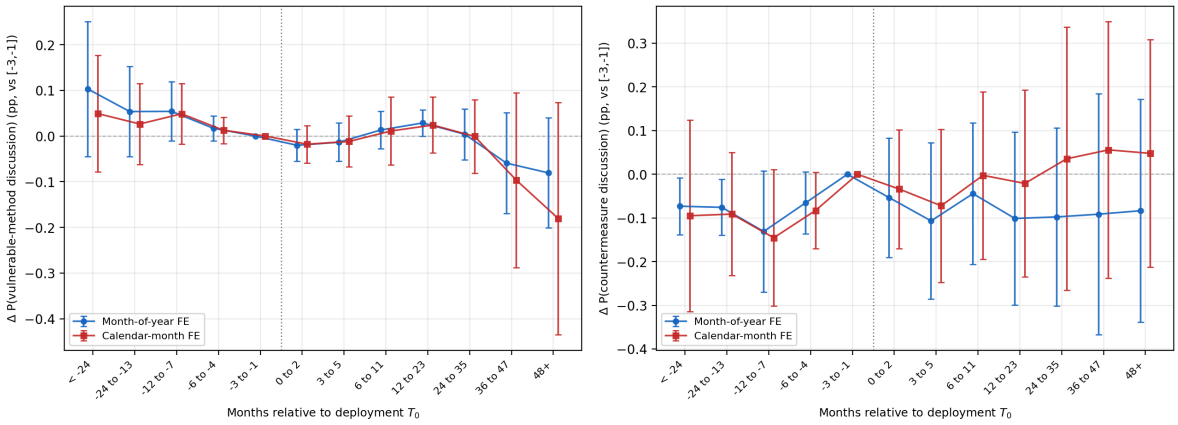


Figure OA3
Criminal Attention Around Anti-Fraud Technology Deployment

Notes: Stacked event-study coefficients, equation (48), pooling the 27 anti-fraud technology deployments in one regression with creator-by-technology fixed effects. The left panel shows the probability of vulnerable-method discussion; the right panel the probability of countermeasure discussion, both in percentage points relative to the $[-3, -1]$ month reference window. Circles: month-of-year time controls (within-creator, seasonally adjusted profile). Squares: calendar-month fixed effects (deployment-specific component, identified from staggered deployment dates). Bars are 95% confidence intervals from standard errors clustered by technology (27 clusters).

Table OA4

Anti-Fraud Technology Deployments, 2013–2023

Category	Technology	Type	Date	Description
Carding Fraud	Device Fingerprinting Maturation	Primary	2018-01-18	ThreatMetrix acquisition entrenches fingerprinting as fraud-detection standard
	reCAPTCHA v3 Invisible Scoring	Secondary	2018-10-29	Behavioral scoring replaces explicit challenges for bot detection
	ML Fraud Scoring at Scale	Secondary	2016-06-01	ML platforms (Sift, Feedzai) reach production scale at tier-1 banks
Account Takeover Fraud	Behavioral Biometrics Scaling	Primary	2017-06-01	BioCatch enables continuous authentication via keystroke/mouse dynamics
	Real-Time Phishing Proxies	Secondary	2018-07-01	EvilGinx 2.0 enables MFA bypass via session hijacking
	Credential Monitoring Services	Secondary	2013-12-04	HIBP aggregates breached credentials for early-warning defense
Banking Fraud	Open Banking APIs	Primary	2018-01-13	PSD2 mandates secure APIs, replacing screen-scraping
	Behavioral Analytics in Banking Apps	Secondary	2019-01-01	Featurespace ARIC reaches 50+ bank deployments
	Device Fingerprinting in Banking	Secondary	2018-01-18	Banks adopt ThreatMetrix to flag suspicious device access
Identity Fraud	ML Identity Verification Platforms	Primary	2019-01-01	Jumio, Onfido, Socure reach critical mass among fintechs
	Liveness Detection in KYC	Secondary	2020-06-01	iProov, FaceTec achieve ISO 30107-3 certification
	Synthetic Identity Detection	Secondary	2021-01-01	Socure Sigma detects synthetic identity construction
Payment Systems Fraud	3D Secure 2.0 Rollout	Primary	2019-04-01	Risk-based dynamic authentication mandated by card networks
	Network Tokenisation (Apple Pay)	Secondary	2014-10-20	Apple Pay pioneers single-use payment tokens
	Real-Time Fraud Scoring APIs	Secondary	2018-01-01	Stripe Radar, Adyen enable millisecond-level risk scoring
SIM and Telecom Fraud	Phone-Centric Identity APIs	Primary	2020-01-01	Prove verifies phone ownership for passwordless login
	SS7 Monitoring Deployment	Secondary	2018-06-01	GSMA guidance drives SS7 abuse detection infrastructure
	eSIM Technology Adoption	Secondary	2018-09-21	iPhone XS eSIM reduces SIM-swap vulnerability
AI Content Detection Fraud	AI Content Detection	Primary	2023-01-02	GPTZero launches first widely-used AI-text detector
	Deepfake Detection Tools	Secondary	2020-09-01	Microsoft Video Authenticator detects manipulated video
	Content Provenance Standards	Secondary	2022-01-26	C2PA v1.0 enables cryptographically-signed content metadata
Circumvention and Evasion Fraud	Advanced Bot Management	Primary	2019-09-01	Cloudflare distinguishes bots via behavioral/TLS fingerprinting
	Browser Fingerprint Hardening	Secondary	2020-01-01	Chrome 80 restricts fingerprinting and third-party cookies
	Threat Intelligence Platform Maturation	Secondary	2020-09-17	Recorded Future IPO signals threat-intel market maturity
Cryptocurrency Fraud	Blockchain Analytics at Scale	Primary	2017-02-01	Chainalysis tracks illicit funds via graph analysis
	Travel Rule Implementation	Secondary	2021-06-01	FATF formalizes counterparty reporting for exchanges
	Exchange KYC Standardisation	Secondary	2019-09-01	Binance mandates KYC; industry-wide standardization follows

Notes: Table lists the 27 anti-fraud technologies (grouped into 9 fraud categories) used to construct deployment-date shocks for the stacked event study in Section OA4. Each technology is classified as *primary* (core defense) or *secondary* (complementary defense) within its category; deployment dates correspond to product launches, regulatory effective dates, or major industry-adoption milestones. Full source citations are provided in the text version of this appendix.

Table OA5
CrimeBB Keyword Dictionaries by Anti-Fraud Technology

Technology (Category)	Keywords
Device fingerprinting maturation (Carding)	*Vuln:* basic carding, beginner carding, bin checker, card checker, card tester, cc checker, fresh socks, gate checker, live cc, luhn check, magnetic stripe, no proxy, simple method, track 2, track2. *Ctr:* anti-detect, antidetect, antidetect browser, browser fingerprint, canvas fingerprint, canvas spoof, device emulat, device id spoof, fingerprint spoof, gologin, iovation, linken sphere, multilogin, residential proxy, threat metrix, threatmetrix, user agent spoof, webgl fingerprint
reCAPTCHA v3 invisible scoring (Carding)	*Vuln:* auto checker, automated card, bot card, brute force, card checker, mass check, scripted check. *Ctr:* 2captcha, anti-captcha, anticaptcha, captcha bypass, captcha farm, captcha solv, captcha token, hcaptcha, invisible captcha, recaptcha bypass, recaptcha v3
ML fraud scoring at scale (Carding)	*Vuln:* basic method, beginner method, easy card, no check site, noob method, simple cashout. *Ctr:* feedzai, forter, fraud detect, fraud score, machine learning, risk engine, risk score, riskified, sift science, signifyd, velocity check
Behavioral biometrics scaling (Account Takeover)	*Vuln:* brute force, checker config, combo list, config, cracked account, credential stuffing, fresh log, hit combo, open bullet, openbullet, password list, sentry mba, sentrymba, simple login, storm. *Ctr:* 2fa bypass, behavioral, biocatch, cookie steal, cookie theft, evil ginx, evilginx, humaniz, mfa bypass, modlishka, mouse emulat, otp bot, real-time phish, realtime phish, reverse proxy phish, session hijack, session token, token steal, typing pattern
Real-time phishing proxies (Account Takeover)	*Vuln:* basic phish, clone site, fake login, phishing kit, phishing page, scam page, static phish. *Ctr:* evilginx, mitm phish, modlishka, muraena, necrobrowser, real-time phish, relay phish, reverse proxy, transparent proxy
Credential monitoring services (Account Takeover)	*Vuln:* database dump, leaked database, old breach, old combo, password reuse, stale log. *Ctr:* exclusive combo, fresh combo, fresh log, haveibeenpwned, hibp, not leaked, private combo, private database, unchecked combo, unshared
Open banking APIs (Banking)	*Vuln:* aggregator exploit, bank account hack, bank drop, bank login, cash out, manual transfer, online banking hack, screen scrap, web inject, wire transfer. *Ctr:* api exploit, api token, behavioral analytics, consent phish, featurespace, oauth, open banking, randomis, session replay, token hijack, transaction pattern
Behavioral analytics in banking (Banking)	*Vuln:* bank login, manual transfer, rdp bank, remote access, teamviewer, vnc bank. *Ctr:* anomaly detect, behavioral analytics, featurespace, pattern evasion, randomis, risk engine, session fingerprint, slow transfer, transaction velocity
Device fingerprinting in banking (Banking)	*Vuln:* bank login, basic access, direct login, no spoof. *Ctr:* anti-detect, antidetect, browser fingerprint, device id, device spoof, emulat, threatmetrix
ML identity verification (Identity)	*Vuln:* counterfeit id, fake birth, fake driver, fake id, fake license, fake passport, forged document, fullz, novelty id, paper id, photoshop id, scan, ssn, template. *Ctr:* 3d mask, ai generated id, camera inject, deepfake id, face swap, injection attack, iproov, jumio, kyc bypass, liveness, liveness bypass, onfido, selfie bypass, socure, synthetic identity, verification bypass, virtual camera
Liveness detection in KYC (Identity)	*Vuln:* photo spoof, printed photo, screenshot selfie, selfie photo, static image. *Ctr:* 3d mask, camera inject, deepfake, face swap, iso 30107, liveness bypass, liveness detect, manycam, obs virtual, presentation attack, virtual camera
Synthetic identity detection (Identity)	*Vuln:* authorized user, cpn, credit privacy number, credit profile, piggybacking, synthetic identity, tradeline. *Ctr:* behavioral link, cross reference, device associ, identity graph, link analysis, socure sigma, synthetic detect
3D Secure 2.0 rollout (Payment Systems)	*Vuln:* 3ds1, mcsc bypass, no 3d secure, no 3ds, no verification, non vbv, payment intercept, replay attack, static password, vbv bypass. *Ctr:* 3d secure 2, 3d secure bypass, 3ds2, 3ds2 bypass, challenge bypass, emv 3ds, frictionless, low value exempt, merchant bypass, non sca merchant, sca exempt, tra exempt
Network tokenisation (Payment Systems)	*Vuln:* card number, magnetic stripe, pan theft, raw card data, stored card. *Ctr:* apple pay exploit, device token, google pay exploit, network token, token, tokeniz, virtual card

Continued on next page

Table OA5 – continued

Technology (Category)	Keywords
Real-time fraud scoring APIs (Payment Systems)	*Vuln:* basic card, high velocity, mass charge, simple cashout. *Ctr:* adyen risk, cool down, drip charge, fraud score, risk score, slow charge, stripe radar, velocity limit
Phone-centric identity APIs (SIM/Telecom)	*Vuln:* call carrier, carrier social, number port, port out, prepaid sim, sim clone, sim jacker, sim swap, social engineer carrier. *Ctr:* carrier insider, corrupt employee, diameter, esim, esim exploit, insider, payfone, prove bypass, silent sms, sim farm, ss7, ss7 exploit, telesign, voip bypass
SS7 monitoring (SIM/Telecom)	*Vuln:* signaling attack, ss7 attack, ss7 exploit, ss7 intercept. *Ctr:* cat2 bypass, diameter security, gtp attack, roaming exploit, signaling firewall, ss7 firewall
eSIM technology (SIM/Telecom)	*Vuln:* blank sim, physical sim, sim card theft, sim clone. *Ctr:* esim exploit, esim hack, esim profile, qr code sim, remote sim
AI content detection (AI Content)	*Vuln:* ai generated, ai phishing, ai voice clone, chatgpt phish, chatgpt scam, copy paste gpt, deepfake, gpt email, gpt social engineer. *Ctr:* ai detector bypass, bypass filter, custom model, dan prompt, fraudgpt, gptzero bypass, humaniz, jailbreak, paraphras, pentagpt, quillbot, uncensored model, undetectable ai, unfiltered llm, watermark remov, wormgpt
Deepfake detection tools (AI Content)	*Vuln:* ai face, ai voice, deepfake, face swap, voice clone. *Ctr:* anti deepfake, deepfake detect, deepware, gan artifact, sensity, undetectable deepfake, video authenticat
Content provenance standards (AI Content)	*Vuln:* ai generated document, ai generated image, synthetic media. *Ctr:* c2pa, content credentials, content provenance, exif remov, metadata strip, watermark remov
Advanced bot management (Circumvention)	*Vuln:* basic bot, basic proxy, curl, headless browser, headless chrome, mechanize, phantomjs, rotate proxy, selenium, simple scrip. *Ctr:* 2captcha, akamai bypass, anti-captcha, anticaptcha, bot protect, captcha solv, cloudflare bypass, cloudflare challenge, datadome bypass, imperva bypass, ja3, ja3 spoof, kasada bypass, perimeterx bypass, playwright stealth, puppeteer stealth, residential proxy, tls fingerprint, undetected chromedriver, undetected_chromedriver
Browser fingerprint hardening (Circumvention)	*Vuln:* cross site cookie, third party cookie, tracking cookie. *Ctr:* cookie bypass, first party set, privacy sandbox, samesite bypass, storage partition
Threat intelligence maturation (Circumvention)	*Vuln:* clearnet forum, open forum, public listing. *Ctr:* counter intelligence, digital shadows, fed, flashpoint, honeypot, intel 471, invite only, law enforcement, le monitoring, operational security, opsec, private forum, recorded future
Blockchain analytics at scale (Crypto)	*Vuln:* bitcoin launder, bitcoin tumbl, btc cash out, btc mixer, clean bitcoin, direct transfer, exchange cashout, no kyc exchange, unregulated exchange. *Ctr:* atomic swap, chain analysis, chain hop, chainalysis, coinjoin, cross chain, crystal blockchain, dex, elliptic, mixer, monero, privacy coin, tornado cash, traced, tumbler, untraceable, wasabi wallet, xmr, zcash
Travel rule implementation (Crypto)	*Vuln:* anonymous transfer, no kyc, p2p crypto, peer to peer, unregulated exchange. *Ctr:* bridge, compliance avoid, decentrali, defi, fatf bypass, non custodial, self custody, travel rule
Exchange KYC standardisation (Crypto)	*Vuln:* anonymous exchange, exchange without id, no kyc exchange, unverified account. *Ctr:* binance verified, bisq, buy kyc account, coinbase verified, dex, exchange account, kyc bypass, localbitcoins, localmonero, paxful, sell verified, verified account

Notes: Keyword dictionaries used to score CrimeBB posts against the 27 technologies in Table OA4. *Vuln* = vulnerable-method terms; *Ctr* = countermeasure terms.

OA5 Monopoly

We analyze the benchmark case of a single PSP, $N = 1$, as the limit of Corollary 1 in which the adaptation externality of Proposition 1(iii) vanishes. At $N = 1$, equation (10) specializes to

$$I^*(\gamma)\big|_{N=1} \;=\; \frac{2\gamma\alpha V}{1 + 2\gamma(\alpha V)^2(1-\theta)^2}, \tag{49}$$

and the optimal liability share of equation (10) specializes to $\gamma^*|_{N=1} = 1$, since the overinvestment wedge in equation (9) carries the factor $(N-1)/N$, which is identically zero at $N = 1$.

Corollary OA.1 (Full liability is optimal under monopoly)**.** *At $N = 1$, the privately optimal detection level coincides with the socially optimal level at $\gamma = 1$:*

$$I^*(1)\big|_{N=1} \;=\; I^W\big|_{N=1} \;=\; \frac{2\alpha V}{1 + 2(\alpha V)^2(1-\theta)^2}. \tag{50}$$

Proof. Setting $N = 1$ in the overinvestment wedge (9) gives a factor of $(N-1)/N = 0$, so the adaptation wedge between the private and social first-order conditions (7) and (8) vanishes; since the private marginal benefit still carries the factor γ, the two conditions coincide exactly at $\gamma = 1$. Setting $\gamma = 1$ and $N = 1$ in the closed-form equilibrium (49) gives $I^*(1)|_{N=1} = 2\alpha V/[1 + 2(\alpha V)^2(1-\theta)^2]$, which by the coincidence of the two first-order conditions equals $I^W|_{N=1}$ evaluated from the social FOC. ■ □

A monopolist fully internalizes the effect of its own detection on model decay, since there are no rivals onto whom the externality can fall: unlike the competitive case of Section IV, where each PSP ignores the market-wide impact of its detection intensity on fraudster adaptation, full liability is exactly optimal here, with no partial-liability correction required.

Most payment sectors are competitive, which motivates our focus on the competitive case, $N \geq 2$, in the main text. The monopoly benchmark is nonetheless of independent policy interest: while a monopolistic PSP internalizes the fraud externality, so too, in principle, could a public central bank digital currency (CBDC) operated by a central bank, achieving the same outcome as (50) without the deadweight loss ordinarily associated with monopoly pricing.

OA6 Targeted Attack and Overinvestment

The baseline model assumes the fraudster observes only aggregate detection $\bar{I}_0$ and adapts its fraud *technology* in response, shifting from Type A to Type B. We now consider an alternative

form of strategic adaptation: the fraudster observes individual detection intensities $I_{1,0}$ and $I_{2,0}$ and adapts its *targeting* rather than its technology. We replace the Type A/Type B portfolio choice with a market-allocation choice: after observing $(I_{1,0}, I_{2,0})$, the fraudster decides what share of its attacks to direct at each PSP's customers.

This extension serves two purposes. First, it isolates the targeting mechanism from the technology-switching mechanism, showing that overinvestment arises from strategic adversarial adaptation regardless of whether that adaptation operates on fraud type or on victim selection. Second, it connects our baseline result to the strategic attack model of Acemoglu et al. (2016), in which an attacker who observes security investments directs attacks toward less protected nodes, generating overinvestment as agents race to avoid being targeted. Our model produces the same qualitative result through a different channel: each PSP ignores how its own detection diverts attacks onto its rival.

Changes to the baseline model. We maintain the two-period structure, $N = 2$ PSPs, risk neutrality, and the quadratic cost $c(I) = \frac{1}{2}I^2$. We make three changes.

(i) *Single fraud type.* There is only one fraud type, which succeeds at PSP i with probability $1 - I_{i,0}$. The Type A/Type B distinction and the adaptation cost $k(\cdot)$ are dropped, isolating the targeting mechanism.

(ii) *Fraudster targeting.* The fraudster observes $(I_{1,0}, I_{2,0})$ at the end of period 0 and chooses $q \in [0, 1]$, the share of the period-1 attack mass α directed at PSP 1's customers; the remainder $1 - q$ is directed at PSP 2's customers. Re-directing attacks between PSPs is costless: targeting is a purely informational response to observed detection, unlike technology switching, which requires learning new methods.

(iii) *Fraudster objective.* The fraudster maximises total expected payoff across both PSPs and both periods:

$$\Pi_{\text{fraudster}} \;=\; \frac{\alpha V}{2}\big[(1 - I_{1,0}) + (1 - I_{2,0})\big] \;+\; \alpha V\big[q(1 - I_{1,0}) + (1 - q)(1 - I_{2,0})\big], \quad (51)$$

where the first term is the period-0 payoff (attack mass split equally before targeting is possible) and the second is the period-1 payoff under the chosen allocation q.

All other elements of the baseline model are unchanged: PSPs simultaneously choose $I_{i,0}$ at $t = 0$, the regulator sets $\gamma \in [0, 1]$, and each PSP minimises its two-period fraud losses plus detection costs, where PSP i's period-t fraud loss is γV times the fraud mass it actually absorbs in period t.

Assumption OA.1 (Interiority)**.** $\alpha V < 1$.

Fraudster best response. Given $(I_{1,0}, I_{2,0})$, the fraudster chooses q to maximise the period-1 term of (51), which is linear in q, so the solution is a corner:

$$q^* \;=\; \begin{cases} 1 & \text{if } I_{1,0} < I_{2,0}, \\ 0 & \text{if } I_{1,0} > I_{2,0}, \\ \frac{1}{2} & \text{if } I_{1,0} = I_{2,0}. \end{cases} \tag{52}$$

The fraudster directs all period-1 attacks at the lower-detection PSP, exactly as in the strategic attack model of Acemoglu et al. (2016). As is standard in that literature, we work with a smooth approximation of q^* that coincides with (52) at the symmetric point and converges to it as sensitivity grows:

$$q^*(I_{1,0}, I_{2,0}) \;=\; \frac{1}{2} + \frac{\beta}{2}(I_{2,0} - I_{1,0}), \qquad \beta \in (0, 1], \tag{53}$$

the range restriction keeping $q^* \in [0, 1]$ for every $(I_{1,0}, I_{2,0}) \in [0, 1]^2$.

PSP problem. Under (53), PSP i's period-1 fraud loss is $\gamma V \cdot \alpha q_i \cdot (1 - I_{i,0})$, where $q_1 = q^*$ and $q_2 = 1 - q^*$, and its period-0 loss is $\gamma V(\alpha/2)(1 - I_{i,0})$, matching the equal split in (51). Writing $A \equiv \gamma\alpha V$, PSP 1's two-period loss is

$$L_1 \;=\; \frac{A}{2}(1-I_{1,0}) + A\Big[\frac{1}{2}+\frac{\beta}{2}(I_{2,0}-I_{1,0})\Big](1-I_{1,0}) + c(I_{1,0}) \;=\; A(1-I_{1,0})\Big[1+\frac{\beta}{2}(I_{2,0}-I_{1,0})\Big] + c(I_{1,0}), \tag{54}$$

with L_2 defined symmetrically.

Proposition OA.1 (Targeted attack and overinvestment)**.** *Under the targeted attack model with $N = 2$, the smooth approximation* (53) *with $\beta > 0$, and Assumption OA.1:*

(i) Equilibrium. *A unique symmetric Nash equilibrium $I^{tgt,*}(\gamma) \in (0, 1)$ exists, is strictly increasing in γ, and equals*

$$I^{tgt,*}(\gamma) \;=\; \frac{\gamma\alpha V\big(1 + \frac{\beta}{2}\big)}{1 + \gamma\alpha V\frac{\beta}{2}}. \tag{55}$$

Best responses are strictly increasing in the rival's investment: the game is one of strategic complements.

(ii) The planner's benchmark. *The socially optimal detection level, $I^{W,tgt} \equiv \alpha V$, is independent of γ.*

(iii) Overinvestment at $\gamma = 1$. $I^{tgt,*}(1) > I^{W,tgt}$*: full liability generates strict overinvestment,*

$$I^{tgt,*}(1) - I^{W,tgt} \;=\; \frac{\alpha V \frac{\beta}{2}(1-\alpha V)}{1+\alpha V \frac{\beta}{2}} \;>\; 0. \tag{56}$$

(iv) Optimal liability. *There exists a unique* $\gamma^{tgt,*} \in (0,1)$ *with* $I^{tgt,*}(\gamma^{tgt,*}) = I^{W,tgt}$*:*

$$\gamma^{tgt,*} \;=\; \frac{1}{1+\frac{\beta}{2}(1-\alpha V)}. \tag{57}$$

$\gamma^{tgt,*}$ *is strictly decreasing in* β*: stronger targeting sensitivity amplifies the diversion externality and requires a lower liability share to prevent overinvestment.*

Proof. Part (i). Differentiating L_1 in (54) with respect to $I_{1,0}$,

$$\frac{\partial L_1}{\partial I_{1,0}} \;=\; -A + \frac{A\beta}{2}\big(-I_{2,0} - 1 + 2I_{1,0}\big) + I_{1,0}.$$

The second derivative is $\partial^2 L_1/\partial I_{1,0}^2 = A\beta + 1 > 0$, so L_1 is strictly convex in $I_{1,0}$ for every $I_{2,0}$: the best response is single-valued. Solving $\partial L_1/\partial I_{1,0} = 0$ for $I_{1,0}$ as a function of $I_{2,0}$,

$$I_{1,0}\big(1+A\beta\big) \;=\; A\Big(1+\frac{\beta}{2}\Big) + \frac{A\beta}{2} I_{2,0} \quad\Longrightarrow\quad \frac{\partial I_{1,0}}{\partial I_{2,0}} \;=\; \frac{A\beta/2}{1+A\beta} \;\in\; (0,1),$$

a strictly positive slope strictly less than one: the game is one of strategic complements, and the best-response map is a contraction, giving existence and uniqueness of a symmetric fixed point $I^{\mathrm{tgt},*}(\gamma)$. Imposing symmetry $I_{1,0} = I_{2,0} = I$ in the first-order condition,

$$-A + \frac{A\beta}{2}(I-1) + I \;=\; 0 \quad\Longrightarrow\quad I\Big(1+\frac{A\beta}{2}\Big) \;=\; A\Big(1+\frac{\beta}{2}\Big),$$

which gives (55) upon substituting $A = \gamma\alpha V$. Writing $I^{\mathrm{tgt},*}(\gamma) = c_1\gamma/(1+c_2\gamma)$ with $c_1 = \alpha V(1+\beta/2) > 0$ and $c_2 = \alpha V\beta/2 > 0$, the derivative $dI^{\mathrm{tgt},*}/d\gamma = c_1/(1+c_2\gamma)^2 > 0$ confirms strict monotonicity in γ.

Part (ii). Aggregate welfare loss sums both PSPs' true resource costs of fraud, which do not depend on γ: total fraud loss in any period is αV regardless of how it is allocated between PSP and consumer. Hence

$$W^{\mathrm{tgt}}(I_1, I_2) \;=\; \frac{\alpha V}{2}\big[(1-I_1)+(1-I_2)\big] \;+\; \alpha V\big[q^*(1-I_1)+(1-q^*)(1-I_2)\big] \;+\; c(I_1)+c(I_2),$$

with q^* as in (53). Write $f(I_1, I_2) \equiv q^*(1 - I_1) + (1 - q^*)(1 - I_2) = q^*(I_2 - I_1) + (1 - I_2)$, so

$$\frac{\partial f}{\partial I_1} \;=\; \frac{\partial q^*}{\partial I_1}(I_2 - I_1) - q^* \;=\; -\frac{\beta}{2}(I_2 - I_1) - q^*.$$

At the symmetric point $I_1 = I_2 = I$, $I_2 - I_1 = 0$ and $q^* = 1/2$, so $\partial f/\partial I_1|_{\text{sym}} = -1/2$, and the β-dependent term vanishes exactly at symmetry. Hence

$$\left.\frac{\partial W^{\text{tgt}}}{\partial I_1}\right|_{\text{sym}} \;=\; -\frac{\alpha V}{2} + \alpha V\Big(-\frac{1}{2}\Big) + I \;=\; -\alpha V + I,$$

so the planner's first-order condition delivers $I^{W,\text{tgt}} = \alpha V$, independent of γ.

Part (iii). Setting $\gamma = 1$ (so $A = \alpha V$) in (55) and subtracting $I^{W,\text{tgt}} = \alpha V$,

$$I^{\text{tgt},*}(1) - I^{W,\text{tgt}} \;=\; \frac{\alpha V(1 + \beta/2)}{1 + \alpha V\beta/2} - \alpha V \;=\; \frac{\alpha V(\beta/2)(1 - \alpha V)}{1 + \alpha V\beta/2},$$

which is (56), strictly positive under Assumption OA.1 and $\beta > 0$.

Part (iv). By part (i), $I^{\text{tgt},*}(\gamma)$ is continuous and strictly increasing on $[0, 1]$, with $I^{\text{tgt},*}(0) = 0 < I^{W,\text{tgt}}$ and, by part (iii), $I^{\text{tgt},*}(1) > I^{W,\text{tgt}}$. The intermediate value theorem delivers a unique $\gamma^{\text{tgt},*} \in (0, 1)$ solving $I^{\text{tgt},*}(\gamma^{\text{tgt},*}) = I^{W,\text{tgt}}$. For the closed form, set $\gamma\alpha V(1 + \beta/2)/(1 + \gamma\alpha V\beta/2) = \alpha V$ and divide both sides by $\alpha V > 0$:

$$\gamma\Big(1 + \frac{\beta}{2}\Big) \;=\; 1 + \gamma\alpha V\frac{\beta}{2} \quad\Longrightarrow\quad \gamma\Big[1 + \frac{\beta}{2}(1 - \alpha V)\Big] \;=\; 1,$$

which gives (57). Since $\beta > 0$ and $\alpha V < 1$, the bracketed term strictly exceeds one, so $\gamma^{\text{tgt},*} \in (0, 1)$. Differentiating,

$$\frac{d\gamma^{\text{tgt},*}}{d\beta} \;=\; \frac{-\frac{1}{2}(1 - \alpha V)}{\big[1 + \frac{\beta}{2}(1 - \alpha V)\big]^2} \;<\; 0,$$

since $\alpha V < 1$. ■ □

Relation to the baseline model. The targeted attack model and the baseline model generate overinvestment through distinct but complementary mechanisms. In the baseline, each PSP ignores how its detection accelerates the fraudster's market-wide adaptation toward harder-to-detect fraud types, degrading every rival's model performance; the game is one of strategic *substitutes*. In the targeted attack model, each PSP ignores how its detection diverts the fraudster's attacks toward rivals, directly raising rivals' period-1 fraud rates; the game is one of strategic *complements*, as in Acemoglu et al. (2016). Despite this difference in

strategic interaction, both models deliver overinvestment at full liability and a strictly partial welfare-maximising liability share. Two caveats bound the claim: the smooth targeting rule (53) is a reduced form, feasible on $[0,1]^2$ only for $\beta \in (0,1]$ and not a limit of the fraudster's discontinuous best response as β grows; the result therefore establishes robustness to reduced-form targeted diversion, not to the exact discontinuous targeting game.

OA7 Detection-Sensitive Market Shares

We relax the assumption that market shares are fixed at $1/N$. Consumers reward PSPs that detect more fraud at $t = 0$ by shifting volume toward them: PSP i's market share is

$$s_i \;=\; \frac{1}{N} + \kappa\big(I_{i,0} - \bar{I}_0\big), \qquad \kappa \geq 0, \tag{58}$$

with $\kappa = 0$ recovering the baseline. Since both u_{psp} revenue and fraud losses scale with volume, PSP i's profit becomes

$$\Pi_i \;=\; u_{\text{psp}}\, s_i \;-\; \gamma\alpha V\, s_i\big(\phi_{i,0} + \phi_{i,1}\big) \;-\; c(I_{i,0}), \tag{59}$$

in place of the baseline L_i in equation (3), where u_{psp}/N no longer drops out as a constant once $\kappa > 0$.

PSP i now internalises two channels through which $I_{i,0}$ raises its payoff: the standard $1/N$ effect on aggregate detection $\bar{I}_0$, operating through fraudster adaptation as in Section IV, and the new direct market-share gain κ. Imposing symmetry $I_{i,0} = \bar{I}_0 = I$ in the first-order condition for (59) yields a quadratic equation in I:

$$a(\kappa)\, I^2 + b(\kappa)\, I + c(\kappa) \;=\; 0, \tag{60}$$

with

$$a(\kappa) = \frac{(1-\theta)^2(\alpha V)^2\gamma\kappa(N-1)}{N}, \tag{61}$$

$$b(\kappa) = 1 - \frac{2\gamma\alpha V\kappa(N-1)}{N} + \frac{(N+1)(1-\theta)^2(\alpha V)^2\gamma}{N^2}, \tag{62}$$

$$c(\kappa) = -\frac{2\gamma\alpha V}{N} - \kappa\,\frac{(N-1)\big[u_{\text{psp}} - 2\gamma\alpha V\big]}{N}. \tag{63}$$

Lemma OA.1 (Equilibrium under market-share sensitivity)**.** *For $\kappa > 0$, $N \geq 2$, and $\theta < 1$, $a(\kappa) > 0$, and since $c(\kappa) < 0$ whenever $u_{psp} > 2\gamma\alpha V$ the quadratic has exactly one positive*

root; the equilibrium is the root of (60)

$$I(\kappa) \;=\; \frac{-b(\kappa) + \sqrt{b(\kappa)^2 - 4a(\kappa)c(\kappa)}}{2a(\kappa)} \tag{64}$$

that is continuous in κ *at* $\kappa = 0$*, where it coincides with the baseline equilibrium* $I^*(\gamma)$ *of equation* (10)*.*

Proof. Setting $\kappa = 0$ collapses $a(\kappa)$ to zero and equation (60) reduces to the baseline linear equation underlying (10), whose unique root on $(0, 1)$ is $I^*(\gamma)$ by Proposition 1(i). Because a, b, and c are continuous in κ and $a(0) = 0$, the two roots of the quadratic (60) vary continuously with κ away from $\kappa = 0$, and by definition of $I(\kappa)$ as the root selected by the sign convention in (64), $I(\kappa) \to I^*(\gamma)$ as $\kappa \to 0$. We verify numerically, over the parameter range underlying Table B2, that $I(\kappa)$ remains the economically relevant root (real-valued and in $(0, 1)$) throughout; we do not establish this analytically for arbitrary parameter values. ■ □

The effect of $\kappa > 0$. Market-share sensitivity introduces a second, purely private motive for detection investment, distinct from fraud reduction: a PSP that detects more than its rivals captures volume from them. This business-stealing motive operates independently of γ, since κ enters $b(\kappa)$ and $c(\kappa)$ even when γ is small, and is unambiguously privately valuable whenever $u_{\text{psp}} > 2\gamma\alpha V$, the franchise value of a stolen unit of volume net of the two periods of liability it carries. Because this channel does not appear in the planner's objective, which cares about aggregate detection and aggregate fraud losses but is indifferent to how volume is reallocated across identical PSPs, $\kappa > 0$ adds a pure business-stealing externality on top of the baseline model-decay externality. The two externalities push in the same direction on the optimal liability: model decay calls for $\gamma^* < 1$ to curb overinvestment, while business-stealing independently inflates equilibrium detection for any γ, including $\gamma = 0$, where $I^*(0) = 0$ in the baseline but $I(\kappa)|_{\gamma=0} > 0$ once $\kappa > 0$, since $c(\kappa)|_{\gamma=0} = -\kappa u_{\text{psp}}(N-1)/N \neq 0$. The regulator must therefore set a strictly lower γ^* as κ rises to offset this additional, liability-independent source of overinvestment.

OA8 Homogeneous Consumers with Endogenous Vigilance

A consumer who exerts vigilance ($\nu = 1$) faces fraud exposure α_L per transaction; a non-vigilant consumer ($\nu = 0$) faces $\alpha_H > \alpha_L$, with $\Delta\alpha \equiv \alpha_H - \alpha_L > 0$. Vigilance costs $c_\nu > 0$.

Consumer utility is

$$U_{\text{con}}(\nu) \;=\; u_{\text{con}} \;-\; (1-\gamma)\,\alpha(\nu)\,V\big[\phi_0+\phi_1\big] \;-\; c_\nu\,\mathbf{1}\{\nu=1\}. \tag{65}$$

The PSP reimburses fraction γ of each fraud loss and, in addition, pays a regulatory fee $\tau \geq 0$ per unit of fraud loss that is not passed to the consumer; the PSP's effective liability is $\Gamma \equiv \gamma + \tau$, entering its problem exactly as γ did in Section IV. Consumer utility depends on γ alone: τ affects the PSP's detection incentive but not the consumer's payoff. This separation is what makes γ and τ non-redundant instruments.

Assumption OA.2 (Social viability)**.** *For all $I \in [0,1]$: (i) $u_{con}+u_{psp}-c(I) > \alpha_L V[\phi_0(I)+\phi_1(I)]$; (ii) $u_{con}+u_{psp}-c(I) < \alpha_H V[\phi_0(I)+\phi_1(I)]$.*

Vigilant consumers generate positive social surplus at every detection level; a market of non-vigilant consumers does not.

The incentive-compatibility condition. The consumer chooses $\nu = 1$ if and only if

$$(1-\gamma)\,\Delta\alpha\,V\big[\phi_0(I^*)+\phi_1(I^*)\big] \;\geq\; c_\nu, \tag{66}$$

where I^* is the equilibrium detection intensity and

$$\phi_0(I^*)+\phi_1(I^*) \;=\; 2(1-I^*)+(1-\theta)^2\alpha_L V (I^*)^2. \tag{67}$$

The left-hand side of (66) is the private benefit of avoiding $\Delta\alpha$ of exposure, net of what the PSP already reimburses; the ICC requires this to cover the vigilance cost c_ν. Differentiating (67),

$$\frac{d[\phi_0+\phi_1]}{dI^*} \;=\; -2+2(1-\theta)^2\alpha_L V I^* \;<\; 0 \tag{68}$$

at every empirically relevant I^*: higher PSP detection lowers total fraud incidence and, with it, the private benefit a consumer captures from its own vigilance. Consumer effort and PSP detection are therefore substitutes, and raising γ alone, which raises both PSP detection and the consumer's reimbursement, crowds out vigilance from two directions at once.

Inducing effort. The equilibrium and socially optimal detection intensities under effective liability Γ are

$$I^*(\Gamma) \;=\; \frac{2\Gamma\alpha_L V N}{N^2+(N+1)\Gamma(1-\theta)^2(\alpha_L V)^2}, \qquad I^W \;=\; \frac{2\alpha_L V}{N+2(\alpha_L V)^2(1-\theta)^2}, \tag{69}$$

exactly Corollary 1's objects at exposure α_L and liability Γ, and $I^*(\Gamma^*) = I^W$ pins the effective liability that implements the social optimum,

$$\Gamma^* \;=\; \frac{N^2}{N^2 + (\alpha_L V)^2(1-\theta)^2(N-1)}, \tag{70}$$

the γ^* of Corollary 1 evaluated at $\alpha = \alpha_L$. Because the ICC left-hand side of (66) is strictly decreasing in γ, there is a largest reimbursement rate compatible with vigilance at any given detection level. Evaluating the ICC at $I^* = I^W$ and setting it to equality gives that ceiling,

$$\tilde{\gamma} \;=\; 1 - \frac{c_\nu}{\Delta\alpha V\big[2(1-I^W) + (1-\theta)^2\alpha_L V (I^W)^2\big]}\,: \tag{71}$$

for $\gamma \le \tilde{\gamma}$ the consumer remains vigilant when detection is set at its socially optimal level; for $\gamma > \tilde{\gamma}$ it does not. The ceiling is defined along the efficient policy path: because the private return to vigilance falls with detection, a policy that tolerated inefficiently low detection could sustain vigilance at reimbursement rates above $\tilde{\gamma}$, but such policies are dominated in what follows.

The regulatory fee. $\tilde{\gamma}$ generally differs from Γ^*: the former is pinned by the consumer's vigilance cost and the exposure gap $\Delta\alpha$, the latter by the curvature of PSP detection costs and the adaptation parameter θ, and nothing ties the two together. When $\Gamma^* > \tilde{\gamma}$, the interesting case, the fee closes the resulting gap,

$$\tau^* \;=\; \Gamma^* - \tilde{\gamma}, \tag{72}$$

so that $\gamma = \tilde{\gamma}$ and $\tau = \tau^*$ jointly deliver effective liability Γ^* while leaving the consumer's incentive problem exactly at its ceiling. When instead $\tilde{\gamma} \ge \Gamma^*$, no fee is needed: $\gamma = \Gamma^*$ and $\tau = 0$ implement I^W with vigilance intact, since $\gamma < \tilde{\gamma}$. And if c_ν is so large that $\tilde{\gamma} < 0$, vigilance cannot be induced at I^W by any nonnegative reimbursement rate; we restrict attention to parameters with $\tilde{\gamma} \in [0, 1]$ throughout.

Proposition OA.2 (Optimal two-instrument rule)**.** *Under Assumption OA.2 and the quadratic specification:*

(i) $\gamma = \tilde{\gamma}$ *is the largest reimbursement rate consistent with vigilance at detection level* I^W*.*

(ii) $\tau = \tau^*$ *implements* $I^* = I^W$ *without altering the consumer's incentive to remain vigilant.*

(iii) A single instrument cannot implement I^W and hold γ at its vigilance ceiling simultaneously unless $\Gamma^* = \tilde{\gamma}$, which fails generically.

(iv) Welfare, evaluated at $\Gamma = \Gamma^*$, is constant in γ on $[0, \tilde{\gamma}]$ and strictly lower for every $\gamma > \tilde{\gamma}$; among welfare-maximizing policies, $(\tilde{\gamma}, \tau^*)$ is the one requiring the smallest fee.

(v) $\tau^* > 0$ if and only if $\Gamma^* > \tilde{\gamma}$. Γ^* is strictly increasing in θ and $\tilde{\gamma}$ is strictly decreasing in θ, so τ^* is strictly increasing in θ on $[0, 1)$ and largest as $\theta \to 1$.

Proof. Fix $N \geq 2$, $\theta < 1$, and $\alpha_L V < 1$.

(i). Substituting (67) at $I^* = I^W$ into (66) and setting the inequality to equality gives (71). Since the ICC left-hand side is strictly decreasing in γ, the inequality holds if and only if $\gamma \leq \tilde{\gamma}$, establishing $\tilde{\gamma}$ as the ceiling.

(ii). With $\gamma = \tilde{\gamma}$ and $\tau = \Gamma^* - \tilde{\gamma}$, $\Gamma = \gamma + \tau = \Gamma^*$, and $I^*(\Gamma^*) = I^W$ by (70). Since U_{con} depends on (γ, τ) only through γ, and $\gamma = \tilde{\gamma}$ is unchanged, part (i) gives that the consumer remains exactly at the margin of vigilance.

(iii). With $\tau \equiv 0$, γ is the sole instrument and must solve $\gamma = \Gamma^*$ (to implement I^W, by (70)) and $\gamma = \tilde{\gamma}$ (to sit at the vigilance ceiling, part i) simultaneously. A single scalar satisfies both equations only if $\Gamma^* = \tilde{\gamma}$. Since Γ^* depends only on (N, α_L, V, θ) while $\tilde{\gamma}$ additionally depends on $(c_\nu, \Delta\alpha)$, the equality $\Gamma^* = \tilde{\gamma}$ imposes one restriction on $(c_\nu, \Delta\alpha)$ for any fixed (N, α_L, V, θ), satisfied only on a measure-zero subset of the parameter space.

(iv). Fix $\Gamma = \Gamma^*$, so $I^* = I^W$ regardless of the split between γ and $\tau = \Gamma^* - \gamma$. Reimbursement is a transfer between PSP and consumer and does not itself consume resources, so welfare varies with γ only through whether the ICC holds. By part (i), the ICC holds for every $\gamma \in [0, \tilde{\gamma}]$, over which welfare is therefore constant at the vigilant-market level given by Assumption OA.2(i). For $\gamma > \tilde{\gamma}$ the consumer sets $\nu = 0$, and Assumption OA.2(ii) gives strictly lower welfare. Every $\gamma \in [0, \tilde{\gamma}]$ paired with $\tau = \Gamma^* - \gamma$ is therefore welfare-maximizing, and $\tau = \Gamma^* - \gamma$ is minimized at $\gamma = \tilde{\gamma}$, the upper end of this interval.

(v). That $\tau^* > 0$ iff $\Gamma^* > \tilde{\gamma}$ is immediate from (72). From (70), $\Gamma^*(\theta) = N/[N + (\alpha_L V)^2(1-\theta)^2]$ is strictly decreasing in $(1-\theta)^2$ and hence strictly increasing in θ. For $\tilde{\gamma}$, write $S(\theta) \equiv \phi_0(I^W) + \phi_1(I^W) = 2(1 - I^W) + (1-\theta)^2 \alpha_L V (I^W)^2$, so that $d\tilde{\gamma}/d\theta$ has the sign of $S'(\theta)$. Differentiating $I^W(\theta)$ in (69) gives

$$\frac{dI^W}{d\theta} \;=\; \frac{8(\alpha_L V)^3(1-\theta)}{\left[N + 2(\alpha_L V)^2(1-\theta)^2\right]^2} \;>\; 0,$$

and

$$S'(\theta) \;=\; \frac{dI^W}{d\theta}\Big[-2 + 2(1-\theta)^2 \alpha_L V I^W\Big] \;-\; 2(1-\theta)\alpha_L V \big(I^W\big)^2.$$

Under $\alpha_L V < 1$, $(1-\theta)^2 \alpha_L V I^W < 1$, so the bracketed term is negative and, multiplied by $dI^W/d\theta > 0$, the first term is negative; the second term is negative outright. Hence $S'(\theta) < 0$ and $\tilde{\gamma}$ is strictly decreasing in θ on $[0,1)$. Since Γ^* is strictly increasing and $\tilde{\gamma}$ strictly decreasing, $\tau^* = \Gamma^* - \tilde{\gamma}$ is strictly increasing on $[0,1)$, with supremum as $\theta \to 1$. □

OA9 Detection When Only Realized Fraud Reveals a Mule

Setup

Let identification depend on the payment through a mule actually clearing rather than merely being attempted. Nest this with the baseline technology of Section VI via $\eta \in [0,1]$: a mule drawn upon is identified with probability

$$\pi(I;\eta) \;\equiv\; \alpha_m\big[\eta + (1-\eta)(1-I)\big] \;=\; \alpha_m(1-\varepsilon I), \qquad \varepsilon \equiv 1-\eta. \tag{73}$$

This is a statement about *whether* a drawn-upon mule is identified, not about how much fraud volume it carries: α_m enters here only as the per-period probability that a mule account is drawn on, and no monetary value appears in any of the definitions below, so there is nothing to convert to per-period terms. $\eta = 1$ recovers Section VI exactly. $\eta = 0$ is the pure realized-fraud technology: $\pi(I;0) = \alpha_m(1-I)$, strictly decreasing in I.

The retained pool as a function of the common symmetric intensity I is, for the no-sharing regime,

$$Z_{\mathrm{NS}}(I;\varepsilon) \;=\; \lambda\big[1-\pi(I;\eta)\big] \;=\; \lambda_1^{\mathrm{NS},0} + \kappa_{\mathrm{NS}}\varepsilon I, \tag{74}$$

and for the sharing regime,

$$Z_{\mathrm{S}}(I;\varepsilon) \;=\; \lambda\big[1-\pi(I;\eta)\big]^{\rho}, \tag{75}$$

where $\lambda_1^{R,0}$ denotes the baseline ($\varepsilon = 0$) value (26) and

$$\kappa_{\mathrm{NS}} \;\equiv\; \lambda\alpha_m, \qquad \kappa_{\mathrm{S}} \;\equiv\; \lambda\rho\alpha_m(1-\alpha_m)^{\rho-1} \;=\; \frac{1}{\varepsilon}Z'_{\mathrm{S}}(0;\varepsilon). \tag{76}$$

PSP i's loss under regime R is (29) with $\lambda_1^R \to Z_R(\bar{I}_0;\varepsilon)$ substituted wherever λ_1^R previously appeared, including inside ϕ_1's calibrating argument, and with the prize taken as $\alpha_m V$ exactly as in the main text.

Result

Proposition OA.3 (Free-Riding on the Identification Margin)**.** *Under quadratic costs, let* $F^R(I,\varepsilon)$ *denote* $\partial L_i^R/\partial I_{i,0}$ *evaluated at the symmetric point* $I_{i,0} = \bar{I}_0 = I$*, maintained increasing in* I *throughout. Then:*

(i) At $I = 0$*, exactly,*

$$F^R(0,\varepsilon) \;=\; -\frac{\gamma\alpha_m V(\lambda+\lambda_1^{R,0})}{N} + \frac{\gamma\alpha_m V\kappa_R\,\varepsilon}{N^2}. \tag{77}$$

For $\gamma > 0$*, it is privately optimal to set* $I^R = 0$ *if and only if* $\varepsilon \geq \varepsilon_R^\dagger \equiv N(\lambda+\lambda_1^{R,0})/\kappa_R$*: the common prefactor* $\gamma\alpha_m V$ *cancels from this condition, so* $\varepsilon_R^\dagger$ *does not depend on whether the prize is* V *or* $\alpha_m V$*.*

(ii) $\varepsilon_{\mathrm{NS}}^\dagger = N(2-\alpha_m)/\alpha_m > N \geq 2$ *for every admissible* $\alpha_m \in (0,1)$*: absent sharing, the corner never binds.*

(iii) $\varepsilon_{\mathrm{S}}^\dagger = N\big[1+(1-\alpha_m)^\rho\big]\big/\big[\rho\alpha_m(1-\alpha_m)^{\rho-1}\big]$ *remains strictly above* 1 *for every* $N \geq 2$ *and every admissible* (α_m,ρ)*.*

(iv) Isolating the mule-specific component of the first-order condition alone gives

$$\tilde{\varepsilon}_{\mathrm{S}}^\dagger \;=\; \frac{N(1-\alpha_m)}{\rho\alpha_m}, \tag{78}$$

strictly decreasing in ρ *and at or below* 1 *whenever* $\rho \geq N(1-\alpha_m)/\alpha_m$*.*

The decomposition into a prize effect and a free-riding effect, used in the main text to explain part (i), is

$$P_R(I) \;\equiv\; -I\big(1-2(1-\theta)^2\alpha_m V I\lambda_1^{R,0}\big), \qquad Q_R(I) \;\equiv\; (1-I)+4(1-\theta)^2\alpha_m V I^2\lambda_1^{R,0}, \tag{79}$$

with $F_\varepsilon^R(I,0) = \gamma\alpha_m V\kappa_R[P_R(I)/N + Q_R(I)/N^2]$ (proof below): these carry $\alpha_m V$, since they describe behaviour away from $I = 0$, where the $\alpha_m V$ prefactor no longer cancels between terms of different order in N; only the exact corner condition at $I = 0$ is $\alpha_m V$-invariant.

Proof of Proposition OA.3

Proof. Write $L_i^R(I_{i,0},\bar{I}_0) = \frac{\gamma\lambda\alpha_m V}{N}(1-I_{i,0}) + \frac{\gamma\alpha_m V}{N}Z(\bar{I}_0)\big[(1-I_{i,0})+(1-\theta)^2 Z(\bar{I}_0)\alpha_m V\bar{I}_0 I_{i,0}\big] + \frac{1}{2}I_{i,0}^2$, $Z = Z_R(\cdot;\varepsilon)$. Differentiating at fixed $\bar{I}_0$, then at fixed $I_{i,0}$, and combining via $F^R =$

$\partial_{I_{i,0}} L_i^R + \frac{1}{N}\partial_{\bar{I}_0} L_i^R$ at $I_{i,0} = \bar{I}_0 = I$,

$$\begin{aligned} F^R(I,\varepsilon) \;=\; & -\frac{\gamma\lambda\alpha_m V}{N} + \frac{\gamma\alpha_m V}{N}\Big[-Z + (1-\theta)^2\alpha_m V I Z^2\Big] + I \\ & + \frac{\gamma\alpha_m V}{N^2}\Big[Z'(1-I) + (1-\theta)^2\alpha_m V I\big(Z^2 + 2IZZ'\big)\Big], \end{aligned} \tag{80}$$

where $Z = Z(I;\varepsilon)$ and $Z' = \partial Z/\partial\bar{I}_0|_{\bar{I}_0=I}$. At $I = 0$: $Z(0;\varepsilon) = \lambda_1^{R,0}$ for every ε, and $Z'(0;\varepsilon) = \kappa_R\varepsilon$ exactly. Substituting into (80) gives (77), in which the common factor $\gamma\alpha_m V$ multiplies both terms and cancels upon setting $F^R(0,\varepsilon) = 0$, leaving the threshold in (i) independent of $\alpha_m V$. Since F^R is increasing in I, $F^R(0,\varepsilon) \geq 0$ is necessary and sufficient for $I^R = 0$ to solve the first-order condition at a boundary optimum. Differentiating (80) in ε at general I gives the decomposition into P_R, Q_R in (79), which retains $\alpha_m V$ since the $\gamma\alpha_m V$ prefactor there multiplies terms of different order in N (P_R/N versus Q_R/N^2) that do not cancel against one another.

For (ii)–(iv), the algebra is exactly as before, since none of it involves V or $\alpha_m V$: $\varepsilon^\dagger_{\mathrm{NS}} = N(2-\alpha_m)/\alpha_m$ exceeds $N \geq 2$ because $\alpha_m \in (0,1)$ gives $(2-\alpha_m)/\alpha_m > 1$; $\varepsilon^\dagger_{\mathrm{S}}$ exceeds 1 for every admissible (N,α_m,ρ): since $\alpha_m(1-\alpha_m)^{\rho-1} < 1$ and $\rho \leq N$, $\varepsilon^\dagger_{\mathrm{S}} > N\big[1+(1-\alpha_m)^\rho\big]/\rho > N/\rho \geq 1$; and dropping the period-0, own-customer term from L_i^R removes $-\gamma\lambda\alpha_m V/N$ from (77), giving $\tilde{\varepsilon}^\dagger_{\mathrm{S}} = N\lambda_1^{\mathrm{S},0}/\kappa_{\mathrm{S}} = N(1-\alpha_m)/(\rho\alpha_m)$, strictly decreasing in ρ. □