Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Signal Quality and the Silence Trap in Collaborative Cybersecurity Governance

Jiguang Wang
Shanxi University
Yi Ren
renyi4@sxu.edu.cn

Shanxi University

Research Article

Keywords: Cybersecurity, Collaborative governance, Whistleblowing, incentives, Verification precision, Evolutionary game theory

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9779128/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Escaping the Silence Trap: Technological Thresholds and Signal Quality in Crowdsourced Cybersecurity Governance

Jiguang Wang, Yi Ren*

[a] *School of Economics and Management, Shanxi University, Taiyuan, 030006*

*Corresponding author.

E-mail addresses: wangjg@sxu.edu.cn (J. Wang); renyi4@sxu.edu.cn (Y. Ren*)

Corresponding author Mailing address：

Yi Ren

No.63, South Central Ring Street, Taiyuan, 030006, China

Tel: +86 17835600131

Signal Quality and the Silence Trap in Collaborative Cybersecurity Governance

Abstract： Collaborative cybersecurity governance relies on “societal sensors” to detect vulnerabilities, yet regulators must filter noisy reports under pronounced information asymmetry. Existing evolutionary game models typically adopt a “perfect monitoring” assumption, obscuring the structural conditions under which incentives remain effective. By incorporating a signal-verification perspective into a tripartite evolutionary framework, this study endogenizes verification precision and false-report penalties as determinants of system stability. The model identifies a critical signal-quality threshold below which the system enters a “Silence Trap”: user participation becomes weakly responsive to incentives, and higher monetary rewards alone fail to restore collaboration. We also identify a “Deterrence Paradox,” whereby harsher penalties for unsubstantiated reporting can suppress valid signals unless paired with sufficiently strong verification infrastructure. These findings shift attention from reward tuning to capacity-first governance, highlighting that the sustainability of crowdsourced regulation is conditional on verification precision as a foundational complement to monetary incentives.

Keywords: Cybersecurity; Collaborative governance; Whistleblowing; incentives; Verification precision; Evolutionary game theory

1. Introduction

In the era of digital governance, the paradigm of regulatory enforcement is undergoing a profound transformation from centralized monitoring to distributed, collaborative oversight. The proliferation of mobile technologies and digital platforms has empowered citizens to act as “societal sensors”, enabling governments to crowdsource the detection of regulatory violations—ranging from environmental pollution and food safety hazards to cybersecurity vulnerabilities (Yang & Wang, 2025; Linders, 2012). By leveraging the “wisdom of crowds”, collaborative governance models promise to overcome the resource constraints of traditional bureaucracy, offering a cost-effective solution to the pervasive problem of information asymmetry between regulators and non-compliant firms (Ansell & Gash, 2008). However, while the potential of crowdsourcing is widely celebrated, the operational stability of such systems remains precarious in practice.

The central challenge facing collaborative governance is not merely incentivizing participation, but ensuring the quality of the information provided. Existing theoretical frameworks, particularly those based on tripartite evolutionary game theory, typically operate under the idealized assumption that citizen reports are accurate and verifiable (Zrahia et al., 2024; Xu et al., 2025). In these “frictionless”

models, increasing monetary rewards linearly translates into higher participation and better compliance. Yet, empirical reality paints a more complex picture. Studies on whistleblower programs and bug bounty platforms reveal that crowdsourcing often generates a deluge of low-quality signals, including false positives, malicious reports, and unverifiable noise, which can overwhelm regulatory capacity and erode the credibility of the enforcement mechanism (Finifter et al., 2013; Givati, 2016). When the cost of verifying these noisy signals (verification friction) becomes prohibitive, the theoretical benefits of collaborative governance may collapse.

Despite the practical urgency of this issue, the theoretical understanding of how signal quality constraints interact with incentive structures remains underdeveloped. Most current models treat verification technology as a black box or assume it to be perfect, thereby failing to capture the "technological threshold" required for incentives to function. This theoretical gap leads to a critical policy dilemma: governments often attempt to boost sluggish participation by simply increasing cash rewards, unaware that without adequate verification precision, such measures may only exacerbate the noise-to-signal ratio and accelerate system failure (Gal-Or et al., 2024). Consequently, a rigorous re-examination of the tripartite dynamics under imperfect information is needed to determine the boundaries of effective governance.

To address these gaps, this study integrates signal detection theory into an evolutionary game framework, explicitly modeling the friction caused by verification errors and the dual risks faced by whistleblowers. Specifically, we seek to answer the following research questions:

RQ1: How do signal frictions, specifically verification precision and false-report penalties, alter the evolutionary stability of collaborative governance systems?

RQ2: Is there a theoretical threshold for signal quality below which monetary incentives become ineffective?

RQ3: How should policymakers design the optimal "incentive-deterrence" portfolio to balance user participation against the risk of system collapse?

Methodologically, we adopt a tripartite evolutionary game theory framework to investigate these questions. Unlike static equilibrium analyses, our model incorporates bounded rationality, allowing agents (government, firms, and users) to dynamically adapt their strategies based on marginal payoffs. We first establish a baseline model under ideal conditions to benchmark performance. Subsequently, we extend the framework by introducing two exogenous parameters: verification precision (q) and false-reporting penalties (f). By solving the replicator dynamics equations and analyzing the Jacobian matrix stability, we mathematically derive the conditions for system evolution. Finally, we employ numerical simulations to visualize the dynamic trajectories and map the policy regimes, thereby verifying the robustness of our theoretical propositions.

By relaxing the "perfect information" assumption, this research makes two distinct contributions to the literature on collaborative governance and mechanism design. First, we identify a "Signal-Quality Threshold" (q^*), a structural boundary previously overlooked in standard evolutionary models. We show analytically that when verification precision falls below this critical level, the system converges to a "governance failure" that cannot be remedied by financial incentives alone, thus challenging the prevailing "monetary solution" hypothesis (Ma et al., 2025). Second, we reveal a "Predator-Prey" oscillation mechanism under high-friction conditions, offering a dynamic explanation for the cyclic instability often observed in real-world regulatory campaigns. Our findings provide a theoretical basis for shifting the focus of governance design from purely economic incentives to a "technology-first" approach, emphasizing that the sustainability of crowdsourced regulation depends fundamentally on the precision of the verification infrastructure.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 constructs the baseline tripartite evolutionary model. Section 4 analyzes the stability of the ideal system. Section 5 extends the model by introducing signal frictions and derives the threshold conditions. Section 6 presents numerical simulations and regime maps to validate the theoretical propositions. Finally, Section 7 summarizes the key findings and discusses both the theoretical contributions and policy implications.

2. Related Literature

This study draws upon and synthesizes three streams of literature: the shift toward multi-stakeholder governance in cybersecurity, the economics of whistleblowing incentives, and the application of evolutionary game theory in regulatory compliance. A central theme across these streams is that incentives only work when verification capacity can separate valid signals from noise, a dimension that is rarely modeled explicitly in dynamic governance settings.

2.1. Collaborative Governance and Multi-Stakeholder Cybersecurity Regulation

Traditional command-and-control regulation has long been criticized for its limited effectiveness in complex and fast-evolving domains characterized by information asymmetry and regulatory lag (Ayres & Braithwaite, 1992; Ostrom, 2010). In response, the literature on collaborative governance emphasizes the involvement of multiple stakeholders, including private actors and civil society, in regulatory processes to enhance information acquisition, monitoring capacity, and adaptive enforcement (Ansell & Gash, 2008; Emerson, Nabatchi, & Balogh, 2012). In digital and cybersecurity contexts, collaborative governance has become particularly salient. Regulators often lack direct access to real-time technical information, while firms and users possess dispersed, experience-based knowledge about vulnerabilities and compliance failures (Kshetri, 2017; Shackelford et al., 2012). As a result, governance arrangements increasingly rely on participatory mechanisms such as crowdsourced

monitoring, third-party audits, and user reporting to supplement formal oversight. Recent European policy guidance similarly frames national cybersecurity strategy implementation as a governance task that requires clear role allocation, cross-sector coordination, and structured public–private cooperation (ENISA, 2024).

However, compared with other policy domains, cybersecurity governance is uniquely verification-intensive. Recent scholarship also emphasizes that cybersecurity governance is inherently multi-stakeholder and institutionally heterogeneous, which makes verification and coordination central rather than peripheral concerns (Savaş & Karataş, 2022). Information provided by users is often noisy, technically complex, and costly to validate, and regulators face binding constraints in distinguishing credible signals from false or low-quality reports (Herley, 2013; Anderson & Moore, 2006). Consequently, the effectiveness of collaborative governance in cybersecurity depends not only on participation levels but also on the institutional capacity to verify and process reported information.

In cybersecurity governance, collaborative arrangements hinge less on encouraging participation per se than on the regulator's ability to verify and filter crowd-generated signals.

2.2. Whistleblowing Incentives and Reporting Frictions

A large body of literature studies whistleblowing and reporting incentives as tools for uncovering hidden misconduct and improving regulatory enforcement (Near & Miceli, 1985; Kaplow & Shavell, 1994). Economic models typically predict that higher rewards and stronger protections increase reporting activity by offsetting reporting costs and retaliation risks (Spagnolo, 2004; Chen & Zhou, 2019). In cybersecurity, bug bounty programs represent a prominent institutionalization of whistleblowing incentives, offering monetary rewards to users who disclose software vulnerabilities (Finifter, Akhawe, & Wagner, 2013; Maillart et al., 2017). Empirical studies show that bounty rewards can substantially increase the volume of reported vulnerabilities, especially from external security researchers (Munaiah & Meneely, 2019; Zhao et al., 2020). At the same time, both empirical and practitioner-oriented research documents a growing noise-to-signal problem. Recent platform-level evidence further shows that external shocks can trigger sharp increases in submissions and duplicates, intensifying triage burdens and making verification capacity a binding constraint for incentive effectiveness (Zrahia et al., 2024). Generous rewards often attract low-quality, duplicate, or unverifiable reports, imposing significant verification costs on platforms and firms (Herley, 2013; Kuehn & Mueller, 2014). Qualitative studies of coordinated vulnerability disclosure programs likewise report persistent operational challenges related to report volume and variable report quality, reinforcing that governance performance depends heavily on verification and triage capacity (Walshe & Simpson, 2022). To mitigate this problem, many programs introduce penalties, strict evidentiary standards, or tiered rewards to discourage frivolous reporting. Related formal analysis in law-and-economics shows that reward schemes can become fragile under

verification errors unless sanctions against fabricated reporting and standards of proof are jointly designed with the system's precision (Buccirossi et al., 2021).

Taken together, this literature suggests that reward-based whistleblowing is fundamentally constrained by verification capacity, turning governance design into a quantity–quality trade-off rather than a simple incentive-maximization problem.

2.3. Evolutionary Games in Compliance and Cybersecurity Governance

Evolutionary game theory (EGT) has been widely applied to analyze governance and compliance problems involving bounded rationality, repeated interaction, and strategic adaptation (Weibull, 1995; Sandholm, 2010). In regulatory settings, EGT models capture how firms, regulators, and social actors adjust strategies over time in response to observed payoffs rather than perfect foresight (Friedman, 1998; Traulsen & Hauert, 2009). Recent studies use EGT to examine collaborative regulation, social supervision, and compliance dynamics in areas such as environmental regulation, platform governance, and information disclosure (Xu et al., 2025; Zhang & Xue, 2023). In closely related digital-security settings, recent work has modeled tripartite governance among platforms, third-party testing organizations, and regulators using replicator dynamics to study stability and policy levers (Tian et al., 2024). These models highlight how incentive schemes and punishment mechanisms can stabilize cooperative outcomes under certain conditions. More recent evolutionary governance models also incorporate dynamically adjusted reward and penalty mechanisms and analyze how these designs alter equilibrium stability and convergence properties (Wang et al., 2025).

However, most EGT-based governance models abstract away verification frictions by assuming deterministic monitoring or perfectly observable signals. Reporting is typically treated as either truthful by assumption or costlessly verifiable, which precludes analysis of how limited verification capacity interacts with incentives and participation over time. By assuming perfect or deterministic monitoring, existing EGT models cannot capture the endogenous degradation of signal quality and participation that arises under verification constraints.

2.4. Research Gap and Positioning

Despite the extensive insights offered by the aforementioned streams of literature, a synthesis reveals a theoretical disconnect that this study aims to bridge.

First, there is a structural mismatch between the complexity of cybersecurity reporting and the "perfect monitoring" assumption that is common in many evolutionary game models of compliance and collaborative governance. Existing frameworks often treat whistleblowing signals as deterministic or costlessly verifiable, which limits their ability to capture the intrinsic noise of crowdsourced reporting. By abstracting away imperfect verification and false or low-evidence reports, such models may overstate

the stability of collaborative governance and cannot explain the trust and participation erosion that may arise under information overload.

Second, while the mechanism-design and whistleblowing literatures acknowledge the noise-to-signal problem, much of this work relies on static equilibrium reasoning. Such analyses provide a useful snapshot but do not characterize the evolutionary feedback loops through which deteriorating signal quality can reduce participation incentives over time. In cybersecurity governance, participation and compliance are shaped by iterative learning and adaptation rather than one-shot transactions.

To bridge these gaps, this study constructs a noise-augmented evolutionary framework. Distinct from prior work that treats verification errors as purely exogenous disturbances, we model verification precision and false-reporting penalties as structural determinants of system stability. This shift moves the focus beyond optimizing reward size and addresses a boundary question that is central to institutional design: what minimum level of verification capability is required for incentive-based participation to remain viable? By doing so, we provide a theoretical instrument for designing resilient collaborative governance mechanisms under imperfect information.

3. Model Construction

3.1. Problem Description

Consider a collaborative regulatory ecosystem focusing on cybersecurity compliance, involving three distinct populations of boundedly rational agents: the Government (G), the Enterprise (F), and the Public User (U).

The Government acts as the regulator, seeking to minimize social risks and maximize public trust. However, due to budget constraints and the technical complexity of cyber threats, it cannot monitor all firms continuously.

The Enterprise (e.g., a digital platform or service provider) faces a trade-off between the high cost of data security compliance and the potential profits from cutting corners (non-compliance).

The Public User (e.g., white-hat hackers or vigilant citizens) acts as a distributed sensor. They possess private information about system vulnerabilities and must decide whether to incur a cost to report this information to the government.

The interactions form a dynamic evolutionary game where strategies are adjusted over time based on marginal payoff comparisons. Figure 1 illustrates the logical relationships and supervision flow among the three parties.

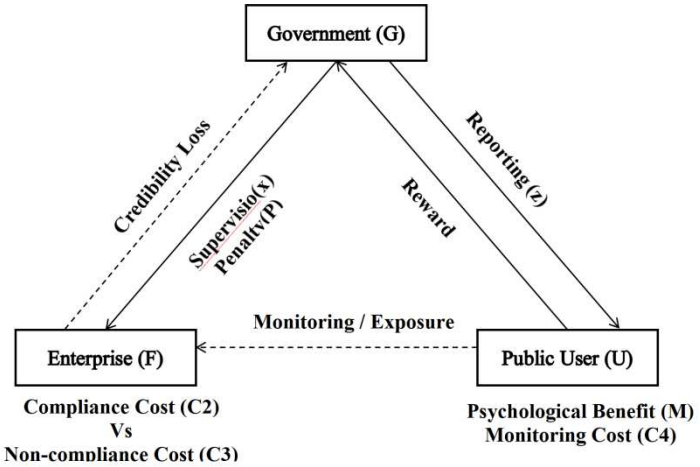

Figure 1: Logical relationship diagram of the three parties

3.2. Basic Assumptions

To construct the mathematical framework, we propose the following five hypotheses based on the standard assumptions of Evolutionary Game Theory (EGT) and the characteristics of collaborative regulation:

Assumption 1.Bounded Rationality. Following the fundamental premise of EGT (Smith, 2012; Friedman, 1998), the Government (*G*), Enterprise (*F*), and Public User (*U*) are boundedly rational agents. Due to information asymmetry and cognitive limitations, they cannot identify the global optimal strategy instantaneously. Instead, they adjust their strategies dynamically through a process of social learning and imitation based on marginal payoff comparisons.

Assumption 2. Strategy Spaces. The strategy spaces for the three parties are defined as binary choices (Zrahia et al., 2024): The Government chooses between *Strict Supervision* (*x*) and *Lax Supervision* (1-*x*).The Enterprise chooses between *Compliance* (*y*) and *Non-compliance* (1-*y*). The Public User chooses between *Participation* (Reporting, *z*) and *Silence* (1-*z*). Here, $x, y, z \in [0,1]$ denote the proportion of the population adopting the active strategy at time *t*.

Assumption 3. Government Payoffs. When choosing Strict Supervision, the government incurs a direct enforcement cost C_1. If a violation is successfully verified (either via strict supervision or user reporting), the government receives a performance reward *E* (e.g., from superior administrations) and collects a fine *P* from the enterprise. Conversely, if the government chooses Lax Supervision and fails to detect the enterprise's violation (1-*y*, 1-*z*), it suffers a reputation loss denoted as *H*. To ensure the regulator has a positive incentive to enforce rules, we assume $E > C_1 + N$,

meaning the performance reward covers both the enforcement cost and the whistleblower bounty.

Assumption 4. Enterprise Payoffs. The enterprise faces a compliance cost C_2 or a non-compliance cost C_3, where the cost saving is $\Delta C = C_2 - C_3 > 0$. If the enterprise is non-compliant and detected, it faces a fine P from the government. Additionally, if the detection is triggered by a user report, the enterprise must pay compensation M to the user. Crucially, to reflect the dilemma of traditional regulation, we assume $P < \Delta C < P + M$. This implies that the government penalty alone (P) is insufficient to deter high-profit violations (ΔC), making the additional cost of user compensation (M) essential for enforcing compliance.

Assumption 5. User Payoffs. Users are victims of non-compliance, suffering a privacy or property loss Q whenever the enterprise violates regulations (1-y). If a user chooses to Participate (z), they incur a monitoring cost C_4. Upon a successful report, the user receives an incentive reward N from the government and compensation M from the enterprise. Thus, the net benefit of a successful report is $(N + M) - C_4 - Q$. If the user remains silent, they solely bear the loss $-Q$. We assume $N + M > C_4$, ensuring that the total incentive package is sufficient to cover the user's participation cost.

To clarify the notation used in the model, we define the specific parameter symbols and their meanings in Table 1.

Table 1. Parameter setting and symbol description

Symbol	Description
x	Probability of the Government choosing *Strict Supervision*
y	Probability of the Enterprise choosing *Compliance*
z	Probability of the Public User choosing *Participation* (Reporting)
C_1	Cost of strict supervision
C_2	Operational cost of compliance
C_3	Operational cost of non-compliance
C_4	Cost of participation (monitoring and reporting)
ΔC	Cost saving from non-compliance ($\Delta C = C_2 - C_3$)
E	Reward received by the government for successfully verifying a violation
P	Fine imposed on the enterprise for verified non-compliance
H	Reputation loss suffered by the government if a violation goes undetected
M	Compensation paid by the enterprise to the user upon verified reporting
N	Incentive reward received from the government for a successful report

Q	Loss (privacy or financial) suffered by the user due to enterprise

In Table 1, $q \in [0,1]$ denotes the regulator/platform's verification capability. Specifically, q is the conditional probability that a valid whistleblowing report is successfully verified (true positive rate). To keep the compliance-status signal structure parsimonious, we assume no erroneous verification of violations when the enterprise is compliant (zero false-positive rate). Hence, under a strict-evidence rule, a report against a compliant enterprise is unsubstantiated and incurs the user penalty *f*. Relaxing this assumption by allowing a small false-positive rate would be a natural extension; however, it does not alter our main mechanism, namely the existence of a verification-quality threshold, as long as false positives remain sufficiently rare.

3.3. Payoff Matrix

Based on the hypotheses, the payoff matrix for the tripartite evolutionary game is presented in Table 2.

Table 2. The payoff matrix of the tripartite evolutionary game

Strategy Profile*(G, F, U)*	Government Payoff U_G	Enterprise Payoff U_F	Public User Payoff U_U
(Strict, Compliance, Report)	$E-N-C_1$	$R-C_2$	$N-C_4$
(Strict, Compliance, Silence)	$E-C_1$	$R-C_2$	0
(Strict, Non-comp, Report)	$E+P-N-C_1$	$R-P-M-C_3$	$N+M-Q-C_4$
(Strict, Non-comp, Silence)	$E+P-C_1$	$R-P-C_3$	$-Q$
(Lax, Compliance, Report)	0	$R-C_2$	$-C_4$
(Lax, Compliance, Silence)	0	$R-C_2$	*0*
(Lax, Non-comp, Report)	*-H*	$R-M-C_3$	$M-Q-C_4$
(Lax, Non-comp, Silence)	*-H*	$R-C_3$	$-Q$

3.4 Replicator Dynamics

According to the Malthusian dynamic equation, we derive the expected payoffs and replicator dynamic equations.

3.4.1 Government

Let U_{11} and U_{12} represent the expected payoffs for Strict Supervision and Lax Supervision from government regulator, respectively, with $\overline{U_1}$ being the average expected payoffs.

$$U_{11} = yz(E - N - C_1) + y(1 - z)(E - C_1) + (1 - y)z(E + P - N - C_1) + (1 - y)(1 - z)(E + P - C_1)$$

$$U_{12} = (1 - y)z(-H) + (1 - y)(1 - z)(-H)$$

$$\overline{U_1} = xU_{11} + (1 - x)U_{12}$$

So, the replicator dynamic equation $F(x)$ is derived as

$$F(x) = \frac{dx}{dt} = x(U_{11} - \overline{U_1}) = x(1 - x)\,[E + P - C_1 + H - y(P + H) - zN] \tag{1}$$

3.4.2 Enterprise

Let U_{21} and U_{22} represent the expected payoffs for the enterprise choosing Compliance operation and Non-Compliance operation, respectively, with $\overline{U_2}$ being the average expected payoffs.

$$U_{21} = xz(R - C_2) + x(1 - z)(R - C_2) + (1 - x)z(R - C_2) + (1 - x)(1 - z)(R - C_2)$$

$$U_{22} = xz(R - P - M - C_3) + x(1 - z)(R - P - C_3) + (1 - x)z(R - M - C_3) + (1 - x)(1 - z)(R - C_3)$$

$$\overline{U_2} = yU_{21} + (1 - y)U_{22}$$

Thus, we obtain the replicator dynamic equation $F(y)$

$$F(y) = \frac{dy}{dt} = y(U_{21} - \overline{U_2}) = y(1 - y)\,[-\Delta C + xP + zM] \tag{2}$$

3.4.3 Public User

Let U_{31} and U_{32} represent the expected payoffs for the public user participating in supervision and reporting and not participating, respectively, with $\overline{U_3}$ being the average expected payoffs.

$$U_{31} = xy(N - C_4) + x(1 - y)(M + N - Q - C_4) + (1 - x)y(-C_4) + (1 - x)(1 - y)(M - Q - C_4)$$

$$U_{32} = x(1 - y)(-Q) + (1 - x)(1 - y)(-Q)$$

$$\overline{U_3} = yU_{31} + (1 - y)U_{32}$$

Thus, we obtain the replicator dynamic equation $F(z)$

$$F(z)=\frac{dz}{dt}=z(U_{31}-\overline{U_3})=z(\ 1\ -z)(M-C_4+xN-yM) \tag{3}$$

4. Stability Analysis of the Baseline System: The Case of Perfect Monitoring

To understand the dynamic behavior of the collaborative governance system under ideal conditions, we analyze the local stability of the equilibrium points derived from the replicator dynamics equations (1)-(3). This establishes a theoretical benchmark for assessing the impact of frictions in the subsequent sections.

4.1. Local Stability Analysis Method

According to Lyapunov's first method, the evolutionary stability of a system can be determined by analyzing the Jacobian matrix (J) of the differential equation system. An equilibrium point is considered an Evolutionarily Stable Strategy (ESS) if all eigenvalues (λ) of the Jacobian matrix evaluated at that point have negative real parts.

The Jacobian matrix J for the tripartite system is given by:

$$J=\begin{pmatrix} \frac{\partial F(x)}{\partial x} & \frac{\partial F(x)}{\partial y} & \frac{\partial F(x)}{\partial z} \\ \frac{\partial F(y)}{\partial x} & \frac{\partial F(y)}{\partial y} & \frac{\partial F(y)}{\partial z} \\ \frac{\partial F(z)}{\partial x} & \frac{\partial F(z)}{\partial y} & \frac{\partial F(z)}{\partial z} \end{pmatrix} \tag{4}$$

Taking the partial derivatives of equations (1), (2), and (3) with respect to *x, y, z*, we derive the specific expressions for the diagonal elements, which determine the local stability signs:

$$\frac{\partial F(x)}{\partial x}=(1-2x)\left[P(1-y)-C_1-(1-y)(z(P-N)-(1-z)(-E))\right] \tag{5}$$

$$\frac{\partial F(y)}{\partial y}=(1-2y)\left[P+L+C_3-C_2-(1-x)(1-z)(P+L)\right] \tag{6}$$

$$\frac{\partial F(z)}{\partial z}=(1-2z)\left[(1-y)(N+M)-C_4\right] \tag{7}$$

(Note: In Eq. 5, the term -E represents the credibility loss avoided when the government regulates strictly or when users report successfully.)

4.2. Stability of the Collaborative Equilibrium

While the tripartite game involves multiple potential equilibrium points, our primary interest lies in the viability of the watchdog mechanism—that is, whether public users are motivated to report violations. We focus on the stability of the user's participation strategy (z) in response to corporate non-compliance ($y \to 0$).

Substituting $y = 0$ (Non-compliance) into Equation (7), the eigenvalue corresponding to the user's participation strategy (λ_z) becomes:

$$\lambda_z = \left.\frac{\partial F(z)}{\partial z}\right|_{y=0} = (1-2z)\left[N + M - C_4\right] \tag{8}$$

This leads to the following critical proposition for the baseline model:

Proposition 1 (Effectiveness of Perfect Whistleblowing). *In a perfect monitoring environment* ($q = 1, f = 0$), *as long as the total incentive exceeds the participation cost* ($N + M > C_4$), *the user population will inevitably evolve towards active participation* ($z = 1$) *whenever the enterprise is non-compliant.*

Proof. Consider the stability of the silence strategy $z = 0$ when the firm is non-compliant ($y = 0$). Substituting $z = 0$ into Eq. (8), we obtain the eigenvalue:

$$\lambda_{z=0} = \left[N + M - C_4\right] \tag{9}$$

Given the basic participation constraint $N + M > C_4$ (from Model Assumption 5), we have $\lambda_{z=0} > 0$. A positive eigenvalue indicates that $z = 0$ is an unstable source point, implying that users will not remain silent. Conversely, checking the participation strategy $z = 1$:

$$\lambda_{z=1} = -\left[N + M - C_4\right] < 0 \tag{10}$$

Since $\lambda_{z=1}$ is negative, the strategy of "Reporting" is locally stable. □

4.3. Regime Interpretation

The analytical results above delineate an "Ideal Governance Regime". In this regime, the effectiveness of collaborative governance is purely a function of economic incentives. Proposition 1 implies that the system possesses a deterministic self-correcting mechanism: if firms violate regulations, users are economically motivated to report, creating a watchdog effect that forces the system away from the lawless state.

Mathematically, this ideal regime is maintained as long as the Ideal Participation Condition holds:

$$N + M > C_4 \tag{11}$$

Equation (11) represents the monetary solution hypothesis often found in traditional literature. It suggests that regulators can always restore system stability simply by increasing the reward N to cover the cost C_4.

However, this conclusion relies heavily on the assumption that every report is accurate and verifiable (i.e., signal noise is zero). This creates a "frictionless" benchmark. As we will demonstrate in the extended model in Section 5, when signal noise is introduced, Equation (11) is no longer sufficient. The deterministic link between rewards and participation will be severed, revealing the "technological threshold" of governance.

5 Extension: Modeling Whistleblowing Frictions and Signal Quality

While the baseline model in Section 4 demonstrates the potential of collaborative governance under ideal conditions, it relies on the simplifying assumption of perfect whistleblowing, where user reports are treated as deterministic signals of non-compliance (Zrahia et al., 2024; Xu et al., 2025). However, in real-world cybersecurity ecosystems, crowdsourced monitoring is inherently noisy. Users may submit inaccurate reports due to technical misunderstandings, or strategically submit malicious reports to exploit reward systems (Gal-Or et al., 2024). Neglecting these frictions can lead to over-incentivization traps, where high rewards generate more noise than value (Givati, 2016).

To address this limitation, this section extends the tripartite framework by incorporating two exogenous friction parameters: verification precision and false-report deterrence.

5.1 Incorporating Imperfect Verification and False-Report Penalties

Following the inspection game framework with imperfect detection technology (Boone, 2025), we relax the assumption that the regulator can perfectly verify all user reports. We introduce two key parameters:

Verification Precision ($q \in [0,1]$): This parameter represents the probability that a legitimate whistleblower report is successfully verified by the regulator as a valid violation. It reflects the technological capacity of the governance system (e.g., automated evidence tracing or blockchain-based auditing). A higher q implies lower signal noise.

False-Report Deterrence ($f > 0$): Drawing on Buccirossi et al. (2021), we introduce a penalty f for unsubstantiated or false reports. This creates a necessary risk-return trade-off for users, curbing opportunistic behavior.

Under these friction conditions, the expected payoff structure for a user choosing to Participate ($z = 1$) changes fundamentally. The payoff is no longer a deterministic value ($N + M - C_4$) but an expected value conditioned on the enterprise's strategy (y) and the system's verification capability (q):

Scenario A: If the Enterprise is Compliant ($y = 1$). Any report submitted by the user is inherently a "False Positive". The verification fails (since there is no violation), and the user incurs the penalty f.

$$U_{z1|y=1} = -f - C_4$$

Scenario B: If the Enterprise is Non-Compliant ($y = 0$). The report is a valid "True Positive", but verification is probabilistic: With probability q, the report is verified. The user receives the reward N and compensation M; With probability 1-q (False Negative), the system fails to verify the violation due to technical noise. The report is deemed unsubstantiated, and the user incurs the penalty f.

$$U_{z1|y=0} = q(N+M) + (1-q)(-f) - C_4$$

Combining these scenarios, the average expected payoff for a user choosing to participate (U_{z1}) is:

$$U_{z1} = y(-f - C_4) + (1-y)\left[q(N+M) + (1-q)(-f) - C_4\right]$$

$$= (1-y)q(N+M) - \left[y + (1-y)(1-q)\right]f - C_4$$

The expected payoff for Silence ($z = 0$) remains normalized to 0 (assuming the baseline loss is sunk). Consequently, the modified replicator dynamics equation for the user population is derived as Equation (12):

$$F(z) = \frac{dz}{dt} = z(1-z)\{(1-y)\cdot q(N+M) - [y + (1-y)(1-q)]f - C_4\} \qquad (12)$$

Note that in replicator dynamics, only payoff differences between strategies matter. Therefore, any baseline loss that affects both reporting and silence identically can be subtracted from both payoffs without changing $F(z)$ in Equation (12). This is why the expected payoff for Silence is normalized to zero, and Equation (12) should be read as the net incentive of whistleblowing relative to silence.

Interpretation of Equation (12):

The term $[y+(1-y)(1-q)]$ represents the aggregate probability that a report is not substantiated under the verification process, and thus triggers the penalty f. It captures two distinct channels of reporting failure faced by a whistleblower:

1.Compliant-enterprise channel (probability y): the enterprise is compliant, so the report lacks a verifiable violation under the strict-evidence rule and is treated as unsubstantiated.

2. Verification-failure channel (probability $(1-y)(1-q)$): the enterprise is noncompliant, but the regulator fails to verify the violation with probability $(1-q)$, so the report is again treated as unsubstantiated.

Equation (12) therefore implies that user participation depends on whether the expected benefit from successful verification, $(1-y)q(N+M)$, outweighs the combined burden of reporting costs C_4 and the expected penalty from unsubstantiated outcomes, $[y+(1-y)(1-q)]f$.

5.2 The Signal-Quality Threshold Effect

A key theoretical contribution of this extension is the identification of a "quality threshold" for effective collaborative governance. Unlike traditional models (Section 4) that suggest a linear relationship between rewards and participation (Ma et al., 2025), we find that signal quality imposes a hard constraint on incentive design.

Proposition 2 (Signal-Quality Threshold). *There exists a critical threshold of verification precision, denoted as* q^*. If the system's verification capability falls below this threshold ($q < q^*$), the user participation strategy z cannot evolve towards the active state ($z \to 1$), regardless of the magnitude of the monetary incentive (N). Under these conditions, the collaborative governance mechanism inevitably fails.

The critical threshold is given by:

$$q^* = \frac{f + C_4}{N + M + f} \tag{13}$$

Proof. According to the replicator dynamics equation for users (Eq. 12), for the participation strategy z to evolve (i.e., for $F(z) > 0$), the expected marginal benefit of reporting must exceed its marginal cost.

We analyze the most favorable condition for reporting: when the enterprise is fully non-compliant ($y = 0$). If public users lack the motivation to report even when

violations are rampant, the governance mechanism is structurally flawed. Substituting $y = 0$ into the bracketed term of Eq. (12) and simplifying the expression, we derive the condition for user participation as follows:

$$q(N+M)-(1-q)f-C_4>0$$

Here, $q(N + M)$ represents the expected reward from a true positive, while $(1-q)f$ represents the expected penalty risk from a false negative (technical failure). Rearranging the terms to solve for q:

$$q\left(N+M+f\right)>f+C_4$$

$$q>\frac{f+C_4}{N+M+f}$$

Thus, we define the lower bound $q^* = \frac{f+C_4}{N+M+f}$. If the actual precision $q \leq q^*$, the net expected payoff for reporting is permanently negative, forcing the user population to converge to the silence strategy (z→0), leading to system collapse. □

5.3 Theoretical Interpretation of the Threshold Mechanism

Proposition 2 offers a counter-intuitive insight for policy design that challenges the prevailing "monetary solution" hypothesis. By synthesizing our mathematical derivation with signal detection theory, we articulate three fundamental mechanisms that govern the stability of collaborative governance:

1.The "Silence Trap" and Noise Risks.. Equation (13) defines a structural boundary we term the "Silence Trap". In environments with immature detection technologies ($q < q^*$), the governance mechanism falls into a state of "technological failure". Here, increasing rewards (N) is not only inefficient but potentially destabilizing. It fails to lower the effective threshold significantly while attracting speculative noise that overwhelms the system. This finding aligns with the "sanction-reward" trade-offs identified in legal theory (Givati, 2016) and explains the empirical observation that successful bug bounty programs require rigorous triage mechanisms to filter low-quality submissions (Finifter et al., 2013). Users rationally choose silence to avoid the "accidental punishment" caused by verification errors, rendering monetary incentives futile.

2. The Paradox of Deterrence. The threshold formula $q^* = \frac{f+C_4}{N+M+f}$ also illuminates a "Deterrence Paradox". While increasing the false-report penalty (f) is theoretically necessary to curb malicious noise, it simultaneously raises the technical threshold q^* required for user participation (since $\frac{\partial q^*}{\partial f} > 0$ when $N+M > C_4$). This implies that stricter penalties require higher verification precision. If a regulator imposes harsh penalties for false reports but lacks the technology to distinguish between "malicious intent" and "technical error" perfectly, they will inadvertently suppress honest whistleblowers, pushing the system back into the Silence Trap.

3. Governance Regime Shift. Consequently, effective governance requires a paradigm shift from purely "Incentive-based Design" to "Technology-constrained Design". The system converges to the collaborative equilibrium (where $x = 1$, $y = 1$, $z = 1$) only within the "Feasible Region" defined by $q > q^*$. Outside this region, policymakers must prioritize upgrading verification infrastructure (increasing q) over increasing budget allocations for rewards (increasing N).

3. The Governance Triad. Consequently, the governance mechanism must be conceptualized not as a linear incentive structure but as a "Governance Triad" consisting of Incentives (N), Verification (q), and Deterrence (f). Effective governance requires a paradigm shift from purely "Incentive-based Design" to "Technology-constrained Design". The system converges to the ideal collaborative equilibrium (where $x = 1$, $y = 1$, $z = 1$) only within the "Feasible Region" defined by $q > q^*$. Outside this region, policymakers must prioritize upgrading verification infrastructure (increasing q) over increasing budget allocations for rewards (increasing N), a condition we will visually demonstrate through regime maps in the following numerical analysis.

6 Numerical Simulations and Regime Analysis

To validate the theoretical propositions derived in Section 5 and to visually characterize the system's evolutionary boundaries, we conduct numerical simulations using MATLAB. Unlike previous studies that focus solely on time-series evolution under ideal conditions, this section specifically analyzes how whistleblowing frictions, specifically verification precision (q) and false-report deterrence (f), alter the governance landscape.

6.1 Parameterization and Experimental Setup

We establish a baseline parameter set (Table 3) derived from the operational realities of cybersecurity regulation. Crucially, these parameter values are calibrated to satisfy

the theoretical stability constraints derived in Section 4, specifically, the participation constraint ($N+M > C_4$) and the regulatory incentive constraint ($E > C_1 + N$). This ensures that the baseline represents a viable collaborative governance system under ideal conditions. In our experimental design, this baseline treats whistleblowing as an "ideal" channel ($q = 1, f = 0$), serving as a control group. Subsequent experiments will introduce frictions ($q < 1, f > 0$) to isolate the impact of signal noise on system stability.

Table 3. Baseline Parameter Values

Parameter	Symbol	Value	Parameter	Symbol	Value
Government			**User (Whistleblower)**		
Regulatory Reward	E	10	Whistleblowing Reward	N	3
Regulatory Cost	C_1	5	Reporting Cost	C_4	2
Enterprise			Compensation	M	6
Penalty	P	12	**Friction Parameters**		
Reputation Loss	H	8	Verification Precision	q	0.9
Compliance Cost Saving	ΔC	15	False-Report Penalty	f	2

(Note: The initial state for all three populations is set to $x_0 = y_0 = z_0 = 0.5$. The system of differential equations is solved numerically using MATLAB's ode45 solver, which employs an explicit Runge-Kutta (4,5) formula, with a time span $t \in [0, 50]$.

6.2. The Impact of Frictions: From Dynamic Oscillation to Systemic Collapse

Building on the baseline parameters, we now introduce signal frictions into the evolutionary framework to simulate how the system deviates from the ideal trajectory. We construct two distinct scenarios to contrast the system's behavior under "High Precision" versus "Low Precision" environments. The simulation results are presented in Figure 2.

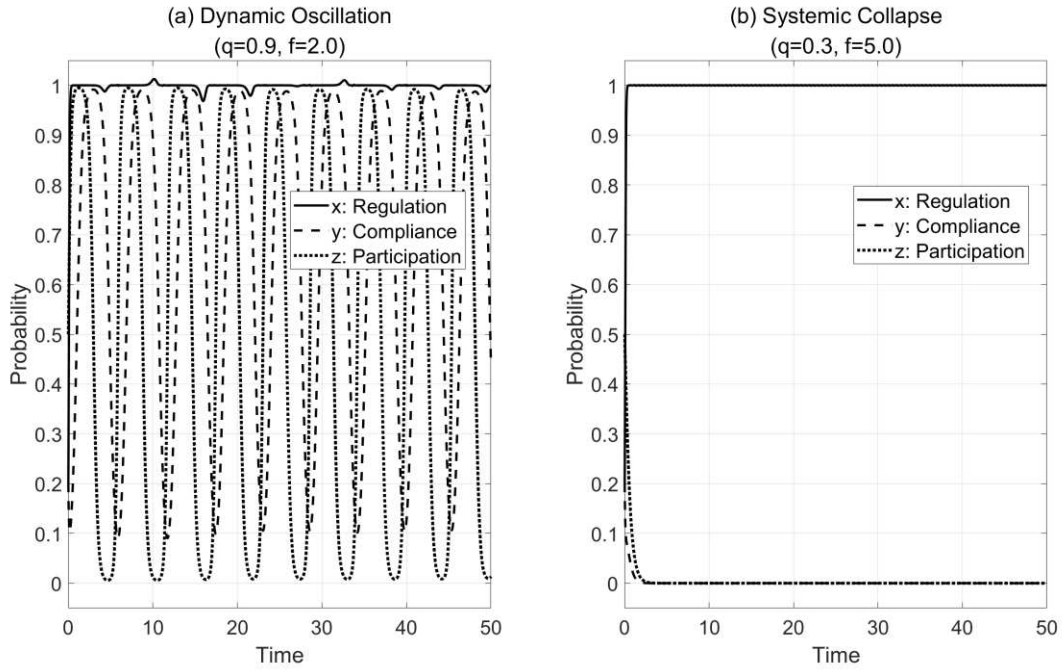

Figure 2. Evolutionary trajectories of tripartite strategies under different signal qualities. (Left Panel) (a) Dynamic Oscillation (High Precision q=0.9, f=2.0): The system maintains a "predator-prey" dynamic equilibrium. User participation (z, dotted line) and firm compliance (y, dashed line) fluctuate cyclically, indicating that high-quality crowdsourcing effectively bounds violations but does not eliminate them. (Right Panel) (b) Systemic Collapse (Low Precision q = 0.3, f = 5.0): The system falls into the "Silence Trap". User participation (z) collapses to 0 immediately, causing the firm (y) to revert to full non-compliance.

As illustrated in Figure 2, the system exhibits periodic oscillations in the high-verification regime rather than converging to a fixed point. These oscillations are not numerical artifacts but arise endogenously from the nonlinear feedback structure of the replicator dynamics. In this parameter region, trajectories cycle around the interior equilibrium, which is consistent with a predator–prey-type interaction between monitoring intensity and corporate compliance. Such behavior is consistent with Hopf-type mechanisms in evolutionary systems, where changes in local stability can lead to closed orbits and sustained cyclical adjustment. Substantively, the result implies that under high technological capacity, effective governance is better understood as a dynamically maintained balance that requires continuous strategic adjustment, rather than a static terminal equilibrium.

6.2.1 Scenario A: Dynamic Stability under High Precision

As illustrated in Figure 2(a), we set the verification precision to a high level (q=0.9) with a moderate penalty for false reports (f = 2.0). Departing from the monotonic

convergence predicted in idealized models, the system here exhibits a “predator-prey” like oscillation. This dynamic behavior arises from the endogenous cyclicality between enforcement and compliance: initially, high user participation (z, dotted line) forces the firm to increase compliance (y, dashed line). However, as the firm becomes compliant ($y \to 1$), the “bounty hunting” opportunities for users diminish, causing participation to drop. This decline subsequently tempts the firm to violate regulations again, restarting the cycle. These findings suggest that under high-quality crowdsourcing, the system maintains a dynamic equilibrium where regulatory pressure effectively bounds the violation rate within a controllable range, rather than settling into a static state.

6.2.2 Scenario B: Systemic Collapse under Low Precision

In contrast, Figure 2(b) depicts the scenario where verification technology is immature ($q = 0.3$), even when a harsher penalty ($f = 5.0$) is imposed to deter noise. The evolutionary paths reveal a rapid systemic collapse. Specifically, the user participation rate (z) precipitates to 0 almost immediately ($t < 10$) because the friction cost derived from false-positive risks outweighs the expected reward. Consequently, bereft of the external threat from crowdsourced supervision, the firm's strategy (y) quickly degrades toward non-compliance ($y \to 0$). This empirical observation validates Proposition 2, confirming that when signal quality falls below the critical threshold (q^*), the system inevitably falls into a “Silence Trap”, leading to complete governance failure regardless of the punitive deterrence employed.

6.3. Regime Maps and Feasible Design Spaces

To provide actionable guidance for mechanism design, we move beyond time-series analysis to conduct a global scan of the parameter space. Figure 3 illustrates the “Policy Regime Map”, visualizing the combined effect of Monetary Incentives (N) and Verification Precision (q) on governance outcomes. The color gradient represents the equilibrium level of user participation (z), ranging from deep blue (Silence Trap, $z \to 0$) to bright yellow (Active Collaboration, $z \to 1$).

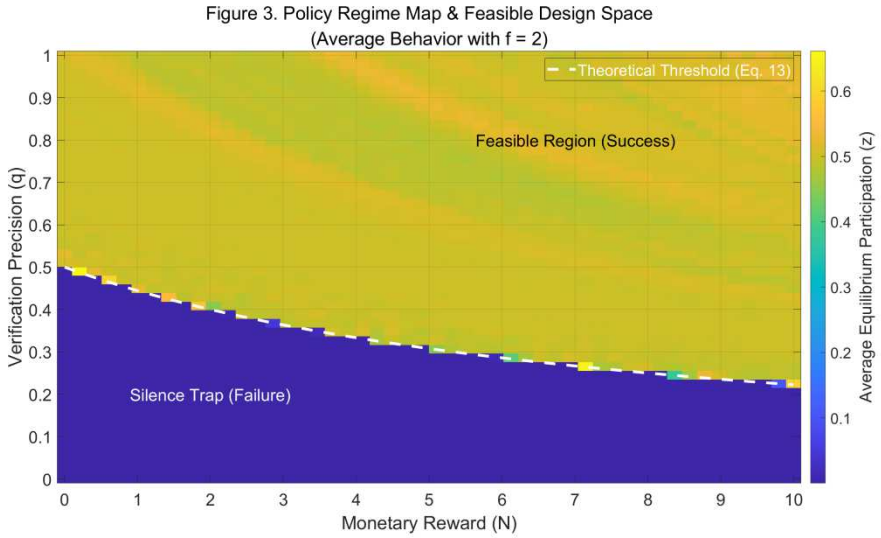

Figure 3. Policy Regime Map & Feasible Design Space.

The X-axis represents the Monetary Reward (N), and the Y-axis represents Verification Precision (q). The white dashed line indicates the theoretical threshold (q^*) derived in Proposition 2. The region below the line is the "Silence Trap" (Failure), while the region above is the "Feasible Region' (Success).

The bifurcation diagram provides a visual empirical validation of Proposition 2, revealing the non-linear structural boundaries of effective governance:

1.The "Hard Floor" of Technology: The white dashed line perfectly delineates the boundary between success and failure. The area below this curve constitutes the "Silence Trap." In this region ($q < q^*$), the system remains trapped in governance failure regardless of how high the monetary reward N is raised (moving rightward along the X-axis). This finding is theoretically significant as it disproves the "monetary solution" hypothesis; it demonstrates that financial incentives cannot substitute for technological capability. When signal quality is insufficient, increasing rewards is not only ineffective but potentially wasteful.

2.The Feasible Region: Effective governance is only attainable in the "Feasible Region" above the threshold curve. Within this zone, higher signal quality (q) allows the system to sustain collaboration with lower monetary costs. The convex shape of the threshold curve implies diminishing returns for rewards: as N increases, the marginal reduction in the required technology threshold (q^*) decreases.

3.Design Implication: Consequently, for policymakers operating in low-tech environments, the most cost-effective strategy to escape the Silence Trap is to invest in upgrading verification precision (moving vertically up the map) rather than merely expanding the bounty budget (moving horizontally right).

6.4. Sensitivity Analysis and Robustness Checks

To ensure that the identified mechanisms, specifically the "Silence Trap" and the "Threshold Effect", are not artifacts of specific parameter choices, we performed two additional robustness checks (results visualized in supplementary materials, summarized here):

6.4.1 Independence from Initial Strategies (Path Independence)

A common concern in evolutionary games is path dependence. We simulated the system evolution under 30 randomly generated initial states ($x_0, y_0, z_0 \in [0,1]$) for the low-quality scenario (q=0.3). The simulation results reveal a striking consistency: regardless of whether the initial public enthusiasm is high or low, all trajectories rapidly converge to the collapse state (z→0). This confirms that the "Silence Trap" is a global attractor. It implies that the governance failure is determined by structural constraints, specifically the lack of verification precision, rather than by the historical starting points of the actors. The path-convergence analysis (see Figure. S1 in Supplementary Materials) demonstrates that the "Silence Trap" is a global attractor. This implies that once the system falls below the precision threshold, no amount of "initial public enthusiasm" can prevent the eventual collapse of collaborative governance.

6.4.2 Robustness of the Threshold Mechanism

We further examined the structural stability of the threshold mechanism by varying the false-report penalty (f) from 2.0 to 5.0. Although strictly punishing false reports shifts the threshold curve upward (implying higher technical precision is required to offset user anxiety), the hyperbolic shape of the boundary curve remains unchanged as predicted by Proposition 2. This persistence indicates that the "hard floor" for signal quality is robust to parameter variations. While incentives (N) and penalties (P) can modulate the specific position of the boundary, they cannot eliminate the fundamental constraint that signal quality (q) must exceed a critical floor (q^*) for the collaborative governance mechanism to function. Further robustness checks (detailed in Supplementary Materials, S2 and S3) reveal that while traditional economic parameters like P and ΔC can shift convergence speeds, they cannot override the structural constraints imposed by signal quality. This implies a diminishing marginal utility of purely monetary incentives (N); when q is low, increasing N fails to alter the system's destination, confirming that technological intervention must precede monetary expansion.

7. Conclusions and Implications

This study constructs an evolutionary game model to explore the dynamics of collaborative cybersecurity governance under imperfect information. By endogenizing verification precision (q) and false-report penalties (f), we move beyond the idealized assumption of perfect monitoring to reveal the structural boundaries of crowdsourced regulation.

7.1. Main Findings

Our mathematical derivation and numerical simulations yield three pivotal findings regarding the stability boundaries of collaborative governance:

1. Structural Bifurcation and the "Silence Trap": We identify a distinct structural bifurcation in the governance mechanism determined by verification precision (q). As proven in Proposition 2 and visualized in the Regime Map (Figure 3), there exists a "hard threshold" (q^*) below which the system collapses into a "Silence Trap". In this regime, user participation exhibits "Incentive Inelasticity": increasing monetary rewards (N) fails to stimulate participation because the expected cost of false-positive penalties outweighs any potential bounty. This proves that governance failure in low-tech environments is a structural inevitability rather than a result of insufficient funding.

2. The Transition from Static to Dynamic Equilibrium: Contrary to the classical assumption that effective regulation leads to a static state of full compliance, our simulations (Section 6.2) reveal that successful governance in high-precision environments manifests as "Dynamic Oscillation". The system maintains stability through a predator-prey cycle: user participation rises to correct corporate violations and recedes when compliance is achieved. This implies that "Success" is not a fixed destination but a continuous, dynamic process of tension and adjustment.

3. The Deterrence Paradox (The Double-Edged Sword of Penalties): We find that the false-report penalty (f) functions as a double-edged sword. While increasing f is theoretically necessary to filter out speculative noise, it simultaneously raises the "barrier to entry" for legitimate whistleblowers. Our sensitivity analysis demonstrates that a high-penalty regime strictly requires higher technological precision to remain viable. If the regulator imposes harsh penalties without matching verification capabilities, they inadvertently suppress valid signals, pushing the system back into the Silence Trap.

7.2. Theoretical Implications

This study contributes to the literature on collaborative governance and evolutionary game theory in two significant ways:

1.Endogenizing Information Friction: unlike previous studies that treat whistleblowing as a frictionless "ideal channel", we prove that signal quality is a structural determinant of governance stability. We challenge the prevailing "monetary solution" hypothesis by demonstrating that incentive mechanisms (N) are effectively "switched off" when the technological infrastructure (q) is immature. This bridges the gap between signal detection theory and institutional design.

2.The "Governance Triad" Framework: We expand the traditional binary focus on "Crime and Punishment" (Becker's model) to a "Governance Triad" consisting of Incentives, Verification, and Deterrence. This framework offers a more holistic lens for analyzing regulatory failures, suggesting that governance effectiveness depends on the geometric balance of these three dimensions rather than the linear increase of any single parameter.

7.3. Policy Implications

Based on the "Governance Triad", we offer actionable recommendations for policymakers:

1.Technology Before Incentives: Regulators must prioritize the modernization of verification technologies over the expansion of reward budgets. In low-tech environments, launching high-reward bounty programs is a resource mismatch that exacerbates noise without improving compliance.

2.Graded Penalty Mechanisms: To resolve the Deterrence Paradox, policymakers should implement a graded penalty system. Distinguishing between "malicious fabrication" (harsh penalty) and "technical error" (low penalty) reduces the psychological cost for honest whistleblowers, thereby lowering the effective threshold for participation.

3.Adaptive Governance: Recognizing the endogenous cyclicality of compliance, regulators should adopt adaptive strategies. During periods of high compliance, when user participation naturally wanes, the government must temporarily increase direct supervision (C_1) to prevent corporate backsliding.

7.4. Limitations and Future Research

This study has several limitations. First, we employ a simplified verification parameter, which captures the probability that a valid report is successfully verified

(the true positive rate). Future research could adopt a fuller confusion-matrix formulation by introducing false-positive errors, thereby examining how wrongful verification outcomes may affect compliant firms and the overall governance dynamics. Second, we assume a homogeneous user population. Subsequent models could incorporate agent heterogeneity, for example by distinguishing expert "white-hat" hackers from novice users, to study how differences in skill and reporting behavior translate into effective signal quality and participation incentives. Third, the enterprise's reputation loss is treated as a static parameter. Embedding dynamic reputation mechanisms, such as cumulative trust erosion and recovery, could yield deeper insights into long-run corporate behavior.

References

Anderson, R., & Moore, T. (2006). The economics of information security. *Science, 314*(5799), 610-613.

Ansell, C., & Gash, A. (2008). Collaborative governance in theory and practice. *Journal of Public Administration Research and Theory*, *18*(4), 543-571.

Ayres, I., & Braithwaite, J. (1992). Responsive regulation: Transcending the deregulation debate. Oxford University Press, USA.

Boone, J. H., & Dahan, M. (2025). Inspection game with imperfect detection technology. *Naval Research Logistics*.*72*, 694-712.

Buccirossi, P., Immordino, G., & Spagnolo, G. (2021). Whistleblower rewards, false reports, and corporate fraud. *European Journal of Law and Economics*, *51*(3), 411-431.

Emerson, K., Nabatchi, T., & Balogh, S. (2012). An integrative framework for collaborative governance. *Journal of Public Administration Research and Theory, 22*(1), 1-29.

Finifter, M., Akhawe, D., & Wagner, D. (2013). An empirical study of vulnerability rewards programs. In *22nd USENIX security symposium (USENIX Security 13)* , 273-288.

Friedman, D. (1998). On economic applications of evolutionary game theory. *Journal of Evolutionary Economics, 8*, 15-43.

Gal-Or, E., Hydari, M. Z., & Telang, R. (2024). Merchants of vulnerabilities: How bug bounty programs benefit software Vendors. *arXiv preprint arXiv:2404.17497*.

Givati, Y. (2016). A theory of whistleblower rewards. *The Journal of Legal Studies, 45*(1), 43-72.

Herley, C. (2013). More is not the answer. *IEEE Security & Privacy, 12*(1), 14-19.

Kaplow, L., & Shavell, S. (1994). Optimal law enforcement with self-reporting of behavior. *Journal of Political Economy, 102*(3), 583-606.

Kshetri, N. (2017). Blockchain's roles in strengthening cybersecurity and protecting privacy. *Telecommunications Policy, 41*(10), 1027-1038.

Kuehn, A., & Mueller, M. (2014, August). Analyzing bug bounty programs: An institutional perspective on the economics of software vulnerabilities. In *2014 TPRC Conference.*

Linders, D. (2012). From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly*, *29*(4), 446-454.

Ma, J., Yu, J., & Cheng, Y. (2025). Dynamic evolution analysis of incentive strategies and symmetry enhancement in the personal-data valorization industry chain. *Symmetry, 17*(10), 1639.

Maillart, T., Zhao, M., Grossklags, J., & Chuang, J. (2017). Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity, 3*(2), 81-90.

Near, J. P., & Miceli, M. P. (1985). Organizational dissidence: The case of whistle-blowing. *Journal of Business Ethics, 4*(1), 1-16.

Ostrom, E. (2017). Polycentric systems for coping with collective action and global environmental change. In *Global justice* (pp. 423-430). Routledge.

Zrahia, A., Gandal, N., Markovich, S., & Riordan, M. (2024). The simple economics of an external shock to a bug bounty platform. *Journal of Cybersecurity*, *10*(1), tyae006.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.

Savaş, S., & Karataş, S. (2022). Cyber governance studies in ensuring cybersecurity: an overview of cybersecurity governance. *International Cybersecurity Law Review, 3*(1), 7-34.

Shackelford, S. J. (2012). Toward cyberpeace: Managing cyberattacks through polycentric governance. *American University Law, 62*, 1273.

Smith, J. M. (Ed.). (2012). *Did Darwin get it right?: Essays on games, sex and evolution*. Springer Science & Business Media.

Tian, Z., Han, M., & Jiang, C. (2024). Research on platform data security governance strategy based on three-party evolutionary game. *Heliyon, 10*(17).e36391.

Traulsen, A., & Hauert, C. (2009). Stochastic evolutionary game dynamics. *Reviews of nonlinear dynamics and complexity, 2*, 25-61.

Walshe, T., & Simpson, A. C. (2022). Coordinated vulnerability disclosure programme effectiveness: Issues and recommendations. *Computers & Security, 123*, 102936.

Wang, S., Zou, K., Zou, Y., & Jiang, Z. (2025). Dynamic tripartite governance in data security: An evolutionary game model with cross-level government supervision. *Plos One, 20*(6), e0325473.

Weibull, J. W. (1995). *Evolutionary game theory*. MIT Press.

Xu, F., Wang, J., & Chen, K. (2025). Mechanism design for public participation in regulatory enforcement: A perfect monitoring perspective. *Computers & Industrial Engineering*, *187*, 109812.

Yang, Q., & Wang, Z. (2025). Digital empowerment and the transformation of modern governance: Trends and challenges. *Public Administration Review*, *85*(1), 12-25.

Zrahia, A., Gandal, N., Markovich, S., & Riordan, M. (2024). The simple economics of an external shock to a bug bounty platform. *Journal of Cybersecurity, 10*(1), tyae006.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx