Accepted at ***Management Science***

Economics of Social Media Fake Accounts

Zihong Huang
Rawls College of Business, Texas Tech University, Lubbock, TX 79409, Zihong.Huang@ttu.edu
De Liu
Carlson School of Management, University of Minnesota, MN 55455, deliu@umn.edu

Amid the rise of the influencer economy, fake social media accounts have become prevalent on many social media platforms. Yet the problem of fake accounts is still poorly understood, and so is the effectiveness of coping strategies. This research models the ecosystem of fake accounts in an influencer economy and obtains insights on fake-account purchasing behaviors, the impact of anti-fake efforts, and the roles of various contextual factors. We show that as the anti-fake effort increases, the equilibrium may transition from a "pooling" equilibrium where a low-quality influencer buys fake accounts to mimic a high-quality one, to a "costly-separating" equilibrium where a high-quality influencer may buy fake accounts to prevent mimicry from a low-quality influencer, and to a "naturally-separating" equilibrium where low- and high-quality influencers are separated without buying fake accounts. We find that increasing anti-fake efforts and increasing social media literacy may sometimes result in more fake accounts. A purely profit-driven platform always prefers a pooling equilibrium with zero anti-fake effort. As a platform puts more weight on consumer welfare, it may exert a positive effort to induce a separating equilibrium, but the platform's preferred anti-fake effort tends to be lower than that of consumers. We also find that the platform sometimes prefers a lower social media literacy and a lower fake-account base price, whereas consumers prefer the opposite. In contrast, improving the anti-fake technology level can benefit both the platform and consumers. Our main insights are applicable to scenarios with more influencer types and repeated interactions.

Key words: Influencer Economy, Fake Accounts, Social Media, Signaling, Social Media Literacy
History: This paper was first submitted on Aug 22, 2022 and is accepted on July 23, 2024

1. Introduction

On Oct 16, 2019, a popular microblogger with 3.8 million followers on Weibo, one of the largest microblogging platforms in China, posted an advertisement. Within 50 minutes, the advertisement garnered 121k views, thousands of likes, and hundreds of comments and shares. The advertiser was thrilled to see the response but surprised by the number of conversions: zero! It turned out that the microblog was infested with fake followers. This incidence is not alone: Facebook, Instagram,

and Twitter have all been reported struggling with fake account problems (Confessore et al. 2018, Moore and Murphy 2019, Freixa 2021, Ortutay 2022).

By "fake accounts," we mean social media accounts designed to impersonate real users with fake personal information and/or behaviors. While a majority of fake accounts are automated (or "bots"), fake accounts may also be created and operated by real humans (Nicas 2020). Fake accounts are created for several reasons. The most common use of fake accounts is to boost influence: social media influencers may buy fake accounts to make their accounts or social media posts appear more influential, which can result in more followers and higher advertising income. Some users may create fake accounts to obtain perks such as signup bonuses and coupons. Others may use fake accounts to spread phishing, scams, and malware. Yet others may use fake accounts to sway public opinions and election outcomes. This paper focuses on influence-boosting fake accounts due to their prevalence.

The demand for influence-boosting fake accounts has surged in the influencer economy (Confessore et al. 2018, Federal Trade Commission 2019), in which large and small social media influencers can get paid for each piece of promotional content (e.g., sponsored posts and product endorsements) they share, depending on their reach (e.g., number of followers) and engagement (e.g., clicks and likes) metrics.[1] As such, influencers have strong incentives to use fake accounts to artificially boost their reach and engagement metrics. The link between the influencer economy and fake accounts is highlighted in the widely publicized case of the Federal Trade Commission (FTC) versus Devumi in 2019. Devumi is a company that made millions of dollars by manufacturing and selling fake accounts and associated services to actors, athletes, musicians, and other high-profile individuals who wanted to appear more popular and influential online (Confessore et al. 2018). Though FTC imposed a fine of $2.5 million on Devumi with the intent of deterring future fake accounts trading, the fake account problem has never abated – for example, in the first quarter of 2022, Facebook shut down 1.6 billion fake accounts and estimated that there were no less than five percent of the Facebook users were still fake after the removal (Warwick 2022).

Because fake accounts can result in wasteful marketing expenditures and mislead social media users, social media platforms have already begun to tackle the fake account problem. A major tool used for fighting fake accounts is automated fake account detection and prevention. For example, Facebook uses machine learning to detect and block fake accounts both before and after they come alive (Hao 2020). Social media platforms also use user verification technologies such as reCAPTCHA and two-factor authentication to deter automated fake account creation. Such platform-led *anti-fake efforts* face considerable technical challenges, however. Not only there is a

[1] `https://www.aspire.io/blog/influencer-budget-calculator`

wide range of user behaviors, making it nearly impossible to automatically differentiate between real and fake accounts (Nicas 2020), but also fake account providers are increasingly adept at emulating genuine users with advanced AI techniques (e.g., a recent report found computer-generated images in fake LinkedIn profiles (Robins 2022). Given these technical challenges, it is unclear how platform-led anti-fake efforts may affect the prevalence of fake accounts.

There is also a concern about whether social media platforms can be trusted to resolve the issue of fake accounts on their own. Social media platforms, due to their inherent interest in maintaining a large number of active users, may be reluctant to reveal a large number of fake accounts or remove them. Moreover, anti-fake efforts such as fake account detection and user verification tend to inconvenience legitimate users. For example, heightened user verification aimed at preventing fake accounts can frustrate legitimate users, potentially driving them away from the platform (ArkoseLabs 2021). In general, we do not yet know whether the platform can be expected to exert the level of anti-fake efforts that is ideal for consumers.

Besides platform-led anti-fake efforts, a few contextual factors may potentially impact the prevalence of fake accounts. For example, schools and online education institutions have taken initiatives to improve *social media literacy* among young and adult social media users (Al Zou'bi 2022). With enhanced social media literacy, users can become better at discerning the veracity of information on social media, which may reduce the influence of fake accounts on consumer decisions (Polanco-Levicán and Salvo-Garrido 2022). Meanwhile, regulators and government agencies can also use tools such as regulations and fines to make it more costly to operate and trade fake accounts. To develop effective strategies for tackling the fake account problem, we need to understand how such factors may impact the prevalence of fake accounts, consumer welfare, and social media platforms' profitability.

Our current understanding of the fake account phenomenon remains highly limited due to the scarcity of reliable data on fake accounts. For example, we do not know what types of influencers are more likely to buy fake accounts or if it is always better to have fewer fake accounts. Existing studies of fake accounts primarily focus on examining fake accounts' activities and developing detection techniques (Raturi 2018, Yuan et al. 2019). Numerous questions surrounding the fake account phenomenon, including the ones highlighted above, remain unanswered. These suggest a need for a systematic examination of the fake account phenomenon so that multiple related questions can be answered holistically.

To address this need, we build a stylized game-theoretical model of fake accounts in the context of an influencer economy. This model comprises several stakeholders, including consumers, a representative influencer, an advertiser, and a social media platform. The influencer can be either

high-quality or low-quality. A proportion of consumers are "informed" about the influencer's quality, whereas the rest are "uninformed." Informed consumers make their following decisions first. Uninformed consumers make their following decisions after seeing the influencer's follower count, which could be inflated by fake accounts. In this economy, the advertiser reaches consumers through the influencer, and both the influencer and the platform get a share of the advertising revenue. The platform can mount an anti-fake effort that increases the price of fake accounts, but such an effort inevitably also increases the nuisance cost of consumers. The *anti-fake technology level* governs the extent to which the anti-fake effort can raise the price of fake accounts without imposing much nuisance cost on consumers. In our model, the platform maximizes a weighted sum of profits and consumer welfare. When the platform assigns zero weight to consumer welfare, it is *purely profit-driven*. As the weight increases, the platform becomes more *consumer-oriented*. We use this model to address a host of questions, such as:

1. What is the social media influencer's equilibrium fake-account purchasing behavior?

2. How does the platform's anti-fake effort affect the number of fake accounts, the platform's profits, and consumer welfare? Is the platform's preferred anti-fake effort aligned with consumers'?

3. How do contextual factors such as the anti-fake technology level, social media literacy (as measured by the proportion of informed consumers), and the base price of fake accounts affect the equilibrium, the platform's profits, and consumer welfare?

Our analyses show that there is a *pooling* equilibrium where a low-quality influencer (L-type for short) *offensively* purchases fake accounts to mimic a high-quality influencer (H-type), a *costly-separating* equilibrium, where an H-type *defensively* purchases fake accounts to prevent an L-type from mimicking, and a *naturally-separating* equilibrium where the two types of influencers separate without needing to purchase fake accounts. As the platform's anti-fake effort increases, the equilibrium generally transitions from a pooling, a costly-separating, and then to a naturally-separating equilibrium. Interestingly, the number of fake accounts may sometimes increase with the anti-fake effort, especially in the pooling equilibrium. This is because, under a pooling equilibrium, a higher anti-fake effort results in a larger gap in the H- and L-type's follower counts, forcing the L-type to purchase more fake accounts despite higher costs. We also observe that the number of fake accounts can jump significantly as the system transitions from a pooling to a separating equilibrium – this is because an L-type is willing to pay more to be seen as an H-type in a separating equilibrium than as an average type in a pooling equilibrium.

We show that a purely profit-driven platform always prefers zero anti-fake effort which results in a pooling equilibrium. The reason is that a pooling equilibrium can result in more uninformed followers and higher advertising revenue. As the platform becomes more consumer-oriented, it may exert a positive effort to induce a costly-separating or a naturally-separating equilibrium. In

general, though, the platform's optimal anti-fake effort tends to be lower than what is optimal for consumers.

Moreover, while consumers benefit from a higher level of social media literacy and a higher fake-account base price, a sufficiently profit-focused platform may prefer the opposite. We also find that, under a pooling equilibrium, the number of fake accounts is unaffected by the base price of fake accounts – this is because, under such an equilibrium, the L-type must make up for the follower gap regardless of the price of fake accounts. This suggests that the social media platform may lack interest in improving social media literacy or raising the price of fake accounts, although such measures can improve consumer welfare. In contrast, both the platform and consumers may benefit from improving anti-fake technology (under a separating equilibrium).

In addition, we show that key insights from the one-shot game with two types of influencers could carry over a repeated setting and a model with more influencer types.

2. Related Literature

To our knowledge, the fake social media account problem has not been formally modeled in the literature. However, the literature has studied several other types of deceptive/manipulative behaviors in commerce and advertising contexts, including deceptive advertising (Piccolo et al. 2018), fake reviews, fake sales (Chen et al. 2022), and click fraud (Wilbur and Zhu 2009). In general, our problem and focus are quite different, but there are similarities in the analytical framework. Below, we discuss the relationship between our research and prior studies of deceptive/manipulative behaviors from three aspects: equilibrium behaviors, coping strategies, and welfare implications.

First, our paper is connected to several studies of equilibrium deceptive/manipulative behaviors that also use the signaling model as the analytical framework. In general, the stream on deceptive advertising as well as fake sales usually studies a game in which sellers compete for buyers using deceptive tactics such as false advertising, fake purchases, fake reviews, and so on, along with pricing decisions. In contrast, influencers in our setting have no pricing decisions – they only need to decide how many fake accounts to purchase to influence consumer and advertiser perceptions of them. Furthermore, the previous studies mainly focus on one type of equilibrium. For instance, Piccolo et al. (2018) characterize a class of pooling equilibria where the L-type sellers deceive a buyer. Similarly, Mayzlin (2006) also finds a pooling equilibrium in which sellers with inferior products lie. In contrast, another paper in the same stream focuses on a separating equilibrium (Corts 2013). Recently, Chen and Papanastasiou (2021) study a game in which the seller manipulates the buyers' beliefs with fake purchases. They find that a bad (i.e., L-type) seller may or may not engage in manipulation (i.e., make fake purchases) while a good (H-type) one never manipulates. We, on the other hand, find a different pattern that not only the L-type influencer makes fake-account

purchasing in the pooling equilibrium, but the H-type manipulates (i.e., buys fake accounts) in the costly-separating equilibrium as well. In addition, we also identify a naturally-separating equilibrium in which neither type buys fake accounts.

Second, our paper is also related to a small literature on the effectiveness of anti-fake strategies. This literature has studied the strategies for helping consumers learn the true quality of products through information disclosure (Papanastasiou et al. 2018, Che and Hörner 2018, Pennycook et al. 2020) and penalizing the information producers for their manipulative behaviors (Papanastasiou 2020, Corts 2014). In particular, Chen and Papanastasiou (2021) study the detection-and-removal strategy against seller manipulation (e.g., via fake purchases and reviews) and observe that more intensive detection-and-removal may lead to more seller manipulation because it increases consumers' trust, which further leads to higher equilibrium prices and greater seller manipulation. We also find that anti-fake efforts may sometimes lead to more fake accounts, but for a different reason: it can increase the gap between H- and L-type influencers, which forces the L-type to buy more fake accounts to make up for the gap. Our model of the anti-fake efforts is also different: they increase the cost of fake accounts but also increase the nuisance costs of consumers. Importantly, we gain insights into other interventions, e.g., increasing the level of anti-fake technology and social media literacy, which are new to this literature.

Third, our study is related to research on the welfare effects of deceptive behaviors. Piccolo et al. (2018) examine how consumer welfare is affected by sellers' deceptive strategies. They suggested that consumer welfare could be higher under the equilibrium with sellers' deceptive advertising. Chen et al. (2022) study the impact of brushing (e.g., fake sales) on consumer welfare and find that brushing can either improve or hurt consumer welfare. Our work on consumer welfare is closest to (Chen and Papanastasiou 2021) which suggests that seller and consumer welfare can be maximized at an intermediate level of anti-fake effort by the platform (e.g., detecting fake reviews) or the government (e.g., law enforcement against fake product endorsements). Different from Chen and Papanastasiou (2021)'s work, our study is set in the context of an influencer economy rather than an e-commerce setting. We study not only the platform's anti-fake effort from the consumer welfare point of view but also the welfare impact of other interventions, such as increasing social media literacy, increasing fake-account base price (which has a similar interpretation as the government's anti-fake effort), and increasing effectiveness of anti-fake technology.

Finally, our paper should be contrasted with the study of click fraud by advertisers in the context of search engine keyword auctions by Wilbur and Zhu (2009). Their focus is on the unfair competition between advertisers in an auction context and its impact on search engine revenue. Their game has a very different structure from ours. In addition, they do not study search engines' strategies for coping with click fraud or consumer welfare implications.

3. The Model

We model fake accounts in the context of influencer marketing, where advertisers pay influencers on social media platforms to promote their brands/products. The ecosystem consists of a social media platform, a representative influencer, a unit mass of consumers, and a representative advertiser. Figure 1 depicts their relationships, which we discuss further below.

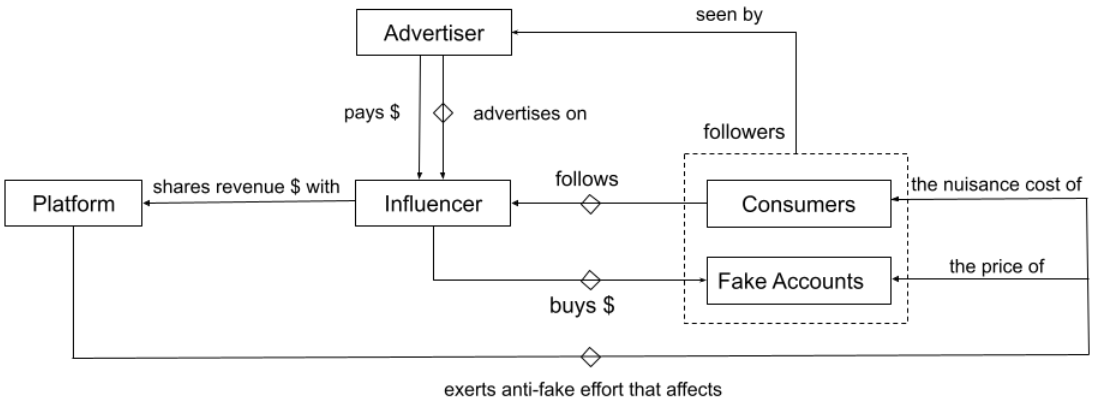


Figure 1 Model Sketch

3.1. The Platform

The social media platform facilitates influencer marketing campaigns, including providing tools for advertisers and influencers to find each other, launch and monitor influencer marketing campaigns, and make payments for such campaigns.[2] In return for such services, the platform gets a share of the revenue generated by influencer marketing campaigns.

The platform chooses the level of anti-fake efforts d $(d \geq 0)$ (*effort* for short), which may include efforts spent on detecting, removing, and preventing fake accounts. For example, the platform may use machine learning to detect fake accounts based on user attributes and behaviors. Such detection can be used to prevent new fake accounts and remove existing ones. It may also deploy technologies such as reCAPTCHA to make it harder to register and operate fake accounts in automated ways. The platform incurs a cost of $\frac{\gamma}{2}d^2$ $(\gamma > 0)$ for its anti-fake effort. The quadratic functional form reflects diminishing returns of such anti-fake effort.

The platform's anti-fake efforts have two effects. First, they increase the *unit price of fake accounts* p_f, which is given by:

$$p_f = p_0 + \frac{1}{1-\tau}d \tag{1}$$

[2] For instance, Facebook/Instagram builds a "Creator Marketplace" to match influencers and brands with a project. It enables influencers to set up profiles, get discovered by brands with a good fit, and negotiate the content details for products/services. The influencer needs to clearly label the content of the product with a tag indicating the partnership with a brand and gets paid when sharing the content of products/services with her followers.

where p_0 is the *base price of fake accounts* and $\tau\,(1>\tau\geq 0)$ is the *anti-fake technology level*. The base price of fake accounts reflects the expenses associated with operating a basic fake account. These costs may be influenced by factors such as the time, effort, and technologies used for creating and maintaining such accounts, as well as the legal risk faced by fake account operators. The anti-fake technology level reflects the effectiveness of current anti-fake technology, which may be understood as the accuracy of machine learning algorithms for detecting fake accounts and the discriminative power of anti-fake tools like reCAPTCHA. This formulation captures the notion that both anti-fake efforts and more effective anti-fake technology can make it more costly to operate fake accounts.

Second, the platform's anti-fake efforts increase the nuisance costs for legitimate users. For example, heightened anti-fake efforts may result in legitimate users answering reCAPTCHA questions more frequently and more likely being misclassified as fake accounts. In general, we assume a consumer incurs a cost c from following the influencer:

$$c = c_0 + c_1\,(1-\tau)\,d \tag{2}$$

where c_0 is the opportunity cost of time and $c_1\,(1-\tau)\,d$ is the nuisance cost imposed by the platform's anti-fake efforts. By this formulation, as the level of anti-fake technology τ increases, consumers' nuisance cost decreases (e.g., they are less likely mistaken as fake accounts).

We assume the platform maximizes a weighted sum of its profit π_p and consumer welfare U:

$$\Pi_p = (1-w)\,\pi_p + wU, \quad w\in[0,1] \tag{3}$$

where the parameter w measures the platform's *consumer orientation*. If $w=0$, the platform is *purely profit-driven*. As w increases, the platform becomes more *consumer-oriented*. Literature shows that firms have goals beyond profit-seeking, such as social responsibility goals (Lankoski and Smith 2018).

3.2. The Influencer

The influencer can be realized as one of the two types: H-type and L-type, with content qualities q_H and q_L $(q_H > q_L)$, respectively. The probability of the influencer being the H-type is ρ. For simplicity, we assume the influencer produces one unit of content. Following prior literature (Shin 2017), we normalize the cost of producing the content to zero. The influencer chooses the number of fake accounts x to maximize its profit π_i (which we will define shortly).

3.3. Consumers

There is a unit mass of risk-neutral consumers. Each consumer decides whether to follow the influencer.[3] A consumer's utility from following an influencer is given by:

$$u(\theta, q) = \theta q - c \tag{4}$$

where q is the quality of the content, θ is the consumer's *valuation* for quality, and c is the cost as defined in (2). We assume θ is random and uniformly distributed on $[0, 1]$.

We further assume that some consumers are better informed than others. For simplicity, we assume a proportion l of consumers are *informed* – they know the influencer's true quality. [4] Such consumers tend to have extensive experience with social media and a stronger ability to tell an influencer's quality. The remaining consumers are *uninformed* – they only know the distribution of the influencer's quality but can observe the influencer's existing followers and form an updated belief about the influencer's quality.[5] Our assumption of uninformed consumers using the influencer's followers to infer quality is supported by prior empirical findings. For example, research shows consumers perceived the influencers with a higher number of followers as being more attractive (Jin and Phua 2014), trustworthy (Jin and Phua 2014), and likable (De Veirman et al. 2017).

We interpret the proportion of informed consumers l as the level of *social media literacy* – the higher the social media literacy, the more consumers are informed and unaffected by fake-account-based manipulations.

3.4. Advertiser

There is one representative advertiser.[6] The advertiser promotes her brand by asking the influencer to share a sponsored post or a product endorsement among the influencer's followers. The advertiser derives a unit revenue μ from advertising to a *real* consumer. The advertiser could be one of the two types: *informed* and *uninformed*. For simplicity, we assume the probability of the advertiser being informed is also l – the same as the proportion of informed consumers. We conduct an analysis with a more general specification in which the advertiser can be informed with a higher probability, and our findings are consistent.

The influencer's type, consumer valuation θ, and the advertiser's type are all private information.

[3] Although we use the term "follow" here, the decision can also be interpreted as a subscription decision or a friendship request, provided that a positive decision allows the consumer to receive the influencer's content.

[4] The driving force of the model remains the same if the informed consumers were only "imperfectly informed" (i.e., they do not know the true quality of the influencer but receive an informative but imperfect signal about it). The key assumption here is that some consumers are *better informed* than others.

[5] We note that, though we use the number of followers as a quality signal, our model can be generalized to other quality signals such as the number of likes and the number of comments. This is because, just as an influencer can purchase fake followers, she can also purchase fake likes/comments generated by fake accounts.

[6] We have also explored an alternative scheme where two or more advertisers compete for the ad slot via a sealed-bid second-price auction, with the auction payment split between the influencer and the platform. The results are quite similar because the driving forces of the model are still the same, although some analyses become less tractable.

3.5. Revenue Sharing

The total advertising revenue generated by the system is $\mu E[n_r]$, where μ is the revenue generated by advertising to a real consumer and $E[n_r]$ is the expected number of real consumers reached. Consistent with common practice in the influencer marketing industry, we assume this revenue is shared between the three parties: the influencer, the advertiser, and the platform. Denote $\lambda_i \in (0,1)$ and $\lambda_p \in (0,1)$ as the influencer's and the platform's shares of revenue, respectively. We denote $\lambda_a \equiv 1-\lambda_i-\lambda_p, \lambda_a \in (0,1)$ as the advertiser's share of the revenue. The parameters λ_i and λ_p are exogenously given and reflect the two parties' relative *bargaining power*.

Given the revenue sharing scheme, the three parties' profits are, respectively

$$\pi_i = \lambda_i \mu E[n_r] - p_f x \tag{5}$$

$$\pi_p = \lambda_p \mu E[n_r] - \frac{\gamma}{2} d^2 \tag{6}$$

$$\pi_a = \lambda_a \mu E[n_r] \tag{7}$$

We note that because the advertiser has no cost, she will always participate in the partnership as long as the platform and the influencer are willing to.

We make the following two assumptions about model parameters. Assumption 1 posits that consumers with the highest valuation for quality (i.e., $\theta = 1$) derive positive utility from following an L-type influencer. Otherwise, the L-type cannot attract any informed followers, and the problem may degenerate. Assumption 2 ensures that the influencer has a high enough bargaining power so that she has incentives to participate in the revenue sharing.

ASSUMPTION 1. $q_L > c$

ASSUMPTION 2. $\lambda_i > \frac{p_f}{\mu + p_f}$

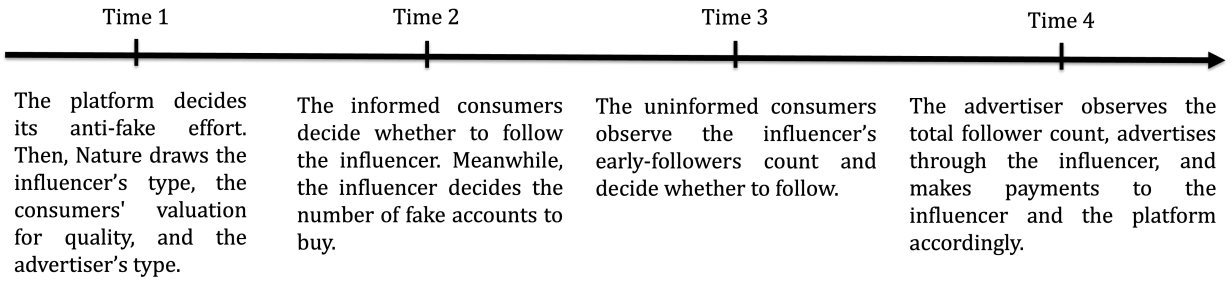


Figure 2 Game Timeline

Decision Variables	
d	The platform's anti-fake effort.
x	The number of fake accounts purchased by the influencer.
n_{in}, n_{un}	The number of informed, uninformed consumers that follow the influencer, respectively.
n_2, n	The number of early and total followers (including both real and fake accounts) of the influencer, respectively.
$n_{r,ia}$,$E[n_{r,ua}]$, $E[n_r]$	The expected number of real followers facing an informed, uninformed, and average advertiser, respectively.
Parameters	
c_o	A consumer's opportunity cost.
c_1	The nuisance cost coefficient.
l	The proportion of informed consumers and the probability of the advertiser being informed (also the level of social media literacy).
p_f, p_0	The unit price and base price of fake accounts, respectively.
q_H, q_L	The H-type and L-type influencer's content quality, respectively.
γ	The cost coefficient for the platform's anti-fake effort.
$\theta \in [0,1]$	Consumers' valuation for the influencer's content quality.
$\lambda_i, \lambda_p, \lambda_a$	The bargaining power of the influencer, the platform, and the advertiser, respectively.
μ	The revenue generated by advertising to a real consumer.
ρ	The probability of drawing an H-type influencer.
$\tau \in [0,1)$	The platform's anti-fake technology level.
$w \in [0,1]$	The platform's consumer orientation.
Outcome Variables	
u_i	Consumer i's expected utility.
U	Consumer welfare.
π_a, π_i, π_p	The expected profit of the advertiser, the influencer, and the platform, respectively.
Π_p	The platform's expected utility: a weighted sum of its profit and consumer welfare.

Table 1 Notation

3.6. Game Timeline

The timeline of the game is as follows. At **time 1**, the platform decides its anti-fake effort, d. Then, Nature draws the influencer's type, the consumers' valuation for quality θ, and the advertiser's type. At **time 2**, informed consumers decide whether to follow the influencer and the influencer decides the number of fake accounts x to buy. Now, the influencer has n_2 *early followers*, which include n_{in} informed consumers and x fake accounts. At **time 3**, uninformed consumers observe the influencer's early followers n_2 and decide whether to follow. After the uninformed consumers' decisions, the influencer has n total followers, which include n_2 early followers and n_{un} uninformed consumers. At **time 4**, the advertiser observes the total follower count n, advertises through the influencer, and makes payments to the influencer and the platform accordingly.

This game timeline captures a few key aspects of different stakeholders' decision environments and the main effects of their decisions. First, we assume informed consumers make their following

decisions before uninformed ones. This is because uninformed consumers rely on the influencer's follower count to infer her quality, and thus it is natural for them to wait for informed consumers' decisions. Second, we assume the influencer purchases fake accounts before uninformed consumers make their decisions. This is to ensure the influencer has a chance to influence their following decisions. Third, we assume the influencer's fake-account purchase and the informed consumers' following decisions occur simultaneously because these decisions are independent of each other, and the model would remain the same if they occur sequentially. Fourth, we assume the platform's anti-fake effort occurs before the influencer's fake-account purchase. This ensures the former can influence the price of fake accounts. It also captures the notion that for a fake account to work, it must survive the platform's anti-fake effort. Finally, we assume that the advertiser moves last because advertisers often begin to advertise with an influencer when she is popular enough, at which point the influencer has already accumulated followers.

4. Equilibrium Analysis

4.1. Preliminaries

Given the consumer utility (4), the number of informed followers for the t-type influencer is:

$$n_{in}^t = l\left(1-\frac{c}{q_t}\right),\ t\in\{H,L\}. \tag{8}$$

It is easy to see that the H-type has more informed followers than the L-type (i.e., $n_{in}^H > n_{in}^L$).

At time 3, the number of uninformed consumers following the t-type influencer is

$$n_{un}^t = (1-l)\left(1-\frac{c}{E\left[q|n_2^t\right]}\right),\ t\in\{H,L\}. \tag{9}$$

where n_2^t is the number of early followers for the t-type, $E\left[q|n_2^t\right] = \Pr\left(H|n_2^t\right)q_H + \Pr\left(L|n_2^t\right)q_L$ is the expected quality of the influencer conditional on the number of early followers n_2^t, and $\Pr\left(t|\cdot\right), t\in\{H,L\}$ is the conditional probability of the influencer being the t-type.

At time 4, an informed advertiser knows the influencer's type and can correctly anticipate the number of real followers, which is

$$n_{r,ia}^t = n_{in}^t + n_{un}^t,\ t\in\{H,L\} \tag{10}$$

An uninformed advertiser forms an expectation of the number of real followers:

$$E[n_{r,ua}] = \Pr\left(H|n\right)\left(n_{in}^H + n_{un}^H\right) + \Pr\left(L|n\right)\left(n_{in}^L + n_{un}^L\right) \tag{11}$$

where $\Pr\left(t|n\right)$ is the probability of the influencer being t-type, conditional on the number of total followers n. Therefore, from the perspective of an t-type influencer, the expected number of real followers that the advertiser would pay for is

$$E\left[n_r^t\right] = ln_{r,ia}^t + (1-l)\,E[n_{r,ua}] \tag{12}$$

4.2. Influencer's Equilibrium Decision

The game between the influencer, consumers, and the advertiser is a variation of the signaling game where the influencer attempts to signal her type to both uninformed consumers and advertiser. Because the number of total followers carries the same information as the number of early followers,[7] without loss of generality, we use the number of early followers as a signal for both uninformed consumers and advertiser.

A strategy profile of this game can be stated as (n_2^H, n_2^L), i.e., the number of early followers for the H- and L-type influencers, respectively. It can be equivalently stated as (x_H, x_L), i.e., the number of fake accounts purchased, given that $n_2^t = n_{in}^t + x_t$ $(t \in \{H, L\})$. Following the signaling game literature, we classify the equilibria as pooling and separating equilibria. If the number of early followers is the same for the two types of influencers, it is a *pooling equilibrium*; otherwise, it is a *separating equilibrium*. We further classify the separating equilibrium into two kinds: (a) a *naturally-separating* equilibrium where neither type of influencer purchases fake accounts and (b) a *costly-separating* equilibrium where at least one type purchases fake accounts. A similar distinction has been made by Guo et al. (2017) in the context of corporate social responsibility.

The signaling game tends to have multiple Perfect Bayesian Equilibria (PBEs). A common way to refine the equilibrium is the *undefeated equilibrium refinement* proposed by Mailath et al. (1993). One benefit of the undefeated refinement is to avoid the global consistency issue associated with the *intuitive criterion* (Mailath et al. 1993), another popular refinement strategy. Moreover, undefeated refinement also tends to select a unique PBE in signaling games. Therefore, we adopt the undefeated refinement in this research. [8]

In the following subsections, we first apply the undefeated refinement within each equilibrium type to obtain a *locally-undefeated* equilibrium and then apply it across equilibrium types to obtain the *globally-undefeated* equilibrium.

4.2.1. Pooling Equilibrium In the pooling equilibrium, H-type and L-type have the same number of early followers. We obtain a range of pooling PBEs with different levels of early followers. However, all higher-level pooling equilibria are defeated by the lowest-level one. The latter is the only locally-undefeated pooling equilibrium as stated in the following lemma.

[7] To see this, note that if an H-type influencer has more (the same, fewer) early followers, she will also have more (the same, fewer) total followers.

[8] In our context, an undefeated equilibrium requires that if an off-equilibrium action in the equilibrium is played by an t-type $(t \in \{H, L\})$ influencer in an alternative equilibrium and both types are better off (at least one type is strictly better off) in the alternative equilibrium, then the belief for the off-equilibrium action in the focal equilibrium should be identical to the belief for the same action under the alternative equilibrium. If the equilibrium fails to hold under such an off-equilibrium belief, we say the equilibrium is *defeated* by the alternative equilibrium.

LEMMA 1. *(Pooling) A strategy profile* $(n_2^H, n_2^L) = (n_{in}^H, n_{in}^H)$ *with the supporting beliefs*

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ \rho, & \text{if } n_2 = n_{in}^H \\ p \in [0, \overline{p}_1], & \text{if } n_{in}^H < n_2 < \overline{n}_2^{pool} \end{cases}$$

is the unique locally-undefeated pooling equilibrium if and only if:

$$d \leq (1-\tau)(\eta_1 - p_0) \equiv d_1 \tag{13}$$

where $\overline{p}_1 \in [\rho, 1]$ *and* $\overline{n}_2^{pool}$ *are given in Appendix A.1 and* $\eta_1 \equiv \lambda_i \mu (1-l) \left[\frac{q_H(E[q]-q_L)}{lE[q](q_H-q_L)} + \rho \right]$.

Lemma 1 describes an equilibrium where the L-type purchases fake accounts while the H-type does not. In this equilibrium, the L-type purchases fake accounts to mimic the H-type's in early followers so that they look identical to uninformed consumers and the uninformed advertiser. We call such purchasing *offensive purchasing*. Specifically, L-type's offensive purchasing x_L^{pool} is:

$$x_L^{pool} = n_{in}^H - n_{in}^L \tag{14}$$

The equilibrium belief holds that anyone who deviates to a lower follower count $n_2 < n_2^H$ (achievable only by an L-type) must be an L-type, whereas anyone who deviates to a higher follower count $n_2 > n_2^H$ is deemed as an H-type with probability $p \in [0, \overline{p}_1]$. The belief p is capped because an overly favorable belief would make deviations to higher follower counts profitable.

Condition (13) is derived from the L-type's incentive compatibility (IC) condition, ensuring that the L-type has no incentive to deviate. To understand this condition, we rewrite (13) as

$$p_f = p_0 + \frac{1}{1-\tau} d \leq \eta_1. \tag{15}$$

The left-hand side is the unit price of fake accounts, and the right-hand side can be interpreted as the revenue gain per fake account for the L-type to stay in the pooling equilibrium, compared with purchasing nothing and being treated as an L-type, her best alternative.

4.2.2. Separating Equilibrium We describe the costly-separating and naturally-separating equilibrium schemes in the next two lemmas.

LEMMA 2. *(costly-separating) A strategy profile* $(n_2^H, n_2^L) = (n_2^{csep}, n_{in}^L)$ *with the support beliefs*

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ p \in [0, \overline{p}_2) & \text{if } n_{in}^H \leq n_2 < n_2^{csep} \\ 1 & \text{if } n_2 = n_2^{csep} \\ p \in [0, 1], & \text{if } n_2^{csep} < n_2 < \overline{n}_2^{csep} \end{cases}$$

is the unique locally-undefeated costly-separating equilibrium if and only if:

$$d \leq (1-\tau)(\eta_2 - p_0) \equiv d_2 \tag{16}$$

where $\overline{n}_2^{csep}$ *is defined in Appendix A.2,* $\eta_2 \equiv \frac{1}{l}\lambda_i\mu\left(1-l^2\right)$*, and*

$$n_2^{csep} = \frac{1}{p_f}\lambda_i\mu\left[(1-l)\left(n_{in}^H - n_{in}^L\right) + \left(n_{un}^H - n_{un}^L\right)\right] + n_{in}^L \tag{17}$$

$$d_1 < d_2 \tag{18}$$

Lemma 2 describes an equilibrium where the H-type purchases fake accounts to deter the L-type from mimicking. We call such purchasing *defensive purchasing*. Condition (16) is derived from the IC condition for the L-type, which ensures that the L-type will mimic the H-type if the latter purchases nothing. To further understand this condition, we rewrite it as

$$p_f = p_0 + \frac{1}{1-\tau}d \leq \eta_2. \tag{19}$$

The left-hand side is the unit price of fake accounts, and the right-hand side η_2 can be interpreted as the revenue gain per fake account for the L-type to deviate from not buying to buying. By ensuring the L-type will deviate when the H-type buys nothing, we expect the H-type to purchase to maintain the separating equilibrium.

The H-type's equilibrium defensive purchase

$$x_H^{csep} = n_2^{csep} - n_{in}^H = \frac{1}{p_f}\left[\lambda_i\mu\left(1-l^2\right) - p_f l\right]\left(\frac{c}{q_L} - \frac{c}{q_H}\right) \tag{20}$$

is the difference between her number of informed followers and the highest follower count the L-type is willing to mimic. The term $\frac{c}{q_L} - \frac{c}{q_H}$ is the gap between the H-type and the L-type's real-follower counts under the separating equilibrium. An L-type's gain from mimicking a non-purchasing H-type is $\lambda_i\mu\left(1-l^2\right)$ times this gap, noting that $\lambda_i\mu\left(1-l^2\right)$ is the influencer's share of the advertising revenue per consumer. The L-type's cost of mimicking is $p_f l$ times the gap, noting that she only needs to fill the gap in the *informed* followers, which is l proportion of all real followers. x_H^{csep} is chosen such that L-type is break-even after having to purchase x_H^{csep} additional fake accounts.

The fact that H-type may purchase fake accounts in equilibrium is interesting in light of the literature's nearly exclusive focus on the L-type's deceptive behaviors (Piccolo et al. 2018, Chen and Papanastasiou 2021). In our study, the H-type may also buy fake accounts to deter the L-type's mimicry, even doing so will not fool the advertiser. Lemma 2 highlights that the H-type influencer may buy fake accounts not to deceive but to establish their superiority. This finding is consistent with the casual observation that high-status influencers frequently purchase fake accounts, as seen in the Devumi case (Mekuli 2021).

Lemma 3. *(naturally-separating) A strategy profile* $(n_2^H, n_2^L) = (n_{in}^H, n_{in}^L)$ *with the support beliefs*

$$P\left(H|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ 1, & \text{if } n_2 = n_{in}^H \\ p \in [0, \overline{p}_4], & \text{if } n_2 > n_{in}^H \end{cases}$$

is the unique locally-undefeated naturally-separating equilibrium if and only if:

$$d > (1-\tau)(\eta_2 - p_0) \equiv d_2 \tag{21}$$

where $\overline{p}_4$ is defined in Appendix A.3.

Lemma 3 states that the two types will separate naturally when condition (21) holds, meaning the highest early-follower count that the L-type can match is smaller than the H-type's informed followers. In other words, the L-type cannot afford to mimic a non-purchasing H-type.

When the base price of fake accounts p_0 is higher than η_2, condition (21) holds regardless of d – *natural separation occurs even if anti-fake effort is zero.* This extreme case seems unrealistic given the prevalence of fake accounts. To rule out such an extreme case, we assume:

ASSUMPTION 3. $p_0 < \eta_2$

4.2.3. Globally Undefeated Equilibrium We note that the pooling, costly-separating, and naturally-separating equilibria may co-exist. The following lemma summarizes the defeating relationship among these equilibria.

LEMMA 4. *Given $d \le d_1$ (so the pooling equilibrium exists), (a) the pooling equilibrium coexists with and defeats the costly-separating equilibrium, (b) the pooling equilibrium cannot coexist with the naturally-separating equilibrium.*

Lemma 4 suggests that the pooling equilibrium may coexist with the costly-separating equilibrium but always defeats it. The next proposition describes the globally-undefeated equilibrium.

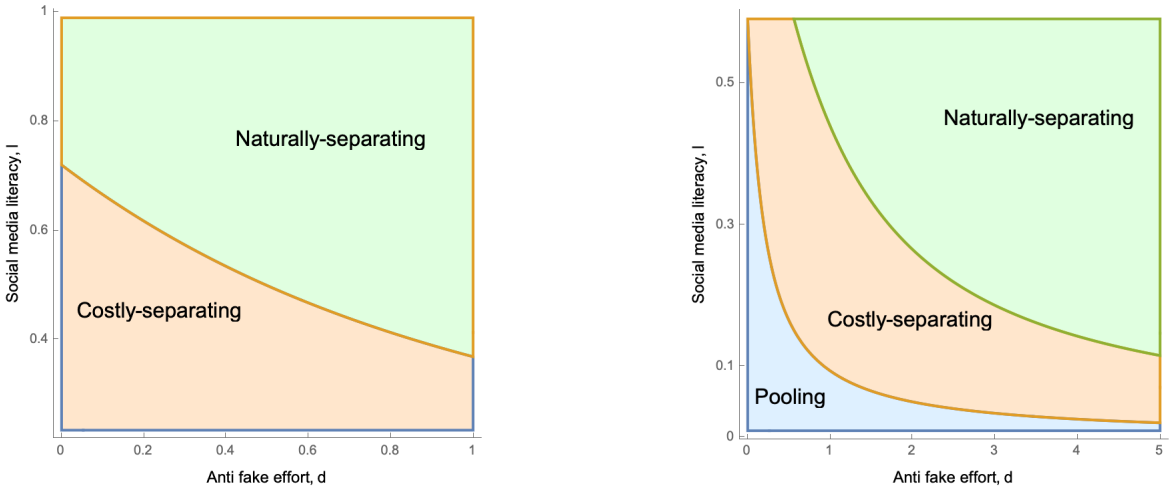
Proposition 1 *a.* **(No-pool)** *If $d_1 \le 0$, the unique globally-undefeated equilibrium is:*

$$\begin{cases} (n_2^{csep}, n_{in}^L), & if\ d \le d_2 \quad \text{(costly-separating)} \\ (n_{in}^H, n_{in}^L), & otherwise \quad \text{(naturally-separating)} \end{cases}$$

b. **(All-eqm)** *If $d_1 > 0$, the unique globally-undefeated equilibrium is:*

$$\begin{cases} (n_{in}^H, n_{in}^H), & if\ d \le d_1 \quad \text{(pooling)} \\ (n_2^{csep}, n_{in}^L), & if\ d_1 < d \le d_2 \quad \text{(costly-separating)} \\ (n_{in}^H, n_{in}^L), & if\ d > d_2 \quad \text{(naturally-separating)} \end{cases}$$

Proposition 1 suggests that the pooling equilibrium cannot exist when the fake-account base price p_0 is relatively high (such that $d_1 \le 0$) (Case **No-pool**). In such a case, when the anti-fake effort d is relatively low, a costly-separating equilibrium prevails; otherwise, a naturally-separating equilibrium would prevail. Figure 3(a) illustrates such a case: with a relatively high fake-account base price p_0, as the anti-fake effort increases, the equilibrium regime transitions from costly-separating to naturally-separating. When the fake-account base price p_0 is relatively low, such that



(a) No-pool ($p_0 = 2$) (b) All-eqm ($p_0 = 0.5$)

Note: $\mu = 10$, $\lambda_i = 0.3$, $\lambda_p = 0.3$, $\rho = 0.1$, $q_H = 20$, $q_L = 10$ $c_0 = 1$, $c_1 = 0.02$,$\tau = 0.8$

Figure 3 Illustration of the Equilibrium Regime Transitions

$d_1 > 0$, Figure 3(b) illustrates such a case (Case **All-eqm**): the equilibrium transitions from pooling to costly-separating, and then to naturally-separating as the anti-fake effort increases.

Proposition 1 and the preceding lemmas suggest that the influencer's equilibrium fake-account purchasing behavior is discontinuous. Specifically, as the equilibrium transitions from pooling to costly separating, the L-type's offensive purchasing first increases and then suddenly drops to zero. The H-type's defensive purchasing first stays at zero and then suddenly jumps to a very high level. Such "rugged" equilibrium behaviors are further illustrated in the next subsection (e.g., Figure 4).

4.3. Comparative Statics

Next, we conduct a set of comparative statistics on how the equilibrium fake-account purchasing changes with the underlying parameters under different equilibrium regimes.

Proposition 2 *Under the pooling equilibrium, the L-type's offensive purchase* x_L^{pool}

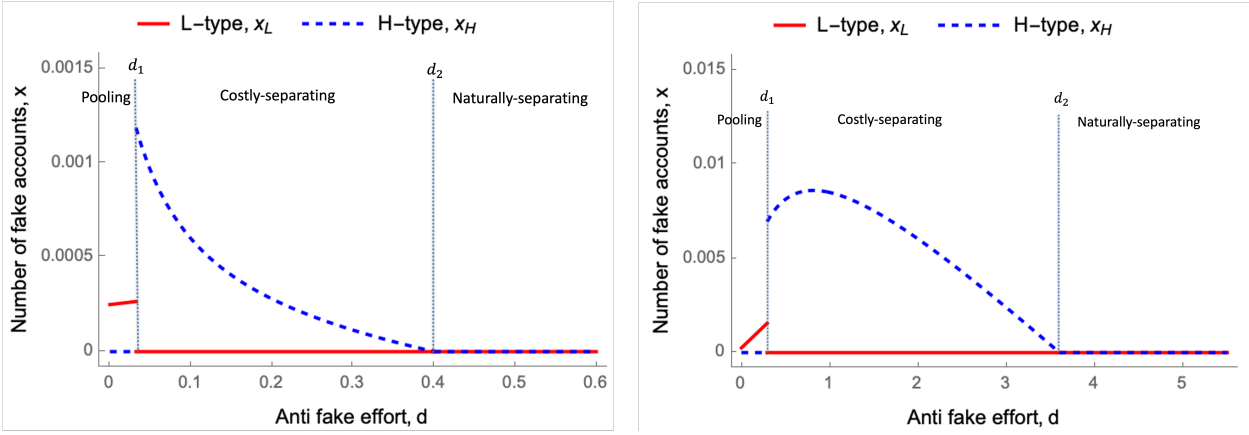
a. increases in the platform's anti-fake effort d and the social media literacy l,

b. decreases in the anti-fake technology level τ *and quality ratio R, and*

c. is unaffected by the fake-account base price p_0 *or the influencer's bargaining power* λ_i*.*

The intuition for Proposition 2 is as follows. As the anti-fake effort increases or the technology level decreases, the consumer nuisance cost increases, causing more informed consumers to drop out. The L-type loses informed consumers more quickly than the H-type,[9] implying the gap between the two types' informed followers increases. Consequently, the L-type must buy more fake accounts to

[9] This is because informed followers have lower valuations for the L-type's content and thus are more likely to drop out when the influencer is the L-type (see Equation (8)).



(a) High anti-fake technology ($\tau = 0.9$) (b) Low anti-fake technology ($\tau = 0.1$)

Note: $\mu = 10, \rho = 0.12, l = 0.5, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 0.01, c_1 = 0.2, p_0 = 0.5$

Figure 4 Impact of anti-fake effort on the number of fake accounts

stay in the pooling equilibrium. This is illustrated in Figure 4 (the pooling regions). The intuition for the effect of social media literacy l is similar: increasing social media literacy leads to more informed followers for both types of influencers and also enlarges the gap between the two types' informed followers, which forces the L-type to buy more to stay in the equilibrium. Increasing the anti-fake technology level decreases the consumers' nuisance cost and shrinks the gap between the two types' informed followers. Increasing the quality ratio R reduces the quality gap between the two types, as a result, the gap between the two types' informed followers. Consequently, the L-type needs fewer fake accounts to stay in the equilibrium. Finally, the number of fake accounts and the influencer's share of advertising revenue doesn't affect the gap between the two types' informed followers. Therefore, the L-type's purchase is unaffected by the fake-account base price or the influencer's bargaining power.

Proposition 3 *Under the costly-separating equilibrium, the H-type's defensive purchase x_H^{csep}*

a. decreases in social media literacy l, fake-account base price p_0, quality ratio R, and anti-fake technology level τ,

b. increases in the influencer's bargaining power λ_i, and

c. decreases in d if the following condition holds; otherwise, first increases then decreases in d.

$$c_1 (1-\tau)^2 \leq \frac{\lambda_i \mu (1-l^2) c_0}{p_0 [\lambda_i \mu (1-l^2) - p_0 l]} \tag{22}$$

The intuition for Proposition 3 is as follows. As seen from Equation (20), H-type's defensive purchasing is the highest number of informed followers the L-type is willing to mimic. Furthermore, x_H^{csep} increases in the L-type's net gain from mimicry $[\lambda_i \mu (1-l^2) - p_f l] \left(\frac{c}{q_L} - \frac{c}{q_H} \right)$ and decreases in the price of fake accounts p_f. The former is determined further by the gap between the L-

and H-type's followers $\left(\frac{c}{q_L}-\frac{c}{q_H}\right)$ and the L-type's marginal gain per follower from mimicry $[\lambda_i\mu(1-l^2)-p_f l]$. Increasing social media literacy l decreases the marginal gain from mimicry (as there will be fewer uninformed consumers), and causes defensive purchasing to decrease. Increasing the fake-account base price p_0 decreases L-type's marginal gain from mimicry and increases the fake-account price. Both effects lead to decreased defensive purchasing. Increasing the quality ratio q_L/q_H reduces the gap between the two types' followers and L-type's marginal gain from mimicry, and thus decreases the H-type's purchasing. Increasing the anti-fake technology level τ decreases the consumers' nuisance cost c, and thus the gap between the two type's followers; it also increases the price of fake accounts p_f. Both reduced follower gap and increased price lead to decreased defensive purchasing. Increasing the influencer's bargaining power λ_i increases the L-type's marginal gain per follower, and thus the H-type's defensive purchasing.

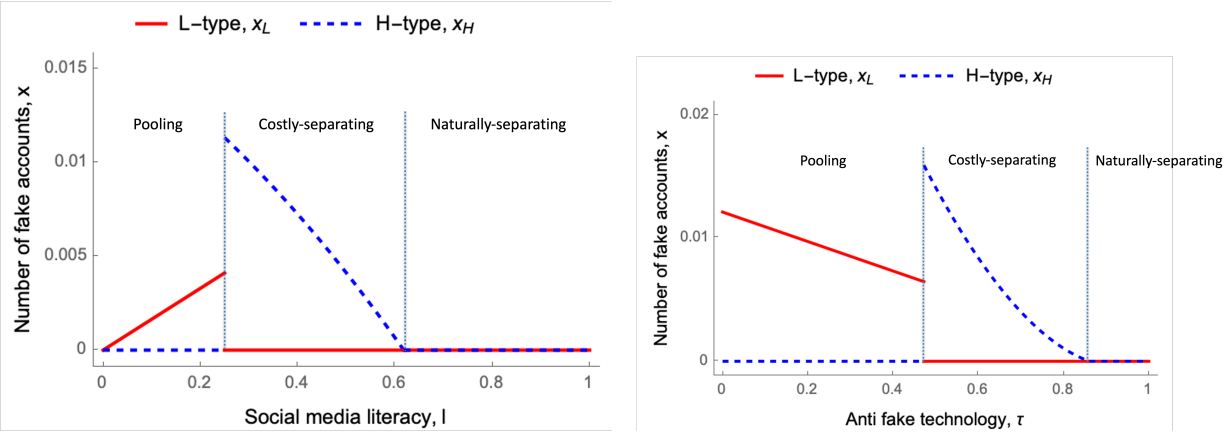
Increasing the anti-fake effort d can produce countervailing effects. As d increases, both the unit price of fake accounts p_f and the consumers' nuisance cost c increase. The former (the "*higher-fake-account-price*" effect) reduces defensive purchasing, with a negative marginal effect of $m_1=-\frac{1}{1-\tau}\frac{\lambda_i\mu\left(1-l^2\right)}{p_f^2}\frac{q_H-q_L}{q_Hq_L}c$. The latter (the "*higher-nuisance-cost*" effect) increases the L-type's marginal gain from mimicry and thus increases defensive purchasing, with a positive marginal effect of $m_2=c_1(1-\tau)\frac{\lambda_i\mu\left(1-l^2\right)-p_f l}{p_f}\frac{q_H-q_L}{q_Hq_L}$. As d increases, fake-account price p_f increases, and the higher-nuisance-cost effect $m_2\to 0$. Therefore, as d increases, the negative effect m_1 will dominate eventually, causing defensive purchasing to decrease. When the technology level τ is relatively high such that (22) holds, the negative effect m_1 always dominates the positive effect m_2, even for a low anti-fake effort. In such a case, the defense purchasing monotonically decreases, as illustrated in Figure 4(a) (the "costly-separating" region). Otherwise, the positive effect dominates for low d but not for high d, causing the defensive purchasing to first increase and then decrease, as illustrated in Figure 4 (b) (the "costly-separating" region).[10]

Figure 5 illustrates the impact of social media literacy and technology level. In panel (a), the L-type's offensive purchasing (in the pooling region) increases in social media literacy, as predicted by Proposition 2, whereas the H-type's defensive purchasing (costly-separating) decreases, as predicted by Proposition 3. Panel (b) shows that increasing the anti-fake technology level can reduce both offensive and defensive fake-account purchasing in their respective equilibrium regions.

5. Anti-Fake Effort and Consumer Welfare

We now turn our attention to the relationship between the platform's anti-fake effort and consumer welfare. We let $U^{pool}(d), U^{csep}(d)$, and $U^{nsep}(d)$ denote consumer welfare under each equilibrium

[10] As we normalize the number of consumers to a unit mass, the number of fake accounts in the figures (i.e., shown on the vertical axis) should be interpreted relatively. For instance, if the number of fake accounts is 0.03, we infer that there are 30 fake accounts per 1,000 consumers.



(a) Impact of social media literacy ($\tau = 0.5, d = 1$) (b) Impact of anti-fake technology level ($l = 0.1, d = 4$)

Note: $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 5, c_0 = 0.01, c_1 = 0.2, p_0 = 1$

Figure 5 Impact of parameters on the equilibrium number of fake-accounts

Equilibrium type	Consumers prefer	Consumer-optimal effort d^C	Condition
No-pool	costly-separating	0	-
All-eqm	pooling	0	$U^{pool}(0) \geq U^{csep}(d_1)$
	costly-separating	d_1	$U^{pool}(0) < U^{csep}(d_1)$

Table 2 Consumer-optimal anti-fake effort and equilibrium type

type, provided that the effort d sustains the equilibrium. The following lemma establishes the effect of anti-fake efforts on consumer welfare:

LEMMA 5. *$U^{pool}(d), U^{csep}(d)$, and $U^{nsep}(d)$ monotonically decrease with anti-fake effort d.*

Intuitively, under each equilibrium type, increasing the anti-fake effort would not impact consumers' inference of influencer quality but increase their nuisance costs, which is welfare-reducing. Lemma 5 implies that consumers always prefer a costly-separating equilibrium to a naturally-separating one because the two equilibria provide the same information about influencer quality but the latter imposes a higher nuisance cost.

Lemma 5 does not imply, however, that consumers would always prefer zero anti-fake effort. To see this, we note that, holding the nuisance cost constant, informed consumers are indifferent between equilibrium types but uninformed ones prefer a separating equilibrium because it leads to better following decisions. When the benefit of separation outweighs the increase in nuisance costs, consumers, as a whole, may prefer a separating equilibrium.

LEMMA 6. *Consumers' preference for equilibrium types and anti-fake effort d^C is summarized in Table 2.*

As seen from Table 2, when the pooling equilibrium is unavailable (*No-pool*), consumers prefer the costly-separating equilibrium with zero anti-fake effort. When all three equilibria are available

Equilibrium type	Platform prefers	Platform-optimal effort d^*	Condition
No-pool	costly-separating	0	-
All-eqm	pooling	0	$\Pi_p^{pool}(0) \geq \Pi_p^{csep}(d_1)$
	costly-separating	d_1	$\Pi_p^{pool}(0) < \Pi_p^{csep}(d_1)$

Table 3 Platform-optimal anti-fake effort and equilibrium type

(*All-eqm*), they may prefer the pooling with zero anti-fake effort or the costly-separating equilibrium with minimum anti-fake effort d_1. The conditions for different cases are provided in the proof.

6. The Platform's Optimal Strategy

Having examined consumers' preferred anti-fake effort, we now turn to the platform's.

6.1. The Platform's Optimal Anti-Fake Effort

We denote $\pi_p^{pool}(d), \pi_p^{csep}(d)$, and $\pi_p^{nsep}(d)$ as the platform's profit under the pooling, costly-separating, and naturally-separating equilibrium, respectively, provided that the chosen d sustains the equilibrium. Lemma 7 summarizes how the platform profit changes with the anti-fake effort.

LEMMA 7. *$\pi_p^{pool}(d)$, $\pi_p^{csep}(d)$, and $\pi_p^{nsep}(d)$ decrease in d.*

The intuition behind Lemma 7 is as follows. In general, *within each equilibrium type*, increasing the anti-fake effort affects the platform's profit in two ways: First, it increases consumer nuisance costs, reducing the number of participating consumers. This *consumer-inconvenience* effect negatively affects the total advertising revenue and, thus, the platform's revenue share. Second, the platform incurs a higher cost for its anti-fake effort. Both effects reduce the platform's profitability.

LEMMA 8. *a.* **(No-pool)** *If $d_1 \leq 0$, $\pi_p^{csep}(0) > \pi_p^{nsep}(d_2)$*

b. **(All-eqm)** *If $d_1 > 0$, (1) $\pi_p^{pool}(0) > \pi_p^{csep}(d_1)$ and (2) $\pi_p^{csep}(d_1) > \pi_p^{nsep}(d_2)$.*

To see Lemma 8 (b.1), we first note that the number of uninformed followers is a concave function of the expected quality of the influencer. Consequently, the number of uninformed followers under the pooling equilibrium – for an influencer with mean quality – is higher than the expected number of uninformed followers under a separating equilibrium – which is the average number of followers between an H- and an L-type influencers. Therefore, the advertising revenue (and the platform's share) is higher under a pooling equilibrium. This, together with the observation that the platform incurs an anti-fake cost under the separating equilibrium, suggests that the platform's profit is higher under the pooling equilibrium than under the separating one.

To see Lemma 8 (a) and (b.2), we note that the two separating equilibria provide the same information about influencer quality but the naturally-separating equilibrium mandates a higher anti-fake effort, which results in fewer participating consumers and a lower revenue.

Parameters	Impact on consumer welfare		
	(1) Pooling with $d^* = 0$	(2) Costly-separating with $d^* = 0$	(3) Costly-separating with $d^* > 0$
Social media literacy	$\uparrow$	$=$	$\uparrow$
Fake-account base price	$=$	$=$	$\uparrow$
Anti-fake technology level	$=$	$=$	$\uparrow$

Table 4 Comparative statistics for consumer welfare with platform-optimal anti-fake effort

Proposition 4 *The platform's optimal anti-fake effort and the associated equilibrium type are summarized in Table 3. Specifically, (a) a purely profit-driven platform (i.e., $w = 0$) optimally chooses zero anti-fake effort, whereas (b) a consumer-oriented platform ($w > 0$) may optimally choose a positive anti-fake effort to induce the costly-separating equilibrium.*

Proposition 4 (a) suggests that a purely profit-driven platform always chooses zero anti-fake effort. This is because, according to Lemma 8, its profit is maximized under the pooling equilibrium which requires no anti-fake effort. Proposition 4 (b) suggests a more consumer-oriented platform may prefer a separating equilibrium. This is because such an equilibrium may yield higher consumer welfare, which the platform values.

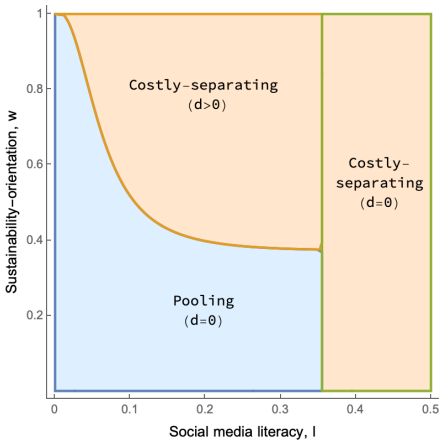
6.2. Comparative Statics

In this subsection, we examine the impact of three model parameters: social media literacy, fake-account base price, and anti-fake technology level. We focus on these parameters because they hold strong managerial implications.

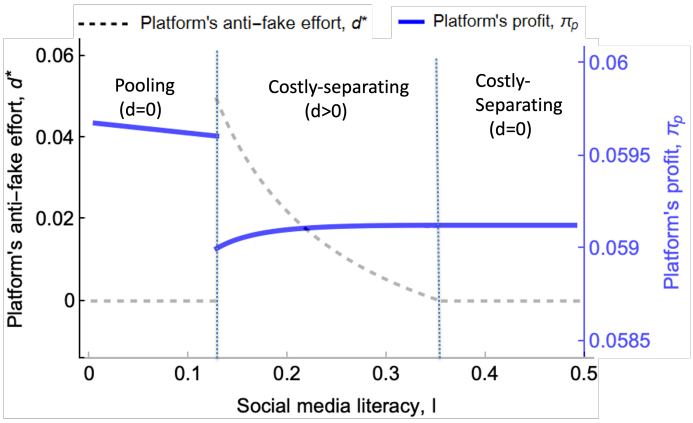
6.2.1. Comparative Statics on Consumer Welfare

Proposition 5 *The impact of parameters on consumer welfare is given in Table 4. Specifically, increasing social media literacy, fake-account base price, and anti-fake technology level weakly improve consumer welfare.*

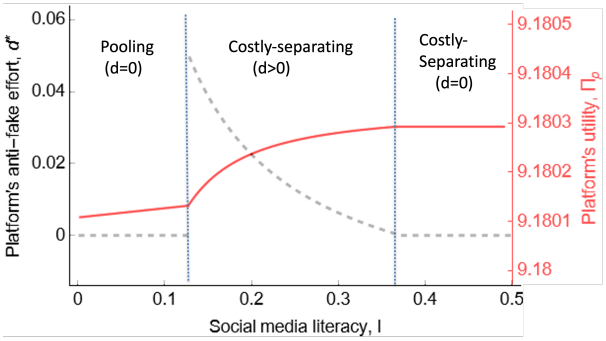
Consumer welfare loss comes from two sources: the nuisance cost of anti-fake efforts and sub-optimal following decisions by uninformed consumers under the pooling equilibrium (or *cost of pooling*). Under the pooling equilibrium (Table 4, case 1), increasing social media literacy will reduce the cost of pooling because there are fewer uninformed consumers. Increasing fake-account base price and anti-fake technology does not affect the cost of pooling or consumer welfare. Under the costly-separating equilibrium with zero anti-fake effort (case 2), as the influencers are already separated and there is no nuisance cost from the anti-fake effort, neither parameter has any impact on consumer welfare. Finally, under the costly-separating equilibrium with positive anti-fake effort (case 3), there is no cost of pooling, but the nuisance cost from the anti-fake effort decreases in



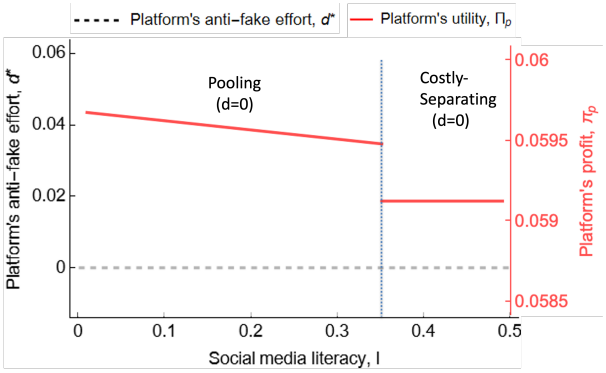
(a) Platform's optimal strategy



(b) Platform's optimal effort and profit ($w = 0.5$)



(c) Platform's optimal effort and utility ($w = 0.5$)



(d) Platform's optimal effort and utility ($w = 0.1$)

Note: $\mu = 0.2, \rho = 0.3, \lambda_i = 0.3, \lambda_p = 0.3, p_0 = 0.1, q_H = 100, q_L = 10, c_0 = 0.2, c_1 = 0.02, \tau = 0.8, \gamma = 0.1$

Figure 6 Impact of social media literacy on the platform

social media literacy, fake-account base price, and anti-fake technology; thus, consumer welfare increases in these parameters.

Next, we examine how each parameter affects the platform's anti-fake effort, profit, and utility. We rely on numeric methods for these analyses because the boundary conditions outlined by Proposition 4 are not analytically tractable. We note that while the patterns we report in the next subsections are illustrated using a specific set of model parameters, they are fairly representative when we conduct robustness checks by systematically varying model parameters (available upon request).

6.2.2. Social Media Literacy Figure 6 illustrates the effect of social media literacy (l) on the platform's optimal effort, profit, and utility. Panel (a) shows that when l is moderate, a more consumer-oriented platform may optimally induce a costly-separating equilibrium with positive anti-fake effort, whereas a more profit-focused platform may prefer a pooling equilibrium with zero anti-fake effort. The higher the l is, the lower the minimum consumer orientation required for the

separating equilibrium, indicating that the platform is more likely to exert anti-fake effort. When l is low enough, consumer welfare is higher under the pooling equilibrium so that even the most consumer-oriented platform would prefer the pooling equilibrium. When l reaches a high level ($\geq$ 0.355 in this example), the pooling equilibrium no longer exists, and the platform optimally induces a costly-separating equilibrium with zero anti-fake effort, *regardless of its consumer orientation*.

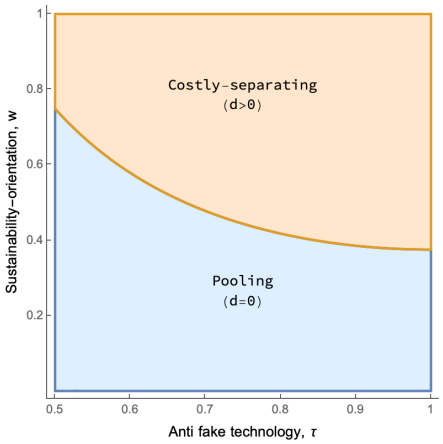
Figure 6(b) further confirms that the platform's anti-fake effort may not be monotonic in social media literacy; a consumer-oriented platform may prefer zero anti-fake effort when social media literacy is low or very high. Furthermore, when the platform does exert a positive anti-fake effort, its effort level decreases as social media literacy increases, suggesting social media literacy education can substitute for the platform's anti-fake effort. The figure also shows that the platform's profit is the highest with zero social media literacy.[11]

Panels (c) and (d) show that the platform's preferred social media literacy depends on its consumer orientation. Recalling that consumer welfare increases with social media literacy, when the platform is highly consumer-oriented, it has sufficient concern for consumer welfare that it may benefit from a higher social media literacy, as shown in Panel (c). Conversely, a more profit-focused platform may prefer a lower social media literacy, as illustrated in Panel (d).

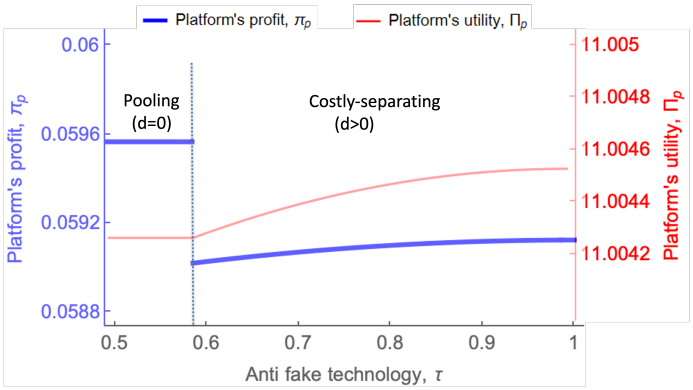
6.2.3. Fake-account Base Price Our findings about the effect of fake-account base price are similar to those about social media literacy (See Figure 8 of Appendix A.15). In particular, (1) when the fake-account base price is relatively low, a more profit-driven platform would not invest in any anti-fake efforts but a more consumer-oriented platform may. (2) When the fake-account base price is high enough, pooling equilibrium no longer exists, and the platform optimally chooses a costly-separating equilibrium with zero anti-fake effort, *regardless of its consumer orientation*. (3) The platform may prefer a lower fake-account base price when it is more profit-focused; the reverse may be true when it is sufficiently consumer-oriented.

6.2.4. Anti-Fake Technology Level Figure 7 illustrates the effect of the anti-fake technology level. Panel (a) shows that as the technology level increases, the minimum consumer orientation to sustain a separating equilibrium decreases, suggesting that better anti-fake technology makes anti-fake efforts more likely. Panel (b) also shows that the platform is unresponsive to changes in anti-fake technology under a pooling equilibrium, regardless of its consumer orientation. When the technology level is high enough, a consumer-oriented platform may prefer a costly-separating equilibrium. In such a case, further improvements in the anti-fake technology would increase the platform's utility and profitability and reduce its anti-fake effort (not plotted). Overall, the platform weakly benefits from better anti-fake technology.

[11] This finding is robust with different model parameters, including different values of consumer orientation w.



(a) Platform's optimal strategy



(b) Platform's profit and utility ($w = 0.6$)

We let $\mu = 0.2, \rho = 0.3, \lambda_i = 0.3, \lambda_p = 0.3, l = 0.2, p_0 = 0.1, q_H = 100, q_L = 10, c_0 = 0.2, c_1 = 0.02, \tau = 0.8, \gamma = 0.1$

Figure 7 Impact of anti-fake technology on the platform's optimal anti-fake effort and profit

		Platform optima		Consumer optima		
Scenario		**Eqm**	d^*	**Eqm**	d^C	**Conditions**
No-pool	1	csep	0	csep	0	-
All-eqm	2	pool	0	pool	0	$U^{csep}(d_1) \leq U^{pool}(0)$
	3	pool	0	csep	d_1	$U^{csep}(d_1) > U^{pool}(0)$ & $\Pi_p^{pool}(0) \geq \Pi_p^{csep}(d_1)$
	4	csep	d_1	csep	d_1	$U^{csep}(d_1) > U^{pool}(0)$ & $\Pi_p^{pool}(0) < \Pi_p^{csep}(d_1)$

Table 5 Consumer-optimal anti-fake effort d^C versus platform-optimal d^*

6.3. Comparing Platform and Consumer Optima

Proposition 6 *The relationship between the platform- and consumer optima is given in Table 5. In general, the platform exerts weakly less anti-fake effort than what is optimal for consumers.*

As seen from Table 5, in some cases (1, 2, and 4), the platform's optimal anti-fake effort is the same as the consumer's. Among these, case 4 exists only when the platform is sufficiently consumer-oriented. In case 3, the platform's optimal anti-fake effort is less than the consumers'. Such a case is more likely when the platform is more profit-driven. In such a case, consumers prefer a separating equilibrium with positive anti-fake effort, whereas the platform prefers a pooling equilibrium with zero anti-fake effort.

7. Extensions

7.1. Three Types of Influencers

Our main model assumes only two types of influencers. One may wonder whether our insights are generalizable to more influencer types. To address this issue, we extend our model to three influencer types. Specifically, we assume an influencer's quality is drawn randomly from three levels, $q_L < q_M < q_H$, with probabilities ρ_L, ρ_M, and ρ_H $(\rho_H = 1 - \rho_M - \rho_L)$, respectively. The

following proposition summarizes the equilibrium types and fake-account purchasing patterns in this extended model.

Proposition 7 *With three influencer types, the unique undefeated equilibrium is given in Table 6 (the corresponding beliefs and conditions are provided in Appendix B).*

Equilibrium Type	**Notation**	H**-Type**	M**-Type**	L**-Type**
Costly Fully Separating	$H^*\|M^*\|L$	$\frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H-n_r^L\right)}{p_f}+n_{in}^L$	$\frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M-n_r^L\right)}{p_f}+n_{in}^L$	n_{in}^L
Naturally Fully Separating	$H\|M\|L$	n_{in}^H	n_{in}^M	n_{in}^L
Fully Pooling	HM^*L^*	n_{in}^H	n_{in}^H	n_{in}^H
Hybrid with M-L pooling	$H\|ML^*$	n_{in}^H	n_{in}^M	n_{in}^M
Hybrid with H-M pooling	$HM^*\|L$	n_{in}^H	n_{in}^H	n_{in}^L

Note: $*$ denotes fake-account purchasing and "|" denotes separation between the two adjacent influencer types.

Table 6 Undefeated Equilibrium with Three Types of Influencers

As seen from Table 6, the three types of equilibria – pooling, costly separating, and naturally separating – are preserved and become the *fully pooling*, *costly fully separating*, and *naturally separating* equilibria, respectively. We also obtain two new "hybrid" equilibrium types where two types pool and separate from the third type: *hybrid with H-M pooling* and *hybrid with M-L pooling*. Similar to the main model, an influencer may purchase fake accounts defensively (e.g., H- and M-types purchase fake accounts under the costly fully separating equilibrium) or offensively (e.g., M- and L-types under the fully pooling equilibrium) in this extended model. Different from the main model, the M-type's purchase may be simultaneously defensive and offensive (e.g., Hybrid with H-M pooling). In addition, under two equilibrium scenarios (i.e., costly fully separating and fully pooling), two of three influencer types purchase fake accounts. This suggests that the extended model can capture the case where fake-account purchasing is prevalent and occurs at multiple levels of influencer quality.

7.2. A Repeated Game

Another potential concern is whether the insights from a one-shot game could be generalized to a repeated setting. When the game is repeated, some existing consumers may remain, and new consumers may join. One may expect that, with repetition, consumers, on average, could become more informed – they may learn the influencer's quality from prior experience with the influencer or from prior revelation of the influencer type (by a separate equilibrium); they may also learn from prior consumers through word of mouth. On the other hand, information asymmetry may

linger because existing uninformed consumers may remain, and new uninformed consumers may join. We similarly expect that the information asymmetry encountered by advertisers will gradually diminish but not disappear immediately. In the following, we use a simplified two-period game to examine how repetition may impact equilibrium behavior and whether the insights from the one-shot game can be applied in the repeated setting.

We consider a repeated game where our original game is repeated for another period, with the influencer type persistent across the two periods. For simplicity, we assume that consumers, fake accounts, and the advertiser are short-lived (i.e., they live for only one period), and the size of consumers stays constant across two periods. To capture the notion that information asymmetry may decrease but will not disappear completely, we assume that the proportion of informed consumers in period 2, l_2, is higher than in period 1 (l) and is a function of the equilibrium type in period 1:

$$l_2 = \begin{cases} l + \delta_{sep} \equiv l_2^{sep}, & \text{if period-1 equilibrium is separating} \\ l + \delta_{pool} \equiv l_2^{pool}, & \text{if period-1 equilibrium is pooling} \end{cases}, \text{ where } \delta_{sep} > \delta_{pool} > 0$$

We interpret consumers in period 2 as a combination of remaining consumers from period 1 and newly joined consumers in period 2. We interpret fake accounts in period 2 as a sum of renewed existing fake accounts and newly purchased fake accounts (if any).

For simplicity, we assume the probability of drawing an informed advertiser in each period is the same as the proportion of informed consumers (relaxing this does not fundamentally change our result). Thus, the advertiser also becomes more informed but may not be fully informed in period 2.

We assume no discounting and that the platform's anti-fake effort remains the same across the two periods – that is, the platform chooses its effort once at the beginning of period 1. For simplicity, we also assume that the influencer only considers the short-term impact of her current-period deviation: i.e., she does not consider any impact of her current-period deviation on the future period. The following proposition describes the equilibrium types across the two periods.

Proposition 8 *a.* **(No-pool in period 1)** *If $d_1 \leq 0$, the equilibrium types in two periods are:*

$$\begin{cases} (\text{costly-separating}, \text{costly-separating}), & if\ d \leq d_2'\left(l_2^{sep}\right) \\ (\text{costly-separating}, \text{naturally-separating}), & if d_2'\left(l_2^{sep}\right) < d \leq d_2 \\ (\text{naturally-separating}, \text{naturally-separating}), & if\ d > d_2 \end{cases}$$

b. **(All-eqm in period 1)** *If $0 < d_1 \leq d_2'\left(l_2^{pool}\right)$, the equilibrium types in two periods are:*

$$\begin{cases} (\text{pooling}, \text{pooling}), & if\ d \leq d_1'\left(l_2^{pool}\right) \\ (\text{pooling}, \text{costly-separating}), & if\ d_1'\left(l_2^{pool}\right) < d \leq d_1 \\ (\text{costly-separating}, \text{costly-separating}), & if\ d_1 < d \leq d_2'\left(l_2^{sep}\right) \\ (\text{costly-separating}, \text{naturally-separating}), & if\ d_2'\left(l_2^{sep}\right) < d \leq d_2 \\ (\text{naturally-separating}, \text{naturally-separating}), & if\ d > d_2 \end{cases}$$

c. **(All-eqm in period-1)** *If* $d_1 > d_2'\left(l_2^{pool}\right)$*, the equilibrium types in two periods are:*

$$\begin{cases} (\text{pooling}, \text{pooling}), & if\ d \le d_1'\left(l_2^{pool}\right) \\ (\text{pooling}, \text{costly-separating}), & if\ d_1'\left(l_2^{pool}\right) < d \le d_2'\left(l_2^{pool}\right) \\ (\text{pooling}, \text{naturally-separating}), & if\ d_2'\left(l_2^{pool}\right) < d \le d_1 \\ (\text{costly-separating}, \text{costly-separating}), & if\ d_1 < d \le d_2'\left(l_2^{sep}\right) \\ (\text{costly-separating}, \text{naturally-separating}), & if\ d_2'\left(l_2^{sep}\right) < d \le d_2 \\ (\text{naturally-separating}, \text{naturally-separating}), & if\ d > d_2 \end{cases}$$

where $d_1'(\cdot)$ *and* $d_2'(\cdot)$ *are defined similarly as* d_1 *and* d_2 *(See Appendix C for details).*

Proposition 8 shows that, given the platform's anti-fake effort in period 1, the equilibrium may stay the same in period 2 or transition to a more "advanced" equilibrium (i.e., pooling → costly-separating, pooling → naturally-separating, or costly-separating → naturally-separating). Therefore, the separating (pooling) equilibrium could become more (less) prevalent in the long run, which is a result of reduced information asymmetry.

An influencer may purchase fake accounts in both periods (e.g., the L-type under (*pooling, pooling*)), just one (e.g., the L-type under (*pooling, costly-separating*)), or none (e.g., the L-type under (*costly-separating, costly-separating*)). In the case of (*pooling, pooling*), the L-type would purchase more fake accounts in period 2 according to Proposition 2, noting that $l_2 > l$. In the case of (*costly-separating, costly-separating*), the H-type would purchase fewer in period 2 by Proposition 3. In general, our one-short game corresponds to a stage game in the repeated setting; consequently, our earlier insights into different equilibrium types can be leveraged to understand how the system will evolve in the repeated setting.

8. Discussion and Conclusion

Motivated by the prevalence of social media fake accounts in the influencer economy and a lack of understanding of this phenomenon, we study a fake account model in which influencers can purchase fake accounts to make them appear more popular to consumers and advertisers, whereas the social platform can mount an anti-fake effort that increases the cost of fake accounts while also increasing the nuisance cost of consumers. We use this model to study the influencer's equilibrium fake-account purchasing behavior, the platform's optimal anti-fake effort, consumer welfare, and how the ecosystem responds to changes in several model parameters.

8.1. Contribution to the literature

Our paper contributes to the literature in three main ways. First, we contribute to the understanding of fake account purchasing behaviors in the influencer economy. We find that not only low-quality influencers may purchase fake accounts to mimic high-quality influencers, but high-quality influencers may also purchase fake accounts to separate themselves. The latter type of equilibrium

has received little attention in the literature on deceptive behaviors, but is practically relevant and holds important implications for how we view and tackle the problem of fake accounts. We also show that as the platform's anti-fake effort increases, the equilibrium regime generally transitions from a pooling equilibrium, where the low-quality influencer purchases fake accounts to mimic a high-type influencer, to a costly-separating one, where the high-quality influencer purchases fake accounts to prevent a pooling equilibrium, and to a naturally-separating one, where the two types separate without purchasing fake accounts. Interestingly, the system behaves "erratically" in the sense that the number of fake accounts may increase, even rise sharply in response to increased anti-fake efforts. For example, the number of fake accounts can increase with the anti-fake effort before eventually decreasing; it may also jump significantly as the equilibrium regime transitions from pooling to separating.

Second, we enhance our understanding of how social media platforms may choose their level of anti-fake efforts and the potential impact of these choices on consumer welfare. We find that a purely profit-driven platform lacks incentive to implement anti-fake efforts. Even if the platform becomes more consumer-oriented, it generally exerts less anti-fake effort than what is optimal for consumers. Consumers may prefer a costly separating equilibrium with a positive anti-fake effort, whereas a profit-driven platform tends to prefer a pooling equilibrium with zero anti-fake effort. The misalignment arises because the platform can attract more uninformed followers (resulting in higher advertising revenue) under a pooling equilibrium than under a separating equilibrium, whereas consumers may prefer the latter.

Finally, we offer novel insights into the effects of different anti-fake measures. These include platform-led anti-fake initiatives, improving consumers' social media literacy, increasing the cost of fake accounts (e.g., through stronger fake-account laws), and advancing more effective anti-fake technologies. We show that while positive anti-fake efforts may be necessary for sustaining a separating equilibrium, neither the platform nor consumers can benefit from additional anti-fake efforts beyond what is necessary for sustaining a desired equilibrium. This is because additional anti-fake efforts increase consumer nuisance costs and may not always reduce information asymmetry about influencer quality.

Consistent with our findings on the effect of anti-fake effort, the system exhibits "erratic" behavior in response to other anti-fake measures: for example, under the pooling equilibrium, the number of fake accounts may increase in social media literacy and not respond to changes in the fake-account base price. Moreover, a profit-focused platform may prefer a lower social media literacy and a lower fake-account base price, even though consumers can benefit from higher social media literacy and a higher fake-account base price. In contrast, both consumers and the platform are weakly better off with better anti-fake technology.

8.2. Managerial implications

Our findings hold managerial implications for platforms, policymakers, and consumer protection agencies. For platforms, one important insight from our analyses is that platforms should not measure the success of their anti-fake efforts using the resulting number of fake accounts. This is because the equilibrium number of fake accounts may increase and even jump as the platform exerts more anti-fake efforts, and because fake accounts can also be "beneficial" – high-quality influencers may also purchase fake accounts to separate themselves. Instead, platforms should focus on whether their anti-fake effort allows consumers and advertisers to better tell apart high- and low-quality influencers. The same also holds for policymakers and regulators who hope to measure the success of their policies and regulations.

Second, the platform should not impose more than necessary anti-fake efforts. Our results show that additional anti-fake efforts can be costly for both the platform and consumers because of added nuisance costs and operating costs of anti-fake efforts.

For policymakers and regulators, we show that platforms generally under-invest in anti-fake efforts at the expense of consumer welfare. This is because anti-fake efforts can expose low-quality influencers and reduce platforms' user base. Moreover, profit-focused platforms may lack incentives to promote social media literacy or implement measures that aim to increase fake account costs, though such measures can increase consumer welfare. In contrast, advancements in anti-fake technologies can benefit both platforms and consumers. Consequently, policymakers and regulators should focus on incentivizing platform investments in advancing anti-fake technologies.

8.3. Limitations and future work

As an initial step in understanding the issue of fake accounts and developing coping strategies, we have specifically examined a type of fake accounts: those created to boost influencers' popularity. Our findings may not generalize to other types of fake accounts, such as those created for spreading scams, malware, and identity theft, or politically motivated fake accounts. Our model assumes that fake accounts distort consumers' and advertisers' perceptions of influencer quality but do not directly harm them. When fake accounts directly harm consumers, we anticipate that the platform will have stronger incentives to implement anti-fake measures. However, we also expect that many of the driving forces outlined in our model will continue to be relevant. Finally, our model follows the signaling game to assume one representative influencer that can be either low or high quality. Future research can also extend this model to directly incorporate competition between influencers.

References

Al Zou'bi RM (2022) The impact of media and information literacy on students' acquisition of the skills needed to detect fake news. *Journal of Media Literacy Education* 14(2):58–71, URL `https://digitalcommons.uri.edu/jmle-preprints/28`.

ArkoseLabs (2021) The Pros and Cons of reCAPTCHA Enterprise. Technical report, Arkose Labs, URL https://www.arkoselabs.com/blog/the-pros-and-cons-of-recaptcha-enterprise/.

Che YK, Hörner J (2018) Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics* 133(2):871–925.

Chen J, Yang L, Hosanagar K (2022) To Brush or Not to Brush: Product Rankings, Consumer Search, and Fake Orders. *Information Systems Research* .

Chen L, Papanastasiou Y (2021) Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation. *Management Science Publication* February:1–17.

Confessore N, Dance GJ, Harris R, Hansen M (2018) The Follower Factory. URL https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html?module=inline.

Corts KS (2013) Prohibitions on false and unsubstantiated claims: Inducing the acquisition and revelation of information through competition policy. *Journal of Law and Economics* 56(2):453–486.

Corts KS (2014) Finite optimal penalties for false advertising. *Journal of Industrial Economics* 62(4):661–681, ISSN 14676451.

De Veirman M, Cauberghe V, Hudders L (2017) Marketing through instagram influencers: The impact of number of followers and product divergence on brand attitude. *International Journal of Advertising* 36(5):798–828.

Federal Trade Commission (2019) Devumi, Owner and CEO Settle FTC Charges They Sold Fake Indicators of Social Media Influence. *ftc.gov* URL https://www.ftc.gov/news-events/news/press-releases/2019/10.

Freixa S (2021) What's Not to Like?: The Growing Problem of Fake Online Reviews & Social Media Accounts. URL https://www.altlegal.com/blog/whats-not-to-like-the-growing-problem-of-fake-online-reviews-amp-social-media-accounts/.

Guo X, Xiao G, Zhang F (2017) Effect of Consumer Awareness on Corporate Social Responsibility under Asymmetric Information. *SSRN Electronic Journal* 1–48.

Hao K (2020) How Facebook uses machine learning to detect fake accounts. URL https://www.technologyreview.com/2020/03/04/905551/how-facebook-uses-machine-learning-to-detect-fake-accounts/.

Jin SA, Phua J (2014) Following celebrities' tweets about brands: The impact of Twitter-based electronic word-of-mouth on consumers source credibility perception, buying intention, and social identification with celebrities. *Journal of Advertising* 43(2):181–195.

Lankoski L, Smith NC (2018) Alternative objective functions for firms. *Organization & Environment* 31(3):242–262.

Mailath GJ, Okuno-Fujiwara M, Postlewaite A (1993) Belief-based refinements in signalling games. *Journal of Economic Theory* 60(2):241–276.

Mayzlin D (2006) Promotional chat on the internet. *Marketing Science* 25(2):155–163.

Mekuli A (2021) Top 10 Instagram Celebs with the Most "Fake" Followers in 2021. URL https://vpnoverview.com/privacy/social-media/instagram-influencers-with-fake-followers/.

Moore E, Murphy H (2019) Facebook's massive fake numbers problem. URL https://www.latimes.com/business/technology/story/2019-11-18/facebooks-massive-fake-numbers-problem.

Nicas J (2020) Why can't the social networks stop fake accounts? URL https://www.nytimes.com/2020/12/08/technology/why-cant-the-social-networks-stop-fake-accounts.html.

Ortutay B (2022) Twitter says it removes 1 million spam accounts a day. URL https://abcnews.go.com/Technology/wireStory/twitter-removes-million-spam-accounts-day-86382214.

Papanastasiou Y (2020) Fake news propagation and detection: A sequential model. *Management Science* 66(5):1826–1846.

Papanastasiou Y, Bimpikis K, Savva N (2018) Crowdsourcing Exploration. *Management Science* 64(4):1727–1746.

Pennycook G, Bear A, Collins ET, Rand DG (2020) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66(11):4944–4957.

Piccolo S, Tedeschi P, Ursino G (2018) Deceptive advertising with rational buyers. *Management Science* 64(3):1291–1310.

Polanco-Levicán K, Salvo-Garrido S (2022) Understanding social media literacy: A systematic review of the concept and its competences. *International Journal of Environmental Research and Public Health* 19(14):8807.

Raturi R (2018) Machine Learning Implementation for Identifying Fake Accounts in Social Network. *International Journal of Pure and Applied Mathematics* 118(20):4785–4797.

Robins MS (2022) LinkedIn has a problem with fake profiles. URL https://www.techradar.com/news/linkedin-has-a-problem-with-fake-profiles.

Shin E (2017) Monopoly pricing and diffusion of social network goods. *Games and Economic Behavior* 102:162–178.

Warwick S (2022) Facebook removed 1.6 billion fake accounts in just three months. URL https://www.imore.com/facebook-removed-16-billion-fake-accounts-just-three-months.

Wilbur KC, Zhu Y (2009) Click fraud. *Marketing Science* 28(2):293–308.

Yuan D, Miao Y, Gong NZ, Yang Z, Li Q, Song D, Wang Q, Liang X (2019) Detecting fake accounts in online social networks at the time of registrations. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 1423–1438.

A. Appendix

A.1. Proof of Lemma 1.

By definition, the two types of influencers under a pooling equilibrium have the same number of early followers, namely, $n_2^H = n_2^L \equiv n_2^*$ ($n_2^* \geq n_{in}^H$). For simplicity, we denote a pooling equilibrium by the number of early followers n_2^*. Under such a pooling equilibrium, we denote $\overline{n}_{in}$, $E[q]$, n_{un}^{pool}, $n_{r,ia}^t$, $n_{r,ua}^{pool}$, and π_t^{pool} as the expected number of informed followers, the expected quality, the expected number of uninformed followers, the informed advertiser's expected number of real followers for t-type influencer ($t \in \{H, L\}$), the uninformed advertiser's expected number of real followers, and the influencer's (type t) expected profit in the pooling equilibrium, respectively. They can be calculated as

$$\overline{n}_{in} = \rho n_{in}^H + (1-\rho) n_{in}^L \tag{23}$$

$$E[q] = \rho q_H + (1-\rho) q_L \tag{24}$$

$$n_{un}^{pool} = (1-l)\left(1 - \frac{c}{E[q]}\right) \tag{25}$$

$$n_{r,ia}^t = n_{in}^t + n_{un}^{pool} \tag{26}$$

$$n_{r,ua}^{pool} = \overline{n}_{in} + n_{un}^{pool} \tag{27}$$

$$\pi_t^{pool} = l\lambda_i \mu n_{r,ia}^t + (1-l)\lambda_i \mu n_{r,ua}^{pool} - p_f x_t^{pool} \tag{28}$$

In the following, we show that (a) there exists a PBE with $n_2^* = n_{in}^H$ and (b) any other PBE with $n_2^* > n_{in}^H$ is defeated by $n_2^* = n_{in}^H$. Finally, we obtain the supporting beliefs for the unique undefeated equilibrium $n_2^* = n_{in}^H$.

First, we establish that any $n_2^* \in \left[n_{in}^H, \overline{n}_2^{pool}\right]$ with the following belief is a PBE (we define the upper bound $\overline{n}_2^{pool}$ in (35)):

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_2^* \\ \rho & \text{if } n_2 \geq n_2^* \end{cases}$$

We can write the equilibrium profits for the two types as

$$\pi_L^{pool} = \lambda_i \mu \left[(1-l)\rho n_{in}^H + (1-\rho+l\rho) n_{in}^L + n_{un}^{pool}\right] - p_f x_L^{pool} \tag{29}$$

$$\pi_H^{pool} = \lambda_i \mu \left[(l+\rho-l\rho) n_{in}^H + (1-l)(1-\rho) n_{in}^L + n_{un}^{pool}\right] - p_f x_H^{pool} \tag{30}$$

An H-type who deviates to $n_2' > n_2^*$ followers will not gain anything comparing to n_2^* but incur an additional fake-account cost. Clearly, the H-type is better off with n_2^*, and thus has no incentive to deviate to $n_2' > n_2^*$. Similarly, the L-type has no incentive to deviate to $n_2' > n_2^*$ either.

For the L-type, if she deviates to $n_2' \in [n_{in}^L, n_2^*)$, she will be seen as an L-type by both the advertiser (informed and uninformed) and uninformed consumers. Thus, the influencer's expected profit is

$$\pi_L^{dev}(n_2') = \lambda_i \mu n_r^L - p_f\left(n_2' - n_{in}^L\right)$$

where $n_r^L = n_{in}^L + n_{un}^L = \left(1 - \frac{c}{q_L}\right)$ is the number of real followers, where $n_{un}^L = (1-l)\left(1 - \frac{c}{q_L}\right)$. Clearly, the L-type is better off with $n_2' = n_{in}^L$, i.e., not purchasing any fake accounts, resulting in a profit of $\pi_L^{dev}(n_{in}^L) = \lambda_i \mu n_r^L$.

For the H-type, if she deviates to $n_2' \in [n_{in}^H, n_2^*)$, she will be seen as an L-type by both the uninformed advertiser and uninformed consumers, and an H-type by the informed advertiser. Thus, the influencer's expected profit is

$$\pi_H^{dev}(n_2') = l\lambda_i \mu n_{r,ia}^H + (1-l)\lambda_i \mu n_{r,ua}^L - p_f\left(n_2' - n_{in}^H\right)$$

where $n_{r,ia}^H = n_{in}^H + n_{un}^L = l\left(1 - \frac{c}{q_H}\right) + (1-l)\left(1 - \frac{c}{q_L}\right)$ and $n_{r,ua}^L = n_{in}^L + n_{un}^L = \left(1 - \frac{c}{q_L}\right)$ are the informed and uninformed advertiser's expected numbers of real followers, respectively. Clearly, the deviation $n_2' = n_{in}^H$ (no purchasing) dominates any other $n_2' \in (n_{in}^H, n_2^*)$. The former results in a profit of $\pi_H^{dev}(n_{in}^H) = l\lambda_i \mu n_{r,ia}^H + (1-l)\lambda_i \mu n_{r,ua}^L$.

Therefore, a sufficient condition for $n_2^* \in \left[n_{in}^H, \overline{n}_2^{pool}\right]$ being a PBE is:

$$\pi_L^{dev}\left(n_{in}^L\right) \leq \pi_L^{pool} \quad \text{and} \quad \pi_H^{dev}\left(n_{in}^H\right) \leq \pi_H^{pool} \tag{31}$$

which is equivalent to

$$\begin{cases} \lambda_i \mu n_r^L \leq \lambda_i \mu\left[(1-l)\rho n_{in}^H + (1-\rho+l\rho) n_{in}^L + n_{un}^{pool}\right] - p_f x_L^{pool} \\ l\lambda_i \mu n_{r,ia}^H + (1-l)\lambda_i \mu n_{r,ua}^L \leq \lambda_i \mu\left[(l+\rho-l\rho) n_{in}^H + (1-l)(1-\rho) n_{in}^L + n_{un}^{pool}\right] - p_f x_H^{pool} \end{cases}$$

$$\iff \begin{cases} \lambda_i \mu n_{un}^L \leq \lambda_i \mu\rho(1-l)\left(n_{in}^H - n_{in}^L\right) + \lambda_i \mu n_{un}^{pool} - p_f x_L^{pool} \\ \lambda_i \mu n_{un}^L \leq \lambda_i \mu\rho(1-l)\left(n_{in}^H - n_{in}^L\right) + \lambda_i \mu n_{un}^{pool} - p_f x_H^{pool} \end{cases}$$

$$\iff \lambda_i \mu n_{un}^L \leq \lambda_i \mu\rho(1-l)\left(n_{in}^H - n_{in}^L\right) + \lambda_i \mu n_{un}^{pool} - p_f x_L^{pool} \tag{32}$$

For the pooling equilibrium to exist, the above condition must hold for the most profitable pooling equilibrium, i.e., when $n_2^* = n_{in}^H$ and $x_L^{pool} = n_{in}^H - n_{in}^L$, or:

$$\lambda_i \mu\left(n_{un}^{pool} - n_{un}^L\right) \geq \left[p_f - \lambda_i \mu\rho(1-l)\right]\left(n_{in}^H - n_{in}^L\right) \tag{33}$$

Noting that $c = c_0 + c_1(1-\tau)d$ and $p_f = p_0 + \frac{1}{1-\tau}d$, we can rewrite (33) to obtain (13), which is $d \leq (1-\tau)(\eta_1 - p_0)$.

We define

$$d_1 \equiv (1-\tau)(\eta_1 - p_0), \text{ where } \eta_1 = \lambda_i \mu(1-l)\left[\frac{1}{l}\frac{q_H E[q] - q_H q_L}{q_H E[q] - E[q] q_L} + \rho\right]. \tag{34}$$

We denote $\overline{n}_2^{pool}$ as the highest n_2 such that (32) holds in equality. This implies:

$$\overline{n}_2^{pool} = \frac{\lambda_i \mu\left[\rho(1-l)\left(n_{in}^H - n_{in}^L\right) + \left(n_{un}^{pool} - n_{un}^L\right)\right]}{p_f} + n_{in}^L \tag{35}$$

In the above, we have found all the pooling PBEs.

Given the continuum of the pooling equilibria, we now show that any pooling equilibrium with $n_2' > n_{in}^H$ is defeated by $n_2^* = n_{in}^H$. We note that under the former equilibrium, it must be that $P(H|n_2) < \rho$ for any off-equilibrium action $n_2 < n_2'$ (because, otherwise, the L-type would have the incentive to lower her n_2 to save fake account costs). However, both types are better off under the equilibrium $n_2^* = n_{in}^H$, suggesting that n_2^* would defeat n_2' unless the latter assigns $P(H|n_{in}^H) = \rho$ to the off-equilibrium action $n_2 = n_{in}^H$ (as in the n_2^* equilibrium) – a contradiction to the requirement that $P(H|n_{in}^H) < \rho$. Therefore, the unique undefeated pooling equilibrium is $n_2^* = n_{in}^H$.

We now obtain the supporting beliefs for the undefeated equilibrium. Consider a general belief

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ \rho & \text{if } n_2 = n_{in}^H \\ p_1 \in [0,1]. & \text{if } n_{in}^H < n_2 < \overline{n}_2^{pool} \end{cases}$$

We have already established that under condition (33), the L-type has no incentive to deviate to any $n_2 < n_{in}^H$. We now consider an H-type's deviation to $n_2 > n_{in}^H$. When the off-equilibrium belief $p_1 \in [0, \rho]$, the H-type is subject to a less favorable belief and incurs an additional fake-account cost. Clearly, the H-type has no incentive to deviate to $n_2 > n_{in}^H$. Similarly, the L-type has no incentive to deviate to $n_2 > n_{in}^H$ either.

We now consider a off-equilibrium belief $p_1 \in (\rho, 1)$. The L- and H-type's incentive for deviating to $n_2 > n_{in}^H$ is identical, so we focus on the H-type's decision. Suppose an H-type deviates to $n_2 > n_{in}^H$. She will need to purchase $x_H^{pool\prime} = n_2 - n_{in}^H$ fake accounts. She will be seen as an average type (with probability p_1) by both the uninformed advertiser and uninformed consumers, and an H-type by the informed advertiser. Thus, the H-type's expected payoff is:

$$\pi_H^{dev\prime}(n_2) = l\lambda_i \mu n_{r,ia}^{H\prime} + (1-l)\lambda_i \mu n_{r,ua}^{pool\prime} - p_f x_H^{pool\prime}$$

where

$$n_{r,ia}^{H\prime} = n_{in}^H + n_{un}^{pool\prime} \tag{36}$$

$$n_{r,ua}^{pool\prime} = \overline{n}_{in}' + n_{un}^{pool\prime} \tag{37}$$

$$\overline{n}_{in}' = p_1 n_{in}^H + (1-p_1) n_{in}^L \tag{38}$$

$$n_{un}^{pool\prime} = (1-l)\left(1 - \frac{c}{E[q]'}\right) \tag{39}$$

$$E[q]' = p_1 q_H + (1-p_1) q_L \tag{40}$$

The IC condition for the H-type requires $\pi_H^{dev\prime}(n_2) \leq \pi_H^{pool}$, which is equivalent to

$$l\lambda_i \mu n_{r,ia}^{H\prime} + (1-l)\lambda_i \mu n_{r,ua}^{pool\prime} - p_f\left(n_2 - n_{in}^H\right) \leq \lambda_i \mu\left[(l+\rho-l\rho) n_{in}^H + (1-l)(1-\rho) n_{in}^L + n_{un}^{pool}\right]$$

$$\iff l n_{r,ia}^{H\prime} + (1-l)\, n_{r,ua}^{pool\prime} \leq \left[(l+\rho-l\rho)\, n_{in}^{H} + (1-l)\,(1-\rho)\, n_{in}^{L} + n_{un}^{pool}\right] + \frac{p_f}{\lambda_i \mu}\left(n_2 - n_{in}^{H}\right)$$

$$\iff z_1 p_1 - z_2 \frac{1}{p_1 + \frac{q_L}{q_H - q_L}} + z_3 \leq 0 \tag{41}$$

where $z_1 = (1-l)\, l \left(\frac{c}{q_L} - \frac{c}{q_H}\right)$, $z_2 = \frac{(1-l)c}{q_H - q_L}$, and $z_3 = (1-l)\,\frac{c}{E[q]} - (1-l)\,\rho l \left(\frac{c}{q_L} - \frac{c}{q_H}\right) - \frac{p_f}{\lambda_i \mu}\left(n_2 - n_{in}^{H}\right)$. We can further rewrite condition (41) as

$$\gamma_1 p_1^2 + \gamma_2 p_1 + \gamma_3 < 0 \tag{42}$$

where $\gamma_1 = z_1$, $\gamma_2 = \frac{q_L}{q_H - q_L} z_1 + z_3$, and $\gamma_3 = \frac{q_L}{q_H - q_L} z_3 - z_2$. Solving the equality corresponding to (42), we obtain:

$$p_1^* \in \left\{ \frac{-\gamma_2 - \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1}, \frac{-\gamma_2 + \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1} \right\}$$

Noting that the first root $\frac{-\gamma_2 - \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1} < \frac{-\left(\frac{q_L}{q_H - q_L} z_1 + z_3\right) - \left(\frac{q_L}{q_H - q_L} z_1 - z_3\right)}{2z_1} = -\frac{q_L}{q_H - q_L} < 0$ and we already know any $p_1 \leq \rho$ can sustain the equilibrium, the condition (42) simplifies to $p_1 \in [0, \overline{p}_1]$, where

$$\overline{p}_1 = max\left\{\rho, min\left\{\frac{-\gamma_2 + \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1}, 1\right\}\right\} \tag{43}$$

A.2. Proof of Lemma 2

We consider a costly-separating equilibrium (n_2^H, n_2^L), where $n_2^H \neq n_2^L$, $n_2^H \geq n_{in}^H$, and $n_2^L \geq n_{in}^L$. In the following, we show that (a) there exists a PBE (n_2^{csep}, n_{in}^L) where $n_2^{csep} > n_{in}^H$ and (b) any other PBE (n_2^H, n_2^L) with $n_2^H > n_2^{csep}$ or $n_{in}^L < n_2^L < n_2^H$ is defeated by (n_2^{csep}, n_{in}^L). Finally, we explore the supporting beliefs for the unique undefeated equilibrium (n_2^{csep}, n_{in}^L).

First, we establish that any (n_2^{H*}, n_2^{L*}) where $n_2^{H*} \in [n_2^{csep}, \overline{n}_2^{csep}]$ and $n_2^{L*} \in [n_{in}^L, n_2^{H*})$ with the following belief is a PBE (we define n_2^{csep} in (17) and $\overline{n}_2^{csep}$ in (47)):

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_2^{H*} \\ 1 & \text{if } n_2 \geq n_2^{H*} \end{cases}$$

Denoting $n_r^L = n_{in}^L + n_{un}^L$, $n_r^H = n_{in}^H + n_{un}^H$, $x_L^{csep} = n_2^{L*} - n_{in}^L$, and $x_H^{csep} = n_2^{H*} - n_{in}^H$, we can write the equilibrium profits for the two types as

$$\pi_L^{csep} = \lambda_i \mu n_r^L - p_f x_L^{csep} \tag{44}$$

$$\pi_H^{csep} = \lambda_i \mu n_r^H - p_f x_H^{csep} \tag{45}$$

First, an L-type deviating to $n_2' < n_2^{H*}$ is always seen as an L-type and better off by not purchasing any fake accounts. In other words, any $n_2^{L\prime} \in (n_{in}^L, n_2^{H*})$ is dominated by $n_2^{L*} = n_{in}^L$.

If an L-type deviates to $n_2^{L\prime} \in [n_2^{H*}, \overline{n}_2^{csep}]$, she will be seen as an H-type by both the uninformed advertiser and uninformed consumers, and an L-type by the informed advertiser. Thus, her expected profit is $\pi_L^{dev}(n_2^{L\prime}) = l\lambda_i\mu n_{r,ia}^L + (1-l)\lambda_i\mu n_{r,ua}^L - p_f(n_2^{L\prime} - n_{in}^H)$, where $n_{r,ia}^L = n_{in}^L + n_{un}^H$ and $n_{r,ua}^L = n_r^H = n_{in}^H + n_{un}^H$. Clearly, the L-type is better off with $n_2^{L\prime} = n_2^{H*}$, resulting in a profit of

$$\pi_L^{dev}\left(n_2^{H*}\right) = l\lambda_i\mu n_{r,ia}^L + (1-l)\lambda_i\mu n_{r,ua}^L - p_f\left(n_2^{H*} - n_{in}^L\right).$$

If an H-type deviates to $n_2^{H\prime} > n_2^{H*}$, she will gain nothing comparing to n_2^{H*} but incur an additional fake-account cost. Clearly, the H-type has no incentive to deviate to $n_2^{H\prime} > n_2^{H*}$. If the H-type deviates to $n_2^{H\prime} \in [n_{in}^H, n_2^{H*})$, she will be seen as an L-type by both the uninformed advertiser and uninformed consumers, and an H-type by the informed advertiser. Thus, her expected profit is

$$\pi_H^{dev}\left(n_2^{H\prime}\right) = l\lambda_i\mu n_{r,ia}^H + (1-l)\lambda_i\mu n_{r,ua}^H - p_f\left(n_2^{H\prime} - n_{in}^L\right)$$

where $n_{r,ia}^H = n_{in}^H + n_{un}^L$ and $n_{r,ua}^H = n_r^L = n_{in}^L + n_{un}^L$. Clearly, the H-type is better off with $n_2^{H\prime} = n_{in}^H$, resulting in a profit of $\pi_H^{dev}(n_{in}^H) = l\lambda_i\mu n_{r,ia}^H + (1-l)\lambda_i\mu n_{r,ua}^H$.

Therefore, a sufficient condition for (n_2^{H*}, n_{in}^L) (where $n_2^{H*} \in [n_2^{csep}, \overline{n}_2^{csep}]$) being an PBE is $\pi_L^{dev}(n_2^{H*}) \le \pi_L^{csep}$ and $\pi_H^{dev}(n_{in}^H) \le \pi_H^{csep}$, or

$$\begin{cases} l\lambda_i\mu n_{r,ia}^L + (1-l)\lambda_i\mu n_{r,ua}^L - p_f(n_2^{H*} - n_{in}^L) \le \lambda_i\mu n_r^L \\ l\lambda_i\mu n_{r,ia}^H + (1-l)\lambda_i\mu n_{r,ua}^H \le \lambda_i\mu n_r^H - p_f(n_2^{H*} - n_{in}^H) \end{cases}$$

$$\iff \begin{cases} \lambda_i\mu\left[(1-l)(n_{in}^H - n_{in}^L) + (n_{un}^H - n_{un}^L)\right] \le p_f(n_2^{H*} - n_{in}^L) \\ p_f(n_2^{H*} - n_{in}^H) \le \lambda_i\mu\left[(1-l)(n_{in}^H - n_{in}^L) + (n_{un}^H - n_{un}^L)\right] \end{cases} \tag{46}$$

The lowest costly-separating equilibrium, denoted as (n_2^{csep}, n_{in}^L), is such that the first condition in (46) holds in equality (the L-type's IC condition binds), which implies (17).

The highest costly-separating equilibrium, denoted as $(\overline{n}_2^{csep}, n_{in}^L)$, is such that the second condition in (46) holds in equality (the H-type's IC condition binds), which implies:

$$\overline{n}_2^{csep} = n_2^{csep} + (n_{in}^H - n_{in}^L) \tag{47}$$

In the above, we have found all the costly-separating PBEs.

For this to be a costly-separating equilibrium, the H-type must purchase fake accounts, that is, $n_2^{csep} > n_{in}^H$, or $\lambda_i\mu\left[(1-l)(n_{in}^H - n_{in}^L) + (n_{un}^H - n_{un}^L)\right] - p_f(n_{in}^H - n_{in}^L) > 0$. We can rearrange the items and obtain

$$d < (1-\tau)(\eta_2 - p_0)$$

where $\eta_2 = \frac{\lambda_i\mu\left(1-l^2\right)}{l}$. We define $d_2 \equiv (1-\tau)(\eta_2 - p_0)$.

Given the continuum of costly-separating equilibria, we now show that any costly-separating equilibrium $(n_2^{H\prime}, n_{in}^L)$ with $n_2^{H\prime} > n_2^{csep}$ is defeated by (n_2^{csep}, n_{in}^L). We note that under the former

equilibrium, it must be that $P(H|n_2) < 1$ for any off-equilibrium action $n_2^H < n_2^{H\prime}$ (because, otherwise, the H-type would have the incentive to lower her n_2 to save fake account costs). However, the H-type is strictly better off under the equilibrium (n_2^{csep}, n_{in}^L) (L-type remains the same), suggesting that (n_2^{csep}, n_{in}^L) would defeat $(n_2^{H\prime}, n_{in}^L)$ unless the latter assigns $P(H|n_2^{csep}) = 1$ to the off-equilibrium action $n_2^H = n_2^{csep}$ (as dictated by the former equilibrium) – a contradiction to the requirement that $P(H|n_2^{csep}) < 1$. Therefore, the unique undefeated costly-separating equilibrium is defined by (n_2^{csep}, n_{in}^L).

We now explore the supporting beliefs for the undefeated equilibrium. We consider the general belief

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ p_2 \in [0,1] & \text{if } n_{in}^H \le n_2 < n_2^{csep} \\ 1 & \text{if } n_2 = n_2^{csep} \\ p_3 \in [0,1], & \text{if } n_2^{csep} < n_2 < \overline{n}_2^{csep} \end{cases}$$

We have already established that under condition (46), the influencer (L- or H-type) has no incentive to deviate to $n_2 > n_2^{csep}$. When the off-equilibrium belief $p_3 \in [0,1)$, the influencer will be seen as an average type; as a result, she gains lower compared to the deviation when the off-equlibirium belief $P(H|n_2^{csep} < n_2 < \overline{n}_2^{csep}) = 1$. Clearly, both H-type and L-type have no incentive to deviate to $n_2 > n_2^{csep}$ when $P(H|n_2 > n_2^{csep}) \in [0,1]$.

We now consider an influencer (L- or H-type)'s deviation to $n_{in}^H \le n_2 < n_2^{csep}$. When $p_2 = 1$, H-type is better off by deviation, thus, $p_2 \in [0,1)$. We already know that the influencer (L- or H-type) has no incentive to deviate when $p_2 = 0$.

When the off-equilibrium belief $P(H|n_{in}^H \le n_2 < n_2^{csep}) = p_2 \in (0,1)$, it's possible for both types' deviation since L-type gains comparing to $n_2^{L*} = n_{in}^L$ meanwhile incur additional fake-account cost, and H-type gets lower profit comparing to $n_2^{H*} = n_2^{csep}$ but incur lower fake-account cost. If the influencer deviates to $n_2 \in (n_{in}^H, n_2^{csep})$, we have $x_L^{csep\prime} = n_2 - n_{in}^L$, and $x_H^{csep\prime} = n_2 - n_{in}^H$ as the fake accounts number, and the influencer is expected by the uninformed advertiser and uninformed consumers to be an H-type with probability p_2 and an L-type with a probability of $(1-p_2)$. However, the informed advertiser knows the influencer's true type. The influencer (L- or H-type)'s expected payoff is

$$\pi_L^{dev}(n_2) = l\lambda_i \mu n_{r,ia}^{L\prime} + (1-l)\lambda_i \mu \overline{n}_{r,ua}^{\prime} - p_f x_L^{csep\prime}$$
$$\pi_H^{dev}(n_2) = l\lambda_i \mu n_{r,ia}^{H\prime} + (1-l)\lambda_i \mu \overline{n}_{r,ua}^{\prime} - p_f x_H^{csep\prime}$$

where

$$\overline{n}_{in}^{\prime} = p_2 n_{in}^H + (1-p_2) n_{in}^L \tag{48}$$
$$E[q]^{\prime} = p_2 q_H + (1-p_2) q_L \tag{49}$$

$$n'_{un} = (1-l)\left(1 - \frac{c}{E[q]'}\right) \tag{50}$$

$$n^{L\prime}_{r,ia} = n^{L}_{in} + n'_{un} \tag{51}$$

$$n^{H\prime}_{r,ia} = n^{H}_{in} + n'_{un} \tag{52}$$

$$\overline{n}'_{r,ua} = \overline{n}'_{in} + n'_{un} \tag{53}$$

To ensure the undefeated costly-separating equilibrium hold, the IC conditions for both H-type and L-type require

$$\begin{cases} \pi^{dev}_{L}(n_2) \leq \pi^{csep}_{L} \\ \pi^{dev}_{H}(n_2) \leq \pi^{csep}_{H} \end{cases}$$

which translates to

$$\begin{cases} l\lambda_i\mu n^{L\prime}_{r,ia} + (1-l)\lambda_i\mu\overline{n}'_{r,ua} - p_f(n_2 - n^{L}_{in}) \leq \lambda_i\mu n^{L}_{r} \\ l\lambda_i\mu n^{H\prime}_{r,ia} + (1-l)\lambda_i\mu\overline{n}'_{r,ua} - p_f(n_2 - n^{H}_{in}) \leq \lambda_i\mu n^{H}_{r} - p_f(n^{csep}_{2} - n^{H}_{in}) \end{cases}$$

Which is equivalent to (binding H-type's IC condition):

$$\lambda_i\mu\left[(1-l)\overline{n}'_{in} + n'_{un}\right] \leq \lambda_i\mu\left(n^{H}_{r} - ln^{H}_{in}\right) + p_f(n_2 - n^{csep}_{2})$$

Thus, we have

$$z_1 p_2 - z_2 \frac{1}{p_2 + \frac{q_L}{q_H - q_L}} + z_3 \leq 0$$

where

$$\begin{cases} z_1 = (1-l)\,l\left(\frac{c}{q_L} - \frac{c}{q_H}\right) \\ z_2 = \frac{(1-l)c}{q_H - q_L} \\ z_3 = (1-l^2)\frac{c}{q_H} - (l-l^2)\frac{c}{q_L} - \frac{p_f}{\lambda_i\mu}(n_2 - n^{csep}_{2}) \end{cases}$$

Further, we translate the above inequality equation as the following general format

$$\gamma_1 p_2^2 + \gamma_2 p_2 + \gamma_3 < 0$$

where

$$\begin{cases} \gamma_1 = z_1 \\ \gamma_2 = \frac{q_L}{q_H - q_L} z_1 + z_3 \\ \gamma_3 = \frac{q_L}{q_H - q_L} z_3 - z_2 \end{cases}$$

$$p_2 \in \left[\frac{-\gamma_2 - \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1}, \frac{-\gamma_2 + \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1}\right]$$

For the lower bound of p_2 above, we have $\frac{-\gamma_2 - \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1} < \frac{-\left(\frac{q_L}{q_H - q_L} z_1 + z_3\right) - \left(\frac{q_L}{q_H - q_L} z_1 - z_3\right)}{2z_1} = -\frac{q_L}{q_H - q_L} < 0$, but for the higher bound of p_2 above, we have either $0 < \frac{-\gamma_2 + \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1} < 1$ or $\frac{-\gamma_2 + \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1} > 1$.

Above all, the off-equilibrium belief p_3 for deviation $n^{csep}_{2} < n_2 < \overline{n}^{csep}_{2}$ can be any between 0 and 1. Let

$$\overline{p}_2 = \min\left\{\frac{-\gamma_2 + \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1}, 1\right\} \tag{54}$$

and any the off-equilibrium belief $p_2 \in [0, \overline{p}_2]$ for deviation $n_{in}^H \leq n_2 < n_2^{csep}$ can support the equilibrium.

Furthermore, we can prove that $d_1 < d_2$. From the proofs for Lemma 1 and Lemma 2, we have

$$d_1 \equiv (1-\tau)(\eta_1 - p_0); d_2 \equiv (1-\tau)(\eta_2 - p_0)$$

$$\eta_1 = \lambda_i \mu (1-l) \left[\frac{1}{l} \frac{q_H E[q] - q_H q_L}{q_H E[q] - E[q] q_L} + \rho \right]; \eta_2 = \lambda_i \mu (1-l) \left(\frac{1}{l} + 1 \right)$$

Since $\frac{q_H E[q] - q_H q_L}{q_H E[q] - E[q] q_L} < 1$ and $\rho < 1$, thus, we have $\eta_1 < \eta_2$, which further leads to $d_1 < d_2$.

A.3. Proof of Lemma 3

By definition, the two types of influencers under the naturally-separating equilibrium should have all informed consumers as their early followers, we denote the naturally-separating equilibrium by $n_2^{H*} = n_{in}^H$ and $n_2^{L*} = n_{in}^L$, which is equivalent to a strategy profile of $(x_H^{nsep}, x_L^{nsep}) = (0, 0)$. In the following we establish that $(n_2^{H*}, n_2^{L*}) = (n_{in}^H, n_{in}^L)$ with the following belief is a PBE.

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ 1 & \text{if } n_2 = n_{in}^H \\ any & \text{if } n_2 > n_{in}^H \end{cases}$$

Denoting $n_r^L = n_{in}^L + n_{un}^L$ and $n_r^H = n_{in}^H + n_{un}^H$, we can write the equilibrium profits for the two types as

$$\pi_L^{nsep} = \lambda_i \mu n_r^L \tag{55}$$

$$\pi_H^{nsep} = \lambda_i \mu n_r^H \tag{56}$$

For the H-type, if she deviates to $n_2^H > n_{in}^H$, she will not gain anything comparing to n_{in}^H but incur an additional fake-account cost. Clearly, the H-type is better off with n_{in}^H, and thus has no incentive to deviate to $n_2^H > n_{in}^H$.

For the L-type, first, she will not purchase fake accounts to deviate to a follower count $n_2^L \in (n_{in}^L, n_{in}^H)$, this is because, with the belief capped at 0 for $n_2 < n_{in}^H$, she will be seen as L-type when $n_2^L \in (n_{in}^L, n_{in}^H)$ and can always be better off by not purchasing any fake accounts (i.e., $n_2^L = n_{in}^L$). If the L-type deviates to n_{in}^H, she will be seen as an H-type by both the uninformed advertiser and uninformed consumers. However, the informed advertiser knows that she is L-type. Thus, her expected profit is

$$\pi_L^{dev}\left(n_{in}^H\right) = l \lambda_i \mu n_{r,ia}^L + (1-l) \lambda_i \mu n_{r,ua}^L - p_f \left(n_{in}^H - n_{in}^L\right)$$

where $n_{r,ia}^L = n_{in}^L + n_{un}^H$, and $n_{r,ua}^L = n_r^H$. Still for the L-type, when the off-equilibrium belief $P(H|n_2 > n_{in}^H) = 0$, she has no incentive to deviate to $n_2^L > n_{in}^H$ since she gains nothing but incur an

additional fake-account cost. When the off-equlibirium belief $P\left(H|n_2 > n_{in}^H\right) = 1$, the deviation to $n_2^L > n_{in}^H$ will be dominated by deviation to $n_2^L = n_{in}^H$, i.e., if the L-type doesn't deviate to $n_2^L = n_{in}^H$, she will not deviate to $n_2^L > n_{in}^H$ either. When the off-equilibrium belief $P\left(H|n_2 > n_{in}^H\right) = p_4 \in (0,1)$ and the L-type deviates to $n_2^L > n_{in}^H$, we have $x_L^{nsep\prime} = n_2 - n_{in}^L$ as the number of fake accounts, and she will be seen as an average type by the uninformed advertisers and the uninformed consumers and is expected to be an H-type with probability p_4 and an L-type with a probability of $(1-p_4)$. Again, the informed advertiser knows that she is L-type. Thus, the L-type's expected payoff is

$$\pi_L^{dev}\left(n_2\right) = l\lambda_i\mu n_{r,ia}^{L\prime} + \left(1-l\right)\lambda_i\mu n_{r,ua}^{L\prime} - p_f x_L^{nsep\prime}$$

where

$$\overline{n}_{in}^{\prime} = p_4 n_{in}^H + \left(1-p_4\right) n_{in}^L \tag{57}$$

$$E\left[q\right]^{\prime} = p_4 q_H + \left(1-p_4\right) q_L \tag{58}$$

$$n_{un}^{\prime} = \left(1-l\right)\left(1 - \frac{c}{E\left[q\right]^{\prime}}\right) \tag{59}$$

$$n_{r,ia}^{L\prime} = n_{in}^L + n_{un}^{\prime} \tag{60}$$

$$n_{r,ua}^{L\prime} = \overline{n}_{in}^{\prime} + n_{un}^{\prime} \tag{61}$$

To ensure the naturally-separating equilibrium holds, the IC conditions for the L-type require

$$\begin{cases} \pi_L^{dev}\left(n_{in}^H\right) \leq \pi_L^{nsep} \\ \pi_L^{dev}\left(n_2\right) \leq \pi_L^{nsep} \end{cases}$$

which translates to

$$\begin{cases} \lambda_i\mu\left[l n_{in}^L + \left(1-l\right) n_{in}^H + n_{un}^H\right] - p_f\left(n_{in}^H - n_{in}^L\right) \leq \lambda_i\mu n_r^L \\ \lambda_i\mu\left[l n_{in}^L + \left(1-l\right)\overline{n}_{in}^{\prime} + n_{un}^{\prime}\right] - p_f\left(n_2 - n_{in}^L\right) \leq \lambda_i\mu n_r^L \end{cases}$$

Which is equivalent to:

$$\begin{cases} \lambda_i\mu\left(1-l^2\right) \leq p_f l \\ \lambda_i\mu\overline{n}_r^{\prime} \leq \lambda_i\mu n_r^L + p_f\left(n_2 - n_{in}^L\right) \end{cases}$$

Similar to the process in the proof of Lemma 2, the second condition translates to

$$z_1 p_4 - z_2 \frac{1}{p_4 + \frac{q_L}{q_H - q_L}} + z_3 \leq 0$$

where

$$\begin{cases} z_1 = \left(1-l\right) l\left(\frac{c}{q_L} - \frac{c}{q_H}\right) \\ z_2 = \frac{\left(1-l\right)c}{q_H - q_L} \\ z_3 = \left(1-l\right)\frac{c}{q_L} - \frac{p_f}{\lambda_i\mu}\left(n_2 - n_{in}^L\right) \end{cases}$$

Further, the off-equilibrium belief p_4 for deviation $n_2 > n_{in}^H$ can be

$$p_4 \in \begin{cases} \left[0, \delta_{sep}\left(n_2\right)\right], & \text{if } \delta_{sep}\left(n_2\right) < 1 \\ \left[0,1\right) & \text{otherwise} \end{cases} \tag{62}$$

where

$$\delta_{sep}(n_2) = \frac{-\left(\frac{lc}{q_H} + \frac{c(1-l)}{q_L} - \frac{p_f\left(n_2 - n_{in}^L\right)}{\lambda_i \mu}\right) + \sqrt{\left(\frac{lc}{q_H} - \frac{c(1-l)}{q_L} + \frac{p_f\left(n_2 - n_{in}^L\right)}{\lambda_i \mu}\right)^2 + 4\frac{l(1-l)c^2}{q_H q_L}}}{2l\left(\frac{c}{q_L} - \frac{c}{q_H}\right)}$$

A.4. Proof of Lemma 4.

By Lemma (1) and Lemma (3), the pooling equilibrium exists when $d < d_1$ and the naturally-separating equilibrium exists when $d > d_2$; also we know $d_1 < d_2$, therefore, the pooling and naturally-separating cannot coexist.

However, the pooling and the costly-separating equilibrium can coexist when $d < d_1$. Comparing the H-type's profits under the two equilibria, we have

$$\pi_H^{csep} - \pi_H^{pool} = \lambda_i \mu n_r^H - p_f x_H^{csep} - \lambda_i \mu \left[(l + \rho - l\rho) n_{in}^H + (1-l)(1-\rho) n_{in}^L + n_{un}^{pool}\right]$$

From L-type's IC condition 13 for the pooling equilibrium, we have

$$\lambda_i \mu n_r^L \leq \lambda_i \mu \left[(1-l)\rho n_{in}^H + (1 - \rho + l\rho) n_{in}^L + n_{un}^{pool}\right] - p_f x_L^{pool}$$

which is equivalent to

$$\lambda_i \mu \left[(l + \rho - l\rho) n_{in}^H + (1-l)(1-\rho) n_{in}^L + n_{un}^{pool}\right] > \lambda_i \mu n_r^L + p_f x_L^{pool} + \lambda_i \mu l \left(n_{in}^H - n_{in}^L\right)$$

Therefore,

$$\begin{aligned}
\pi_H^{csep} - \pi_H^{pool} &= \lambda_i \mu n_r^H - p_f x_H^{csep} - \lambda_i \mu \left[(l + \rho - l\rho) n_{in}^H + (1-l)(1-\rho) n_{in}^L + n_{un}^{pool}\right] \\
&< \lambda_i \mu n_r^H - \left[\lambda_i \mu n_r^L + p_f x_L^{pool} + \lambda_i \mu l \left(n_{in}^H - n_{in}^L\right)\right] - p_f x_H^{csep} \\
&= \lambda_i \mu \left[n_r^H - n_r^L - l\left(n_{in}^H - n_{in}^L\right)\right] - p_f \left(n_{in}^H - n_{in}^L\right) - p_f \frac{\left[\lambda_i \mu (1 - l^2) - p_f l\right]\left(\frac{c}{q_L} - \frac{c}{q_H}\right)}{p_f} \\
&= \left[\lambda_i \mu \left(1 - l^2\right) - p_f l\right]\left(\frac{c}{q_L} - \frac{c}{q_H}\right) - \left[\lambda_i \mu \left(1 - l^2\right) - p_f l\right]\left(\frac{c}{q_L} - \frac{c}{q_H}\right) = 0
\end{aligned}$$

Above all, we have

$$\pi_H^{csep} - \pi_H^{pool} < 0$$

The H-type gets a higher payoff in the pooling equilibrium than in the costly-separating equilibrium, thus, she has an incentive to deviate from the costly-separating equilibrium to the alternative pooling equilibrium. The L-type is also better off in the pooling equilibrium; in addition, the uninformed consumers' and advertisers' beliefs in the costly-separating equilibrium about such a pooling deviation are inconsistent with that in the alternative pooling equilibrium. Thus, the costly-separating is defeated, and the pooling will be the only undefeated equilibrium when $d < d_1$.

A.5. Proof of Lemma 5.

We define consumer welfare as $U = U_{in} + U_{un}$ which includes the welfare of the informed and uninformed consumers.

(1) Under the pooling equilibrium, consumer welfare is:

$$\begin{aligned} U^{pool}(d) &= lU_{in}^{pool} + (1-l)\,U_{un}^{pool} \\ &= \rho l \int_{\frac{c}{q_H}}^{1} (\theta q_H - c)\,d\theta + (1-\rho)\,l \int_{\frac{c}{q_L}}^{1} (\theta q_L - c)\,d\theta + (1-l)\,E\left[\int_{\frac{c}{E[q]}}^{1} (\theta q - c)\,d\theta\right] \\ &= l\rho\frac{(q_H - c)^2}{2q_H} + l\,(1-\rho)\,\frac{(q_L - c)^2}{2q_L} + \\ &\quad (1-l)\left[\rho\left(\frac{q_H}{2} - c + \frac{2E[q] - q_H}{2E^2[q]}c^2\right) + (1-\rho)\left(\frac{q_L}{2} - c + \frac{2E[q] - q_L}{2E^2[q]}c^2\right)\right] \end{aligned}$$

Because we assume $q_H > q_L > c$, $U^{pool}(d)$ monotonically decreases in d

(2) Under the costly- and naturally-separating equilibria, consumer welfare is

$$\begin{aligned} U^{csep}(d) = U^{nsep}(d) &= \rho \int_{\frac{c}{q_H}}^{1} (\theta q_H - c)\,d\theta + (1-\rho) \int_{\frac{c}{q_L}}^{1} (\theta q_L - c)\,d\theta \\ &= \rho\frac{(q_H - c)^2}{2q_H} + (1-\rho)\,\frac{(q_L - c)^2}{2q_L} \end{aligned}$$

Again, $U^{csep}(d)$ and $U^{nsep}(d)$ monotonically decrease in d.

Additionally, we have

$$\begin{aligned} U^{csep}(d) - U^{pool}(d) &= (1-l)\left\{\rho\left[\frac{(q_H - c)^2}{2q_H} - \left(\frac{q_H}{2} - c + \frac{2E[q] - q_H}{2E^2[q]}c^2\right)\right] + \right. \\ &\qquad \left. (1-\rho)\left[\frac{(q_L - c)^2}{2q_L} - \left(\frac{q_L}{2} - c + \frac{2E[q] - q_L}{2E^2[q]}c^2\right)\right]\right\} \\ &= (1-l)\left[\frac{(E[q] - q_H)^2}{2E[q]\,q_H} + \frac{(E[q] - q_L)^2}{2E[q]\,q_L}\right] > 0 \end{aligned}$$

Thus, when the consumer's nuisance cost is the same, consumers have a higher welfare when the influencers are separated than when they are pooled together.

A.6. Proof of Lemma 6.

a) If $d_1 \leq 0$, the pooling equilibrium doesn't exist, and we only need to compare the consumer's local optimum with the costly- and naturally-separating equilibria to determine optimal d^C. Noting that because $U^{csep}(d)$ and $U^{nsep}(d)$ decrease in d, thus, the maximum consumer welfare under the costly- and naturally-separating equilibrium are $U^{csep}(0)$ and $U^{nsep}(d_2)$, respectively. Moreover, we have $U^{csep}(0) > U^{csep}(d_2) = U^{nsep}(d_2)$. So the consumer optimum is $U^{csep}(0)$, achieved through a costly-separating equilibrium with $d^C = 0$.

b) If $d_1 > 0$, the platform can induce any of the three types of equilibria. Given that consumer welfare under the pooling, costly and naturally-separating equilibria are maximized at $d = 0$, $d = d_1$ and $d = d_2$, respectively, we have $U^{csep}(d_1) > U^{csep}(d_2) = U^{nsep}(d_2)$, and we only need to compare consumer welfare under pooling and costly-separating equilibria. Thus, consumer welfare can be either $U^{pool}(0)$, achieved through a pooling equilibrium with $d^C = 0$, or $U^{csep}(d_1)$, achieved through a costly-separating equilibrium with $d^C = d_1$, whichever yields a higher consumer welfare.

The consumer-optimal anti-fake effort and equilibrium type are summarized in Table 2.

A.7. Proof of Lemma 7.

(1) Under the pooling equilibrium, the platform profit is:

$$\begin{aligned}
\pi_p^{pool} &= \rho\pi_{p,H}^{pool} + (1-\rho)\pi_{p,L}^{pool} \\
&= \rho\left[l\lambda_p\mu n_{r,ia}^H + (1-l)\lambda_p\mu n_{r,ua}^{pool} - \frac{\gamma}{2}d^2\right] + (1-\rho)\left[l\lambda_p\mu n_{r,ia}^L + (1-l)\lambda_p\mu n_{r,ua}^{pool} - \frac{\gamma}{2}d^2\right] \\
&= l\lambda_p\mu\left[\rho n_{r,ia}^H + (1-\rho)n_{r,ia}^L\right] + (1-l)\lambda_p\mu n_{r,ua}^{pool} - \frac{\gamma}{2}d^2 \\
&= l\lambda_p\mu\left[\rho n_{in}^H + (1-\rho)n_{in}^L + n_{un}^{pool}\right] + (1-l)\lambda_p\mu\left(\overline{n}_{in} + n_{un}^{pool}\right) - \frac{\gamma}{2}d^2 \\
&= \lambda_p\mu\left(\overline{n}_{in} + n_{un}^{pool}\right) - \frac{\gamma}{2}d^2 \\
&= \lambda_p\mu\left[l\rho\left(1-\frac{c}{q_H}\right) + l(1-\rho)\left(1-\frac{c}{q_L}\right) + (1-l)\left(1-\frac{c}{E[q]}\right)\right] - \frac{\gamma}{2}d^2 \\
&= \lambda_p\mu - \lambda_p\mu\left(\frac{l\rho}{q_H} + \frac{l(1-\rho)}{q_L} + \frac{1-l}{E[q]}\right)(c_0 + c_1(1-\tau)d) - \frac{\gamma}{2}d^2 \\
&= -\frac{\gamma}{2}d^2 - \lambda_p\mu c_1(1-\tau)\left(\frac{l\rho}{q_H} + \frac{l(1-\rho)}{q_L} + \frac{1-l}{E[q]}\right)d + \lambda_p\mu - \lambda_p\mu c_0\left(\frac{l\rho}{q_H} + \frac{l(1-\rho)}{q_L} + \frac{1-l}{E[q]}\right)
\end{aligned}$$

We can rewrite π_p^{pool} as

$$\pi_p^{pool}(d) = \omega_1 d^2 + \omega_2 d + \omega_3 \tag{63}$$

where

$$\begin{cases}
\omega_1 = -\frac{\gamma}{2} \\
\omega_2 = -\lambda_p\mu c_1(1-\tau)\left(\frac{l\rho}{q_H} + \frac{l(1-\rho)}{q_L} + \frac{1-l}{E[q]}\right) \\
\omega_3 = \lambda_p\mu - \lambda_p\mu c_0\left(\frac{l\rho}{q_H} + \frac{l(1-\rho)}{q_L} + \frac{1-l}{E[q]}\right)
\end{cases}$$

Because $\omega_1 < 0$ and $\omega_2 < 0$, we have $\frac{\partial\pi_p^{pool}}{\partial d} < 0$ and $\pi_p^{pool}(d)$ monotonically decreases in d.

(2)Under the costly-separating equilibrium, the platform profit is

$$\begin{aligned}
\pi_p^{csep} &= \rho\pi_{p,H}^{csep} + (1-\rho)\pi_{p,L}^{csep} \\
&= \rho\lambda_p\mu n_r^H + (1-\rho)\lambda_p\mu n_r^L - \frac{\gamma}{2}d^2 \\
&= \lambda_p\mu\left[1 - \left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right)c\right] - \frac{\gamma}{2}d^2
\end{aligned}$$

We can rewrite π_p^{csep} as

$$\pi_p^{csep}(d) = \alpha_1 d^2 + \alpha_2 d + \alpha_3 \tag{64}$$

where

$$\begin{cases} \alpha_1 = -\frac{\gamma}{2} \\ \alpha_2 = -c_1(1-\tau)\left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right) \\ \alpha_3 = \lambda_p \mu - c_0\left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right) \end{cases}$$

Noting that $\alpha_1 < 0$ and $\alpha_2 < 0$, we have $\frac{\partial \pi_p^{csep}}{\partial d} < 0$, thus, $\pi_p^{csep}(d)$ monotonically decreases in d.

(3)Under the naturally-separating equilibrium, the platform's profit is

$$\begin{aligned} \pi_p^{nsep} &= \rho \pi_{p,H}^{nsep} + (1-\rho)\pi_{p,L}^{nsep} \\ &= \rho \lambda_p \mu n_r^H + (1-\rho)\lambda_p \mu n_r^L - \frac{\gamma}{2} d^2 \\ &= \lambda_p \mu \left[1 - \left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right) c\right] - \frac{\gamma}{2} d^2 \end{aligned}$$

Noting that $c = c_0 + c_1(1-\tau)d$, $\pi_p^{nsep}(d)$ monotonically decreases in d.

A.8. Proof of Lemma 8

The number of total uninformed consumers under the pooling, costly-separating, and naturally-separating equilibrium are, respectively,

$$\begin{aligned} n_{un}^{pool} &= (1-l)\left(1 - \frac{1}{\rho q_H + (1-\rho) q_L}\right) \\ n_{un}^{csep} &= (1-l)\rho\left(1 - \frac{c}{q_H}\right) + (1-l)(1-\rho)\left(1 - \frac{c}{q_L}\right) \end{aligned}$$

Then, we have

$$\begin{aligned} n_{un}^{pool} - n_{un}^{csep} &= (1-l)\left(1 - \frac{1}{\rho q_H + (1-\rho) q_L}\right) - (1-l)\rho\left(1 - \frac{c}{q_H}\right) - (1-l)(1-\rho)\left(1 - \frac{c}{q_L}\right) \\ &= (1-l)\left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L} - \frac{1}{\rho q_H + (1-\rho) q_L}\right) c > 0 \end{aligned}$$

Thus, the number of total uninformed consumers under the pooling equilibrium is always larger than that under the costly-separating/naturally-separating equilibrium.

(a) If $d_1 \leq 0$, the pooling equilibrium doesn't exist. Note that the platform's profit under the costly and naturally-separating equilibria decreases with d. Thus, $\pi_p^{csep}(0) > \pi_p^{csep}(d_2) = \pi_p^{nsep}(d_2)$.

(b) If $d_1 > 0$, the platform can induce any of the three types of equilibria. According to the proof of Lemma 7, we have

$$\pi_p^{pool}(0) - \pi_p^{csep}(d_1) > \pi_p^{pool}(d_1) - \pi_p^{csep}(d_1)$$

$$
\begin{aligned}
&= \lambda_p\mu\left[1-\left(\frac{l\rho}{q_H}+\frac{l\left(1-\rho\right)}{q_L}+\frac{1-l}{E\left[q\right]}\right)c\right]-\lambda_p\mu\left[1-\left(\frac{\rho}{q_H}+\frac{1-\rho}{q_L}\right)c\right]\\
&= \lambda_p\mu\left[\left(\frac{\rho}{q_H}+\frac{1-\rho}{q_L}\right)-\left(\frac{l\rho}{q_H}+\frac{l\left(1-\rho\right)}{q_L}+\frac{1-l}{E\left[q\right]}\right)\right]c\\
&= \lambda_p\mu\left(1-l\right)\left(\frac{\rho}{q_H}+\frac{1-\rho}{q_L}-\frac{1}{\rho q_H+\left(1-\rho\right)q_L}\right)c>0
\end{aligned}
$$

In addition, we know that $\pi_p^{csep}=\pi_p^{nsep}$ when the anti-fake effort d is the same. As the profits under both equilibria monotonically decreases in d, thus, we have $\pi_p^{csep}\left(d_1\right)>\pi_p^{csep}\left(d_2\right)=\pi_p^{nsep}\left(d_2\right)$.

A.9. Proof of Proposition 1.

If $\eta_1\leq p_0$, the condition for the pooling equilibrium is not met, so the pooling equilibrium cannot exist. The only equilibrium is either costly-separating, if $d\leq\left(1-\tau\right)\left(\eta_2-p_0\right)$, or naturally-separating, otherwise. Since there is only one equilibrium under any condition, it is also an undefeated equilibrium, which is summarized in case (a).

When $\eta_1>p_0$, if $d\leq d_1$, by Lemma (4), the pooling coexists with and defeats the costly-separating equilibrium. When $d_1<d\leq d_2$, the costly-separating equilibrium is the sole equilibrium and thus undefeated equilibrium. When $d>d_2$, the naturally-separating equilibrium is the only remaining equilibrium and thus undefeated. Case (b) summarizes these undefeated equilibrium refinement outcomes.

A.10. Proof of Proposition 2.

The conclusions follow from the signs of the first-order derivatives: $\frac{\partial x_L^{pool}}{\partial l}=\frac{q_H-q_L}{q_Hq_L}c>0$; $\frac{\partial x_L^{pool}}{\partial p_0}=0$; $\frac{\partial x_L^{pool}}{\partial R}=-\frac{l}{q_L}c<0$; $\frac{\partial x_L^{pool}}{\partial\tau}=-\frac{q_H-q_L}{q_Hq_L}dlc_1<0$; $\frac{\partial x_L^{pool}}{\partial\lambda_i}=0$; $\frac{\partial x_L^{pool}}{\partial d}=l\frac{q_H-q_L}{q_Hq_L}c_1\left(1-\tau\right)>0$.

A.11. Proof of Proposition 3.

We firstly note that $\frac{\partial x_H^{csep}}{\partial l}=-\left(\frac{2\lambda_i\mu l}{p_f}+1\right)\frac{q_H-q_L}{q_Hq_L}c<0$; $\frac{\partial x_H^{csep}}{\partial p_0}=-\frac{\lambda_i\mu\left(1-l^2\right)}{p_f^2}\frac{q_H-q_L}{q_Hq_L}c<0$; $\frac{\partial x_H^{csep}}{\partial R}=-\frac{\lambda_i\mu\left(1-l^2\right)-p_fl}{p_f}\frac{1}{q_L}c\leq0$ (noting that $\frac{\lambda_i\mu\left(1-l^2\right)}{p_f}-l>0$ under costly-separating equilibrium); $\frac{\partial x_H^{csep}}{\partial\lambda_i}=\frac{\mu\left(1-l^2\right)}{p_f}\frac{q_H-q_L}{q_Hq_L}c>0$, and $\frac{\partial x_H^{csep}}{\partial\tau}=-\frac{q_H-q_L}{q_Hq_L}\left(\frac{\lambda_i\mu\left(1-l^2\right)dc}{p_f^2\left(1-\tau\right)^2}+\left(\frac{\lambda_i\mu\left(1-l^2\right)}{p_f}-l\right)c_1d\right)<0$.

$$
\frac{\partial x_H^{csep}}{\partial d}=\frac{q_H-q_L}{q_Hq_L}\frac{1}{p_f^2\left(1-\tau\right)}\left\{\left(\lambda_i\mu\left(1-l^2\right)-p_fl\right)c_1\left(1-\tau\right)^2p_f-\lambda_i\mu\left(1-l^2\right)c\right\}
$$

If $p_0\left[\lambda_i\mu\left(1-l^2\right)-p_0l\right]c_1\left(1-\tau\right)^2<\lambda_i\mu\left(1-l^2\right)c_0$, i.e., $c_1\left(1-\tau\right)^2\leq\frac{\lambda_i\mu\left(1-l^2\right)c_0}{p_0\left[\lambda_i\mu\left(1-l^2\right)-p_0l\right]}$ we have $\frac{\partial x_H^{csep}}{\partial d}<0$ for all $d\geq0$, thus, x_H^{csep} is monotonically decreasing with d.

If $c_1\left(1-\tau\right)^2>\frac{\lambda_i\mu\left(1-l^2\right)c_0}{p_0\left[\lambda_i\mu\left(1-l^2\right)-p_0l\right]}$, when $d\to0$, we have $\frac{\partial x_H^{csep}}{\partial d}>0$,

but when $d \to (1-\tau)\left[\frac{\lambda_i \mu\left(1-l^2\right)}{l} - p_0\right]$, we have $\frac{\partial x_H^{csep}}{\partial d} = -\frac{q_H - q_L}{q_H q_L}\frac{1}{p_f^2(1-\tau)}\lambda_i \mu\left(1-l^2\right)c < 0$. Thus, x_H^{csep} is not a monotonic function of d.

A.12. Proof of Proposition 4.

a) If $d_1 \leq 0$, the pooling equilibrium doesn't exist, and we only need to compare the platform's local optimal strategies under the costly- and naturally-separating equilibria to determine optimum d^*. Noting that because $\pi_p^{nsep}(d)$ and $U^{nsep}(d)$ decrease in d, the maximum naturally-separating equilibrium utility is $\Pi_p^{nsep}(d_2)$, which is the same as the $\Pi_p^{csep}(d_2)$. Similarly, $\Pi_p^{csep}(d)$ decreases in d, thus, the maximum costly-separating equilibrium utility is $\Pi_p^{csep}(0)$ and we have $\Pi_p^{csep}(0) > \Pi_p^{csep}(d_2) = \Pi_p^{nsep}(d_2)$. So, the platform's optimum utility is $\Pi_p^{csep}(0)$, achieved through a costly-separating equilibrium with $d^* = 0$.

b) If $d_1 > 0$, the platform can induce any of the three types of equilibria. Given that the utilities under the pooling, costly and naturally-separating equilibria are maximized at $d = 0$, $d = d_1$ and $d = d_2$, respectively. We have $\pi_p^{csep}(d_1) > \pi_p^{nsep}(d_2)$, and $U^{csep}(d_1) > U^{csep}(d_2) = U^{nsep}(d_2)$. Thus, we have $\Pi_p^{csep}(d_1) > \Pi_p^{csep}(d_2) = \Pi_p^{nsep}(d_2)$, so we only need to compare the platform's utility under pooling and costly-separating equilibria. Noting that both $\Pi_p^{pool}(d)$ and $\Pi_p^{csep}(d)$ decrease in d, the platform's optimum utility can be either $\Pi_p^{pool}(0)$, achieved through a pooling equilibrium with $d^* = 0$, or $\Pi_p^{csep}(d_1)$, achieved through a costly-separating equilibrium with $d^* = d_1$, whichever yields a higher utility.

The platform-optimal anti-fake effort and equilibrium types are summarized in Table 3.

A.13. Proof of Proposition 5.

1) When the platform-optimal effort is 0, and the pooling equilibrium arises,

$$U_{pool}^* = l\rho\frac{(q_H - c_0)^2}{2q_H} + l(1-\rho)\frac{(q_L - c_0)^2}{2q_L} + (1-l)\left[\rho\left(\frac{q_H}{2} - c_0 + \frac{2E[q] - q_H}{2E^2[q]}c_0^2\right) + (1-\rho)\left(\frac{q_L}{2} - c_0 + \frac{2E[q] - q_L}{2E^2[q]}c_0^2\right)\right]$$

We then have $\frac{\partial U_{pool}^*}{\partial l} = \rho\frac{(q_H - c_0)^2}{2q_H} + (1-\rho)\frac{(q_L - c_0)^2}{2q_L} - \rho\left(\frac{q_H}{2} - c_0 + \frac{2E[q] - q_H}{2E^2[q]}c_0^2\right) - (1-\rho)\left(\frac{q_L}{2} - c_0 + \frac{2E[q] - q_L}{2E^2[q]}c_0^2\right) > 0$ and $\frac{\partial U_{pool}^*}{\partial p_0} = 0$.

Because the platform exerts no effort, consumer welfare is unaffected by the technology level τ.

2) When the platform-optimal effort is 0, and the costly-separating equilibrium arises,

$$U_{csep}^* = \rho\frac{(q_H - c_0)^2}{2q_H} + (1-\rho)\frac{(q_L - c_0)^2}{2q_L}$$

We then have $\frac{\partial U_{csep}^*}{\partial l} = 0$ and $\frac{\partial U_{csep}^*}{0} = 0$. Again, the platform exerts no effort, and consumer welfare is unaffected by the technology level τ.

3) When the Platform-optimum is d_1 and the costly-separating equilibrium arises,

$$U^*_{csep} = U^*_{nsep} = \rho\frac{(q_H - c)^2}{2q_H} + (1-\rho)\frac{(q_L - c)^2}{2q_L}$$

where $c = c_0 + c_1(1-\tau)^2(\eta_1 - p_0)$. We have

$$\begin{aligned}\frac{\partial U^*_{csep}}{\partial l} &= \frac{\partial \pi^*_{csep}}{\partial c}\frac{\partial c}{\partial l} = -\left[\rho\frac{(q_H - c)}{q_H} + (1-\rho)\frac{(q_L - c)}{q_L}\right]c_1(1-\tau)^2\frac{\partial \eta_1}{\partial l} \\ &= \left[\rho\frac{(q_H - c)}{q_H} + (1-\rho)\frac{(q_L - c)}{q_L}\right]c_1(1-\tau)^2\frac{\lambda_i\mu\rho q_H}{E[q]\,l^2} > 0,\end{aligned}$$

$\frac{\partial U^*_{csep}}{\partial p_0} = \frac{\partial \pi^*_{csep}}{\partial c}\frac{\partial c}{\partial p_0} = \left[\rho\frac{(q_H - c)}{q_H} + (1-\rho)\frac{(q_L - c)}{q_L}\right]c_1(1-\tau)^2 > 0$, and $\frac{\partial U^*_{csep}}{\partial \tau} = \frac{\partial \pi^*_{csep}}{\partial c}\frac{\partial c}{\partial \tau} = 2\left[\rho\frac{(q_H - C)}{q_H} + (1-\rho)\frac{(q_L - c)}{q_L}\right]c_1(1-\tau)(\eta_1 - p_0) > 0$.

A.14. Proof of Proposition 6.

a) If $d_1 \le 0$, $d^C = 0$. By Proposition 4, the platform's optimal anti-fake effort is $d^* = 0$ as well. So, we have $d^* = d^C$.

b) If $d_1 > 0$, the consumer-optimal d^C can be either 0 when $U^{csep}(d_1) \le U^{pool}(0)$, or d_1 otherwise. (1) When $U^{csep}(d_1) \le U^{pool}(0)$, the equilibrium obtained under the consumer-optimal $d^C = 0$ is pooling equilibrium. In this case, by Lemma 8, we have $\pi_p^{csep}(d_1) \le \pi_p^{pool}(0)$, thus, $\Pi_p^{csep}(d_1) \le \Pi_p^{pool}(0)$, and the equilibrium obtained under the platform-optimal $d^* = 0$ is pooling equilibrium as well. Thus, we have $d^* = d^C = 0$. (2) When $U^{csep}(d_1) > U^{pool}(0)$, the equilibrium obtained under the consumer-optimal $d^C = d_1$ is the costly-separating equilibrium. By numeric simulation, we can find examples for both cases $\Pi_p^{pool}(0) \ge \Pi_p^{nsep}(d_2)$, and $\Pi_p^{pool}(0) < \Pi_p^{nsep}(d_2)$. Thus, we have $d^* = 0 < d^C = d_1$ when $\Pi_p^{pool}(0) \ge \Pi_p^{csep}(d_1)$, or $d^* = d^C = d_1$ when $\Pi_p^{pool}(0) < \Pi_p^{csep}(d_1)$.

A.15. Impact of fake-account base price on platform profit, anti-fake effort, and utility

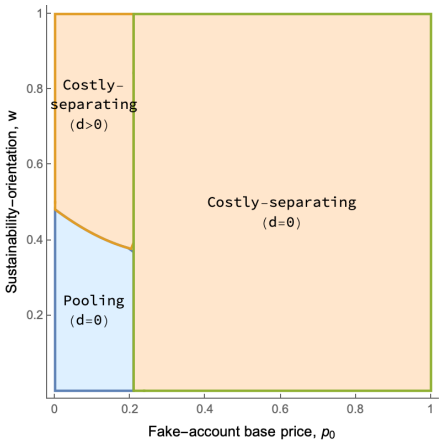
See Figure 8

B. Proof of Proposition 7 (Three Types of Influencers).

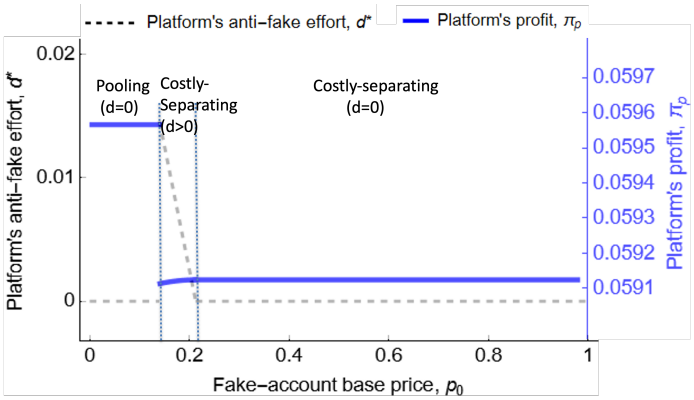
B.1. Costly Fully Separating

For the fully separating equilibrium, we show that $(n_2^{H*}, n_2^{M*}, n_2^{L*}) = \left(n_2^{sepH}, n_2^{sepM}, n_{in}^L\right)$ with $n_2^{sepM} > n_{in}^M$ and $n_2^{sepH} > n_{in}^H$ (i.e., the L-type does not buy, the M- and H-type buy enough to maintain separation) and the following belief is a PBE.

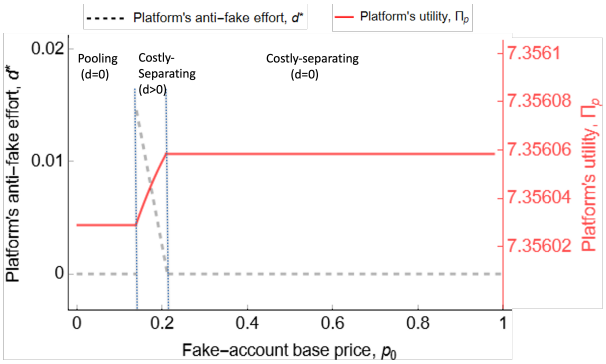
$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_2^{sepH} \\ 1, & \text{if } n_2 \ge n_2^{sepH} \end{cases}$$



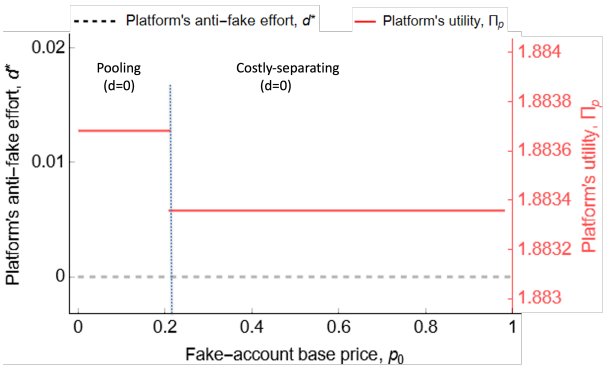
(a) Platform's optimal strategy



(b) Platform's optimal effort and profit ($w = 0.4$)



(c) Platform's optimal effort and utility ($w = 0.4$)



(d) Platform's optimal effort and utility ($w = 0.1$)

Note: $\mu = 0.2, \rho = 0.3, \lambda_i = 0.3, \lambda_p = 0.3, l = 0.2, q_H = 100, q_L = 10, c_0 = 0.2, c_1 = 0.02, \tau = 0.8, \gamma = 0.1$

Figure 8 Impact of fake-account base price on the platform

$$P(M|n_2) = \begin{cases} 0, & \text{if } n_2 < n_2^{sepM} \\ 1, & \text{if } n_2^{sepM} \le n_2 < n_2^{sepH} \\ 0, & \text{if } n_2 \ge n_2^{sepH} \end{cases}$$

$$P(L|n_2) = \begin{cases} 1, & \text{if } n_2 < n_2^{sepM} \\ 0, & \text{if } n_2 \ge n_2^{sepM} \end{cases}$$

Similar to the argument made in the Proof of Lemma 1, we argue that the H-type will not deviate to $n_2^H > n_2^{sepH}$. Similarly, the M-type will not deviate to $n_2^M \in (n_2^{sepM}, n_2^{sepH})$, the L-type will not deviate to $n_2^L \in (n_{in}^L, n_2^{sepM})$.

1) H-type Influencer

If the H-type deviates to $n_2^{H\prime} \in [n_2^{sepM}, n_2^{sepH})$, she will be viewed as an M-type by both the uninformed consumers and the uninformed advertiser (and an H-type by an informed advertiser). Thus, her expected profit is

$$\pi_H^{dev}\left(n_2^{H\prime}\right) = l\lambda_i\mu n_{r,ia}^{H\prime} + (1-l)\lambda_i\mu n_{r,ua}^{H\prime} - p_f\left(n_2^{H\prime} - n_{in}^H\right)$$

where $n_{r,ia}^{H\prime} = n_{in}^{H} + n_{un}^{M}$, $n_{r,ua}^{H\prime} = n_{r}^{M} = n_{in}^{M} + n_{un}^{M}$. Clearly, the H-type is better off with $n_2^{H\prime} = n_2^{sepM}$, resulting in a profit of $\pi_H^{dev}\left(n_2^{sepM}\right) = l\lambda_i\mu\left(n_{in}^{H} + n_{un}^{M}\right) + (1-l)\,\lambda_i\mu n_r^M - p_f\left(n_2^{sepM} - n_{in}^{H}\right)$.

The IC condition requires $\pi_H^{dev}\left(n_2^{sepM}\right) \leq \pi_H^{sep*} = \lambda_i\mu n_r^H - p_f\left(n_2^{sepH} - n_{in}^{H}\right)$, which translates to:

$$l\lambda_i\mu\left(n_{in}^{H} + n_{un}^{M}\right) + (1-l)\,\lambda_i\mu\left(n_{in}^{M} + n_{un}^{M}\right) - p_f\left(n_2^{sepM} - n_{in}^{H}\right) \leq \lambda_i\mu n_r^H - p_f\left(n_2^{sepH} - n_{in}^{H}\right) \tag{65}$$

After simplification, we have

$$p_f\left(n_2^{sepH} - n_2^{sepM}\right) \leq \lambda_i\mu\left[(1-l)\left(n_{in}^{H} - n_{in}^{M}\right) + \left(n_{un}^{H} - n_{un}^{M}\right)\right]$$

If the H-type purchases fewer than x_H^* such that $n_2^{H\prime\prime} < n_2^{sepM}$ (say $x_H'' < n_2^{sepM} - n_{in}^{H}$), she will be viewed as an L-type by both the uninformed consumers and the uninformed advertiser (and an H-type by an informed advertiser). Thus, her expected profit is

$$\pi_H^{dev}\left(n_2^{H\prime\prime}\right) = l\lambda_i\mu n_{r,ia}^{H\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{H\prime\prime} - p_f\left(n_2^{H\prime\prime} - n_{in}^{H}\right)$$

where $n_{r,ia}^{H\prime\prime} = n_{in}^{H} + n_{un}^{L}$, $n_{r,ua}^{H\prime\prime} = n_{r}^{L} = n_{in}^{L} + n_{un}^{L}$. Clearly, her best deviation of this type is not to purchase any fake account and the resulting expected profit is $\pi_H^{dev}\left(n_2^{H\prime\prime}\right) = l\lambda_i\mu n_{r,ia}^{H\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{H\prime\prime}$. The IC condition requires $\pi_H^{dev}\left(n_2^{H\prime\prime}\right) \leq \pi_H^{*sep}$, which translates to:

$$l\lambda_i\mu n_{r,ia}^{H\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{H\prime\prime} \leq \lambda_i\mu n_r^H - p_f\left(n_2^{sepH} - n_{in}^{H}\right) \tag{66}$$

By simplification, we have

$$p_f\left(n_2^{sepH} - n_{in}^{H}\right) \leq \lambda_i\mu\left[(1-l)\left(n_{in}^{H} - n_{in}^{L}\right) + \left(n_{un}^{H} - n_{un}^{L}\right)\right]$$

where $n_r^H = n_{in}^H + n_{un}^H$, $n_r^M = n_{in}^M + n_{un}^M$, and $n_r^L = n_{in}^L + n_{un}^L$

This IR condition for the H-type can be naturally satisfied under condition (66).

Thus, we have

$$\begin{cases} n_2^{sepH} \leq \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right)}{p_f} + n_{in}^H \\ n_2^{sepH} - n_2^{sepM} \leq \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^M\right)}{p_f} \end{cases}$$

2) M-type influencers

If the M-type purchases more than x_M^* such that $n_2^{M\prime} \geq n_2^{sepH}$(say $x_M' \geq n_2^{sepH} - n_{in}^M$), she will be viewed as an H-type by both the uninformed consumers and the uninformed advertiser (and an M-type by an informed advertiser). Thus, her expected profit is

$$\pi_M^{dev}\left(n_2^{M\prime}\right) = l\lambda_i\mu n_{r,ia}^{M\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{M\prime} - p_f\left(n_2^{M\prime} - n_{in}^{M}\right)$$

where $n_{r,ia}^{M\prime} = n_{in}^{M} + n_{un}^{H}$, $n_{r,ua}^{M\prime} = n_{r}^{H} = n_{in}^{H} + n_{un}^{H}$. Clearly, the M-type is better off with $n_2^{M\prime} = n_2^{sepH}$, resulting in a profit of $\pi_M^{dev}\left(n_2^{sepH}\right) = l\lambda_i\mu\left(n_{in}^{M} + n_{un}^{H}\right) + (1-l)\,\lambda_i\mu n_r^H - p_f\left(n_2^{sepH} - n_{in}^{M}\right)$.

The IC condition requires $\pi_M^{dev}\left(n_2^{sepH}\right) \leq \pi_M^{*sep} = \lambda_i \mu n_r^M - p_f\left(n_2^{sepM} - n_{in}^M\right)$, which translates to:

$$l\lambda_i\mu\left(n_{in}^M + n_{un}^H\right) + (1-l)\,\lambda_i\mu n_r^H - p_f\left(n_2^{sepH} - n_{in}^M\right) \leq \lambda_i\mu n_r^M - p_f\left(n_2^{sepM} - n_{in}^M\right) \tag{67}$$

By simplification, we have

$$p_f\left(n_2^{sepH} - n_2^{sepM}\right) \geq \lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^M\right)$$

If the M-type purchases fewer than x_M^* such that $n_2^{M\prime\prime} < n_2^{sepM}$ (say $x_M'' < n_2^{sepM} - n_{in}^M$), she will be viewed as an L-type by both the uninformed consumers and the uninformed advertiser (and an M-type by an informed advertiser). Thus, her expected profit is

$$\pi_M^{dev}\left(n_2^{M\prime\prime}\right) = l\lambda_i\mu n_{r,ia}^{M\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{M\prime\prime} - p_f\left(n_2^{M\prime\prime} - n_{in}^M\right)$$

where $n_{r,ia}^{M\prime\prime} = n_{in}^M + n_{un}^L$, $n_{r,ua}^{M\prime\prime} = n_r^L = n_{in}^L + n_{un}^L$. Clearly, her best deviation of this type is not to purchase any fake account and the resulting expected profit is $\pi_M^{dev}\left(n_2^M\right) = l\lambda_i\mu n_{r,ia}^{M\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{M\prime\prime}$. The IC condition requires $\pi_M^{dev}\left(n_2^M\right) \leq \pi_M^{*sep}$, which translates to:

$$p_f\left(n_2^{sepM} - n_{in}^M\right) \leq \lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right) \tag{68}$$

Again, the IR condition for the M-type can be naturally satisfied under condition (68)

Thus, we have

$$\begin{cases} n_2^{sepM} \leq \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right)}{p_f} + n_{in}^M \\ n_2^{sepH} - n_2^{sepM} \geq \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^M\right)}{p_f} \end{cases}$$

3) L-type influencer

If the L-type purchases more than x_L^* such that $n_2^{sepM} \leq n_2^{L\prime} < n_2^{sepH}$(say $n_2^{sepM} - n_{in}^L \leq x_L' < n_2^{sepH} - n_{in}^L$), she will be viewed as an M-type by both the uninformed consumers and the uninformed advertiser (and an L-type by an informed advertiser). Thus, her expected profit is

$$\pi_L^{dev}\left(n_2^{L\prime}\right) = l\lambda_i\mu n_{r,ia}^{L\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{L\prime} - p_f\left(n_2^{L\prime} - n_{in}^L\right)$$

where $n_{r,ia}^{L\prime} = n_{in}^L + n_{un}^M$, $n_{r,ua}^{L\prime} = n_r^M = n_{in}^M + n_{un}^M$. Clearly, the L-type is better off with $n_2^{L\prime} = n_2^{sepM}$, resulting in a profit of $\pi_L^{dev}\left(n_2^{sepM}\right) = l\lambda_i\mu\left(n_{in}^L + n_{un}^M\right) + (1-l)\,\lambda_i\mu n_r^M - p_f\left(n_2^{sepM} - n_{in}^L\right)$.

the IC condition requires $\pi_L^{dev}\left(n_2^{sepM}\right) \leq \pi_L^{*sep} = \lambda_i\mu n_r^L$, which translates to:

$$p_f\left(n_2^{sepM} - n_{in}^L\right) \geq \lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right) \tag{69}$$

If the L-type purchases more than x_L^* such that $n_2^{L\prime\prime} \geq n_2^{sepH}$(say $x_L'' \geq n_2^{sepH} - n_{in}^L$), she will be viewed as an H-type by both the uninformed consumers and the uninformed advertiser (and an L-type by an informed advertiser). Thus, her expected profit is

$$\pi_L^{dev}\left(n_2^{L\prime\prime}\right) = l\lambda_i\mu n_{r,ia}^{L\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{L\prime\prime} - p_f\left(n_2^{M\prime\prime} - n_{in}^M\right)$$

where $n_{r,ia}^{L\prime\prime} = n_{in}^{L} + n_{un}^{H}$, $n_{r,ua}^{L\prime\prime} = n_{r}^{H} = n_{in}^{H} + n_{un}^{H}$. Clearly, the L-type is better off with $n_2^{L\prime\prime} = n_2^{sepH}$, resulting in a profit of $\pi_L^{dev}\left(n_2^{sepH}\right) = l\lambda_i\mu\left(n_{in}^{L} + n_{un}^{H}\right) + (1-l)\,\lambda_i\mu n_r^H - p_f\left(n_2^{sepH} - n_{in}^{L}\right)$. the IC condition requires $\pi_L^{dev}\left(n_2^{sepH}\right) \le \pi_L^{*sep} = \lambda_i\mu n_r^L$, which translates to:

$$p_f\left(n_2^{sepH} - n_{in}^{L}\right) \ge \lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right) \tag{70}$$

The IR condition for the L-type can be naturally satisfied.

Combing the IC conditions for the L-type, we have

$$\begin{cases} n_2^{sepM} \ge \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right)}{p_f} + n_{in}^{L} \\ n_2^{sepH} \ge \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right)}{p_f} + n_{in}^{L} \end{cases}$$

4) Combining all conditions above

Combing the IC and IR conditions for H-type, M-type, and L-type, we have

$$\begin{cases} n_2^{sepH} \in \left[\frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right)}{p_f} + n_{in}^{L}, \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right)}{p_f} + n_{in}^{H}\right] \\ n_2^{sepM} \in \left[\frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right)}{p_f} + n_{in}^{L}, \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right)}{p_f} + n_{in}^{M}\right] \\ n_2^{sepH} - n_2^{sepM} = \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^M\right)}{p_f} \end{cases}$$

As a range of separating equilibria exist, we apply the *undefeated equilibrium refinement* and obtain

$$\begin{aligned} \left\{n_2^{H*}, n_2^{M*}, n_2^{L*}\right\} &= \left\{n_2^{*sepH}, n_2^{*sepM}, n_2^{L*}\right\} \\ &= \left\{\frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right)}{p_f} + n_{in}^{L}, \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right)}{p_f} + n_{in}^{L}, n_{in}^{L}\right\} \end{aligned}$$

under the belief system we have defined for this case.

As $n_2^{*sepH} \ge n_{in}^{H}$ and $n_2^{*sepM} \ge n_{in}^{M}$, in this case, thus, we also should have

$$\begin{cases} \lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right) \ge p_f\left(n_{in}^{H} - n_{in}^{L}\right) \\ \lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right) \ge p_f\left(n_{in}^{M} - n_{in}^{L}\right) \end{cases}$$

Finally, we can obtain

$$\begin{cases} n_2^{H*} = \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right)}{p_f} + n_{in}^{L} \\ n_2^{M*} = \frac{\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right)}{p_f} + n_{in}^{L} \\ n_2^{L*} = n_{in}^{L} \end{cases} \tag{71}$$

under the condition $\lambda_i\mu\left(1-l^2\right) \ge p_f l$

B.2. Naturally Fully Separating

In this case, we show that $(n_2^{H*}, n_2^{M*}, n_2^{L*}) = (n_{in}^H, n_{in}^M, n_{in}^L)$ (i.e., none of the three types buys fake accounts) with the following belief is a PBE.

$$P(H|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ 1, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

$$P(M|n_2) = \begin{cases} 0, & \text{if } n_2 < n_{in}^M \\ 1, & \text{if } n_{in}^M \leq n_2 < n_{in}^H \\ 0, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

$$P(L|n_2) = \begin{cases} 1, & \text{if } n_2 < n_{in}^M \\ 0, & \text{if } n_2 \geq n_{in}^M \end{cases}$$

First, we argue that the H-type will not buy more than x_H^* to make $n_2^H > n_{in}^H$. Also, the M-type will not purchase more than x_M^* such that $n_{in}^M < n_2^M < n_{in}^H$, the L-type will not purchase more than x_L^* such that $n_{in}^L < n_2^L < n_{in}^M$.

1) M-type influencer

If the M-type purchases more than x_M^* such that $n_2^{M\prime} \geq n_{in}^H$ (say $x'_M \geq n_{in}^H - n_{in}^M$), she will be viewed as an H-type by both the uninformed consumers and the uninformed advertiser (and an M-type by an informed advertiser). Thus, her expected profit is

$$\pi_M^{dev}\left(n_2^{M\prime}\right) = l\lambda_i\mu n_{r,ia}^{M\prime} + (1-l)\lambda_i\mu n_{r,ua}^{M\prime} - p_f\left(n_2^{M\prime} - n_{in}^M\right)$$

where $n_{r,ia}^{M\prime} = n_{in}^M + n_{un}^H$, $n_{r,ua}^{M\prime} = n_r^H = n_{in}^H + n_{un}^H$. Clearly, the M-type is better off with $n_2^{M\prime} = n_{in}^H$, resulting in a profit of $\pi_M^{dev}(n_{in}^H) = l\lambda_i\mu(n_{in}^M + n_{un}^H) + (1-l)\lambda_i\mu n_r^H - p_f(n_{in}^H - n_{in}^M)$.

The IC condition requires $\pi_M^{dev}(n_{in}^H) \leq \pi_M^{*sep} = \lambda_i\mu n_r^M$, which translates to:

$$\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^M\right) \leq p_f\left(n_{in}^H - n_{in}^M\right) \tag{72}$$

2) L-type influencer

If the L-type purchases more than x_L^* such that $n_{in}^M \leq n_2^{L\prime} < n_{in}^H$ (say $n_{in}^M - n_{in}^L \leq x'_L < n_{in}^H - n_{in}^L$), she will be viewed as an M-type by both the uninformed consumers and the uninformed advertiser (and an L-type by an informed advertiser). Thus, her expected profit is

$$\pi_L^{dev}\left(n_2^{L\prime}\right) = l\lambda_i\mu n_{r,ia}^{L\prime} + (1-l)\lambda_i\mu n_{r,ua}^{L\prime} - p_f\left(n_2^{L\prime} - n_{in}^L\right)$$

where $n_{r,ia}^{L\prime} = n_{in}^L + n_{un}^M$, $n_{r,ua}^{L\prime} = n_r^M = n_{in}^M + n_{un}^M$. Clearly, the L-type is better off with $n_2^{L\prime} = n_{in}^M$, resulting in a profit of $\pi_L^{dev}(n_{in}^M) = l\lambda_i\mu(n_{in}^L + n_{un}^M) + (1-l)\lambda_i\mu n_r^M - p_f(n_{in}^M - n_{in}^L)$. the IC condition requires $\pi_L^{dev}(n_{in}^M) \leq \pi_L^{*sep} = \lambda_i\mu n_r^L$, which translates to:

$$\lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right) \leq p_f\left(n_{in}^M - n_{in}^L\right) \tag{73}$$

If the L-type purchases more than x_L^* such that $n_2^{L\prime\prime} \geq n_{in}^H$ (say $x_L'' \geq n_{in}^H - n_{in}^L$), she will be viewed as an H-type by both the uninformed consumers and the uninformed advertiser (and an L-type by an informed advertiser). Thus, her expected profit is

$$\pi_L^{dev}\left(n_2^{L\prime\prime}\right) = l\lambda_i\mu n_{r,ia}^{L\prime\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{L\prime\prime} - p_f\left(n_2^{M\prime\prime} - n_{in}^M\right)$$

where $n_{r,ia}^{L\prime\prime} = n_{in}^L + n_{un}^H$, $n_{r,ua}^{L\prime\prime} = n_r^H = n_{in}^H + n_{un}^H$. Clearly, the L-type is better off with $n_2^{L\prime\prime} = n_{in}^H$, resulting in a profit of $\pi_L^{dev}\left(n_{in}^H\right) = l\lambda_i\mu\left(n_{in}^L + n_{un}^H\right) + (1-l)\,\lambda_i\mu n_r^H - p_f\left(n_{in}^H - n_{in}^L\right)$. the IC condition requires $\pi_L^{dev}\left(n_{in}^H\right) \leq \pi_L^{*sep} = \lambda_i\mu n_r^L$, which translates to:

$$\lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right) \leq p_f\left(n_{in}^H - n_{in}^L\right) \tag{74}$$

3) Combining all conditions above

Combing the IC conditions for H-type, M-type, and L-type, we have

$$\begin{cases} \lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^M\right) \leq p_f\left(n_{in}^H - n_{in}^M\right) \\ \lambda_i\mu\left(1-l^2\right)\left(n_r^M - n_r^L\right) \leq p_f\left(n_{in}^M - n_{in}^L\right) \\ \lambda_i\mu\left(1-l^2\right)\left(n_r^H - n_r^L\right) \leq p_f\left(n_{in}^H - n_{in}^L\right) \end{cases}$$

The three conditions can be simplified as one condition: $\lambda_i\mu\left(1-l^2\right) < p_f l$

For this equilibrium to hold, the individual rational (IR) condition for the influencers and the advertiser can be naturally satisfied.

Finally, we can obtain

$$\begin{cases} n_2^{H*} = n_{in}^H \\ n_2^{M*} = n_{in}^M \\ n_2^{L*} = n_{in}^L \end{cases} \tag{75}$$

under the condition $\lambda_i\mu\left(1-l^2\right) < p_f l$

B.3. Hybrid with $M-L$ Pooling

In this case, we show that $\left(n_2^{H*}, n_2^{M*}, n_2^{L*}\right) = \left(n_{in}^H, n_{in}^M, n_{in}^M\right)$ (i.e., neither H- nor M-type buys fake accounts, and the L-type buys to mimic the M-type) with the following belief is a PBE.

$$P\left(H|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ 1, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

$$P\left(M|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^M \\ \frac{\rho_M}{\rho_M+\rho_L}, & \text{if } n_{in}^M \leq n_2 < n_{in}^H \\ 0, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

$$P\left(L|n_2\right) = \begin{cases} 1, & \text{if } n_2 < n_{in}^M \\ \frac{\rho_L}{\rho_M+\rho_L}, & \text{if } n_{in}^M \leq n_2 < n_{in}^H \\ 0, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

We can write the equilibrium profits for the three types as

$$\pi_H^{sep} = \lambda_i \mu n_r^H \tag{76}$$

$$\pi_M^{pool} = l\lambda_i \mu n_{r,ia}^M + (1-l)\,\lambda_i \mu n_{r,ua}^{ML} \tag{77}$$

$$\pi_L^{pool} = l\lambda_i \mu n_{r,ia}^L + (1-l)\,\lambda_i \mu n_{r,ua}^{ML} - p_f x_L^{pool} \tag{78}$$

where

$$\begin{aligned}
n_{r,ia}^M &= n_{in}^M + n_{un}^{ML} \\
n_{r,ia}^L &= n_{in}^L + n_{un}^{ML} \\
n_{r,ua}^{ML} &= n_{in}^{ML} + n_{un}^{ML} \\
n_{in}^{ML} &= \frac{\rho_M}{\rho_M + \rho_L} n_{in}^M + \frac{\rho_L}{\rho_M + \rho_L} n_{in}^L \\
E\left[q_{ML}\right] &= \frac{\rho_M}{\rho_M + \rho_L} q_M + \frac{\rho_L}{\rho_M + \rho_L} q_L \\
n_{un}^{ML} &= (1-l)\left(1 - \frac{c}{E\left[q_{ML}\right]}\right)
\end{aligned}$$

We argue that the H-type will not deviate to $n_2^H > n_{in}^H$. Also, the M-type and L-type will not deviate to $n_2 \in (n_{in}^M, n_{in}^H)$.

1) M-type influencer

If the M-type purchases more than x_M^* such that $n_2^{M\prime} \geq n_{in}^H$ (say $x_M' \geq n_{in}^H - n_{in}^M$), she will be viewed as an H-type by both the uninformed consumers and the uninformed advertiser (and an M-type by an informed advertiser). Thus, her expected profit is

$$\pi_M^{dev}\left(n_2^{M\prime}\right) = l\lambda_i \mu n_{r,ia}^{M\prime} + (1-l)\,\lambda_i \mu n_{r,ua}^{M\prime} - p_f\left(n_2^{M\prime} - n_{in}^M\right)$$

where $n_{r,ia}^{M\prime} = n_{in}^M + n_{un}^H$, $n_{r,ua}^{M\prime} = n_r^H = n_{in}^H + n_{un}^H$. Clearly, the M-type is better off with $n_2^{M\prime} = n_{in}^H$, resulting in a profit of $\pi_M^{dev}\left(n_{in}^H\right) = l\lambda_i \mu\left(n_{in}^M + n_{un}^H\right) + (1-l)\,\lambda_i \mu n_r^H - p_f\left(n_{in}^H - n_{in}^M\right)$.

The IC condition requires $\pi_M^{dev}\left(n_{in}^H\right) \leq \pi_M^{pool} = l\lambda_i \mu n_{r,ia}^M + (1-l)\,\lambda_i \mu n_{r,ua}^{ML}$, which translates to:

$$\lambda_i \mu\left[(1-l)\left(n_{in}^H - n_{in}^{ML}\right) + \left(n_{un}^H - n_{un}^{ML}\right)\right] \leq p_f\left(n_{in}^H - n_{in}^M\right) \tag{79}$$

The IR condition for the M-type can be automatically satisfied.

2) L-type influencer

If the L-type purchases more than x_L^* such that $n_2^{L\prime} \geq n_{in}^H$ (say $x_L' \geq n_{in}^H - n_{in}^L$), she will be viewed as an H-type by both the uninformed consumers and the uninformed advertiser (and an L-type by an informed advertiser). Thus, her expected profit is

$$\pi_L^{dev}\left(n_2^{L\prime}\right) = l\lambda_i \mu n_{r,ia}^{L\prime} + (1-l)\,\lambda_i \mu n_{r,ua}^{L\prime} - p_f\left(n_2^{L\prime} - n_{in}^L\right)$$

where $n_{r,ia}^{L\prime} = n_{in}^{L} + n_{un}^{H}$, $n_{r,ua}^{L\prime} = n_{r}^{H} = n_{in}^{H} + n_{un}^{H}$. Clearly, the L-type is better off with $n_2^{L\prime} = n_{in}^{H}$, resulting in a profit of $\pi_L^{dev}\left(n_{in}^{H}\right) = l\lambda_i\mu\left(n_{in}^{L} + n_{un}^{H}\right) + (1-l)\,\lambda_i\mu n_r^H - p_f\left(n_{in}^{H} - n_{in}^{L}\right)$. The IC condition requires $\pi_L^{dev}\left(n_{in}^{H}\right) \leq \pi_L^{pool} = l\lambda_i\mu n_{r,ia}^{L} + (1-l)\,\lambda_i\mu n_{r,ua}^{ML} - p_f\left(n_{in}^{M} - n_{in}^{L}\right)$, which translates to:

$$\lambda_i\mu\left[(1-l)\left(n_{in}^{H} - n_{in}^{ML}\right) + \left(n_{un}^{H} - n_{un}^{ML}\right)\right] \leq p_f\left(n_{in}^{H} - n_{in}^{M}\right) \tag{80}$$

If the L-type purchases fewer than x_L^* such that $n_2^{L\prime\prime} < n_{in}^{M}$ (say $0 \leq x_L'' < n_{in}^{M} - n_{in}^{L}$), she will be viewed as an L-type by both informed and uninformed consumers/advertisers. Her best deviation of this type is not to buy any fake account, and the resulting expected profit is

$$\pi_L^{dev}\left(n_{in}^{L}\right) = \lambda_i\mu n_r^L$$

The IC condition requires $\pi_L^{dev}\left(n_{in}^{L}\right) \leq \pi_L^{pool} = l\lambda_i\mu n_{r,ia}^{L} + (1-l)\,\lambda_i\mu n_{r,ua}^{ML} - p_f\left(n_{in}^{M} - n_{in}^{L}\right)$, which translates to:

$$\lambda_i\mu\left[(1-l)\left(n_{in}^{ML} - n_{in}^{L}\right) + \left(n_{un}^{ML} - n_{un}^{L}\right)\right] \geq p_f\left(n_{in}^{M} - n_{in}^{L}\right) \tag{81}$$

3) Combing the IC and IR conditions for the L-type, we have

$$\begin{cases} \lambda_i\mu\left[(1-l)\left(n_{in}^{H} - n_{in}^{ML}\right) + \left(n_{un}^{H} - n_{un}^{ML}\right)\right] \leq p_f\left(n_{in}^{H} - n_{in}^{M}\right) \\ \lambda_i\mu\left[(1-l)\left(n_{in}^{ML} - n_{in}^{L}\right) + \left(n_{un}^{ML} - n_{un}^{L}\right)\right] \geq p_f\left(n_{in}^{M} - n_{in}^{L}\right) \end{cases}$$

When the L-type's IC condition holds, her IR condition and M-type's IC condition are automatically satisfied.

Finally, we can obtain

$$\begin{cases} n_2^{H*} = n_{in}^{H} \\ n_2^{M*} = n_{in}^{M} \\ n_2^{L*} = n_{in}^{M} \end{cases} \tag{82}$$

under the condition

$$\begin{cases} \lambda_i\mu\left[(1-l)\left(n_{in}^{H} - n_{in}^{ML}\right) + \left(n_{un}^{H} - n_{un}^{ML}\right)\right] \leq p_f l\left(\frac{c}{q_M} - \frac{c}{q_H}\right) \\ \lambda_i\mu\left[(1-l)\left(n_{in}^{ML} - n_{in}^{L}\right) + \left(n_{un}^{ML} - n_{un}^{L}\right)\right] \geq p_f l\left(\frac{c}{q_L} - \frac{c}{q_M}\right) \end{cases}$$

B.4. Hybrid with $H - M$ Pooling

In this case, we show that $\left(n_2^{H*}, n_2^{M*}, n_2^{L*}\right) = \left(n_{in}^{H}, n_{in}^{H}, n_{in}^{L}\right)$ (i.e., neither H- nor L-type buys, but the M-type buys to mimic the H-type) with the following belief is a PBE.

$$P\left(H|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^{H} \\ \frac{\rho_H}{\rho_H + \rho_M}, & \text{if } n_2 \geq n_{in}^{H} \end{cases}$$

$$P\left(M|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^{H} \\ \frac{\rho_M}{\rho_H + \rho_M}, & \text{if } n_2 \geq n_{in}^{H} \end{cases}$$

$$P\left(L|n_2\right)=\begin{cases}1, & \text{if } n_2<n_{in}^H\\ 0, & \text{if } n_2\geq n_{in}^H\end{cases}$$

We can write the equilibrium profits for the three types as

$$\pi_H^{pool}=l\lambda_i\mu n_{r,ia}^H+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HM} \tag{83}$$

$$\pi_M^{pool}=l\lambda_i\mu n_{r,ia}^M+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HM}-p_f x_M^{pool} \tag{84}$$

$$\pi_L^{sep}=\lambda_i\mu n_r^L \tag{85}$$

where

$$\begin{aligned}
n_{r,ia}^H &= n_{in}^H+n_{un}^{HM}\\
n_{r,ia}^M &= n_{in}^M+n_{un}^{HM}\\
n_{r,ua}^{HM} &= n_{in}^{HM}+n_{un}^{HM}\\
n_{in}^{HM} &= \frac{\rho_H}{\rho_H+\rho_M}n_{in}^H+\frac{\rho_M}{\rho_H+\rho_M}n_{in}^M\\
E\left[q_{HM}\right] &= \frac{\rho_H}{\rho_H+\rho_M}q_H+\frac{\rho_M}{\rho_H+\rho_M}q_M\\
n_{un}^{HM} &= \left(1-l\right)\left(1-\frac{c}{E\left[q_{HM}\right]}\right)
\end{aligned}$$

First, we argue that the H-type will not deviate to $n_2^H>n_{in}^H$. The M-type will not deviate to $n_2^M>n_{in}^H$. Also, the L-type will not deviate to $n_2^L\in(n_{in}^L,n_{in}^H)$.

1) M-type influencer

If the M-type purchases fewer than x_M^* such that $n_{in}^M\leq n_2^{M\prime}<n_{in}^H$ (say $0\leq x_M'\leq n_{in}^H-n_{in}^M$), she will be viewed as an L-type by both the uninformed consumers and the uninformed advertiser (and an M-type by an informed advertiser). Thus, her expected profit is

$$\pi_M^{dev}\left(n_2^{M\prime}\right)=l\lambda_i\mu n_{r,ia}^{M\prime}+\left(1-l\right)\lambda_i\mu n_{r,ua}^{M\prime}-p_f\left(n_2^{M\prime}-n_{in}^M\right)$$

where $n_{r,ia}^{M\prime}=n_{in}^M+n_{un}^L$, $n_{r,ua}^{M\prime}=n_r^L=n_{in}^L+n_{un}^L$. Clearly, her best deviation of this type is not to purchase any fake account and the resulting expected profit is $\pi_M^{dev}\left(n_2^M\right)=l\lambda_i\mu n_{r,ia}^{M\prime}+\left(1-l\right)\lambda_i\mu n_{r,ua}^{M\prime}$. The IC condition requires

$$\pi_M^{dev}\left(n_2^M\right)\leq\pi_M^{pool}=l\lambda_i\mu n_{r,ia}^M+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HM}-p_f\left(n_{in}^H-n_{in}^M\right)$$

which translates to:

$$\lambda_i\mu\left[\left(1-l\right)\left(n_{in}^{HM}-n_{in}^L\right)+\left(n_{un}^{HM}-n_{un}^L\right)\right]\geq p_f\left(n_{in}^H-n_{in}^M\right) \tag{86}$$

The IR condition for the M-type can be naturally satisfied under condition (86)

2) L-type influencer

If the L-type purchases more than x_L^* to achieve a follower count $n_2^{L\prime}$ higher than n_{in}^H (say $x_L' \geq n_{in}^H - n_{in}^L$), she will be seen as an average type of H- and M-type by both the uninformed consumers and advertisers (and an L-type by an informed advertiser). Thus, her expected profit is

$$\pi_L^{dev}\left(n_2^{L\prime}\right) = l\lambda_i\mu n_{r,ia}^{L\prime} + (1-l)\,\lambda_i\mu n_{r,ua}^{L\prime} - p_f\left(n_2^{L\prime} - n_{in}^L\right)$$

where $n_{r,ia}^{L\prime} = n_{in}^L + n_{un}^{HM}$, $n_{r,ua}^{L\prime} = n_{r,ua}^{HM} = n_{in}^{HM} + n_{un}^{HM}$. Clearly, the L-type is better off with $n_2^{L\prime} = n_{in}^H$, resulting in a profit of $\pi_L^{dev}\left(n_{in}^H\right) = l\lambda_i\mu\left(n_{in}^L + n_{un}^{HM}\right) + (1-l)\,\lambda_i\mu\left(n_{in}^{HM} + n_{un}^{HM}\right) - p_f\left(n_{in}^H - n_{in}^L\right)$.

The IC condition requires $\pi_L^{dev}\left(n_{in}^H\right) \leq \pi_L^{sep} = \lambda_i\mu n_r^L$, which translates to:

$$\lambda_i\mu\left[(1-l)\left(n_{in}^{HM} - n_{in}^L\right) + \left(n_{un}^{HM} - n_{un}^L\right)\right] \leq p_f\left(n_{in}^H - n_{in}^L\right) \tag{87}$$

The IR condition for the L-type is naturally satisfied.

3) Combining all conditions above

Combing the IC and IR conditions for the M-type and L-type, we have

$$p_f\left(n_{in}^H - n_{in}^M\right) \leq \lambda_i\mu\left[(1-l)\left(n_{in}^{HM} - n_{in}^L\right) + \left(n_{un}^{HM} - n_{un}^L\right)\right] \leq p_f\left(n_{in}^H - n_{in}^L\right)$$

Finally, we can obtain

$$\begin{cases} n_2^{H*} = n_{in}^H \\ n_2^{M*} = n_{in}^H \\ n_2^{L*} = n_{in}^L \end{cases} \tag{88}$$

under the conditions

$$p_f l\left(\frac{c}{q_M} - \frac{c}{q_H}\right) \leq \lambda_i\mu\left[(1-l)\left(n_{in}^{HM} - n_{in}^L\right) + \left(n_{un}^{HM} - n_{un}^L\right)\right] \leq p_f l\left(\frac{c}{q_L} - \frac{c}{q_H}\right)$$

B.5. Fully Pooling

In this case, we first show that $(n_2^{H*}, n_2^{M*}, n_2^{L*}) = (n_{in}^H, n_{in}^H, n_{in}^H)$ (i.e., the H-type does not buy and the M- and L-type buy to mimic the H-type) with the following belief is a PBE.

$$P\left(H|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ \rho_H, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

$$P\left(M|n_2\right) = \begin{cases} 0, & \text{if } n_2 < n_{in}^H \\ \rho_M, & \text{if } n_2 \geq n_{in}^H \end{cases}$$

$$P\left(L|n_2\right)=\begin{cases}1, & \text{if } n_2<n_{in}^H\\ 1-\rho_H-\rho_M, & \text{if } n_2\geq n_{in}^H\end{cases}$$

We can write the equilibrium profits for the three types as

$$\pi_H^{pool}=l\lambda_i\mu n_{r,ia}^H+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HML} \tag{89}$$

$$\pi_M^{pool}=l\lambda_i\mu n_{r,ia}^M+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HML}-p_f x_M^{pool} \tag{90}$$

$$\pi_L^{pool}=l\lambda_i\mu n_{r,ia}^L+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HML}-p_f x_L^{pool} \tag{91}$$

where

$$\begin{aligned}n_{r,ia}^H&=n_{in}^H+n_{un}^{HML}\\ n_{r,ia}^M&=n_{in}^M+n_{un}^{HML}\\ n_{r,ia}^L&=n_{in}^L+n_{un}^{HML}\\ n_{r,ua}^{HML}&=n_{in}^{HML}+n_{un}^{HML}\\ n_{in}^{HML}&=\rho_H n_{in}^H+\rho_M n_{in}^M+\left(1-\rho_H-\rho_M\right)n_{in}^L\\ E\left[q_{HML}\right]&=\rho_H q_H+\rho_M q_M+\left(1-\rho_H-\rho_M\right)q_L\\ n_{un}^{HML}&=\left(1-l\right)\left(1-\frac{c}{E\left[q_{HML}\right]}\right)\end{aligned}$$

Clearly, neither type will deviate to a higher follower count than n_{in}^H.

1) M-type Influencer

If the M-type purchases fewer than x_M^* such that $n_2^{M\prime}<n_{in}^H$ (say $x_M'<n_{in}^H-n_{in}^M$), she will be viewed as an L-type by both the uninformed consumers and the uninformed advertiser (and an M-type by an informed advertiser). Thus, her expected profit is

$$\pi_M^{dev}\left(n_2^{M\prime}\right)=l\lambda_i\mu n_{r,ia}^{M\prime}+\left(1-l\right)\lambda_i\mu n_{r,ua}^{M\prime}-p_f\left(n_2^{M\prime}-n_{in}^M\right)$$

where $n_{r,ia}^{M\prime}=n_{in}^M+n_{un}^L$, $n_{r,ua}^{M\prime}=n_r^L=n_{in}^L+n_{un}^L$. Clearly, her best deviation of this type is not to purchase any fake account and the resulting expected profit is $\pi_M^{dev}\left(n_2^M\right)=l\lambda_i\mu n_{r,ia}^{M\prime}+\left(1-l\right)\lambda_i\mu n_{r,ua}^{M\prime}$. The IC condition requires

$$\pi_M^{dev}\left(n_2^M\right)\leq\pi_M^{pool}=l\lambda_i\mu n_{r,ia}^M+\left(1-l\right)\lambda_i\mu n_{r,ua}^{HML}-p_f\left(n_{in}^H-n_{in}^M\right)$$

which translates to:

$$\lambda_i\mu\left[\left(1-l\right)\left(n_{in}^{HML}-n_{in}^L\right)+\left(n_{un}^{HML}-n_{un}^L\right)\right]\geq p_f\left(n_{in}^H-n_{in}^M\right) \tag{92}$$

When the IC condition holds, the IR condition for the M-type can be naturally satisfied.

2) For the L-type influencer

If the L-type purchases fewer than x_L^* such that $n_2^{L\prime} < n_{in}^H$ (say $x_L' < n_{in}^H - n_{in}^L$), she will be viewed as an L-type by both the informed and uninformed advertisers/consumers. Her best deviation of this type is not to buy any fake account and the resulting expected profit is

$$\pi_L^{dev}\left(n_{in}^L\right) = \lambda_i \mu n_r^L$$

The IC condition requires

$$\pi_L^{dev}\left(n_{in}^L\right) \leq \pi_L^{pool} = l\lambda_i \mu n_{r,ia}^L + (1-l)\,\lambda_i \mu n_{r,ua}^{HML} - p_f\left(n_{in}^H - n_{in}^L\right)$$

which translates to:

$$\lambda_i \mu\left[(1-l)\left(n_{in}^{HML} - n_{in}^L\right) + \left(n_{un}^{HML} - n_{un}^L\right)\right] \geq p_f\left(n_{in}^H - n_{in}^L\right) \tag{93}$$

When the L-type's IC condition holds, her IR condition and M-type's IC condition is automatically satisfied.

Finally, we can obtain

$$\begin{cases} n_2^{H*} = n_{in}^H \\ n_2^{M*} = n_{in}^H \\ n_2^{L*} = n_{in}^H \end{cases} \tag{94}$$

under the condition $\lambda_i \mu\left[(1-l)\left(n_{in}^{HML} - n_{in}^L\right) + \left(n_{un}^{HML} - n_{un}^L\right)\right] \geq p_f l\left(\frac{c}{q_L} - \frac{c}{q_H}\right)$.

C. Proof of Proposition 8 (A Repeated Game)

First, we construct the new d_1' and d_2' for the period 2.

$$\begin{cases} d_1'\left(l_2^{pool}\right) = (1-\tau)\left(\eta_{1,pool}' - p_0\right) \\ d_1'\left(l_2^{sep}\right) = (1-\tau)\left(\eta_{1,sep}' - p_0\right) \\ d_2'\left(l_2^{pool}\right) = (1-\tau)\left(\eta_{2,pool}' - p_0\right) \\ d_2'\left(l_2^{sep}\right) = (1-\tau)\left(\eta_{2,sep}' - p_0\right) \end{cases} \tag{95}$$

where

$$\begin{cases} \eta_{1,pool}' \equiv \lambda_i \mu\left(1 - l_2^{pool}\right)\left[\frac{q_H(E[q]-q_L)}{l_2^{pool}E[q](q_H-q_L)} + \rho\right] \\ \eta_{1,sep}' \equiv \lambda_i \mu\left(1 - l_2^{sep}\right)\left[\frac{q_H(E[q]-q_L)}{l_2^{sep}E[q](q_H-q_L)} + \rho\right] \\ \eta_{2,pool}' \equiv \frac{\lambda_i \mu\left(1-l_{pool}'^2\right)}{l_2^{pool}} \\ \eta_{2,sep}' \equiv \frac{\lambda_i \mu\left(1-l_{sep}'^2\right)}{l_2^{sep}} \end{cases} \tag{96}$$

Similar to the Proof of Lemma 4, we can prove that $d_1'\left(l_2^{pool}\right) < d_2'\left(l_2^{pool}\right)$, and $d_1'\left(l_2^{sep}\right) < d_2'\left(l_2^{sep}\right)$. In addition, we know that the η_1' and η_2' decrease with l, meanwhile, we know $l < l_2^{pool} < l_2^{sep}$, thus we have $d_1 > d_1'\left(l_2^{pool}\right) > d_1'\left(l_2^{sep}\right)$ and $d_2 > d_2'\left(l_2^{pool}\right) > d_2'\left(l_2^{sep}\right)$.

Case (a): when $d_1 \leq 0$.

In period 1, the condition for the pooling equilibrium is not met, so the pooling equilibrium cannot exist. The only equilibrium is either costly-separating, if $d \leq d_2$, or naturally-separating, otherwise. In period 2, $d_1'\left(l_2^{sep}\right) < d_1 < 0$, thus, the condition for the pooling equilibrium is not met either. The only equilibrium in period 2 is either costly-separating if $d \leq d_2'\left(l_2^{sep}\right)$, or naturally-separating, otherwise.

In addition, we know that $d_2 > d_2'\left(l_2^{sep}\right)$, if $d \leq d_2'\left(l_2^{sep}\right)$, the equilibrium in both periods is the costly separating; if $d_2'\left(l_2^{sep}\right) \leq d \leq d_2$, the equilibrium in period 1 is costly-separating while that in period 2 is naturally-separating; if $d > d_2$, the equilibrium in both periods is naturally-separating.

Since there is only one combination of equilibria in period 1 and 2 under any condition, it is also an undefeated equilibrium, which is summarized in case (a).

Case (b): When $0 < d_1 \leq d_2'\left(l_2^{pool}\right)$.

In period 1, if $d \leq d_1$, by Lemma (4), the pooling coexists with and defeats the costly-separating equilibrium. Given the equilibrium in period 1 is pooling and $d_1 > d_1'\left(l_2^{pool}\right)$, if $d \leq d_1'\left(l_2^{pool}\right)$, the equilibrium in period 2 is pooling as well; if $d_1'\left(l_2^{pool}\right) < d \leq d_1$, the equilibrium in period 2 is costly-separating.

In period 1, if $d_1 < d \leq d_2$, the costly-separating equilibrium is the sole equilibrium and thus undefeated equilibrium. Given the equilibrium in period 1 is costly-separating, $d_1 > d_1'\left(l_2^{sep}\right)$, and $d_2 > d_2'\left(l_2^{sep}\right)$, if $d_1 < d \leq d_2'\left(l_2^{sep}\right)$, the equilibrium in period 2 is costly separating; if $d_2'\left(l_2^{sep}\right) < d \leq d_2$, the equilibrium in period 2 is naturally separating;

In period 1, if $d > d_2$, the naturally-separating equilibrium is the only remaining equilibrium and thus undefeated. Given $d_2 > d_2'\left(l_2^{sep}\right)$, the equilibrium in period 2 is naturally separating as well.

The undefeated equilibrium is summarized in case (b).

Case (c): When $d_1 > d_2'\left(l_2^{pool}\right)$.

In period 1, if $d \leq d_1$, the pooling is the undefeated equilibrium. Given the equilibrium in period 1 is pooling and $d_1 > d_2'\left(l_2^{pool}\right) > d_1'\left(l_2^{pool}\right)$, if $d \leq d_1'\left(l_2^{pool}\right)$, the equilibrium in period 2 is pooling as well; if $d_1'\left(l_2^{pool}\right) < d \leq d_2'\left(l_2^{pool}\right)$, the equilibrium in period 2 is costly-separating; if $d_2'\left(l_2^{pool}\right) < d \leq d_1$, the equilibrium in period 2 is naturally-separating.

In period 1, if $d_1 < d \leq d_2$, the costly-separating equilibrium is the sole equilibrium and thus undefeated equilibrium. Given the equilibrium in period 1 is costly-separating, $d_1 > d_1'\left(l_2^{sep}\right)$, and $d_2 > d_2'\left(l_2^{sep}\right)$, if $d_1 < d \leq d_2'\left(l_2^{sep}\right)$, the equilibrium in period 2 is costly separating; if $d_2'\left(l_2^{sep}\right) < d \leq d_2$, the equilibrium in period 2 is naturally separating;

In period 1, if $d > d_2$, the naturally-separating equilibrium is the only remaining equilibrium and thus undefeated. Given $d_2 > d_2'\left(l_2^{sep}\right)$, the equilibrium in period 2 is naturally separating as well.

The undefeated equilibrium is summarized in case (c).