# Economics of Social Media Fake Accounts

Zihong Huang[1], De Liu[2]

Carlson School of Management, University of Minnesota

[1]huan0707@umn.edu, [2]deliu@umn.edu

Aug 31, 2022

**Abstract**

Amid the rise of the influencer economy, fake social media accounts have become a prevalent problem 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 social media literacy, anti-fake technology, and costs of fake accounts. We show that not only low-quality influencers may buy fake accounts to mimic high-quality ones in a "pooling" equilibrium, high-quality influencers may also buy to prevent mimicry in a "costly-separating" equilibrium. There is also a "naturally-separating" equilibrium where the two types are separated without buying fake accounts. We find that increasing anti-fake efforts and social media literacy may cause more fake accounts. The platform generally prefers either a zero-effort pooling equilibrium or a high-effort naturally-separating equilibrium. Compared to the level of anti-fake efforts preferred by consumers, the platform may be overly or insufficiently aggressive. Some anti-fake strategies, such as increasing social media literacy and fake-account costs, may benefit consumers but not the platform. One exception is increasing the effectiveness of anti-fake technology, which benefits both the platform and consumers and reduces the number of fake accounts.

***Keywords***-Influencer Economy, Fake Accounts, Social Media, Signaling, Social Media Literacy.

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

By "fake accounts", we mean social media accounts designed to impersonate real users with fake personal information and/or behaviors. While reports suggest that 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 a few different purposes. One type of fake accounts aims to help businesses, individuals, and topics gain influence (e.g., by following, liking, sharing, and mentioning). Another type is created to obtain perks of social media accounts (e.g., signup bonuses and coupons). There are also malicious fake accounts created to spread phishing, scam, malware, and politically-motivated fake accounts aiming to sway opinions, voting, and election outcomes. This paper mainly focuses on influence-boosting fake accounts and leaves other types of fake accounts for future investigation.

The demand for influence-boosting fake accounts is boosted by the rise of the influencer economy (Confessore et al., 2018; Federal Trade Commission, 2019), which allows large and small social media influencers to get paid for promoting products among their followers (e.g., through sponsored posts). Because influencers' pay is tied to their influence, which is often measured by reach (e.g., number of followers) and engagement (e.g., clicks and likes), they have strong incentives to boost their influence, sometimes by buying fake accounts and associated services. The link between the influencer economy and fake accounts is illustrated in the widely-publicized case of Federal Trade Commission (FTC) versus Devumi in 2019. Devumi is a company that made millions of dollars by manufacturing and selling fake accounts/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 is never abated – e.g., in the first quarter of 2022, Facebook shut down 1.6 billion fake accounts and estimates that there are no less than five percent of the Facebook users are still fake after the removal (Warwick, 2022).

While some fake accounts may seem harmless, fake accounts have been associated with several problems. First, they pose a major threat to the influencer economy, causing distorted outcome metrics and wasteful marketing spending. Second, as many consumers use influence indicators in their decisions (Federal Trade Commission, 2019), fake accounts can mislead consumers. Furthermore, consumer experiences may degrade as a result of their interactions with fake accounts, especially with automated fake accounts. Therefore, there is an urgent need among campaign managers, social media platforms, consumers, and policymakers to understand the fake account problem and develop effective coping strategies.

Social media platforms have already begun to tackle the fake account problem. A major tool used for fighting fake accounts is automated fake account identification and prevention (which we refer to as *anti-fake efforts*). For example, social media platforms often use user verification technologies such as reCAPTCHA and two-factor authentication to deter automated fake-account creation. Facebook uses machine-learning systems to block and detect fake accounts both before and after they become alive (Hao, 2020; Condliffe, 2020). While social media platforms have become better at automated fake-account detection and prevention, there are still considerable challenges in catching fake accounts reliably using automated approaches. This is not only because there are a large variety of user behaviors making the automatic separation of real and fake accounts nearly impossible, especially with manually-operated fake accounts(Nicas, 2020), but also because fake account providers are also getting better at mimicking real users, sometimes with advanced AI (e.g., a recent report found computer-generated images in fake LinkedIn profiles (Robins, 2022)). The observation that fake accounts remain prevalent and appear to be increasing on social media platforms raises the question of whether the platforms' anti-fake efforts are effective in reducing fake accounts.

A further issue is whether social media platforms have adequate incentives to combat fake accounts. Social media platforms have an inherent interest in attracting and reporting a large number of users (Chen, 2022; Stolzoff, 2018), and may be reluctant to remove a large number of fake accounts. Moreover, anti-fake efforts can cause inconvenience among real consumers. For

example, increased user verification and misclassifying real consumers as fake accounts can cause user frustration and drive them away (ArkoseLabs, 2021; Kaudelka, 2021). So, it is unclear whether platforms will devote the amount of anti-fake efforts that are preferred by consumers.

Another ongoing effort to combat the fake account problem is social media literacy education. Schools and other online education platforms (MediaLiteracyNow, 2022) have already begun to provide social media literacy education to young and adult social media users (Taibi et al., 2021; Al Zou'bi, 2022). The idea is that by promoting social media literacy, social media users can better differentiate between high and low-quality influencers. The question remains, however, whether increased social media literacy will lead to fewer fake accounts. Moreover, would social media platforms be incentivized to improve the social media literacy of their users?

Besides the above questions, there are also unanswered basic questions about social media fake accounts. For example, while low-quality influencers can buy fake accounts to impersonate high-quality ones, high-quality influencers can also buy fake accounts to signal their superiority. It is not yet clear what types of influencers are more likely to buy fake accounts. A related issue is whether fewer fake accounts are always better for the platform or consumers.

The above discussion suggests that there is a need for a systematic examination of the fake account problem so that multiple connected issues can be studied in a holistic framework. To our knowledge, this need has not been met by existing research. Existing studies of fake accounts primarily focus on examining fake accounts' activities (Stringhini et al., 2013) and developing detection techniques (Raturi, 2018; Yuan et al., 2019). Motivated by this gap, we build a game-theoretical model of fake accounts in the context of the influencer economy. This model comprises a unit mass of consumers, an influencer, an advertiser, and a social media platform. The influencer's quality can be either high or low. A proportion of consumers are "informed" about the influencer's quality whereas the rest are "uninformed." We use the proportion of "informed" consumers to capture the social media literacy level. The uninformed consumers can draw inferences from the influencer's followers when deciding whether to follow the influencer. The influencer, who receives a share of the advertising surplus, can purchase fake accounts to attract more uninformed consumers. The advertiser is also uninformed about the influencer's quality and bases her participation decision on the expected number of real followers. The platform, who receives a share of the total advertising surplus, can mount an anti-fake effort that increases the cost of fake accounts, but also increases

the nuisance cost of consumers. The impacts of the anti-fake efforts are governed by the *anti-fake technology* – a higher technology implies a higher cost of fake accounts and lower nuisance costs for consumers. We use this model to address a host of questions, such as:

1. What types of equilibrium does the game have? How many fake accounts a high- or low-quality influencer will buy in equilibrium?

2. How does the platform's anti-fake effort affect the number of fake accounts? What is the platform's optimal anti-fake effort?

3. How do model parameters such as the anti-fake technology level, social media literacy, and the base cost of fake accounts, which can be targets of fake-account interventions, affect the equilibrium number of fake accounts, platform profits, and consumer welfare?

4. What level of anti-fake effort is optimal for consumers? How do the platform's optimal anti-fake effort and preferred fake-account interventions compare with those of consumers?

Answers to the above questions are of broad interest to campaign managers, social media platforms, consumers, and policymakers.

Our analyses suggest that there is a "*pooling*" equilibrium where a low-quality ($L$-type) influencer purchases fake accounts to mimic a high-quality ($H$-type) influencer (termed as "offensive purchasing"), a "*costly separating*" equilibrium, where an $H$-type purchases fake accounts to prevent an $L$-type from mimicking (termed as "defensive purchasing"), and a "*naturally separating*" equilibrium where the two types of influencers separate without purchasing.

As the platform's anti-fake effort increases, the equilibrium generally transitions from pooling, to costly separating, and then to naturally separating. Within each equilibrium, however, the number of fake accounts may increase with the anti-fake effort. For example, in the pooling equilibrium, the $L$-type's offensive purchasing increases in the platform's anti-fake effort. This is because the latter causes a larger follower gap between the two types of influencers, forcing the $L$-type to purchase more to make up for the gap.

The platform optimally chooses between zero anti-fake effort, which results in a pooling equilibrium (or a costly separating one when pooling does not exist), and a high-effort naturally-separating equilibrium with no fake accounts. The platform generally prefers a zero-effort pooling equilibrium

when the anti-fake technology is low (so nuisance costs of anti-fake efforts are high) and social media literacy is low (so the proportion of uninformed consumers is high), because, in such cases, the benefit of separation is low and the nuisance costs of anti-fake efforts are high.

Interestingly, some "intuitive" anti-fake interventions may have adverse or no effects on the number of fake accounts and/or platform profits. For example, under the pooling equilibrium, increasing social media literacy can lead to *more* fake accounts, whereas increasing the base cost of fake accounts has no impact. The former occurs because higher social media literacy can lead to a large follower gap. The latter is because the $L$-type must make up for the follower gap despite the increased cost of fake accounts. Both higher social media literacy and a higher base cost of fake accounts can lead to lower platform profits. In fact, our numerical simulations suggest that firms have no incentive to invest in either social media literacy or increasing fake-account costs. In contrast, the platform has incentives to improve its anti-fake technology, which can reduce the number of fake accounts and increase platform profits.

We find that consumers may prefer an intermediate-effort separating equilibrium or a zero-effort pooling equilibrium, as they trade-off between getting the benefit of influencer separation but enduring the associated nuisance cost in the former case, and not suffering from any nuisance cost but enduring a lack of influencer separation in the latter case. Consumers always prefer costly separating equilibrium over naturally separating one, however, because of the higher nuisance cost imposed by the latter for no additional benefit to consumers.

We also find that the misalignment between platform profits and consumer welfare can go both ways. On the one hand, the platform can be insufficiently aggressive in tackling fake accounts than the preferred level by consumers, especially when social media literacy is low. On the other hand, the platform may also be overly aggressive in reducing fake accounts out of concern for the high costs of fake accounts imposed on the advertiser and the influencer, which consumers do not care about. Furthermore, consumers benefit from increased social media literacy and a higher base cost of fake accounts, whereas the platform may not.

## 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, Tedeschi, and Ursino, 2018), fake reviews, fake sales (Chen, Yang, and Hosanagar, 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 connections and differences 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, Tedeschi, and Ursino (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 assume that an $H$-type seller never cheats (i.e., does not make fake purchases). We, on the other hand, study both separating and pooling equilibria, where the $H$-type and the $L$-type influencers buy fake accounts, respectively. 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 of helping consumers learn the true quality of products through information disclosure (Papanastasiou, Bimpikis, and Savva, 2018; Che and Hörner, 2018; Pennycook et al., 2020) and penalizing the information producers for their manipulative behav-

iors (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 sometime 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 on 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, Tedeschi, and Ursino (2018) examine how consumers' welfare is affected by sellers' deceptive strategies. They suggested that consumer welfare could be higher under the equilibrium with sellers' deceptive advertising. Chen, Yang, and Hosanagar (2022) study the impact of brushing (i.e. 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) who find 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 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 the cost of fake accounts (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

The ecosystem for fake accounts consists of four types of players: a social media platform, one representative influencer, a unit mass of consumers, and one advertiser. The platform hosts the influencer who produces social media content. Consumers choose whether to follow the influencer and consume her content.[1] The advertiser is interested in reaching the consumers by asking the influencer to share sponsored posts (or product placement). The advertiser, the platform, and the influencer share the surplus of the system according to their bargaining power.

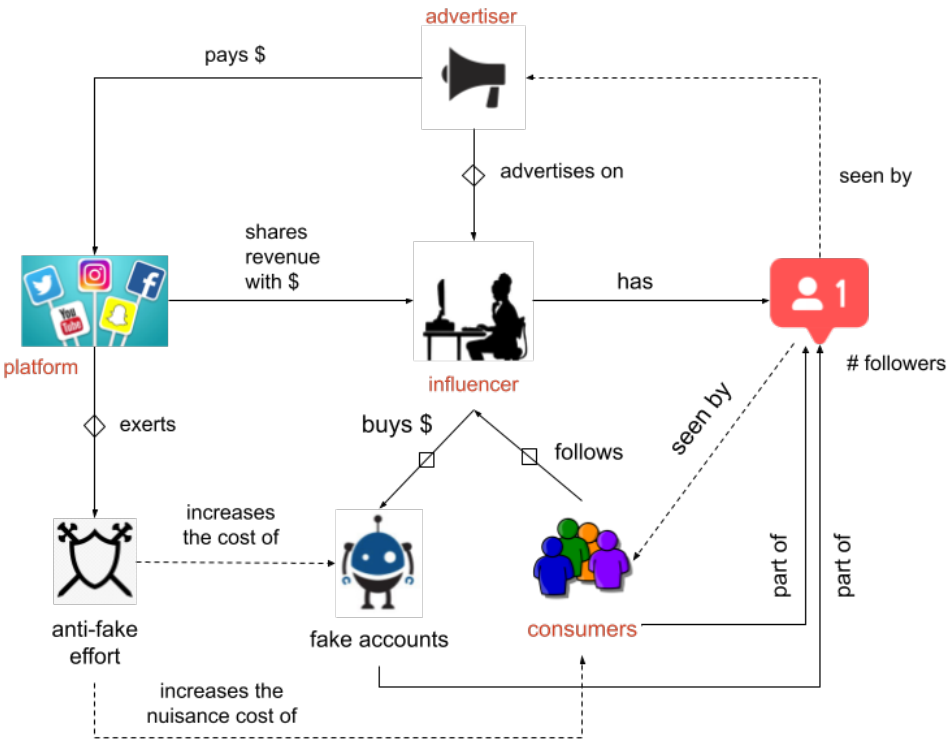



Figure 1: Model Sketch

**Influencer**: For simplicity, we assume the influencer produces one unit of content. Following prior literature (Shin, 2017; Guo et al., 2019), we normalize the production cost of the content to zero. The quality of the content $q$, which can also be interpreted as the quality of the influencer, is

[1] 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 access the influencer's content.

a random draw from two levels, $q_H$ and $q_L$ ($q_H > q_L$), with probabilities $\rho$ and $1-\rho$, respectively. We call an influencer with $q_H$ ($q_L$) content quality an $H$-type *($L$-type)* influencer. The influencer's type is private information. We denote

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

as the *ex-ante expected quality* of the influencer and $r_q \equiv q_L/q_H$ as the *quality ratio*. We will discuss the influencer's fake-account purchase decision separately below.

**Consumers**: Consumers are risk neutral. A consumer derives utility $\theta q$ for consuming a unit of content with quality $q$, where $\theta$ is the consumer's taste for content quality. We assume $\theta$ is uniformly distributed on $[0, 1]$. Consumers are differently informed about the influencer's quality. We assume that a proportion $l$ of consumers are *informed* – they know the true quality of the influencer.[2] We use informed consumers to capture the case where consumers have the knowledge and ability to judge the quality of the influencer (e.g., by searching for and evaluating the influencer's prior content). The remaining $1-l$ consumers are *uninformed* – they do not know the true quality, but know the distribution of $q$. Uninformed consumers can use the influencer's number of followers to update their beliefs about the influencer's quality.[3] Our assumption of consumers using popularity indicators to infer quality is supported by prior empirical work. 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, Cauberghe, and Hudders, 2017). We interpret the parameter $l$ as the level of *social media literacy* among the consumers – the higher the social media literacy, the larger the proportion of informed consumers. We further assume that a consumer's informedness is independent of the consumer's taste $\theta$ for quality.

Consumers decide whether to follow the influencer. If a consumer chooses not to follow the influencer, she will not see the content and effectively quit the platform and thus get zero utility. If a consumer follows the influencer, she gains access to the influencer's content but also incurs a cost $c$, which consists of an opportunity cost of time $c_0$ and a nuisance cost $c_d$ due to the platform's

[2]This simplifying assumption is generalizable to a good but imperfect signal of quality.

[3]Though we use the number of followers as a popularity indicator, our insights are generalizable to other popularity indicators such as the number of likes, the number of forwards, and the number of comments. For example, when the fake followers behave like real followers in liking, the number of likes is proportional to the number of followers.

anti-fake effort (more details later). Formally, a consumer $\theta$'s utility of following an influencer with quality $q$ is given by:

$$u(\theta, q) = \theta q - \underbrace{(c_0 + c_d)}_{c} \tag{2}$$

By this formulation, consumers follow the influencer if and only if they expect a positive utility from following.

**Assumption 1.** $q_L > c$

By this assumption, an $L$-type influencer's quality is high enough so that informed consumers with the highest taste for quality could still choose to follow her so that the problem will not degenerate.

**Advertiser**: The advertiser derives a unit value $\mu$ from advertising to a real consumer. The advertiser cannot tell whether an influencer's followers are real or fake. The advertiser does not know the influencer's true quality either, but knows the distribution of the influencer's quality and can update her belief after seeing the number of followers.

There is a unit *transaction cost* $\varphi$ for advertising to a fake account. This is because fake accounts not only create zero value from advertisers, but also cause a wastage (e.g. computing and human costs associated with tracking, bookkeeping, auditing, etc). In sum, the advertising revenue and cost are $\mu n_r$ and $\varphi x$, respectively, where $n_r$ is the number of real consumers reached.

**Fake accounts**: The influencer can purchase fake social media accounts to impress uninformed consumers and the advertiser. The unit cost of fake social media accounts is $c_f$, which is a function of the platform's anti-fake effort (more details below).

The total surplus generated by advertising is as follows.

$$\text{(Total surplus) } \pi = \mu n_r - (\varphi + c_f)\, x \tag{3}$$

We assume that the advertiser, the influencer, and the platform share the total expected surplus $E[\pi]$. Specifically, for $E[\pi]$ dollars of expected surplus generated by advertising, the influencer, the platform, and the advertiser receive $\lambda_i E[\pi]$, $\lambda_p E[\pi]$, and $\lambda_a E[\pi]$ dollars respectively. We assume that $\lambda_i, \lambda_p$,and $\lambda_a \geq 0$ and $\lambda_a \equiv 1 - \lambda_i - \lambda_p$. [4] The parameters $\lambda_i$ and $\lambda_p$ are exogenously fixed,

[4]We have also explored an alternative scheme where two or more advertisers compete for the ad slot via a sealed-

reflecting the influencer and the platform's relative *bargaining power*, respectively. The surplus-sharing scheme is common in the influencer marketing industry. For example, YouTube shares 55% of the advertising revenue with influencers and Facebook shares 45% of ad revenue with influencers (Pahwa, 2022; Spangler, 2021).

By the surplus sharing scheme, the advertiser is expected to pay $\lambda_p E[\pi] + \lambda_i E[\pi] + c_f E[x]$ to the platform. So, the advertiser's expected profit is

$$\begin{aligned} \pi_a &= \mu E[n_r] - \varphi E[x] - (\lambda_p E[\pi] + \lambda_i E[\pi] + c_f E[x]) \\ &= \lambda_a E[\pi] = \lambda_a \{\mu E[n_r] - (\varphi + c_f) E[x]\} \end{aligned} \tag{4}$$

The advertiser decides whether to advertise on this platform based on whether her expected profit is positive.

The influencer is expected to receive a payment of $\lambda_i E[\pi] + c_f E[x]$. Her profit is, therefore,

$$\pi_i = \lambda_i E[\pi] + c_f (E[x] - x) \tag{5}$$

The influencer chooses the number of fake accounts $x$ to maximize her expected profit.

**Platform**: The platform chooses an anti-fake effort $d$ ($d \geq 0$) (*effort* for short). The platform's anti-fake effort may include both detection and prevention of fake social media accounts. For example, the platform may use machine learning to detect fake social media accounts based on abnormal account profiles and behaviors. It may also deploy technologies such as reCAPTCHA and two-factor authentication to make it harder to register and operate fake accounts. A zero effort $d = 0$ means the platform does nothing about fake accounts. A higher effort $d$ implies more aggressive detection, more frequent scans, and/or more rigorous user verification.

The platform's anti-fake effort holds implications for both fake-account operators and consumers. On one hand, increasing anti-fake efforts can lead to the prevention, catching, and removal of more fake accounts, raising the cost of operating fake accounts. On the other hand, it can also result in a negative externality of increased nuisance costs for legitimate users. For example, increased detection may lead to more legitimate users being misclassified as fake accounts and users spending

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.

more time proving themselves legitimate users (e.g. answering reCAPTCHA questions). The size of the impact is a function of the *anti-fake technology level* $\tau\,(1 > \tau \geq 0)$. A higher technology level $\tau$ is associated with more effective detection algorithms and prevention technologies, which can increase the cost of fake accounts and reduce the nuisance costs of consumers from anti-fake efforts (Confessore et al., 2018; Dhawan and Ekta, 2016).

Specifically, we let consumers' nuisance cost be $c_d = c_1\,(1-\tau)\,d$ and the consumer's cost becomes

$$c = c_0 + c_1\,(1-\tau)\,d \tag{6}$$

where $c_1$ is the *coefficient of nuisance cost*. By (6), consumers' nuisance cost increases with the platform's effort $d$ and decreases with technology level $\tau$. We let the unit cost of fake accounts $c_f$ be

$$c_f = \kappa + \frac{1}{1-\tau}d \tag{7}$$

where $\kappa$ is the fake-account base cost. By (7), the unit cost of fake accounts increases with platform effort $d$ and technology level $\tau$.

We normalize the platform's marginal cost of operation to zero.[5] The platform's expected profit is:

$$\pi_p = \lambda_p E\left[\pi\right] = \lambda_p\left\{\mu E\left[n_r\right] - \left(\varphi + c_f\right) E\left[x\right]\right\}. \tag{8}$$

The platform chooses the anti-fake effort $d$ to maximize its expected profit $\pi_p$.

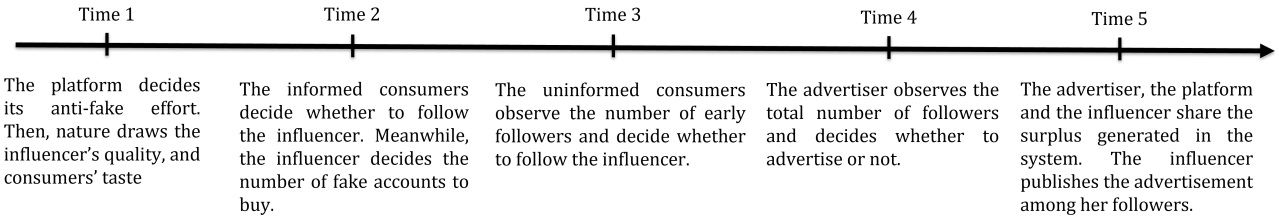



Figure 2: Game Timeline

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 content quality $q$, and the consumers' tastes $\theta$. At **time**

[5] As we will show, the platform's anti-fake effort has a complex effect on the platform's profitability. Assuming away the cost of anti-fake effort allows us to more clearly see the trade-offs facing the platform when it comes to anti fake accounts.

| *Notation* | *Interpretation* |
|---|---|
| $d$ | The platform's decision variable, the anti-fake effort. |
| $x$ | The number of fake accounts purchased by the influencer. |
| $l$ | The proportion of informed consumers, which also represents the social media literacy level. |
| $\lambda_i, \lambda_p, \lambda_a$ | The bargaining power of the influencer, the platform, and the advertiser, respectively. |
| $\mu$ | The value generated by advertising to a real consumer. |
| $n_{in}, n_{un}, n_r, n_2, n$ | The number of informed, uninformed, real, early (including informed and fake), and total followers the influencer has. |
| $\bar{n}_{in}$ | The ex-ante expected number of informed consumers. |
| $q_H, q_L, \bar{q}$ | The content quality of $H$- and $L$-type influencers and the unconditional expected quality of an influencer, respectively. |
| $\rho$ | The probability of drawing an $H$-type influencer. |
| $c_d$ | A consumer's nuisance cost from the platform's anti-fake effort. |
| $c_f$ | The unit cost of fake accounts. |
| $c_o$ | A consumer's opportunity cost. |
| $c_1$ | The coefficient of nuisance cost. |
| $\kappa$ | The fake-account base cost. |
| $\varphi$ | The transaction cost of advertising to a fake account. |
| $u_i$ | Consumer $i$'s expected utility. |
| $r_q$ | The quality ratio $r_q \equiv q_L/q_H$. |
| $\pi_a$, $\pi_i, \pi_p$ | The expected profit of the advertiser, the influencer, and the platform, respectively. |
| $\pi_p^{pool}, \pi_p^{csep}, \pi_p^{nsep}$ | The platform's profit in pooling, costly separating, and naturally separating equilibrium, respectively. |
| $U$ | Consumer welfare. |
| $\theta$ | Consumers' taste for quality, $\theta \in [0, 1]$ |
| $\tau$ | The platform's anti-fake technology level, $\tau \in [0, 1)$ |

Table 1: Notations

**2**, informed consumers decide whether to follow the influencer. Meanwhile, the influencer decides the number of fake accounts $x$ to buy. After the influencer and the informed consumers' decisions, the influencer has $n_2$ *early followers*, which include $n_{in}$ informed consumers and $x$ fake followers. At **time 3**, uninformed consumers observe the number of early followers $n_2$ and decide whether to follow the influencer. After the uninformed consumers' decision, 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 number of followers $n$ and decides whether to advertise or not. If yes, the influencer shares the advertisement among her followers. At **time 5**, if the advertiser decides to advertise, she pays a proportion of the advertising surplus, $(\lambda_i + \lambda_p)\,\pi$, to the platform, who then shares $\lambda_i \pi$ with the

influencer.

Our model is a simplified, discrete version of the real-world fake-account ecosystem. We argue that our stylized model can capture the key aspects of different stakeholders' decision environments and the main effects of their decisions. The rationales for the decision sequence of our model are as follows:

- First, we assume informed consumers make their following decisions before uninformed ones. We make this assumption because uninformed consumers rely on the influencer's popularity to infer her quality, and thus it is natural for them to wait for the popularity signal to materialize before deciding whether to follow the influencer. In contrast, informed consumers already know the true quality and thus have no reason to wait.

- Second, we assume the influencer purchases fake accounts before uninformed consumers make their decisions. One of the benefits of purchasing fake accounts is to convince uninformed consumers to follow the influencer. Therefore, the influencer prefers to purchase fake accounts before uninformed consumers make their decisions, which is what we model.

- 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. The model would remain the same if these decisions occur sequentially.

- Fourth, we assume the platform's anti-fake effort occurs before the influencer's fake-account purchase decision. We use this decision order to allow the platform's anti-fake effort to influence the cost of fake accounts. This decision order also captures the notion that for a fake account to work, it must have survived the platform's anti-fake effort.

- Finally, we assume that the advertiser moves after the uninformed consumers. This assumption reflects the observation that advertisers often begin to advertise with an influencer when she is popular enough, at which point the influencer has already attracted both informed and uninformed consumers, and may have already bought fake accounts.

# 4 Equilibrium Analysis

## 4.1 Preliminaries

Given the consumer utility function (2), the number of informed consumers following the $H$- and $L$-type influencers at time 2 are, respectively

$$n_{in}^{H} = l\left(1 - \frac{c}{q_H}\right); n_{in}^{L} = l\left(1 - \frac{c}{q_L}\right). \tag{9}$$

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, let $\Pr(H|n_2)$ and $\Pr(L|n_2)$ denote the probability of the influencer being $H$-type and $L$-type, respectively, conditional on the number of early followers being $n_2$ and $E[q|n_2] = \Pr(H|n_2)\, q_H + \Pr(L|n_2)\, q_L$ be the expected quality of the influencer conditional on the number of early followers $n_2$. Let $n_2^H$ and $n_2^L$ be the number of earlier followers for the $H$-type and $L$-type, respectively. The number of uninformed followers for the $H$-type and $L$-type influencers at time 3 are, respectively

$$n_{un}^{H} = (1-l)\left(1 - \frac{c}{E\left[q|n_2^H\right]}\right); n_{un}^{L} = (1-l)\left(1 - \frac{c}{E\left[q|n_2^L\right]}\right). \tag{10}$$

At time 4, $\Pr(H|n)$ and $\Pr(L|n)$ denote the probability of the influencer being $H$-type and $L$-type, respectively, conditional on the total number of followers $n$. The expected number of real followers at time 4 is

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

## 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 the advertiser. Though, technically, uninformed consumers use the number of early followers as a signal whereas the advertiser uses the number of total followers, the two signals contain identical information – compared to an $L$-type influencer, if an $H$-type influencer has more (the same, fewer) early followers, she will also have more (the same, fewer) total followers. Therefore, without loss of generality, we use the number of early followers as the signal for both uninformed consumers and the advertiser.

A strategy profile of this game can be represented by $(x_H, x_L)$, i.e., the number of fake accounts purchased by the $H$- and $L$-type influencers, respectively. 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, we say 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 influencers purchases fake accounts and they separate naturally; (b) a *costly separating* equilibrium where at least one type purchases fake accounts. A similar distinction has been made by Guo, Xiao, and Zhang (2017) in the context of corporate social responsibility.

The signaling game tends to have multiple Perfect Bayesian Equilibria (PBEs) because the out-of-equilibrium beliefs can be arbitrary. A popular strategy is to refine PBEs using the *lexicographically maximum sequential equilibrium* (LMSE) concept (Mailath, Okuno-Fujiwara, and Postlewaite, 1993). Following prior literature (Mailath, Okuno-Fujiwara, and Postlewaite, 1993), we adopt LMSE as our equilibrium concept and define an LMSE strategy profile for our context as:

**Definition 1.** The strategy profile $(x_H, x_L)$ *lexicographically dominates (l-dominates)* $(x'_H, x'_L)$ if (a) $U_H(x_H, x_L) > U_H(x'_H, x'_L)$ or (b) $U_H(x_H, x_L) = U_H(x'_H, x'_L)$ and $U_L(x_H, x_L) > U_L(x'_H, x'_L)$, where $U_\omega(\cdot, \cdot)$ denotes $\omega$-type ($\omega \in \{H, L\}$) influencer's equilibrium profit. Then, a strategy profile $(x_H, x_L)$ is a *lexicographically maximum sequential equilibrium* (LMSE) if there does not exist another PBE $(x'_H, x'_L)$ that l-dominates $(x_H, x_L)$.

Intuitively, LMSE requires that there are no other PBEs that yield a higher payoff for the $H$-type or yield the same payoff for the $H$-type but a higher payoff for the $L$-type. One of the advantages of LMSE is that it avoids the global consistency issue associated with Intuitive Criterion (Mailath, Okuno-Fujiwara, and Postlewaite, 1993), another popular equilibrium refinement strategy. Moreover, LMSE tends to select a unique PBE in signaling games.

In the following subsections, we first analyze each equilibrium type separately and apply the LMSE refinement within the equilibrium type to obtain *locally-refine*d LMSEs, and then apply the LMSE refinement across equilibrium types to obtain the *globally-refined* LMSEs.

### 4.2.1 Pooling Equilibrium

As mentioned above, in the pooling equilibrium, $H$-type and $L$-type influencers have the same number of early followers. Using the LMSE refinement, we obtain a unique pooling equilibrium as described in Lemma 1.

**Lemma 1.** *(Pooling) Under the belief*

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

*a strategy profile* $\left(x_H^{pool}, x_L^{pool}\right) = \left(0, l\frac{q_H - q_L}{q_H q_L}c\right)$ *is the unique pooling equilibrium if and only if the following condition holds:*

$$d \leq (1-\tau)\left(\eta_1 - \kappa\right) \tag{12}$$

*where*

$$\begin{aligned} \eta_1 &\equiv \frac{\lambda_i \mu}{\lambda_i + (1-\lambda_i)\rho} \frac{\overline{n}_{in} + \overline{n}_{un} - n_{in}^L - n_{un}^L}{n_{in}^H - n_{in}^L} - \frac{\lambda_i (1-\rho)\varphi}{\lambda_i + (1-\lambda_i)\rho} \\ &= \frac{\lambda_i \mu}{\lambda_i + (1-\lambda_i)\rho} \frac{l\rho\left(\frac{1}{q_L} - \frac{1}{q_H}\right) + (1-l)\left(\frac{1}{q_L} - \frac{1}{E[q]}\right)}{l\left(\frac{1}{q_L} - \frac{1}{q_H}\right)} - \frac{\lambda_i (1-\rho)\varphi}{\lambda_i + (1-\lambda_i)\rho} \end{aligned} \tag{13}$$

Lemma 1 describes an equilibrium where the $L$-type influencer purchases fake accounts while the $H$-type doesn't. In this equilibrium, the $L$-type purchases enough fake accounts to make up the gap between the two types' informed followers so that they look identical to uninformed consumers and the advertiser. We call this type of fake-account purchasing *offensive purchasing*.

Condition (12) ensures that the $L$-type influencer achieves the highest payoff at the equilibrium, i.e., the $L$-type's incentive compatibility (IC) condition. To understand this condition, we note that (12) can also be written as

$$c_f = \kappa + \frac{1}{1-\tau} d \leq \eta_1. \tag{14}$$

The left-hand side is the unit cost of fake accounts and the right-hand side $\eta_1$, defined in Equation (13), can be interpreted as the revenue gain per fake account when the $L$-type moves from

purchasing nothing and being treated as an $L$-type to buying enough fake accounts to pool with the $H$-type. As long as the unit cost of fake accounts is lower than the revenue gain, the $L$-type is incentivized to pool with the $H$-type.

We note that the $L$-type's individual rationality (IR) condition is automatically satisfied because she prefers pooling to separating and her separating payoff is nonnegative (noting she incurs no cost under separating). The advertiser's IR condition is automatically satisfied also because the total surplus from advertising is positive when the $L$-type's IR condition holds. The $H$-type has no incentive to purchase fake accounts because a higher follower count is not rewarded under the current belief. The IR condition for the $H$-type is automatically satisfied because she incurs no cost.

### 4.2.2 Separating Equilibrium

We describe the costly separating and naturally separating equilibrium schemes in the next two lemmas.

**Lemma 2.** *(Costly Separating) Under the belief*

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

*a strategy profile* $\left(x_H^{csep}, x_L^{csep}\right) = \left(\frac{\left(\lambda_i\mu - c_f l\right)\left(\frac{c}{q_L} - \frac{c}{q_H}\right)}{\lambda_i\left(\varphi + c_f\right)}, 0\right)$ *is the unique costly separating equilibrium if and only if the following condition holds:*

$$d \leq \left(1-\tau\right)\left(\eta_2 - \kappa\right) \tag{15}$$

*where*

$$\eta_2 \equiv \frac{\lambda_i\mu\left[\left(n_{in}^H + n_{un}^H\right) - \left(n_{in}^L + n_{un}^L\right)\right]}{n_{in}^H - n_{in}^L} = \frac{\lambda_i\mu}{l}. \tag{16}$$

Lemma 2 describes an equilibrium where the $H$-type influencer purchases fake accounts to deter the $L$-type (who does not purchase) from mimicking. We call such fake-account purchasing *defensive purchasing*.

To understand the intuition behind this Lemma, we rewrite condition (15) as

$$c_f = \kappa + \frac{1}{1-\tau} d \leq \eta_2. \tag{17}$$

The left-hand side is the unit cost of fake accounts, and the right-hand side $\eta_2$ can be interpreted as the marginal revenue gain per fake account if the $L$-type moves from purchasing nothing to buying enough fake accounts to mimic a non-purchasing $H$-type and being treated as one. The condition (15) is the IC condition for the $L$-type. It ensures that the $L$-type influencer has an incentive to mimic the $H$-type if the latter purchases nothing. In other words, the $L$-type is willing to match the $H$-type's informed followers. Therefore, the $H$-type must purchase enough fake accounts to make the $L$-type indifferent between mimicking and not mimicking[6].

The $H$-type's equilibrium defensive purchase $x_H^{csep}$ is the difference between her number of informed followers and the highest number the $L$-type is willing to mimic (see proof of this Lemma for more details). The expression for $x_H^{csep}$ has the following interpretation. $\frac{c}{q_L} - \frac{c}{q_H}$ is the gap between an $H$-type and an $L$-type's number of real followers, if all consumers were informed. An $L$-type influencer's gain from mimicking is $\lambda_i \mu$ times of this gap (noting that $\lambda_i$ is the influencer's share and $\mu$ is the advertising value generated by each consumer). The $L$-type's cost of mimicking is $c_f l$, noting that she only needs to pay for the gap in the *informed* followers, which is a proportion $l$ of the gap in real followers (informed and uninformed). The denominator $\lambda_i (\varphi + c_f)$ is the unit cost of a fake account shouldered by the influencer. In sum, the $H$-type would purchase the number of fake accounts such that the $L$-type is indifferent between mimicking or not.

The fact that $H$-type may purchase fake accounts in equilibrium is interesting as one may intuitively think that $L$-type has the most incentives to buy fake accounts. For example, the entire deceptive advertising literature focuses on the deceptive behavior of $L$-type sellers (Piccolo, Tedeschi, and Ursino, 2018; Chen and Papanastasiou, 2021). In our study, when the highest early-

[6]The $H$-type's IC condition is automatically satisfied for the following reason. As we know, $H$-type's revenue loss from her best deviation – purchasing nothing and being treated as an $L$-type – is the same as $L$-type's revenue gain from purchasing nothing to mimicking the $H$-type. However, the $H$-type has to purchase fewer fake accounts to maintain the separating equilibrium than what the $L$-type must do to mimic the $H$-type, since the $H$-type has more informed followers to begin with. Therefore, when the $L$-type is break-even from mimicking the $H$-type, the $H$-type must lose profits by deviating to the purchase-nothing strategy. In other words, when the $L$-type's IC condition holds, the IC condition for $H$-type is naturally satisfied. We note that the advertiser's IR condition is naturally satisfied when the $H$-type's IC condition holds because the total surplus from advertising is positive. The $L$-type's IR condition is also naturally satisfied because she incurs no cost. Like before, the IR condition for the $H$-type is automatically satisfied when her IC condition holds.

follower count that the $L$-type is willing to mimic exceeds the $H$-type's informed consumers count, but less than the maximum number of fake accounts the $H$-type is willing to purchase, the $H$-type also has an incentive to buy fake accounts to deter the $L$-type's mimicry, knowing that this would not fool the advertiser – in a separating equilibrium, when the advertiser would infer the influencer's true type and therefore the number of real followers. The Lemma highlights that $H$-type influencers may buy fake accounts not to deceive, but to signal their high quality. This finding is consistent with the observation that high-status influencers also frequently purchase fake accounts (Mekuli, 2021) as in the Devumi case.

**Lemma 3.** *(Naturally Separating) Under the belief*

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

*the strategy profile* $\left(x_H^{nsep}, x_L^{nsep}\right) = (0,0)$ *is a unique naturally separating equilibrium if and only if:*

$$d > (1-\tau)\left(\eta_2 - \kappa\right) \tag{18}$$

*where* $\eta_2$ *is defined in Lemma 2.*

Lemma 3 states that the two types will be separated naturally when condition (18) holds. We have an intuitive explanation for this: from Lemma 2, we know that condition (18) indicates that the highest number of early followers that the $L$-type is willing to match is smaller than the $H$-type's informed followers. Therefore, the $L$-type can't afford to mimic the $H$-type, and the $H$-type can deter the $L$-type without purchasing fake accounts. It is straightforward to show that the IR conditions for both types of influencers and the advertiser are satisfied when there are no fake account purchasing.

We note that when the fake-account base cost $\kappa$ is above a threshold $\eta_2$, natural separation occurs without any platform anti-faking effort (i.e. $d = 0$). This extreme case is unrealistic given the prevalence of fake accounts in the current influencer marketing platforms. To rule out such a case and simplify the subsequent analysis, we assume:

**Assumption 2.** $\eta_2 > \kappa$.

This assumption ensures that the fake-account base cost is low enough so that some anti-faking effort is required to achieve natural separation.

### 4.2.3 Globally-refined Equilibrium

Lemmas 1, 2, and 3 characterize the unique locally-refined pooling and separating equilibria and the conditions for their existence, but multiple types of equilibria may still co-exist. We again use *LMSE* to select a unique equilibrium across equilibrium types (See the proof in the Appendix). The following Lemma summarizes the l-dominance relationship between different types of equilibria.

**Lemma 4.** *Suppose* $d \leq (1-\tau)(\eta_1 - \kappa)$, *so the pooling equilibrium exists. (a) If* $\eta_2 \leq \eta_1$, *when* $d \leq (1-\tau)(\eta_2 - \kappa)$, *the pooling equilibrium coexists with and dominates the costly separating equilibrium; when* $d \in [(1-\tau)(\eta_2 - \kappa), (1-\tau)(\eta_1 - \kappa)]$, *the pooling equilibrium coexists with and dominates the naturally separating equilibrium. (b) When* $\eta_2 > \eta_1$ *and* $d \leq (1-\tau)(\eta_1 - \kappa)$, *the pooling and costly separating equilibria coexist. The pooling equilibrium l-dominates the costly separating equilibrium.*

For notational convenience, we denote

$$\begin{aligned} d_1 &\equiv (1-\tau)(\eta_1 - \kappa) \\ d_2 &\equiv (1-\tau)(\eta_2 - \kappa) \end{aligned}$$

By Lemmas 1, 2, and 4, we can interpret $d_1$ as the minimum effort to break down a pooling equilibrium, $d_2$ as the minimum effort for the naturally separating to overtake costly separating. The next Proposition describes the globally-refined LMSE equilibrium and its existence conditions.

**Proposition 1.** *The influencer's globally-refined equilibrium strategy is as follows.*

*a.* ***(No-pool)*** *If* $d_1 \leq 0$, *there is no pooling equilibrium and*

$$(x_H^*, x_L^*) = \begin{cases} \left(x_H^{csep}, x_L^{csep}\right), & if\ d \leq d_2 \\ \left(x_H^{nsep}, x_L^{nsep}\right), & otherwise \end{cases}$$

*b.* ***(No-csep)*** *If $d_1 > 0$ and $d_2 \leq d_1$, there is no costly separating equilibrium and*

$$(x_H^*, x_L^*) = \begin{cases} \left(x_H^{pool}, x_L^{pool}\right), & if\ d \leq d_1 \\ \left(x_H^{nsep}, x_L^{nsep}\right), & otherwise \end{cases}$$

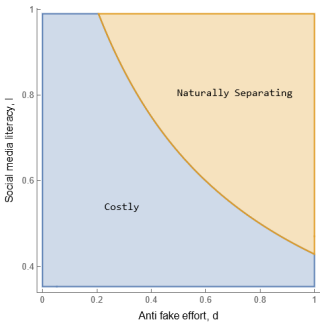
*c.* ***(All-eqm)*** *If $d_1 > 0$ and $d_2 > d_1$,each equilibrium is likely and*

$$(x_H^*, x_L^*) = \begin{cases} \left(x_H^{pool}, x_L^{pool}\right), & if\ d \leq d_1 \\ \left(x_H^{csep}, x_L^{csep}\right), & if d_1 < d \leq d_2 \\ \left(x_H^{nsep}, x_L^{nsep}\right), & if\ d > d_2 \end{cases}$$

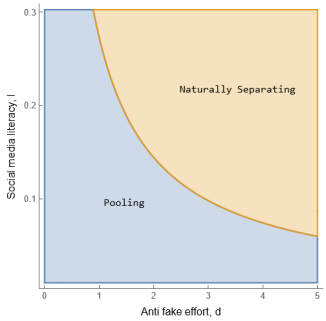
Proposition 1 suggests that the pooling equilibrium can not exist if the fake-account base cost $\kappa$ is relatively high (such that $d_1 = (1-\tau)(\eta_1 - \kappa) \leq 0$) (Case No-pool). In such a case, when the anti-fake effort $d$ is relatively low, a costly separating equilibrium is achieved; when $d$ is relatively high, a naturally separating equilibrium holds. These equilibrium scenarios are illustrated in Figure 3 (a), which shows that, for a relatively high fake-account base cost $\kappa$, as the anti-fake effort increases, the equilibrium regime transitions from costly separating to naturally separating.

When the fake-account base cost $\kappa$ is relatively low (such that $d_1 = (1-\tau)(\eta_1 - \kappa) > 0$), there are two scenarios: the costly separating equilibrium does not exist (Case No-csep) and does (Case All-eqm). The costly separating equilibrium does not exist when it is relatively easy to achieve natural separation (e.g., when the anti-fake technology level is high and the proportion of informed consumers is high). Specifically, when naturally separating overtakes costly separating (which requires $d > d_2$) when pooling is still sustainable (which requires $d \leq d_1$), or when the $H$-type still prefers pooling over costly separating (which requires $d < d_3$), the costly separating equilibrium cannot exist. This scenario is illustrated in Figure 3 (b), where the equilibrium transitions from pooling to naturally separating as the anti-fake effort $d$ increases. Otherwise, the costly separating equilibrium exists and we would observe the equilibrium transitions from pooling to costly separating and then to naturally separating as the anti-fake effort increases, as illustrated in Figure 3 (c).

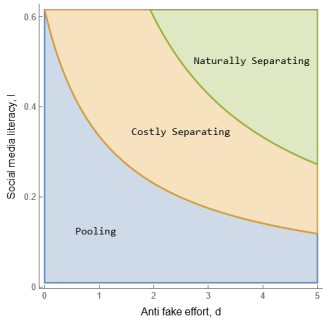
Taken together, Proposition 1 and the three Lemmas before it suggests that the influencer's




(a) Costly to naturally separating when $d_1 \leq 0$




(b) Pooling to naturally separating when $d_1 > 0$ and $d_2 \leq d_1$




(c) Pooling to costly separating then to naturally separating when $d_1 > 0$ and $d_2 > d_1$

(a) *Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 15, c_0 = 1, c_1 = 0.02, \kappa = 2, \varphi = 1, \tau = 0.8$.
(b) *Note:* $\mu = 10, \rho = 0.9, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 1, c_1 = 0.02, \kappa = 1, \varphi = 1, \tau = 0.9$.
(c) *Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 1, c_1 = 0.02, \kappa = 1, \varphi = 1, \tau = 0.5$.

Figure 3: Illustration of the equilibrium regime transitions

equilibrium behavior is quite "rugged." Specifically, as the equilibrium transitions from pooling to costly separating, the $L$-type's offensive purchasing first increases and then suddenly drops to zero. Meanwhile, the $H$-type's defensive purchasing first stays at zero and then suddenly jumps to a high level. This becomes more evident in the next section's comparative-static plots (Figure 4).

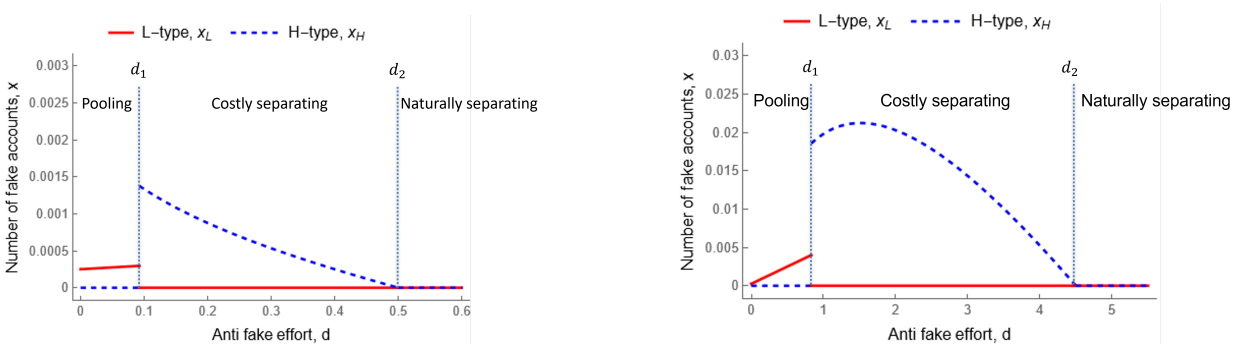
### 4.3 Comparative Statics

Having characterized the equilibrium strategy profile, we next 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* $x_L^{pool}$

- *increases in the platform's anti-fake effort* $d$ *and the social media literacy* $l$,
- *decreases in the anti-fake technology level* $\tau$ *and quality ratio* $r_q$, *and*
- *does not change with the fake-account base cost* $\kappa$, *the influencer's bargaining power* $\lambda_i$, *or the platform's bargaining power* $\lambda_p$.

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




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

*Note:* $\mu = 10, \rho = 0.12, l = 0.5, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 0.01, c_1 = 0.2, \kappa = 1, \varphi = 1$

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

out. Moreover, the $L$-type loses informed consumers more quickly than the $H$-type[7], implying the gap between the two types' informed followers increases. Consequently, the $L$-type must buy more fake accounts to fill in the gap. This is illustrated in Figure 4 (the pool equilibrium region in both panels). The intuition for the effect of social media literacy $l$ is similar: increasing the social media literacy leads to more informed followers for both types of influencers, and also enlarges the gap between the two types' informed followers. This forces the $L$-type to buy more to keep up. 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 $r_q$ makes the quality differential between the two types' smaller, as a result, the gap between two types' information followers shrinks. Consequently, the $L$-type needs fewer fake accounts to make up the difference. Finally, the number of fake accounts and the influencer's share of advertising surplus doesn't affect the gap between the two type's informed follower. Therefore, $L$-type's purchase is unaffected by the fake-account base cost or the influencer's bargaining power.

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

- *decreases in social media literacy $l$, the fake-account base cost $\kappa$, the quality ratio $r_q$, and the anti-fake technology level $\tau$,*
- *increases in the influencer's bargaining power $\lambda_i$, and*

[7]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 (9)).

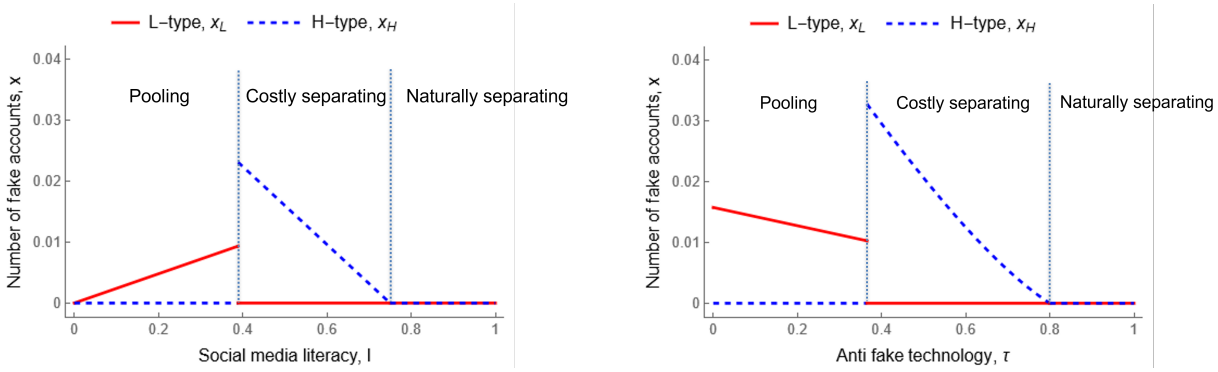
- *either decreases with the anti-fake effort $d$, when the following condition holds, or first increases then decreases in $d$, otherwise.*

$$c_1 (1-\tau)^2 \leq \frac{\lambda_i (\mu + \varphi l) c_0}{(\kappa + \varphi)(\lambda_i \mu - \kappa l)} \tag{19}$$

The intuition for Proposition 3 is as follows. As we learn from Lemma (2), the number of fake accounts purchased by the $H$-type $x_H^{csep}$ is tied to the highest number of informed followers the $L$-type is willing to mimic. Furthermore, the $H$-type's defensive purchasing $x_H^{csep}$ increases with the $L$-type's net gain from mimicry $(\lambda_i \mu - c_f l)\left(\frac{c}{q_L} - \frac{c}{q_H}\right)$ and decreases with the influencer's share of costs associated with fake accounts $\lambda_i (\varphi + c_f)$. 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 $(\lambda_i \mu - c_f l)$. Increasing social media literacy $l$ decreases the marginal gain from mimicry (as there will be fewer uninformed consumers), and thus the defensive purchasing $x_H^{csep}$. Increasing the fake-account base cost $\kappa$ leads to an increase in fake account cost $c_f$, which decreases $L$-type's marginal gain from mimicry and increases influencer's share of fake-account costs. Both effects lead to decreased defensive purchasing. Increasing the quality ratio $q_L/q_H$ shrinks the gap between the two types' followers and thus decreases the $H$-type's purchasing. Increasing the anti-fake technology level $\tau$ decreases the consumers' nuisance cost $c$, and thus the gap between the two type's followers; it also increases the cost of fake accounts $c_f$. As argued earlier, both lead to decreased defensive purchasing. Increasing the influencer's bargaining power $\lambda_i$ 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 cost of fake accounts $c_f$ and the consumers' nuisance cost $c$ increase. The former (the "*higher-fake-account-cost*" effect) has a *negative* impact on defensive purchasing, with a marginal effect of $m_1 = -\frac{1}{1-\tau}\frac{\lambda_i \mu + \varphi l}{(c_f+\varphi)^2}\frac{q_H - q_L}{\lambda_i q_H q_L} c$. The latter (the "*higher-nuisance-cost*" effect) has a *positive* effect on the $L$-type's marginal gain from mimicry and thus a positive effect on the defensive purchasing, with a marginal effect of $m_2 = c_1 (1-\tau)\frac{\lambda_i \mu - c_f l}{c_f + \varphi}\frac{q_H - q_L}{\lambda_i q_H q_L}$. As $d$ increases, fake account cost $c_f$ increases, and the higher-nuisance-cost effect $m_2 \to 0$. Therefore, as $d$ increases, we expect the negative effect $m_1$ to dominate eventually, causing defensive purchasing to decrease. Moreover, when the technology level $\tau$ is relatively high such that $c_1 (1-\tau)^2 \leq \frac{\lambda_i(\mu+\varphi l)c_0}{(\kappa+\varphi)(\lambda_i\mu - \kappa l)}$, 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, we expect the positive effect to dominate 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).[8]




(a) Impact of social media literacy (b) Impact of anti-fake technology level

(a) *Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 5, c_0 = 0.01, c_1 = 0.2, \kappa = 1, \varphi = 1, \tau = 0.5, d = 1.5$
(b) *Note:* $\mu = 10, \rho = 0.1, l = 0.5, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 5, c_0 = 0.01, c_1 = 0.2, \kappa = 1, \varphi = 1, d = 1.$

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

Figure 5 illustrates the impact of social media literacy ($l$) and technology level ($\tau$). In panel (a), we observe that $L$-type's offensive purchasing increases in social media literacy, as predicted by Proposition 2, whereas the $H$-type's defensive purchasing under the costly separating regime decreases, as predicted by Proposition 3 (case b). Figure 5 (b) tells us that increasing the anti-fake technology level can reduce both offensive and defensive fake-account purchasing.

## 5 The Platform's Optimal Anti-fake Effort

Having analyzed the influencer's equilibrium fake-account purchasing, we turn our attention to the platform's optimal anti-fake effort. First, the following Lemma establishes how the platform profit changes with the anti-fake effort under each type of equilibrium. We denote $\pi_p^{pool}(d), \pi_p^{csep}(d)$, and $\pi_p^{nsep}(d)$ as the platform profit as a function of anti-fake effort $d$ under the pooling, costly separating, and naturally separating equilibria, respectively, provided that the chosen $d$ supports the equilibrium.

[8] 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.

| Conditions | | Case | $d^*$ | Equilibrium |
|---|---|---|---|---|
| No-pool: $d_1 \leq 0$ | $\pi_p^{csep}(0) \geq \pi_p^{nsep}(d_2)$ | 1 | 0 | Costly separating |
| | otherwise | 2 | $d_2$ | Naturally separating |
| No-csep: $d_1 > 0$ and $d_2 \leq d_1$ | $\pi_p^{pool}(0) \geq \pi_p^{nsep}(d_1)$ | 3 | 0 | Pooling |
| | otherwise | 4 | $d_1$ | Naturally separating |
| All-eqm: $d_1 > 0$ and $d_2 > d_1$ | $\pi_p^{pool}(0) \geq \pi_p^{nsep}(d_2)$ | 5 | 0 | Pooling |
| | otherwise | 6 | $d_2$ | Naturally separating |

Table 2: Platform's optimal anti-fake effort

**Lemma 5.** *$\pi_p^{pool}(d)$ and $\pi_p^{nsep}(d)$ decrease in $d$. $\pi_p^{csep}(d)$ is a convex function of $d$.*

The intuitions behind these findings are as follows. In general, the anti-fake effort affects platform profits in two ways. First, an increase in anti-fake effort can increase consumer nuisance cost and thus reduce the number of participating consumers. This "*consumer-inconvenience*" effect negatively affects the total advertising surplus and the platform's share of that surplus. Second, the anti-fake effort may affect the costs of fake accounts (the "*fake-account-cost*" effect), which consists of the cost of purchasing fake accounts and the cost of advertising to them. Reducing the number of fake accounts helps reduce such costs and thus increase platform's profitability under the surplus sharing scheme. Under the pooling equilibrium, besides increasing user inconvenience, increasing the anti-fake effort also leads to more fake accounts (Proposition 2). Both effects reduce the platform profitability. So, $\pi_p^{pool}(d)$ decreases with the anti-fake effort overall. Under a naturally separating equilibrium, there are no fake-account purchases, so the only effect is the negative consumer-inconvenience effect. Thus, $\pi_p^{nsep}(d)$ also decreases with the anti-fake effort. Under the costly separating equilibrium, from Proposition 3, we know that the anti-fake effort may either increase or decrease $H$-type's defensive purchasing. In the former case, $\pi_p^{csep}(d)$ decreases in the anti-fake effort $d$, but in the latter case, $\pi_p^{csep}(d)$ may increase or decrease in $d$. In the proof, we show that $\pi_p^{csep}(d)$ is a convex function of $d$, so that $\pi_p^{csep}(d)$ may increase, decrease, or first decrease and then increase in $d$.

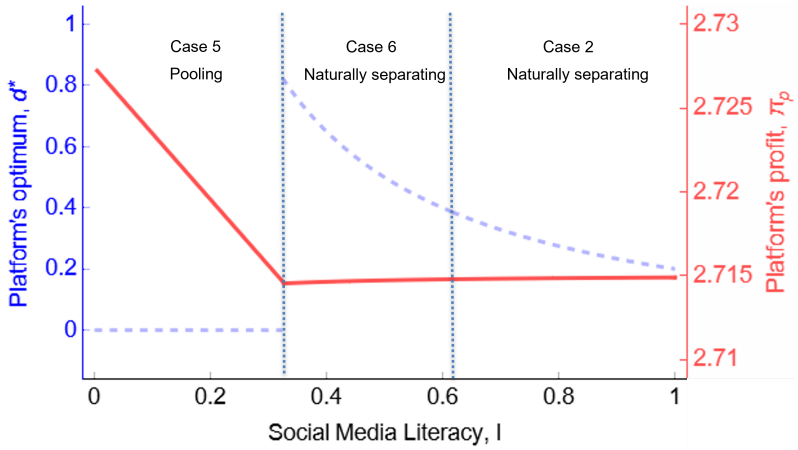
**Lemma 6.** *The platform's profit under the pooling equilibrium with $d = 0$ dominates that under the costly separating equilibrium with $d = d_1$, i.e., $\pi_p^{pool}(0) > \pi_p^{csep}(d_1)$.*

Given Lemma 5 and 6, we can obtain the platform's optimal anti-fake effort as summarized in Proposition 4.

**Proposition 4.** *The platform's optimal anti-fake effort is given in Table 2. The platform may optimally induce pooling, costly separating, or naturally separating equilibrium.*

This proposition shows that each type of equilibrium is possible. The platform tends to choose two extremes: on one hand, the platform may do nothing about the fake accounts (i.e. $d^* = 0$), resulting in either a pooling equilibrium or a costly separating one (when the former does not exist). In such a case, the platform avoids the loss of consumers due to nuisance costs of anti-fake efforts, but suffers from loss of profits due to fake accounts. On the other hand, the platform may optimally exert a high anti-fake effort to eliminate all fake accounts (thus achieving a naturally separating equilibrium). In the latter case, the platform avoids loss of profits due to fake accounts but loses some consumers due to nuisance costs from anti-fake efforts and fewer uninformed followers.[9]

## 5.1 Comparative Statics for Platform's Optimal Strategy and Profit



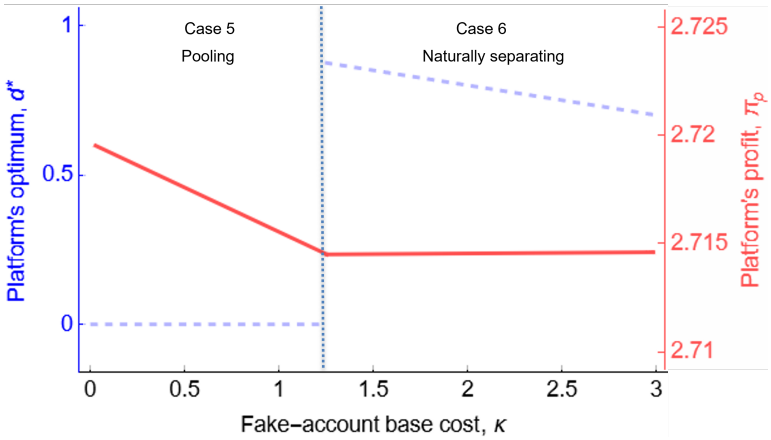

*Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 1, c_1 = 0.02, \kappa = 1, \tau = 0.9, \varphi = 1$

Figure 6: Impact of social media literacy on platform's optimal anti-fake effort and profit

As the conditions for Proposition 4 are not analytically tractable, we rely on numeric methods to obtain further insights on how underlying parameters drive the platform's optimal anti-fake effort. We focus on the effect of social media literacy, fake-account base cost, and anti-fake technology level as these are of high interest.

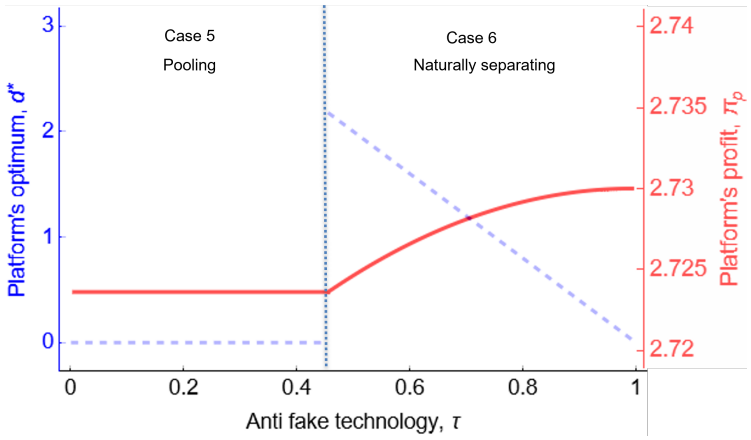
In the case of social media literacy, Figure 6 shows that the platform profit first decreases then increases, but the maximum profit is achieved at the lowest level of social media literacy.

[9] From (10), one can observe that the total number of uninformed followers under a pooling equilibrium $(1-l)\left[1-\frac{c}{(1-\rho)q_L+\rho q_H}\right]$ is higher than that under a separating equilibrium $(1-l)\left[(1-\rho)\left(1-\frac{c}{q_L}\right)+\rho\left(1-\frac{c}{q_H}\right)\right]$.




*Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 1, c_1 = 0.02, l = 0.3, \tau = 0.9, \varphi = 1$

Figure 7: Impact of fake-account base cost on platform's optimal anti-fake effort and profit




*Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 1, c_1 = 0.02, \kappa = 1, \varphi = 1, l = 0.6$

Figure 8: Impact of anti-fake technology on platform's optimal anti-fake effort and profit

Therefore, the social media platform is not incentivized to improve consumers' social media literacy. Specifically, when social media literacy is relatively low, it is optimal for the platform to do nothing about fake accounts. In such a case, the platform's profit decreases because the $L$-type's offensive purchasing increases with social media literacy (Proposition 2). As social media literacy exceeds a certain threshold, the platform finds it optimal to switch to the naturally separating equilibrium with a high anti-fake effort. As social media literacy further increases, the optimal effort decreases, which reduces the "consumer-inconvenience" effect and causes the platform's profit to increase.

From Figure 7, we find a similar pattern in the platform's optimal anti-fake effort and profit as a function of the fake-account base cost. Therefore, measures aimed at raising fake-account base costs (such as imposing a legal penalty on fake account trading) may harm the platform's profit.

Increasing the anti-fake technology level has different effects on the anti-fake effort and platform profits. From Figure 8, the anti-fake effort and platform profit are firstly unresponsive to increases in the anti-fake technology under the pooling equilibrium. Once the technology level exceeds a certain threshold, the platform switches to a naturally-separating equilibrium, under which the platform's profit increases and its anti-fake effort decreases with the anti-fake technology level. Therefore, the platform is incentivized to invest in improving its anti-fake technology.

In sum, these numerical examples highlight that the platform's profit function and equilibrium anti-fake effort are also "rugged" – they behave quite differently under different equilibrium regimes, making it difficult to implement the optimal anti-fake effort and evaluate different anti-fake strategies.

# 6 Consumer Welfare Analysis

So far, we have examined the equilibrium outcomes and anti-fake effort from the platform's point of view. It is also important to examine the same from a consumer welfare point of view as it is the focus of stakeholders such as policymakers and consumer protection agencies. We measure consumer welfare as the sum of expected payoffs of informed and uninformed consumers. We use $U^{pool}(d), U^{csep}(d)$, and $U^{nsep}(d)$ to denote consumer welfare as a function of the anti-fake effort under each equilibrium, respectively. We denote $d^C$ as the consumer-optimal anti-fake effort.

**Lemma 7.** *$U^{pool}(d), U^{csep}(d)$, and $U^{nsep}(d)$ all monotonically decrease in anti-fake effort $d$.*

Intuitively, under each equilibrium, the perceived quality of the influencer does not change with the anti-fake effort $d$ and so the only way anti-fake effort affects consumer welfare is by affecting their nuisance cost, which decreases with $d$. Consequently, the total consumer welfare monotonically decreases in the anti-fake effort, $d$. This also implies that consumers always prefer costly separating to naturally separating because the latter imposes a higher nuisance cost.

Lemma 7 does not mean that consumers always prefer zero anti-fake effort, however. This is because, ceteris paribus, uninformed consumers would benefit from higher total welfare when the influencer types are separated than when they are pooled.[10] When the benefit of separation is high

[10]This can be seen from Proof of Lemma 7 in Appendix A, the total welfare of uninformed followers under a separating equilibrium is higher than that under a pooling equilibrium.

and the nuisance cost from anti-fake efforts is low, consumers may prefer a separating equilibrium. The following Proposition summarizes when the optimal consumer welfare is achieved and how the consumer-optimal anti-fake effort compares to the platform-optimal.

**Proposition 5.** *The consumer-optimal anti-fake effort $d^C$ and its relationship with the platform optimal $d^*$ are given by Table 3.*

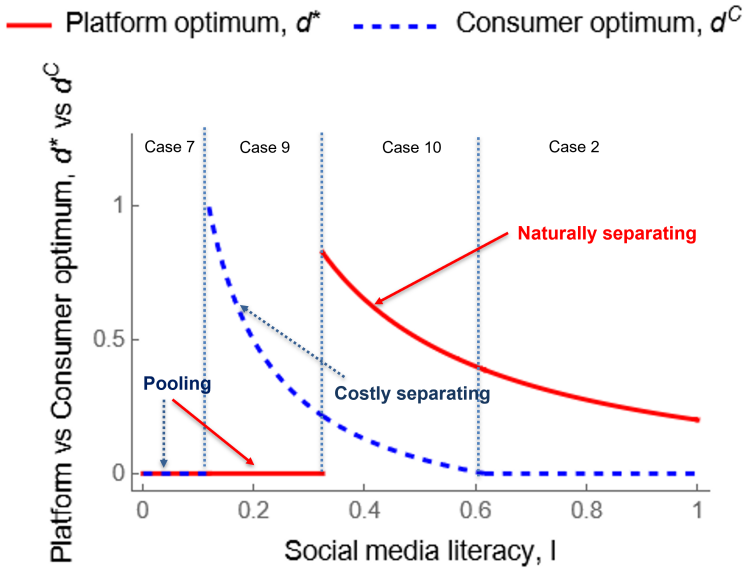
Proposition 5 shows that consumers may prefer either zero anti-fake effort or a high anti-fake effort $d_1$. The former usually corresponds to a pooling equilibrium (cases 3-4 and 7-8), but also it can also be a costly separating equilibrium (with zero anti-fake effort) when the pooling equilibrium does not exist (cases 1-2). Latter usually means a costly separating equilibrium (case 9-10), but it can also be a naturally separating equilibrium when the costly separating one does not exist (cases 5-6). As discussed earlier, the advantage of zero anti-fake effort lies in no nuisance costs, and that of a high anti-fake effort lies in the benefit of separation for uninformed consumers when the nuisance cost is low. When the anti-fake technology is high (so a high anti-fake effort does not add much nuisance cost) and the proportion of uninformed consumers is high (so the benefit of separation is even greater), consumers are more likely to prefer a high anti-fake effort.

By overlapping the conditions for the consumers' and the platform's preferences, we can explain all the cases in Table 3. For example, if the anti-fake technology level, $\tau$, is high but the proportion of informed consumers, $l$, is low, the consumers prefer separating to pooling due to the low nuisance cost, and the platform will prefer pooling to naturally separating as there is a big loss of uninformed consumers if the equilibrium switch, which is the case 9.

To illustrate the relationship between consumer- and platform-optimal anti-fake efforts, we plot them as a function of social media literacy. As seen in Figure 9, when social media literacy is relatively low, both the platform's and consumers' optimal anti-fake efforts are zero. As social media literacy increases, the platform's optimal anti-fake effort falls below consumers' optimum. As social media literacy further increases, the platform may over-invest in anti-fake efforts, ending up hurting consumers' welfare.

| Conditions | | | Case | Platform Optimal | | Consumer Optimal | | Comparison |
|---|---|---|---|---|---|---|---|---|
| | | | | $d^*$ | Eqm | $d^C$ | Eqm | |
| No-pool: $d_1 \leq 0$ | NA | $\pi_p^{nsep}(d_2) \leq \pi_p^{csep}(0)$ | 1 | 0 | csep | 0 | csep | $d^* = d^C$ |
| | | otherwise | 2 | $d_2$ | nsep | | | $d^* > d^C$ |
| No-csep: $d_1 > 0$ and $d_2 \leq d_1$ | $U^{nsep}(d_1) \leq U^{pool}(0)$ | $\pi_p^{nsep}(d_1) \leq \pi_p^{pool}(0)$ | 3 | 0 | pool | 0 | pool | $d^* = d^C$ |
| | | otherwise | 4 | $d_1$ | nsep | | | $d^* > d^C$ |
| | otherwise | $\pi_p^{nsep}(d_1) \leq \pi_p^{pool}(0)$ | 5 | 0 | pool | $d_1$ | nsep | $d^* < d^C$ |
| | | otherwise | 6 | $d_1$ | nsep | | | $d^* = d^C$ |
| All-eqm: $d_1 > 0$ and $d_2 > d_1$ | $U^{csep}(d_1) \leq U^{pool}(0)$ | $\pi_p^{nsep}(d_2) \leq \pi_p^{pool}(0)$ | 7 | 0 | pool | 0 | pool | $d^* = d^C$ |
| | | otherwise | 8 | $d_2$ | nsep | | | $d^* > d^C$ |
| | otherwise | $\pi_p^{nsep}(d_2) \leq \pi_p^{pool}(0)$ | 9 | 0 | pool | $d_1$ | csep | $d^* < d^C$ |
| | | otherwise | 10 | $d_2$ | nsep | | | $d^* > d^C$ |

Table 3: Consumer-optimal anti-fake effort and comparison with the platform-optimal




*Note:* $\mu = 10, \rho = 0.1, \lambda_i = 0.3, \lambda_p = 0.3, q_H = 20, q_L = 10, c_0 = 1, c_1 = 0.02, \kappa = 1, \tau = 0.9, \varphi = 1$

Figure 9: Impact of social media literacy on platform's vs. consumer optimal anti-fake effort

## 6.1 Welfare Comparative Statics

In this section, we conduct comparative static analyses of consumer welfare under the platform's optimal strategy. We analyze how consumer welfare will change over the parameters (social media literacy $l$, the fake-account base cost $\kappa$, anti-fake technology level $\tau$) under different equilibrium scenarios(See technical details in Appendix).

**Proposition 6.** *The welfare impact of parameters is given in Table 4. Increasing social media literacy, fake-account base cost, and anti-fake technology level weakly improves consumer welfare.*

| **Perturbed Parameter** | Pooling $d^* = 0$ | Costly Separating $d^* = 0$ | Naturally Separating $d^* = d_1$ *or* $d_2$ |
|---|---|---|---|
| Social media literacy, $l$ | $\uparrow$ | $=$ | $\uparrow$ |
| The fake-account base cost, $\kappa$ | $=$ | $=$ | $\uparrow$ |
| Anti-fake technology level, $\tau$ | $=$ | $=$ | $\uparrow$ |

Table 4: How three model parameters impact consumer welfare

Consumer welfare loss comes from two sources: the nuisance cost of anti-fake efforts and the cost of pooling which causes uninformed consumers to make suboptimal decisions. In the pooling equilibrium, increasing social media literacy will reduce the cost of pooling because there are fewer uninformed consumers. Increasing the other two parameters, fake account base cost and anti-fake technology, does not affect the cost of pooling or consumer welfare. In the costly separating with zero anti-fake effort, as the influencers are already separated and there is no nuisance cost from the anti-fake effort, thus, neither parameter has any impact on consumer welfare. Finally, in the naturally separating equilibrium, there is no cost of pooling. Meanwhile, the anti-fake effort ($d_1$ or $d_2$) required for natural separation decreases in social media literacy, fake-account base cost, and anti-fake technology (see the definitions of $d_1$ and $d_2$), thus, consumer welfare improves with the three parameters.

# 7 Extension to Three Types of Influencers

Our main model assumes only two levels of influencer quality. In reality, there are many influencer quality levels and most influencers are neither the lowest-quality type nor the highest-quality type. It would be important to examine whether the insights obtained using two influencer types would generalize to more influencer types. To address this question, we extend our model to three influencer types.

In the extended model, we assume an influencer's quality is drawn randomly from three levels, $\{q_L, q_M, q_H\}$ ($q_L < q_M < q_H$), with probabilities $\rho_L, \rho_M, \rho_H$, respectively, where $\rho_H = 1 - \rho_M - \rho_L$. We call an influencer $M$-type ($H$-type, $L$-type) influencer if her quality is $q_M$ ($q_H, q_L$). The following Proposition summarizes the types of LMSE equilibria (see Appendix B for technical details).

**Proposition 7.** *With three influencer types, the LMSE strategy profile is given by Table 5 (the corresponding beliefs and conditions are provided in the Appendix).*

| *Equilibrium* | *H-Type* | *M-Type* | *L-Type* |
|---|---|---|---|
| Costly fully separating (H\|M\|L) | $\frac{\lambda_i\mu - c_f l}{\lambda_i(\varphi + c_f)}\left(\frac{c}{q_L} - \frac{c}{q_H}\right)$ | $\frac{\lambda_i\mu - c_f l}{\lambda_i(\varphi + c_f)}\left(\frac{c}{q_L} - \frac{c}{q_M}\right)$ | 0 |
| Naturally fully separating (H\|M\|L) | 0 | 0 | 0 |
| Fully pooling (HML) | 0 | $l\left(\frac{c}{q_M} - \frac{c}{q_H}\right)$ | $l\left(\frac{c}{q_L} - \frac{c}{q_H}\right)$ |
| H\|ML Hybrid | 0 | 0 | $l\left(\frac{c}{q_L} - \frac{c}{q_M}\right)$ |
| HM\|L Hybrid | 0 | $l\left(\frac{c}{q_M} - \frac{c}{q_H}\right)$ | 0 |

Table 5: Equilibria of fake accounts purchasing strategy for three types of influencers

As seen from Proposition 7 and Table 5 , the pooling (HML) and separating (both costly and naturally) equilibria (H|M|L) extend to the three-type case, but we also obtain two new "hybrid" equilibrium types: i.e., some types pool together while separating from other type(s). Specifically, the $M$-type influencer may pool with the $H$-type and separate from the $L$-type (HM|L) or pool with the $L$-type and separate from the $H$-type (H|ML). Similar to the main model, an influencer may purchase fake accounts defensively (i.e. to separate from a lower type) or offensively (i.e. to pool with a higher type). Different from the main model, the $M$-type's purchase may be simultaneously defensive and offensive (i.e. in the case of HM|L). In addition, we note that in several equilibrium scenarios, two of the three-influencer types purchase fake accounts. This indicates that as the number of influencer types increases, fake-account purchasing can become more prevalent. This finding may explain why fake-account purchasing appears to be prevalent and occurs among both high- and low-profile influencers.

# 8 Discussion and Conclusion

Motivated by the prevalence of fake social media 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 has dual effects: increasing the cost of fake accounts and 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 optimization, and the effects of a few parameters (namely, the level of the anti-

fake technology, social media literacy, and base costs of fake accounts) on the equilibrium number of fake accounts, platform profits, and consumer welfare.

### 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, which has not been examined in the literature. Different from previous studies of other deceptive behaviors in related settings, we find that purchasing fake accounts can also be a signaling device used by high-quality influencers to differentiate from low-quality ones. This equilibrium scenario receives little attention in the prior literature, and yet it holds important implications on how we view and tackle the problem of fake accounts. In addition, we also find a scenario where the platform can induce a natural separation of two types of influencers without either type purchasing fake accounts. We hope that our new equilibrium findings can inspire follow-up research, including empirical validation of our findings and extension of our findings to other settings.

Second, we contribute to the underdeveloped literature on coping strategies for deceptive behaviors by studying a host of underexplored interventions including mounting an "imperfect" anti-fake effort, increasing social media literacy, and increasing fake-account costs. Our analyses provide several novel insights into how these interventions may fare. We report several negative findings but also bring some encouraging news. On the negative side, a few anti-fake interventions may have adverse or no effects on the number of fake accounts. For example, under the pooling equilibrium, increasing anti-fake efforts and social media literacy can increase the number of fake accounts, whereas increasing the base cost of fake accounts has no impact. On the positive side, we find that many interventions are more likely to have the intended effects under the costly separating equilibrium. Moreover, increasing the anti-fake technology level can consistently reduce the number of fake accounts.

Thirdly, our analyses of consumer welfare also produce several interesting insights. First, consumers may not always prefer the highest anti-fake effort. In fact, they may sometimes prefer no anti-fake effort at all and they also always prefer an intermediate anti-fake effort that merely induces costly separation rather than a high effort that eliminates all fake accounts. Moreover, the platform may be inadequately aggressive or overly aggressive in tackling fake accounts, compared

to the level desired by consumers. For example, we found cases where consumers prefer a costly separating equilibrium with an intermedia anti-fake effort whereas the platform either prefers a zero-effort pooling equilibrium or a high-effort naturally-separating equilibrium. The two parties also diverge on what interventions are beneficial to them. For example, the platform often has no incentive to invest in social media literacy or increase the base cost of fake accounts, whereas doing so generally improves consumer welfare.

### 8.2 Managerial implications

Our findings hold managerial implications for both platforms and policymakers/consumer protection agencies. For platforms, it is important to recognize that fake accounts may be used by low-quality influencers to distort popularity signals or a constructive purpose or by high-quality influencers to signal their superior quality. As such, depending on how fake accounts are used in the current equilibrium, the platform strategies may be different.

In general, the platform may find it optimal to either do nothing about the fake account problem, or to exert a high anti-fake effort to induce a natural separation among influencers with different qualities. The former is more profitable when the anti-fake technology is relatively ineffective and the proportion of uninformed consumers is high on the platform. In such a case, it is too costly for the platform to induce a natural separation of influencer types due to the high nuisance costs imposed by anti-fake efforts. Exerting an inadequate level of anti-fake effort could exacerbate the fake account problem as low-quality influencers fight back by buying even more fake accounts. In general, it is not optimal for platforms to exert more anti-fake effort than what is necessary to induce the sorting of influencers by their quality. The platform generally benefits from better anti-fake technology, thus should invest in improving such technology. In contrast, the platform may not benefit from increased consumer social literacy or tougher regulations on fake account trading.

For consumer protection agencies and policymakers, we note that it is never optimal to eliminate all fake accounts. This is because an intermediate level of anti-fake effort may already achieve the goal of "separating" influencers by quality, though such a goal may require high-quality influencers to buy fake accounts to reinforce their superiority. Pushing anti-fake efforts beyond this point can cause greater inconvenience among consumers without providing further benefits to consumers. In fact, when the anti-fake technology is very ineffective and the proportion of informed consumers is

high, consumers may prefer zero anti-fake effort and tolerate the fake accounts in the system and the distortion of quality signals they cause. Consumer protection agencies should note that platforms may not voluntarily adopt anti-fake strategies that benefit consumers: for example, platforms may be unwilling to invest in social media literacy or adopt harsher penalties for fake account trading. One common ground between consumer protection agencies and social media platforms is the anti-fake technology: increasing the effectiveness of such technologies can benefit both the platforms and the consumers.

### 8.3 Limitations and future work

As a first step toward understanding the fake account problem and coping strategies, we have simplified our analysis by focusing on the fake accounts created to help influencers gain popularity. Further work should examine other types of fake accounts, such as ones that profit from spreading scams, malware, and identity theft or politically motivated ones. Relatively, our model assumes that the fake accounts impose costs on advertisers and influencers, but not directly on consumers (other than distorting the quality signals and rendering their following decisions suboptimal). Future research could relax this assumption by allowing fake accounts to have a negative spillover effect on consumer experiences. In the latter case, we expect the platform to have stronger incentives to reduce the number of fake accounts, but conjecture that many of the intuitions developed in our model to still hold. Finally, because fake accounts are at the root of many other deceptive behaviors, e.g. fake reviews. it would be interesting to jointly consider the problem of fake accounts and other deceptive behaviors that depend on it.

## References

Al Zou'bi, Reem M (2022). "The impact of media and information literacy on students' acquisition of the skills needed to detect fake news". In: *Journal of Media Literacy Education* 14.2, pp. 58–71. URL: https://digitalcommons.uri.edu/jmle-preprints/28.

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

Che, Yeon-Koo and Johannes Hörner (2018). "Recommender systems as mechanisms for social learning". In: *The Quarterly Journal of Economics* 133.2, pp. 871–925.

Chen, Jenn (2022). *How to measure the value of social media*. URL: https://sproutsocial.com/insights/social-media-value/.

Chen, Jin, Luyi Yang, and Kartik Hosanagar (2022). "To Brush or Not to Brush: Product Rankings, Consumer Search, and Fake Orders". In: *Information Systems Research*.

Chen, Li and Yiangos Papanastasiou (2021). "Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation". In: *Management Science Publication* February, pp. 1–17.

Condliffe, Jamie (2020). *Stopping fake accounts is a cat-and-mouse game. Can Facebook win with AI?* URL: https://www.protocol.com/facebook-machine-learning-fake-accounts.

Confessore, Nicholas et al. (2018). *The Follower Factory*. URL: https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html?module=inline.

Corts, Kenneth S. (2013). "Prohibitions on false and unsubstantiated claims: Inducing the acquisition and revelation of information through competition policy". In: *Journal of Law and Economics* 56.2, pp. 453–486.

— (2014). "Finite optimal penalties for false advertising". In: *Journal of Industrial Economics* 62.4, pp. 661–681.

De Veirman, Marijke, Veroline Cauberghe, and Liselot Hudders (2017). "Marketing through instagram influencers: The impact of number of followers and product divergence on brand attitude". In: *International Journal of Advertising* 36.5, pp. 798–828.

Dhawan, Sanjeev and Ekta (2016). "Implications of Various Fake Profile Detection Techniques in Social Networks". In: *IOSR Journal of Computer Engineering (IOSR-JCE)*, pp. 49–55.

Federal Trade Commission (2019). "Devumi, Owner and CEO Settle FTC Charges They Sold Fake Indicators of Social Media Influence". In: *ftc.gov*. URL: https://www.ftc.gov/news-events/news/press-releases/2019/10/devumi-owner-ceo-settle-ftc-charges-they-sold-fake-indicators-social-media-influence-cosmetics-firm.

Freixa, Sara (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, Hong et al. (2019). "Economic Analysis of Reward Advertising". In: *Production and Operations Management* 28.10, pp. 2413–2430.

Guo, Xiaomeng, Guang Xiao, and Fuqiang Zhang (2017). "Effect of Consumer Awareness on Corporate Social Responsibility under Asymmetric Information". In: *SSRN Electronic Journal*, pp. 1–48.

Hao, Karen (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, Seung A.Annie and Joe Phua (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". In: *Journal of Advertising* 43.2, pp. 181–195.

Kaudelka, Lauren (2021). *THE PROS AND CONS OF USING CAPTCHA ON YOUR SENIOR LIVING WEBSITE.* URL: https://blog.growmarkentum.com/pros-and-cons-using-captcha.

Mailath, George J., Masahiro Okuno-Fujiwara, and Andrew Postlewaite (1993). "Belief-based refinements in signalling games". In: *Journal of Economic Theory* 60.2, pp. 241–276.

Mayzlin, Dina (2006). "Promotional chat on the internet". In: *Marketing Science* 25.2, pp. 155–163.

MediaLiteracyNow (2022). *What is Media Literacy?* URL: https://medialiteracynow.org/what-is-media-literacy/.

Mekuli, Arta (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, Elaine and Hannah Murphy (2019). *Facebook's massive fake numbers problem.* URL: https://www.latimes.com/business/technology/story/2019-11-18/facebooks-massive-fake-numbers-problem.

Nicas, Jack (2020). *Why cant 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, Barbara (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.

Pahwa, Aashish (2022). *Influencer Business Model — How Do Influencers Make Money?* URL: https://www.feedough.com/influencer-business-model-make-money/.

Papanastasiou, Yiangos (2020). "Fake news propagation and detection: A sequential model". In: *Management Science* 66.5, pp. 1826–1846.

Papanastasiou, Yiangos, Kostas Bimpikis, and Nicos Savva (2018). "Crowdsourcing Exploration". In: *Management Science* 64.4, pp. 1727–1746.

Pennycook, Gordon et al. (2020). "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings". In: *Management Science* 66.11, pp. 4944–4957.

Piccolo, Salvatore, Piero Tedeschi, and Giovanni Ursino (2018). "Deceptive advertising with rational buyers". In: *Management Science* 64.3, pp. 1291–1310.

Raturi, Rohit (2018). "Machine Learning Implementation for Identifying Fake Accounts in Social Network". In: *International Journal of Pure and Applied Mathematics* 118.20, pp. 4785–4797.

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

Shin, Euncheol (2017). "Monopoly pricing and diffusion of social network goods". In: *Games and Economic Behavior* 102, pp. 162–178. URL: http://dx.doi.org/10.1016/j.geb.2016.12.004.

Spangler, Todd (2021). *YouTube Tops 2 Million Creators in Ad-Revenue Sharing Program.* URL: https://variety.com/2021/digital/news/youtube-partner-program-2-million-creators-1235045674/?sub_action=logged_in.

Stolzoff, Simone (2018). *The problem with social media has never been about bots. It's always been about business models.* URL: https://qz.com/1449402/how-to-solve-social-medias-bot-problem/.

Stringhini, Gianluca et al. (2013). "Follow the Green: Growth and Dynamics in Twitter Follower Markets". In: *Proceedings of the 2013 conference on Internet measurement conference*, pp. 163–176.

Taibi, Davide et al. (2021). "An innovative platform to promote social media literacy in school contexts". In: *Proceedings of the European Conference on e-Learning, ECEL*, pp. 460–470.

Warwick, Stephen (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, Kenneth C. and Yi Zhu (2009). "Click fraud". In: *Marketing Science* 28.2, pp. 293–308.

Wong, Queenie (2019). *TikTok is filled with adult-dating scams and fake accounts*. URL: https://www.cnet.com/news/privacy/tiktok-is-filled-with-adult-dating-scams-and-fake-accounts-report-says/.

Yuan, Dong et al. (2019). "Detecting fake accounts in online social networks at the time of registrations". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1423–1438.

# A Appendix

## A.1 Proof of Lemma 1.

*Proof.* By the definition of the pooling equilibrium, the two types of influencers should have an identical number of early followers, namely, $n_2^H = n_2^L \equiv n_2^*$. We first show that $n_2^H = n_2^L = n_{in}^H$ (i.e., the $H$-type does not buy fake accounts and the $L$-type buys enough to make up the difference) with the following belief is a PBE:

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

The equilibrium influencer strategy profile is represented as

$$\left(x_H^{pool}, x_L^{pool}\right) = \left(0, n_{in}^H - n_{in}^L\right) \tag{20}$$

and the equilibrium profits are, respectively

$$\pi_L^{pool} = \lambda_i \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - \left(1-\rho\right)\left(\varphi + c_f\right) x_L^{pool}\right] + c_f \left[\left(1-\rho\right) x_L^{pool} - x_L^{pool}\right] \tag{21}$$

$$\pi_H^{pool} = \lambda_i \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - \left(1-\rho\right)\left(\varphi + c_f\right) x_L^{pool}\right] + c_f \left(1-\rho\right) x_L^{pool} \tag{22}$$

We first argue that $x_H = 0$ is optimal for the $H$-type. This is because, with the belief capped at $\rho$ for $n_2 \geq n_2^*$ (which is the same belief if she stays in the equilibrium), she is worse off by purchasing any fake accounts. Similarly, the $L$-type would also be worse off by purchasing more than $x_L^{pool}$. Obviously, the $H$-type cannot purchase fewer than zero fake account. If the $L$-type purchases fewer than $x_L^{pool}$ (say $x_L' < x_L^{pool}$), she will be seen as an $L$-type and will attract $n_{un}^L = (1-l)[1 - \frac{c}{q_L}]$ uninformed followers. In such a case, the advertiser correctly considers the influencer as a $L$-type and expects her not to purchase fake accounts, the $L$-type influencer knows the advertiser's consideration, thus, her best strategy is not to purchase any fake account, thus, her expected payoff is

$$\pi_L' = \lambda_i \mu \left( n_{in}^L + n_{un}^L \right)$$

The IC condition requires

$$\pi_L' \leq \pi_L^{pool} = \lambda_i \left[ \mu \left( \overline{n}_{in} + \overline{n}_{un} \right) - (1-\rho) \left( \varphi + c_f \right) x_L^{pool} \right] + c_f \left[ (1-\rho) x_L^{pool} - x_L^{pool} \right]$$

which translates to:

$$\lambda_i \mu \left( \overline{n}_{in} + \overline{n}_{un} - n_{in}^L - n_{un}^L \right) \geq \left[ \lambda_i (1-\rho) \left( \varphi + c_f \right) + c_f \rho \right] \left( n_{in}^H - n_{in}^L \right) \tag{23}$$

When the $L$-type's IC condition holds, her IR condition is automatically satisfied, and the total surplus from the advertising must be non-negative, thus, the $H$-type influencer's and the advertiser's IR condition can be naturally satisfied as well.

By (9),

$$x_L^{pool} = n_{in}^H - n_{in}^L = l \left( 1 - \frac{c}{q_H} \right) - l \left( 1 - \frac{c}{q_L} \right) = l \left( \frac{c}{q_L} - \frac{c}{q_H} \right).$$

Furthermore,

$$\begin{cases} \overline{n}_{un} = (1-l) \left( 1 - \frac{c}{\overline{q}} \right) \\ \overline{n}_{in} = \rho n_{in}^H + (1-\rho) n_{in}^L \end{cases} \tag{24}$$

Substituting (9) in the main paper and the above into (20) and (23), we obtain Lemma **1**. □

## A.2 Proof of Proposition 2.

*Proof.* The conclusions follow from the signs of the first order derivatives: $\frac{\partial x_L^{pool}}{\partial l} = \frac{q_H - q_L}{q_H q_L} c > 0$; $\frac{\partial x_L^{pool}}{\partial \kappa} = 0$; $\frac{\partial x_L^{pool}}{\partial r_q} = -\frac{l}{q_L} c < 0$ ; $\frac{\partial x_L^{pool}}{\partial \tau} = -\frac{q_H - q_L}{q_H q_L} dl c_1 < 0$; $\frac{\partial x_L^{pool}}{\partial \lambda_i} = 0$; $\frac{\partial x_L^{pool}}{\partial d} = l \frac{q_H - q_L}{q_H q_L} c_1 (1-\tau) > 0$. □

## A.3 Proof of Lemma 2.

*Proof.* For the separating equilibrium, we have $n_2^H \neq n_2^L$. In this case, we show that $n_2^{*L} = n_{in}^L, n_2^{*H} = n_2^{sep} > n_{in}^H$ (i.e., the $L$-type does not buy fake accounts and the $H$-type buys enough to keep a leading status) with the following belief is a PBE.

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

The corresponding strategy profile is

$$\left(x_H^{csep}, x_L^{csep}\right) = \left(n_2^{sep} - n_{in}^H, 0\right) \tag{25}$$

and the equilibrium profits are, respectively

$$\pi_L^{csep} = \lambda_i \mu \left(n_{in}^L + n_{un}^L\right) \tag{26}$$

$$\pi_H^{csep} = \lambda_i \left[\mu \left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^{csep}\right] \tag{27}$$

Similar to the arguments made in the Proof of Lemma 1, the $H$-type will not purchase more than $x_H^{csep}$ to achieve a higher follower count than $n_2^{sep}$. Similarly, the $L$-type will purchase fake followers to achieve a follower count $n_2^L \in (0, n_2^{sep}) \cup (n_2^{sep}, \infty)$.

If the $H$-type purchases fewer than $x_H^{csep}$ (say $x'_H < n_2^{sep} - n_{in}^H$), she will be viewed as an $L$-type. Her best deviation of this type is not to purchase any fake account and the resulting expected payoff is $\pi'_H = \lambda_i \mu \left(n_{in}^L + n_{un}^L\right)$. The IC condition for the $H$-type requires $\pi'_H \leq \pi_H^{csep}$, which translates to:

$$\lambda_i \mu \left(n_{in}^L + n_{un}^L\right) \leq \lambda_i \left[\mu \left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) \left(n_2^{sep} - n_{in}^H\right)\right] \tag{28}$$

When the $H$-type's IC condition holds, her IR condition and the advertiser's IR condition are automatically satisfied.

We now consider whether the $L$-type has incentives to mimic the $H$-type. In doing so, she must purchase $x'_L = n_2^{sep} - n_{in}^L$. In such a case, the advertiser mistakenly considers the $L$-type influencer as a $H$-type and expects her to purchase $x_H^{csep}$ fake accounts, and her profit is

$$\pi'_L = \lambda_i \left[\mu \left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^{csep}\right] + c_f x_H^{csep} - c_f \left(n_2^{sep} - n_{in}^L\right) \tag{29}$$

The IC condition requires $\pi'_L \leq \pi_L^{csep}$, which translates to:

$$\lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) \leq \lambda_i \left(\varphi + c_f\right) \left(n_2^{sep} - n_{in}^H\right) + c_f \left(n_{in}^H - n_{in}^L\right) \tag{30}$$

The IR condition for the $L$-type can be naturally satisfied since $\pi_L^{csep} > 0$.

For the case of $L$-type under this equilibrium, the IR condition for the advertiser is automatic since $L$-type doesn't buy fake accounts.

Combing the IC conditions for $H$-type and $L$-type, and IR condition for the advertiser, we have

$$\underline{n}_2^{sep} \equiv \frac{\lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^H - n_{in}^L\right)}{\lambda_i \left(\varphi + c_f\right)} + n_{in}^H \quad \leq \quad n_2^{sep} \leq \frac{\mu}{\varphi + c_f} \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) + n_{in}^H \equiv \bar{n}_2^{sep}$$

where $\underline{n}_2^{sep}$ and $\bar{n}_2^{sep}$ denotes the lower and upper bounds of the $H$-type's early followers respectively. Specifically, $\underline{n}_2^{sep}$ is the highest early-follower count that the $L$-type is willing to mimic and $\bar{n}_2^{sep}$ is the highest early-follower count that the $H$-type is willing to maintain (beyond which she would prefer not buying and being treated as the $L$-type).

Given the continuum of separating equilibria, we can apply the LMSE refinement. We note from (26) and (27) that the $H$-type is strictly worse off under a higher $n_2^{sep}$. In other words, the equilibria associated with $\underline{n}_2^{sep}$ lexicographically dominate all other separating equilibria. Therefore, the unique separating LMSE is defined by $n_2^{*sep} = \underline{n}_2^{sep}$. The condition for this equilibrium is $\underline{n}_2^{sep} \geq n_{in}^H$

$$\begin{aligned}
& \frac{\lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^H - n_{in}^L\right)}{\lambda_i \left(\varphi + c_f\right)} + n_{in}^H \geq n_{in}^H \\
\iff \quad & \lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^H - n_{in}^L\right) \geq 0 \\
\iff \quad & c_f \leq \frac{\lambda_i \mu (n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L)}{n_{in}^H - n_{in}^L} \equiv \eta_2 \\
\iff \quad & d \leq (1 - \tau) \left(\frac{\lambda_i \mu}{l} - \kappa\right)
\end{aligned}$$

where the last step is due to $\frac{n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L}{n_{in}^H - n_{in}^L} = \frac{1}{l}$ (which can be verified by substituting (9) in

the main paper) and $c_f = \kappa + \frac{d}{1-\tau}$. Thus, we have

$$\begin{aligned} x_H^{csep} &= n_2^{*sep} - n_{in}^H \\ &= \underline{n}_2^{sep} - n_{in}^H \\ &= \frac{\lambda_i\mu\left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^H - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} \\ &= \frac{\left(\lambda_i\mu - c_f l\right)\left(\frac{c}{q_L} - \frac{c}{q_H}\right)}{\lambda_i\left(\varphi + c_f\right)} = \frac{\lambda_i\mu - c_f l}{\lambda_i\left(\varphi + c_f\right)}\frac{q_H - q_L}{q_H q_L}c \end{aligned}$$

□

## A.4 Proof of Lemma 3.

*Proof.* Under the naturally separating equilibrium, the number of early followers for the $H$-type and $L$-type influencers should be $n_{in}^H$ and $n_{in}^L$, respectively. The equilibrium strategy profile is simply $(x_H^*, x_L^*) = (0, 0)$. We first show that this strategy profile with the following belief can be a PBE:

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

with corresponding payoffs

$$\pi_L^{nsep} = \lambda_1\mu\left(n_{in}^L + n_{un}^L\right) \tag{31}$$

$$\pi_H^{nsep} = \lambda_1\mu\left(n_{in}^H + n_{un}^H\right) \tag{32}$$

Similar to the proof of Lemma 2, we can first establish that neither $H$-type nor $L$-type has an incentive to achieve a follower count higher than $n_{in}^H$, and the $L$-type has no incentive to achieve a follower count $n_2^L \in \left(n_{in}^L, n_{in}^H\right)$. Similar to the proof of Lemma 2, the $L$-type has an incentive to mimic the $H$-type (by purchasing $x_L' = n_{in}^H - n_{in}^L$) if and only if $\pi_L' = \lambda_i\mu\left(n_{in}^H + n_{un}^H\right) - c_f\left(n_{in}^H - n_{in}^L\right) < \pi_L^{nsep} = \pi_L^{csep}$ (or equivalently $\underline{n}_2^{sep} < n_{in}^H$, where $\underline{n}_2^{sep}$ is defined in the proof of Lemma 2 as the highest early-follower count that the $L$-type is willing to mimic), this condition

translates to:

$$\begin{aligned} & \lambda_i\mu\left(n_{in}^H+n_{un}^H\right)-c_f\left(n_{in}^H-n_{in}^L\right)<\lambda_i\mu\left(n_{in}^L+n_{un}^L\right) \\ \iff \quad & c_f>\frac{\lambda_i\mu(n_{in}^H+n_{un}^H-n_{in}^L-n_{un}^L)}{n_{in}^H-n_{in}^L}\equiv\eta_2 \\ \iff \quad & d>(1-\tau)\left(\frac{\lambda_i\mu}{l}-\kappa\right). \end{aligned}$$

We finally note the boundary between the costly and naturally separating equilibria is defined by whether $\underline{n}_2^{sep}$ (the highest early-follower count the $L$-type can mimic) exceeds $n_{in}^H$. If it does, a costly separating equilibrium exists where the $H$-type purchases fake accounts to maintain a separating equilibrium. Conversely, a naturally separating equilibrium exists. □

## A.5 Proof of Proposition 3.

*Proof.* We firstly note that $\frac{\partial x_H^{csep}}{\partial l}=-\frac{c_f}{\varphi+c_f}\frac{q_H-q_L}{\lambda_i q_H q_L}c<0$, $\frac{\partial x_H^{csep}}{\partial \kappa}=-\frac{l\varphi+\lambda_i\mu}{\left(c_f+\varphi\right)^2}\frac{q_H-q_L}{\lambda_i q_H q_L}c<0$, $\frac{\partial x_H^{csep}}{\partial r_q}=-\frac{\lambda_i\mu-c_f l}{\lambda_i\left(\varphi+c_f\right)}\frac{1}{q_L}c\leq 0$ (noting that $\frac{\lambda_i\mu}{c_f}-l$ under costly separating equilibrium), and $\frac{\partial x_H^{csep}}{\partial \lambda_i}=\frac{c_f l}{\lambda_i^2\left(\varphi+c_f\right)}\frac{q_H-q_L}{q_H q_L}c>0$.

We also have $\frac{\partial x_H^{csep}}{\partial \tau}=\frac{q_H-q_L}{\lambda_i q_H q_L}\left(-\frac{d\left(\lambda_i\mu-c_f l\right)c_1}{\left(c_f+\varphi\right)}-\frac{d\left(\lambda_i\mu-c_f l\right)c}{\left(c_f+\varphi\right)^2(1-\tau)^2}-\frac{dlc}{\left(c_f+\varphi\right)(1-\tau)^2}\right)<0$, where the last step is because $\frac{\lambda_i\mu}{c_f}-l\geq 0$.

$$\begin{aligned} \frac{\partial x_H^{csep}}{\partial d} &= \frac{q_H-q_L}{\lambda_i q_H q_L}\frac{1}{\left(c_f+\varphi\right)^2(1-\tau)}\left\{\left[\left(\lambda_i\mu-c_f l\right)c_1(1-\tau)^2-lc\right]\left(c_f+\varphi\right)-\left(\lambda_i\mu-c_f l\right)c\right\} \\ &= \frac{q_H-q_L}{\lambda_i q_H q_L}\frac{1}{\left(c_f+\varphi\right)^2(1-\tau)}\left\{\left[\left(\lambda_i\mu-\kappa l\right)c_1(1-\tau)^2-lc_0-2lc_1(1-\tau)d\right]\left(c_f+\varphi\right)-\left(\lambda_i\mu-c_f l\right)c\right\} \end{aligned}$$

If $\left(\lambda_i\mu-\kappa l\right)c_1(1-\tau)^2-lc_0<\frac{(\lambda_i\mu-\kappa l)c_0}{\kappa+\varphi}$, i.e., $c_1(1-\tau)^2<\frac{\lambda_i(\mu+\varphi l)c_0}{(\kappa+\varphi)(\lambda_i\mu-\kappa l)}$ we have $\frac{\partial x_H^{csep}}{\partial d}<0$ for all $d\geq 0$, i.e., $x_H^{csep}$ is monotonically decreasing with $d$.

If $c_1(1-\tau)^2>\frac{\lambda_i(\mu+\varphi l)c_0}{(\kappa+\varphi)(\lambda_i\mu-\kappa l)}$, when $d\to 0$, we have $\frac{\partial x_H^{csep}}{\partial d}>0$,

but when $d\to(1-\tau)\left(\frac{\lambda_i\mu}{l}-\kappa\right)$, we have $\frac{\partial x_H^{csep}}{\partial d}=-\frac{q_H-q_L}{\lambda_i q_H q_L}\frac{1}{\left(c_f+\varphi\right)^2(1-\tau)}\frac{l\left[c_0+c_1\left(\frac{\lambda_i\mu}{l}-\kappa\right)(1-\tau)^2\right]}{\left(\frac{\lambda_i\mu}{l}+\lambda_i\varphi\right)(1-\tau)}<0$.

Thus, $x_H^{csep}$ is not a monotonic function of $d$. □

### A.6 Proof of Lemma (4).

*Proof.* By Lemma (1) and Lemma (3), the pooling and naturally separating equilibria can coexist when $\eta_2 < \eta_1$ and $(1-\tau)(\eta_2-\kappa) \leq d \leq (1-\tau)(\eta_1-\kappa)$. Comparing the $H$-type's profits under the two equilibria, we have

$$\pi_{i,H}^{nsep} - \pi_{i,H}^{pool} = \lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - \left\{\lambda_i \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - (1-\rho)(\varphi + c_f) x_L^{pool}\right] + c_f (1-\rho) x_L^{pool}\right\}$$

From $L$-type's IC condition 12, we have

$$\lambda_i \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - (1-\rho)(\varphi + c_f) x_L^{pool}\right] + c_f (1-\rho) x_L^{pool} > \lambda_i \mu \left(n_{in}^L + n_{un}^L\right) + c_f x_L^{pool}$$

Therefore,

$$\begin{aligned}
\pi_{i,H}^{nsep} - \pi_{i,H}^{pool} &= \lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - \left\{\lambda_i \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - (1-\rho)(\varphi + c_f) x_L^{pool}\right] + c_f (1-\rho) x_L^{pool}\right\} \\
&< \lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - \left[\lambda_i \mu \left(n_{in}^L + n_{un}^L\right) + c_f x_L^{pool}\right] \\
&= \lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^H - n_{in}^L\right) \\
&= (\lambda_i \mu - c_f l) \left(\frac{c}{q_L} - \frac{c}{q_H}\right)
\end{aligned}$$

We can translate the condition in this case $(1-\tau)(\eta_2-\kappa) \leq d$ to $c_f > \eta_2 = \frac{\lambda_i \mu}{l}$ , i.e., $\lambda_i \mu - c_f l \leq 0$, thus,

$$\pi_{i,H}^{nsep} - \pi_{i,H}^{pool} = (\lambda_i \mu - c_f l) \left(\frac{c}{q_L} - \frac{c}{q_H}\right) < 0$$

Thus, the pooling equilibrium $l$-dominates the naturally separating equilibrium when $\eta_2 < \eta_1$ and $(1-\tau)(\eta_2-\kappa) \leq d \leq (1-\tau)(\eta_1-\kappa)$

By Lemma (1) and Lemma (2), pooling and costly separating equilibria coexist when 1) $\eta_2 < \eta_1$

and $d \leq (1-\tau)(\eta_2 - \kappa)$, or 2) $\eta_2 > \eta_1$ and $d \leq (1-\tau)(\eta_1 - \kappa)$ . Note that

$$\begin{aligned}
\pi_{i,H}^{csep} - \pi_{i,H}^{pool} &= \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f) x_H^{csep}\right] - \\
&\quad \left\{\lambda_i \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - (1-\rho)(\varphi + c_f) x_L^{pool}\right] + c_f (1-\rho) x_L^{pool}\right\} \\
&< \lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - \left[\lambda_i \mu \left(n_{in}^L + n_{un}^L\right) + c_f x_L^{pool}\right] - \lambda_i (\varphi + c_f) x_H^{csep} \\
&= \lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^H - n_{in}^L\right) - \lambda_i (\varphi + c_f) \frac{(\lambda_i \mu - c_f l)\left(\frac{c}{q_L} - \frac{c}{q_H}\right)}{\lambda_i (\varphi + c_f)} \\
&= (\lambda_i \mu - c_f l)\left(\frac{c}{q_L} - \frac{c}{q_H}\right) - (\lambda_i \mu - c_f l)\left(\frac{c}{q_L} - \frac{c}{q_H}\right) = 0
\end{aligned}$$

Therefore, the pooling equilibrium l-dominates the costly separating equilibrium when 1) $\eta_2 < \eta_1$ and $d \leq (1-\tau)(\eta_2 - \kappa)$, or 2) $\eta_2 > \eta_1$ and $d \leq (1-\tau)(\eta_1 - \kappa)$

Above all, as long as the pooling equilibrium exists, it dominates the other two equilibria: costly and naturally separating equilibrium. □

### A.7 Proof of Proposition 1.

*Proof.* If $\eta_1 \leq \kappa$, the condition for the pooling equilibrium is not met, so the pooling equilibrium cannot exist. The only equilibrium is either costly separating, if $d \leq (1-\tau)(\eta_2 - \kappa)$, or naturally separating, otherwise. Since there is only one equilibrium under any condition, it is also LMSE.

When $\eta_1 > k$, we discuss two scenarios. If $\eta_2 \leq \eta_1$, by Lemma (4), the pooling dominates the costly separating equilibrium when $d \leq (1-\tau)(\eta_2 - \kappa)$ and the naturally separating equilibrium when $(1-\tau)(\eta_2 - \kappa) \leq d \leq (1-\tau)(\eta_1 - \kappa)$, thus, the pooling equilibrium is the only remaining equilibrium whenever it exists. leading to case (b).

Turning to the case $\eta_2 > \eta_1$. Under such a case, when $d \leq (1-\tau)(\eta_1 - \kappa) < (1-\tau)(\eta_2 - \kappa)$, the costly separating equilibrium exists but is l-dominated by the pooling equilibrium. When $(1-\tau)(\eta_1 - \kappa) < d \leq (1-\tau)(\eta_2 - \kappa)$, the costly separating equilibrium is the sole equilibrium and thus LMSE. When $d > (1-\tau)(\eta_2 - \kappa)$, the naturally separating equilibrium is the only remaining equilibrium and thus LMSE. Case (c) summarizes these LMSE refinement outcomes. □

## A.8 Proof of Lemma 5.

*Proof.* Recall that $\pi_p = \lambda_p \pi = \lambda_p \{\mu E[n_r] - (\varphi + c_f) E[x]\}$, the platform profit under the pooling equilibrium is:

$$\begin{aligned}
\pi_p^{pool} &= \lambda_p \left[\mu\left(\overline{n}_{in} + \overline{n}_{un}\right) - (1-\rho)(\varphi + c_f) x_L^{pool}\right] \\
&= \lambda_p \left\{\mu\left[l\rho\left(1 - \frac{c}{q_H}\right) + l(1-\rho)\left(1 - \frac{c}{q_L}\right) + (1-l)\left(1 - \frac{c}{\bar{q}}\right)\right]\right. \\
&\quad \left. - (1-\rho) l \left(\frac{1}{q_L} - \frac{1}{q_H}\right) c(\varphi + c_f)\right\} \\
&= \lambda_p \left\{\mu - \left(\left[\mu\delta + (1-\rho) l \frac{q_H - q_L}{q_H q_L}\varphi\right] c + (1-\rho) l \frac{q_H - q_L}{q_H q_L} c_f c\right)\right\}
\end{aligned}$$

where $\delta \equiv \frac{l\rho}{q_H} + \frac{l(1-\rho)}{q_L} + \frac{1-l}{\bar{q}}$.

We can rewrite $\pi_p^{pool}$ as

$$\pi_p^{pool}(d) = \omega_1 d^2 + \omega_2 d + \omega_3 \tag{33}$$

where

$$\begin{cases}
\omega_1 = & -\lambda_p (1-\rho) l \frac{q_H - q_L}{q_H q_L} c_1 \\
\omega_2 = & -\lambda_p \left\{c_1 (1-\tau)\left[\mu\delta + (1-\rho) l \frac{q_H - q_L}{q_H q_L}(\varphi + \kappa)\right] + (1-\rho) l \frac{q_H - q_L}{q_H q_L} \frac{c_0}{1-\tau}\right\} \\
\omega_3 = & \lambda_p \mu - \lambda_p \left[\mu\delta + (1-\rho) l \frac{q_H - q_L}{q_H q_L}(\varphi + \kappa)\right] c_0
\end{cases}$$

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 $d$.

Under the costly separating equilibrium.

$$\begin{aligned}
\pi_p^{csep} &= \rho \pi_{p,H}^{csep} + (1-\rho) \pi_{p,L}^{csep} \\
&= \rho\lambda_p \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f) x_H^{csep}\right] + (1-\rho)\left[\lambda_p \mu\left(n_{in}^L + n_{un}^L\right)\right] \\
&= \lambda_p \mu \left[1 - \left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right) c\right] - \frac{\lambda_p}{\lambda_i} \rho\left(\lambda_i \mu - c_f l\right) \frac{q_H - q_L}{q_H q_L} c
\end{aligned}$$

We can rewrite $\pi_p^{csep}$ as

$$\pi_p^{csep}(d) = \alpha_1 d^2 + \alpha_2 d + \alpha_3 \tag{34}$$

where

$$\begin{cases} \alpha_1 = \frac{\lambda_p}{\lambda_i}\rho l \frac{q_H - q_L}{q_H q_L} c_1 \\ \alpha_2 = \frac{\lambda_p}{\lambda_i}\rho \frac{q_H - q_L}{q_H q_L}\left[\frac{c_0 l}{1-\tau} - (\lambda_i \mu + \kappa l) c_1\right] - c_1 \lambda_p \mu \left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right) \\ \alpha_3 = \lambda_p \mu \left[1 - c_0 \left(\frac{\rho}{q_H} + \frac{1-\rho}{q_L}\right)\right] - \frac{\lambda_p}{\lambda_i}\rho \frac{q_H - q_L}{q_H q_L} c_0 (\lambda_i \mu + \kappa l) \end{cases}$$

Obviously, $\alpha_1 > 0$, therefore $\frac{\partial^2 \pi_p^{csep}(d)}{\partial d^2} > 0$ and $\pi_p^{csep}(d)$ is a convex function of $d$.

Under the naturally separating equilibrium

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

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

## A.9 Proof of Lemma 6.

*Proof.* According to Lemma 5, we have

$$
\begin{aligned}
\pi_{p,H}^{pool}(0)-\pi_{p,H}^{csep}(d_1) &> \pi_{p,H}^{pool}(d_1)-\pi_{p,H}^{csep}(d_1)\\
&= \lambda_p\left[\mu\left(\overline{n}_{in}+\overline{n}_{un}\right)-(1-\rho)\left(\varphi+c_f\right)x_L^{pool}\right]\\
&\quad -\lambda_p\left[\mu\left(n_{in}^H+n_{un}^H\right)-\left(\varphi+c_f\right)x_H^{csep}\right]\\
&> \frac{\lambda_p}{\lambda_i}\left\{\lambda_i\mu\left(n_{in}^L+n_{un}^L\right)-c_f\left[(1-\rho)\,x_L^{pool}-x_L^{pool}\right]\right\}\\
&\quad -\lambda_p\left[\mu\left(n_{in}^H+n_{un}^H\right)-\left(\varphi+c_f\right)x_H^{csep}\right]\\
&= \lambda_p\mu\left(n_{in}^L+n_{un}^L\right)-\lambda_p\mu\left(n_{in}^H+n_{un}^H\right)+\frac{\lambda_p}{\lambda_i}c_f\rho x_L^{pool}+\lambda_p\left(\varphi+c_f\right)x_H^{csep}\\
&= \lambda_p\mu\left(n_{in}^L+n_{un}^L-n_{in}^H-n_{un}^H\right)+\frac{\lambda_p}{\lambda_i}c_f\rho l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)+\frac{\lambda_p}{\lambda_i}\left(\lambda_i\mu-c_f l\right)\left(\frac{c}{q_L}-\frac{c}{q_H}\right)\\
&= -\lambda_p\mu\left(\frac{c}{q_L}-\frac{c}{q_H}\right)+\frac{\lambda_p}{\lambda_i}c_f\rho l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)-\frac{\lambda_p}{\lambda_i}c_f l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)+\lambda_p\mu\left(\frac{c}{q_L}-\frac{c}{q_H}\right)\\
&= (\rho-1)\frac{\lambda_p}{\lambda_i}c_f l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)
\end{aligned}
$$

$$
\begin{aligned}
\pi_{p,L}^{pool}(0)-\pi_{p,L}^{csep}(d_1) &> \pi_{p,L}^{pool}(d_1)-\pi_{p,L}^{csep}(d_1)\\
&= \lambda_p\left[\mu\left(\overline{n}_{in}+\overline{n}_{un}\right)-(1-\rho)\left(\varphi+c_f\right)x_L^{pool}\right]-\lambda_p\mu\left(n_{in}^L+n_{un}^L\right)\\
&> \frac{\lambda_p}{\lambda_i}\left\{\lambda_i\mu\left(n_{in}^L+n_{un}^L\right)-c_f\left[(1-\rho)\,x_L^{pool}-x_L^{pool}\right]\right\}-\lambda_p\mu\left(n_{in}^L+n_{un}^L\right)\\
&= \lambda_p\mu\left(n_{in}^L+n_{un}^L\right)-\lambda_p\mu\left(n_{in}^L+n_{un}^L\right)+\frac{\lambda_p}{\lambda_i}c_f\rho x_L^{pool}\\
&= \rho\frac{\lambda_p}{\lambda_i}c_f l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\rho\left[\pi_{p,H}^{pool}(0)-\pi_{p,H}^{csep}(d_1)\right]+(1-\rho)\left[\pi_{p,L}^{pool}(0)-\pi_{p,L}^{csep}(d_1)\right]>\\
&\rho(\rho-1)\frac{\lambda_p}{\lambda_i}c_f l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)+(1-\rho)\rho\frac{\lambda_p}{\lambda_i}c_f l\left(\frac{c}{q_L}-\frac{c}{q_H}\right)=0\\
&\qquad\qquad\qquad\qquad\qquad\qquad\iff\quad \pi_p^{pool}(0)-\pi_p^{csep}(d_1)>0
\end{aligned}
$$

Therefore, the platform's profit at $d=0$ with pooling equilibrium obtained dominates the

platform's profit at $d = d_1$ with costly separating equilibrium obtained. □

### A.10 Proof of Proposition 4.

*Proof.* a) If $d_1 \leq 0$, the pooling equilibrium doesn't exist. we only need to compare the platform's local optimal strategies between the costly and naturally separating equilibria to decide optimum $d^*$. Noting that because $\pi_p^{nsep}(d)$ decreases in $d$, the maximum naturally-separating equilibrium profit is $\pi_p^{nsep}(d_2)$, which is the same as the $\pi_p^{csep}(d_2)$. Moreover, because $\pi_p^{csep}(d)$ is convex, its local optimum is achieved at either 0 or $d_2$. So the platform's optimum payoff can be either $\pi_p^{csep}(0)$, achieved through a costly separating equilibrium with $d^* = 0$, or $\pi_p^{nsep}(d_2)$, achieved through a naturally separating equilibrium with $d^* = d_2$, whichever yields higher payoff.

b) If $d_1 > 0$ and $d_2 \leq d_1$, the costly separating equilibrium doesn't exist, we only need to compare the pooling equilibrium for $d \in [0, d_1]$ with a naturally separating equilibrium with for $d > d_1$. Noting that both $\pi_p^{pool}(d)$ and $\pi_p^{nsep}(d)$ decrease in $d$, the platform's optimum payoff can be either $\pi_p^{pool}(0)$, achieved through a pooling equilibrium with $d^* = 0$, or $\pi_p^{nsep}(d_1)$, achieved through a naturally separating equilibrium with $d^* = d_1$, whichever yields higher payoff.

c) If $d_1 > 0$ and $d_2 > d_1$, the platform can induce either of the three types of equilibria. Given that the payoffs under the pooling and naturally separating equilibria are maximized at $d = 0$ and $d = d_2$, respectively, and the payoff under the costly separating equilibrium reaches the maximum at either $d = d_1$or $d = d_2$ (due to Lemma 5), we need to compare the platform's profits at $d = 0$, $d = d_1$, and $d = d_2$ respectively.

According to Lemma 6, we only need to compare the platform's profits at $d = 0$ and pooling equilibrium obtained, and $d = d_2$ and naturally separating equilibrium obtained, respectively, which leads to the result in Table 2. □

### A.11 Proof of Lemma 7.

*Proof.* We define consumer welfare as $U = U_{in} + U_{un}$ which includes the welfare of the informed and uninformed consumers. Consumer welfare under the pooling equilibrium 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 $d$

Under the costly and naturally separating equilibrium.

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

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

Additional, we have

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

Thus, when the consumer's nuisance cost is the same, consumers have a higher welfare when the influencers are separated than that in the pooling. □

## A.12 Proof of Proposition 5.

*Proof.* a) If $d_1 \leq 0$, the pooling equilibrium doesn't exist. we only need to compare consumer welfare under the costly and naturally separating equilibria to decide consumer-optimal $d^C$. Noting that, by Lemma 7, both $U^{csep}(d)$ and $U^{nsep}(d)$ decrease in $d$, we have $U^{csep}(0) > U^{csep}(d_2) = U^{nsep}(d_2)$. Thus, the optimal consumer welfare is $U^{csep}(0)$, achieved through a costly separating equilibrium with $d^C = 0$.

b) If $d_1 > 0$ and $d_2 \leq d_1$, the costly separating equilibrium doesn't exist, we only need to compare the pooling equilibrium for $d \in [0, d_1]$ and a naturally separating equilibrium with for $d > d_1$. Similarly, the optimal consumer welfare can be either $U^{pool}(0)$, achieved through a pooling equilibrium with $d^C = 0$, or $U^{nsep}(d_1)$, achieved through a naturally separating equilibrium with $d^C = d_1$, whichever yields higher consumer welfare.

c) If $d_1 > 0$ and $d_2 > d_1$, all three types of equilibria can exist. Given Proposition 1 and the consumer welfare decreases in $d$ under each type of equilibrium, we infer that the consumer optimum is the largest of $U^{pool}(0)$, $U^{csep}(d_1)$, and $U^{nsep}(d_2)$. Because $d_1 < d_2$ and $U^{csep}(d) = U^{csep}(d)$, we have $U^{csep}(d_1) > U^{nsep}(d_2)$. So, the optimal 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 higher consumer welfare.

We now turn to compare $d^C$ and $d^*$.

a) If $d_1 \leq 0$, $d^C = 0$. By Proposition 4, the platform's optimal anti-fake effort is $d^* = 0$ or $d^* = d_2$. So, we have $d^* = d^C$, when $\pi_p^{csep}(0) \geq \pi_p^{nsep}(d_2)$ or $d^* > d^C$, otherwise.

b) If $d_1 > 0$ and $d_2 \leq d_1$, the consumer-optimal $d^C$ can be either 0 when $U^{nsep}(d_1) \leq U^{pool}(0)$, or $d_1$ otherwise. 1) When$U^{nsep}(d_1) \leq U^{pool}(0)$, the equilibrium obtained under the consumer-optimal $d^C$ is pooling equilibrium. By numeric simulation, we find examples for both cases $\pi_p^{nsep}(d_1) \leq \pi_p^{pool}(0)$ and $\pi_p^{nsep}(d_1) > \pi_p^{pool}(0)$. So we have $d^* = d^C = 0$ when $U^{nsep}(d_1) \leq U^{pool}(0)$ and $\pi_p^{nsep}(d_1) \leq \pi_p^{pool}(0)$, or $d^* = d_1 > d^C$ when $U^{nsep}(d_1) \leq U^{pool}(0)$ and $\pi_p^{nsep}(d_1) > \pi_p^{pool}(0)$. Thus, we have $d^* \geq d^C$, when $U^{nsep}(d_1) \leq U^{pool}(0)$ and pooling equilibrium is obtained. 2) When$U^{nsep}(d_1) > U^{pool}(0)$, the equilibrium obtained under the consumer-optimal $d^C$ is the naturally separating equilibrium. By numeric simulation, we still can find examples for both cases $\pi_p^{nsep}(d_1) \leq \pi_p^{pool}(0)$ and $\pi_p^{nsep}(d_1) > \pi_p^{pool}(0)$. So we have $d^* = 0 < d^C$ when $U^{nsep}(d_1) > U^{pool}(0)$

and $\pi_p^{nsep}(d_1) \leq \pi_p^{pool}(0)$, or $d^* = d^C = d_1$ when $U^{nsep}(d_1) > U^{pool}(0)$ and $\pi_p^{nsep}(d_1) > \pi_p^{pool}(0)$. Thus, we have $d^* \leq d^C$, when $U^{nsep}(d_1) > U^{pool}(0)$ and naturally separating equilibrium is obtained.

c) If $d_1 > 0$ and $d_2 > d_1$,the consumer-optimal $d^C$ can be either 0 when $U^{csep}(d_1) \leq U^{pool}(0)$, or $d_1$ otherwise. 1) When$U^{csep}(d_1) \leq U^{pool}(0)$, the equilibrium obtained under the consumer-optimal $d^C$ is pooling equilibrium. By numeric simulation, we find examples for both cases $\pi_p^{pool}(0) \geq \pi_p^{nsep}(d_2)$, and $\pi_p^{pool}(0) < \pi_p^{nsep}(d_2)$. So we have $d^* = d^C = 0$ when $U^{csep}(d_1) \leq U^{pool}(0)$ and $\pi_p^{pool}(0) \geq \pi_p^{nsep}(d_2)$, or $d^* > d^C$ when $U^{nsep}(d_1) \leq U^{pool}(0)$ and $\pi_p^{pool}(0) < \pi_p^{nsep}(d_2)$. Thus, we have $d^* \geq d^C$, when $U^{csep}(d_1) \leq U^{pool}(0)$ and pooling equilibrium is obtained. 2) When$U^{csep}(d_1) > U^{pool}(0)$, the equilibrium obtained under the consumer-optimal $d^C$ is the costly separating equilibrium. By numeric simulation, we still can find examples for both cases $\pi_p^{pool}(0) \geq \pi_p^{nsep}(d_2)$, and $\pi_p^{pool}(0) < \pi_p^{nsep}(d_2)$. So we have $d^* < d^C$ when $U^{csep}(d_1) > U^{pool}(0)$ and $\pi_p^{pool}(0) \geq \pi_p^{nsep}(d_2)$, or $d^* > d^C$ when $U^{nsep}(d_1) > U^{pool}(0)$ and $\pi_p^{pool}(0) < \pi_p^{nsep}(d_2)$. Thus, we have $d^* \lesseqgtr d^C$, when $U^{csep}(d_1) > U^{pool}(0)$ and the costly separating equilibrium is obtained. □

### A.13 Welfare Comparative Statics.

*Proof.* 1) Platform's Optimum is 0 and Pooling Equilibrium Obtained

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

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

$\frac{\partial U_{pool}^*}{\partial \kappa} = 0$

The platform exerts no anti-fake effort, at that time, the surplus and welfare are not affected by the anti-fake technology level $\tau$

2) Platform's Optimum is 0 and Costly Separating Equilibrium Obtained

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

$\frac{\partial U_{csep}^{*}}{\partial l}=0$, $\frac{\partial U_{csep}^{*}}{\partial \kappa}=0$

Again, the platform exerts no anti-fake effort, at that time, the surplus and welfare are not affected by the anti-fake technology level $\tau$

3) Platform's Optimum is $d_1$ Naturally Separating Equilibrium Obtained

$$U_{nsep}^{*}=\rho\frac{\left(q_H-C\right)^2}{2q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)^2}{2q_L}$$

where $C=c_0+c_1\left(1-\tau\right)^2\left(\eta_1-\kappa\right)$

$$\begin{aligned}\frac{\partial U_{nsep}^{*}}{\partial l}&=\frac{\partial \pi_{nsep}^{*}}{\partial C}\frac{\partial C}{\partial l}=-\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)^2\frac{\partial \eta_1}{\partial l}\\&=\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)^2\frac{\lambda_i\mu}{\lambda_i+\left(1-\lambda_i\right)\rho}\frac{\left(\frac{1}{q_L}-\frac{1}{E[q]}\right)}{\left(\frac{1}{q_L}-\frac{1}{q_H}\right)}\frac{1}{l^2}>0\end{aligned}$$

$\frac{\partial U_{nsep}^{*}}{\partial \kappa}=\frac{\partial \pi_{nsep}^{*}}{\partial C}\frac{\partial C}{\partial \kappa}=\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)^2>0$
$\frac{\partial U_{nsep}^{*}}{\partial \tau}=\frac{\partial \pi_{nsep}^{*}}{\partial C}\frac{\partial C}{\partial \tau}=2\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)\left(\eta_1-\kappa\right)>0$

4) Platform's Optimum is $d_2$ and Naturally Separating Equilibrium Obtained

$$U_{nsep}^{*}=\rho\frac{\left(q_H-C\right)^2}{2q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)^2}{2q_L}$$

Where $C=c_0+c_1\left(1-\tau\right)^2\left(\eta_2-\kappa\right)$

$$\begin{aligned}\frac{\partial U_{nsep}^{*}}{\partial l}&=\frac{\partial \pi_{nsep}^{*}}{\partial C}\frac{\partial C}{\partial l}=-\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)^2\frac{\partial \eta_2}{\partial l}\\&=\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)^2\frac{\lambda_i\mu}{l^2}>0\end{aligned}$$

$\frac{\partial U_{nsep}^{*}}{\partial \kappa}=\frac{\partial \pi_{nsep}^{*}}{\partial C}\frac{\partial C}{\partial \kappa}=\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)^2>0$
$\frac{\partial U_{nsep}^{*}}{\partial \tau}=\frac{\partial \pi_{nsep}^{*}}{\partial C}\frac{\partial C}{\partial \tau}=2\left[\rho\frac{\left(q_H-C\right)}{q_H}+\left(1-\rho\right)\frac{\left(q_L-C\right)}{q_L}\right]c_1\left(1-\tau\right)\left(\eta_1-\kappa\right)>0$ □

# B Proofs of Equilibria with Three Types of Influencers.

## B.1 Fully Separating: costly separating.

*Proof.* For the fully separating equilibrium, we have $n_2^H \neq n_2^M$, $n_2^H \neq n_2^L$, and $n_2^M \neq n_2^L$. In this case, we show that $n_2^{*L} = n_{in}^L$ ,$n_2^{*M} = n_2^{sepM} > n_{in}^M$, and $n_2^{*H} = n_2^{sepH} > n_{in}^H$. (i.e., the $L$-type does not buy fake accounts, the $M$- and $H$-type buy enough to keep a leading status) with the following belief is a PBE.

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

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

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

The corresponding strategy profile is

$$\left(x_H^*, x_M^*, x_L^*\right) = \left(n_2^{sepH} - n_{in}^H, n_2^{sepM} - n_{in}^M, 0\right) \tag{35}$$

Similar to the argument made in the Proof of Lemma 1, we argue that the $H$-type will not buy more than $x_H^*$ to achieve a higher follower count than $n_2^{sepH}$. Similarly, the $M$-type will not buy more than $x_M^*$ to achieve a follower count that satisfies $n_2^{sepM} < n_2^M < n_2^{sepH}$, the $L$-type will not buy more than $x_L^*$ to achieve a follower count that meets $n_{in}^L < n_2^L < n_2^{sepM}$.

1) For $H$-type Influencer

If the $H$-type purchases fewer than $x_H^*$ to make $n_2^{sepM} \leq n_2^H < n_2^{sepH}$ (say $n_2^{sepM} - n_{in}^H \leq x_H' < n_2^{sepH} - n_{in}^H$), she will be viewed as a $M$-type. Her best deviation of this type is to purchase $x_H' = \left(n_2^{sepM} - n_{in}^H\right)$ fake accounts and the resulting expected profit is

$$\pi'_H = \lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - (\varphi + c_f)\, x_M^*\right] + c_f x_M^* - c_f\left(n_2^{sepM} - n_{in}^H\right)$$

The IC condition requires $\pi'_H \leq \pi_H^{*sep} = \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f)\, x_H^*\right]$, which translates to:

$$\lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - (\varphi + c_f)\left(n_2^{sepM} - n_{in}^M\right)\right] + c_f\left(n_{in}^H - n_{in}^M\right) \leq \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f)\left(n_2^{sepH} - n_{in}^H\right)\right] \tag{36}$$

By simplification, we have

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

If the $H$-type purchases fewer than $x_H^*$ to make $n_2^H < n_2^{sepM}$ (say $x''_H < n_2^{sepM} - n_{in}^H$), she will be viewed as an $L$-type. Her best deviation of this type is not to purchase any fake account and the resulting expected profit is $\pi''_H = \lambda_i \mu\left(n_{in}^L + n_{un}^L\right)$. The IC condition requires $\pi''_H \leq \pi_H^{*sep} = \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f)\, x_H^*\right]$, which translates to:

$$\mu\left(n_{in}^L + n_{un}^L\right) \leq \mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f)\, x_H^* \tag{37}$$

This IR condition for the $H$-type can be naturally satisfied under condition 37

Thus, we have

$$\begin{cases} n_2^{sepH} \leq \frac{\mu}{\varphi + c_f}\left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) + n_{in}^H \\ n_2^{sepH} - n_2^{sepM} \leq \frac{\lambda_i\left[\mu\left(n_{in}^H + n_{un}^H - n_{in}^M - n_{un}^M\right) + \left(\varphi + c_f\right)\left(n_{in}^H - n_{in}^M\right)\right] - c_f\left(n_{in}^H - n_{in}^M\right)}{\lambda_i\left(\varphi + c_f\right)} \end{cases}$$

2) For $M$-type influencers

If the $M$-type purchases more than $x_M^*$ to make $n_2^M \geq n_2^{sepH}$(say $x'_M \geq n_2^{sepH} - n_{in}^M$), she will be viewed as a $H$-type. Her best deviation of this type is to purchase $x'_M = n_2^{sepH} - n_{in}^M$ fake accounts and the resulting expected profit is

$$\pi'_M = \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f)\, x_H^*\right] + c_f x_H^* - c_f\left(n_2^{sepH} - n_{in}^M\right)$$

The IC condition requires $\pi'_M \leq \pi_M^{*sep} = \lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - (\varphi + c_f)\left(n_2^{sepM} - n_{in}^M\right)\right]$, which translates to:

$$\lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - (\varphi + c_f)\, x_H^*\right] - c_f\left(n_{in}^H - n_{in}^M\right) \leq \lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - (\varphi + c_f)\left(n_2^{sepM} - n_{in}^M\right)\right] \tag{38}$$

By simplification, we have

$$c_f\left(n_{in}^H - n_{in}^M\right) \geq \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H - n_{in}^M - n_{un}^M\right) - (\varphi + c_f)\left[\left(n_2^{sepH} - n_2^{sepM}\right) - \left(n_{in}^H - n_{in}^M\right)\right]\right]$$

If the $M$-type purchases fewer than $x_M^*$ to make $n_2^M < n_2^{sepM}$ (say $x''_M < n_2^{sepM} - n_{in}^M$), she will be viewed as an $L$-type. Her best deviation of this type is not to purchase any fake account and the resulting expected profit is

$$\pi''_M = \lambda_i \mu\left(n_{in}^L + n_{un}^L\right)$$

The IC condition requires $\pi''_M \leq \pi_M^{*sep} = \lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - (\varphi + c_f)\left(n_2^{sepM} - n_{in}^M\right)\right]$, which translates to:

$$\lambda_i \mu\left(n_{in}^L + n_{un}^L\right) \leq \lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - (\varphi + c_f)\left(n_2^{sepM} - n_{in}^M\right)\right] \tag{39}$$

Again, the IR condition for the $M$-type can be naturally satisfied under condition 39

Thus, we have

$$\begin{cases} n_2^{sepM} \leq \frac{\mu}{\varphi + c_f}\left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) + n_{in}^M \\ n_2^{sepH} - n_2^{sepM} \geq \frac{\lambda_i\left[\mu\left(n_{in}^H + n_{un}^H - n_{in}^M - n_{un}^M\right) + \left(\varphi + c_f\right)\left(n_{in}^H - n_{in}^M\right)\right] - c_f\left(n_{in}^H - n_{in}^M\right)}{\lambda_i\left(\varphi + c_f\right)} \end{cases}$$

3) For $L$-type influencer

If the $L$-type purchases more than $x_L^*$ to make $n_2^{sepM} \leq n_2^L < n_2^{sepH}$(say $n_2^{sepM} - n_{in}^L \leq x'_L < n_2^{sepH} - n_{in}^L$), she will be viewed as a $M$-type. Her best deviation of this type is to purchase $x'_L = \left(n_2^{sepM} - n_{in}^L\right)$ fake accounts and the resulting expected payoff is

$$\pi'_L = \lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - \left(\varphi + c_f\right) x_M^*\right] + c_f x_M^* - c_f\left(n_2^{sepM} - n_{in}^L\right)$$

the IC condition requires $\pi'_L \leq \pi_L^{*sep} = \lambda_i \mu\left(n_{in}^L + n_{un}^L\right)$, which translates to:

$$\lambda_i \left[\mu\left(n_{in}^M + n_{un}^M\right) - \left(\varphi + c_f\right) x_M^*\right] - c_f\left(n_{in}^M - n_{in}^L\right) \leq \lambda_i \mu\left(n_{in}^L + n_{un}^L\right) \tag{40}$$

If the $L$-type purchases more than $x_L^*$ to make $n_2^L \geq n_2^{sepH}$(say $x''_L \geq n_2^{sepH} - n_{in}^L$), she will be viewed as a $H$-type. Her best deviation of this type is to purchase $x''_L = \left(n_2^{sepH} - n_{in}^L\right)$ fake accounts and the resulting expected payoff is

$$\pi''_L = \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^*\right] + c_f x_H^* - c_f\left(n_2^{sepH} - n_{in}^L\right)$$

the IC condition requires $\pi''_L \leq \pi_L^{*sep} = \lambda_i \mu\left(n_{in}^L + n_{un}^L\right)$, which translates to:

$$\lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^*\right] - c_f\left(n_{in}^H - n_{in}^L\right) \leq \lambda_i \mu\left(n_{in}^L + n_{un}^L\right) \tag{41}$$

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

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

$$\begin{cases} n_2^{sepM} \geq \frac{\lambda_i \mu\left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^M - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} + n_{in}^M \\ n_2^{sepH} \geq \frac{\lambda_i \mu\left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^H - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} + n_{in}^H \end{cases}$$

4) Combining all conditions above

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

$$\begin{cases} n_2^{sepH} \in \left[\frac{\lambda_i \mu\left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^H - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} + n_{in}^H, \frac{\mu}{\varphi + c_f}\left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) + n_{in}^H\right] \\ n_2^{sepM} \in \left[\frac{\lambda_i \mu\left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^M - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} + n_{in}^M, \frac{\mu}{\varphi + c_f}\left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) + n_{in}^M\right] \\ n_2^{sepH} - n_2^{sepM} = \frac{\lambda_i\left[\mu\left(n_{in}^H + n_{un}^H - n_{in}^M - n_{un}^M\right) + \left(\varphi + c_f\right)\left(n_{in}^H - n_{in}^M\right)\right] - c_f\left(n_{in}^H - n_{in}^M\right)}{\lambda_i\left(\varphi + c_f\right)} \end{cases}$$

As a range of separating equilibria exist, we still apply the *LMSE* to conduct equilibrium

refinements, and we have

$$\begin{aligned}\left\{n_2^{H*}, n_2^{M*}, n_2^{L*}\right\} &= \left\{n_2^{*sepH}, n_2^{*sepM}, n_2^{L*}\right\} \\ &= \left\{\begin{array}{l}\dfrac{\lambda_i\mu\left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^H - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} + n_{in}^H, \\ \dfrac{\lambda_i\mu\left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) - c_f\left(n_{in}^M - n_{in}^L\right)}{\lambda_i\left(\varphi + c_f\right)} + n_{in}^M, n_{in}^L\end{array}\right\}\end{aligned}$$

under the belief system we defined for this case.

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

$$\begin{cases}\lambda_i\mu(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L) \geq c_f\left(n_{in}^H - n_{in}^L\right) \\ \lambda_i\mu\left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) \geq c_f\left(n_{in}^M - n_{in}^L\right)\end{cases}$$

Under the uniform distribution,

$$\begin{cases}n_{in}^H = l\left(1 - \frac{c}{q_H}\right) \\ n_{in}^M = l\left(1 - \frac{c}{q_M}\right) \\ n_{in}^L = l\left(1 - \frac{c}{q_L}\right) \\ n_{un}^H = (1-l)\left(1 - \frac{c}{q_H}\right) \\ n_{un}^M = (1-l)\left(1 - \frac{c}{q_M}\right) \\ n_{un}^L = (1-l)\left(1 - \frac{c}{q_L}\right)\end{cases} \tag{42}$$

Finally, we can obtain

$$\begin{cases}x_H^* = \frac{\lambda_i\mu - c_f l}{\lambda_i(\varphi + c_f)}\left(\frac{c}{q_L} - \frac{c}{q_H}\right) \\ x_M^* = \frac{\lambda_i\mu - c_f l}{\lambda_i(\varphi + c_f)}\left(\frac{c}{q_L} - \frac{c}{q_M}\right) \\ x_L^* = 0\end{cases} \tag{43}$$

under the condition $\lambda_i\mu \geq c_f l$ □

## B.2 Fully Separating: naturally separating.

*Proof.* By the definition of fully separating equilibrium, the $H$-type, $M$-type and $L$-type influencers should differ in the number of early followers, namely, $n_2^H \neq n_2^M$, $n_2^H \neq n_2^L$, and $n_2^M \neq n_2^L$. In this case, we show that $n_2^{*L} = n_{in}^L$ ,$n_2^{*M} = n_{in}^M$, and $n_2^{*H} = n_{in}^H$. (i.e., none of the three types buys fake accounts) with the following belief is a PBE.

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

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

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

The corresponding strategy profile is

$$\left(x_H^*, x_M^*, x_L^*\right) = (0, 0, 0) \tag{44}$$

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 $x_M^*$ to make $n_{in}^M < n_2^M < n_{in}^H$, the $L$-type will not purchase more than $x_L^*$ to make $n_{in}^L < n_2^L < n_{in}^M$.

1) For $M$-type influencer

If the $M$-type purchases more than $x_M^*$ to make $n_2^M \geq n_{in}^H$ (say $x'_M \geq n_{in}^H - n_{in}^M$), she will be viewed as a $H$-type. Her best deviation of this type is to purchase $x'_M = \left(n_{in}^H - n_{in}^M\right)$ fake accounts and the resulting expected profit is

$$\pi'_M = \lambda_i \left[\mu\left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^*\right] + c_f x_H^* - c_f\left(n_{in}^H - n_{in}^M\right)$$

The IC condition requires $\pi'_M \leq \pi_M^{*sep} = \lambda_i \mu\left(n_{in}^M + n_{un}^M\right)$, which translates to:

$$\lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^M - n_{un}^M\right) - c_f \left(n_{in}^H - n_{in}^M\right) \leq 0 \tag{45}$$

2) For $L$-type influencer

If the $L$-type purchases more than $x_L^*$ to make $n_{in}^M \leq n_2^L < n_{in}^H$ (say $n_{in}^M - n_{in}^L \leq x_L' < n_{in}^H - n_{in}^L$), she will be viewed as a $H$-type. Her best deviation of this type is to purchase $x_L' = \left(n_{in}^M - n_{in}^L\right)$ fake accounts and the resulting expected payoff is

$$\pi_L' = \lambda_i \left[\mu \left(n_{in}^M + n_{un}^M\right) - \left(\varphi + c_f\right) x_M^*\right] + c_f x_M^* - c_f \left(n_{in}^M - n_{in}^L\right)$$

the IC condition requires $\pi_L' \leq \pi_L^{*sep} = \lambda_i \mu \left(n_{in}^L + n_{un}^L\right)$, which translates to:

$$\lambda_i \mu \left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^M - n_{in}^L\right) \leq 0 \tag{46}$$

If the $L$-type purchases more than $x_L^*$ to make $n_2^L \geq n_{in}^H$ (say $x_L'' \geq n_{in}^H - n_{in}^L$), she will be viewed as a $H$-type. Her best deviation of this type is to purchase $x_L'' = \left(n_{in}^H - n_{in}^L\right)$ fake accounts and the resulting expected payoff is

$$\pi_L'' = \lambda_i \left[\mu \left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^*\right] + c_f x_H^* - c_f \left(n_{in}^H - n_{in}^L\right)$$

the IC condition requires $\pi_L'' \leq \pi_L^{*sep} = \lambda_i \mu \left(n_{in}^L + n_{un}^L\right)$, which translates to:

$$\lambda_i \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) - c_f \left(n_{in}^H - n_{in}^L\right) \leq 0 \tag{47}$$

4) Combining all conditions above

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

$$\begin{cases} \lambda_1 \mu \left(n_{in}^H + n_{un}^H - n_{in}^M - n_{un}^M\right) < c_f \left(n_{in}^H - n_{in}^M\right) \\ \lambda_1 \mu \left(n_{in}^M + n_{un}^M - n_{in}^L - n_{un}^L\right) < c_f \left(n_{in}^M - n_{in}^L\right) \\ \lambda_1 \mu \left(n_{in}^H + n_{un}^H - n_{in}^L - n_{un}^L\right) < c_f \left(n_{in}^H - n_{in}^L\right) \end{cases}$$

The three conditions can be simplified as one condition: $\lambda_i \mu < c_f l$

For this equilibrium to hold, the individual rational condition for the influencers and the ad-

vertiser can be naturally satisfied.

Finally, we can obtain

$$\begin{cases} x_H^* = 0 \\ x_M^* = 0 \\ x_L^* = 0 \end{cases} \tag{48}$$

under the condition $\lambda_i \mu < c_f l$ □

### B.3 Fully Pooling.

*Proof.* For this fully pooling equilibrium case, we have $n_2^H = n_2^M = n_2^L$. In this case, we first show that $n_2^H = n_2^M = n_2^L = n_{in}^H$ (i.e., the $H$-type does not buy fake accounts and the $M$-, $L$-type buys enough to make up the difference) with the following belief is a PBE.

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

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

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

The corresponding strategy profile is

$$(x_H^*, x_M^*, x_L^*) = \left(0, n_{in}^H - n_{in}^M, n_{in}^H - n_{in}^L\right) \tag{49}$$

All three types will not buy more than $x_H^*$,$x_M^*$, $x_L^*$ respectively to achieve a higher follower count than $n_{in}^H$.

1) For $M$-type Influencer

If the $M$-type purchases fewer than $x_M^*$ to make $n_2^M < n_{in}^H$ (say $x'_M < n_{in}^H - n_{in}^M$), she will

be viewed as an $L$-type. Her best deviation of this type is not to buy any fake account and the resulting expected profit is

$$\pi'_M = \lambda_i \mu \left( n_{in}^L + n_{un}^L \right)$$

The IC condition requires

$$\begin{aligned} \pi'_M \leq \pi_M^{*pool} &= \lambda_i \left[ \mu \left( \bar{n}_{in}^{HML} + \bar{n}_{un}^{HML} \right) - \left[ \rho_M \left( n_{in}^H - n_{in}^M \right) + (1 - \rho_H - \rho_M) \left( n_{in}^H - n_{in}^L \right) \right] (\varphi + c_f) \right] \\ &\quad + c_f \left[ \rho_M \left( n_{in}^H - n_{in}^M \right) + (1 - \rho_H - \rho_M) \left( n_{in}^H - n_{in}^L \right) - \left( n_{in}^H - n_{in}^M \right) \right] \end{aligned}$$

which translates to:

$$\begin{aligned} &\left\{ \lambda_i (\varphi + c_f) \left[ \rho_M \left( n_{in}^H - n_{in}^M \right) + (1 - \rho_H - \rho_M) \left( n_{in}^H - n_{in}^L \right) \right] \right. \\ &\left. - c_f \left[ (1 - \rho_H - \rho_M) \left( n_{in}^H - n_{in}^L \right) - (1 - \rho_M) \left( n_{in}^H - n_{in}^M \right) \right] \right\} \leq \lambda_i \mu \left( \bar{n}_{in}^{HML} + \bar{n}_{un}^{HML} - n_{in}^L - n_{un}^L \right) \end{aligned} \tag{50}$$

This IR condition for the $M$-type can be naturally satisfied

2) For the $L$-type influencer

Similar to the proof for $M$-type above, we have

$$\begin{aligned} \lambda_i (\varphi + c_f) \left[ \rho_M \left( n_{in}^H - n_{in}^M \right) + (1 - \rho_H - \rho_M) \left( n_{in}^H - n_{in}^L \right) \right] &- c_f \left[ \rho_M \left( n_{in}^H - n_{in}^M \right) - (\rho_H + \rho_M) \left( n_{in}^H - n_{in}^L \right) \right] \\ &\leq \lambda_i \mu \left( \bar{n}_{in}^{HML} + \bar{n}_{un}^{HML} - n_{in}^L - n_{un}^L \right) \end{aligned}$$

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

Finally, we can obtain

$$\begin{cases} x_H^* = 0 \\ x_M^* = n_{in}^H - n_{in}^M = l\left(\frac{c}{q_M} - \frac{c}{q_H}\right) \\ x_L^* = n_{in}^H - n_{in}^L = l\left(\frac{c}{q_L} - \frac{c}{q_H}\right) \end{cases} \tag{51}$$

□

### B.4 *H*- type Separating, *M*- and *L*- types Pooling.

*Proof.* For the hybrid case, we have $n_2^H \neq n_2^M = n_2^L$. In this case, we show that $n_2^{*M} = n_2^{*L} = n_{in}^M, n_2^{*H} = n_{in}^H$ (i.e., the $M$-type does not buy fake accounts, the $H$-type buys enough to keep a leading status, and the $L$-type buys enough to make up the difference) with the following belief is a PBE.

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

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

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

Again, we argue that the $H$-type will not buy more than $x_H^*$ to achieve a higher follower count than $n_{in}^H$.

The corresponding strategy profile is

$$\left(x_H^*, x_M^*, x_L^*\right) = \left(0, 0, n_{in}^M - n_{in}^L\right) \tag{52}$$

1) For $M$-type influencer

If the$M$-type purchases more than $x_M^*$ to make $n_2^M \geq n_{in}^H$(say $x_M' \geq n_{in}^H - n_{in}^M$), she will be viewed as an $H$-type. Her best deviation of this type is to purchase $x_M' = \left(n_{in}^H - n_{in}^M\right)$ fake accounts

and the resulting expected payoff is

$$\pi'_M = \lambda_i \left[\mu \left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^*\right] + c_f x_H^* - c_f \left(n_{in}^H - n_{in}^M\right)$$

The IC condition requires $\pi'_M \leq \pi_M^{*pool} = \lambda_i \left[\mu \left(\bar{n}_{in}^{ML} + \bar{n}_{un}^{ML}\right) - \frac{\rho_L}{\rho_M+\rho_L} \left(\varphi + c_f\right) x_L^*\right] + c_f \frac{\rho_L}{\rho_M+\rho_L} x_L^*$, which translates to:

$$\lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - c_f \left(n_{in}^H - n_{in}^M\right) \leq \lambda_i \left[\mu \left(\bar{n}_{in}^{ML} + \bar{n}_{un}^{ML}\right) - \frac{\rho_L}{\rho_M + \rho_L} \left(\varphi + c_f\right) x_L^*\right] + c_f \frac{\rho_L}{\rho_M + \rho_L} x_L^* \tag{53}$$

The IR condition for the $M$-type requires that $\lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - c_f \left(n_{in}^H - n_{in}^M\right) > 0$,

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

$$\begin{cases} \frac{\lambda_i \mu}{c_f} \geq \frac{\left(n_{in}^H - n_{in}^M\right)}{\left(n_{in}^H + n_{un}^H\right)} \\ \lambda_i \mu \left(n_{in}^H + n_{un}^H - \bar{n}_{in}^{ML} - \bar{n}_{un}^{ML}\right) \leq c_f \left(n_{in}^H - n_{in}^M\right) + \left[c_f - \lambda_i \left(\varphi + c_f\right)\right] \frac{\rho_L}{\rho_M+\rho_L} \left(n_{in}^M - n_{in}^L\right) \end{cases}$$

2) For $L$-type influencer

If the $L$-type purchases more than $x_L^*$ to make $n_2^L \geq n_{in}^H$ (say $x'_L \geq n_{in}^H - n_{in}^L$), she will be viewed as an $H$-type. Her best deviation of this type is to purchase $x'_L = \left(n_{in}^H - n_{in}^L\right)$ fake accounts and the resulting expected profit is

$$\pi'_L = \lambda_i \left[\mu \left(n_{in}^H + n_{un}^H\right) - \left(\varphi + c_f\right) x_H^*\right] + c_f x_H^* - c_f \left(n_{in}^H - n_{in}^L\right)$$

The IC condition requires $\pi'_L \leq \pi_L^{*pool} = \lambda_i \left[\mu \left(\bar{n}_{in}^{ML} + \bar{n}_{un}^{ML}\right) - \frac{\rho_L}{\rho_M+\rho_L} \left(\varphi + c_f\right) x_L^*\right] + c_f \left[\frac{\rho_L}{\rho_M+\rho_L} x_L^* - x_L^*\right]$, which translates to:

$$\lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - c_f \left(n_{in}^H - n_{in}^L\right) \leq \lambda_i \left[\mu \left(\bar{n}_{in}^{ML} + \bar{n}_{un}^{ML}\right) - \frac{\rho_L}{\rho_M + \rho_L} \left(\varphi + c_f\right) x_L^*\right] - c_f \frac{\rho_M}{\rho_M + \rho_L} x_L^* \tag{54}$$

The IR condition for the $L$-type requires that $\lambda_i \mu \left(n_{in}^H + n_{un}^H\right) - c_f \left(n_{in}^H - n_{in}^L\right) \geq 0$,

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

$$\begin{cases} \frac{\lambda_i \mu}{c_f} \geq \frac{\left(n_{in}^H - n_{in}^L\right)}{\left(n_{in}^H + n_{un}^H\right)} \\ \lambda_i \mu \left(n_{in}^H + n_{un}^H - \bar{n}_{in}^{ML} - \bar{n}_{un}^{ML}\right) \leq c_f \left(n_{in}^H - n_{in}^M\right) + \left[c_f \frac{\rho_M}{\rho_M + \rho_L} - \lambda_i \frac{\rho_L}{\rho_M + \rho_L} \left(\varphi + c_f\right)\right] \left(n_{in}^M - n_{in}^L\right) \end{cases}$$

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

Finally, we can obtain

$$\begin{cases} x_H^* = 0 \\ x_M^* = 0 \\ x_L^* = l \left(\frac{c}{q_L} - \frac{c}{q_M}\right) \end{cases} \tag{55}$$

□

### B.5 *H*- and *M*- types pooling, *L*- type separating.

*Proof.* For the hybrid case, we have $n_2^H = n_2^M \neq n_2^L$. In this case, we show that $n_2^{*H} = n_2^{*M} = n_{in}^H$, and $n_2^{*L} = n_{in}^L$ (i.e., neither $H$- nor $L$-type buys fake accounts, and the $M$-type buys enough fake accounts to make up the difference with the $H$-type's) with the following belief is a PBE.

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

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

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

The corresponding strategy profile is

$$(x_H^*, x_M^*, x_L^*) = \left(0, n_{in}^H - n_{in}^M, 0\right) \tag{56}$$

First, we argue that the $H$-type will not buy more than $x_H^*$ to achieve a higher follower count than $n_{in}^H$. The $M$-type will not buy more than $x_M^*$ to achieve a higher follower count than $n_{in}^H$. Also, the $L$-type will not buy more than $x_L^*$ to make $n_{in}^L < n_2^L < n_{in}^H$.

1) For $M$-type influencer

If the $M$-type purchases fewer than $x_M^*$ to make $n_2^M < n_{in}^H$ (say $x_M' \leq n_{in}^H - n_{in}^M$), she will be viewed as an $L$-type. Her best deviation of this type is not to buy any fake account and the resulting expected profit is

$$\pi_M' = \lambda_i \mu \left(n_{in}^L + n_{un}^L\right)$$

The IC condition requires

$$\pi_M' \leq \pi_M^{*pool} = \lambda_i \left[\mu \left(\bar{n}_{in}^{HM} + \bar{n}_{un}^{HM}\right) - \frac{\rho_M}{\rho_H + \rho_M} \left(\varphi + c_f\right) x_M^*\right] + c_f \left[\frac{\rho_M}{\rho_H + \rho_M} x_M^* - x_M^*\right]$$

which translates to:

$$\left[\lambda_i \frac{\rho_M}{\rho_H + \rho_M} \left(\varphi + c_f\right) + c_f \frac{\rho_H}{\rho_H + \rho_M}\right] \left(n_{in}^H - n_{in}^M\right) \leq \lambda_i \mu \left(\bar{n}_{in}^{HM} + \bar{n}_{un}^{HM} - n_{in}^L - n_{un}^L\right) \tag{57}$$

This IR condition for the $M$-type can be naturally satisfied under condition 57

2) For $L$-type influencer

If the $L$-type purchases more than $x_L^*$ to achieve a higher follower count than $n_{in}^H$ (say $x_L' \geq n_{in}^H - n_{in}^L$), she will be pooled with $H$- and $M$-type. Her best deviation of this type is to buy $x_L' = \left(n_{in}^H - n_{in}^L\right)$ fake accounts and the resulting expected payoff is

$$\pi_L' = \lambda_i \left[\mu \left(\bar{n}_{in}^{HM} + \bar{n}_{un}^{HM}\right) - \frac{\rho_M}{\rho_H + \rho_M} \left(\varphi + c_f\right) x_M^*\right] + c_f \frac{\rho_M}{\rho_H + \rho_M} x_M^* - c_f \left(n_{in}^H - n_{in}^L\right)$$

The IC condition requires $\pi'_L \leq \pi_L^{*sep} = \lambda_i \mu \left(n_{in}^L + n_{un}^L\right)$, which translates to:

$$\lambda_i \mu \left(\bar{n}_{in}^{HM} + \bar{n}_{un}^{HM} - n_{in}^L - n_{un}^L\right) \leq c_f \left(n_{in}^H - n_{in}^L\right) + \left[\lambda_i \left(\varphi + c_f\right) - c_f\right] \frac{\rho_M}{\rho_H + \rho_M} \left(n_{in}^H - n_{in}^M\right) \quad (58)$$

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

4) Combining all conditions above

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

$$\left[\lambda_i \frac{\rho_M}{\rho_H + \rho_M} \left(\varphi + c_f\right) + c_f \frac{\rho_H}{\rho_H + \rho_M}\right] \left(n_{in}^H - n_{in}^M\right) \leq \lambda_i \mu \left(\bar{n}_{in}^{HM} + \bar{n}_{un}^{HM} - n_{in}^L - n_{un}^L\right) \leq$$
$$c_f \left(n_{in}^H - n_{in}^L\right) + \left[\lambda_i \left(\varphi + c_f\right) - c_f\right] \frac{\rho_M}{\rho_H + \rho_M} \left(n_{in}^H - n_{in}^M\right)$$

Finally, we can obtain

$$\begin{cases} x_H^* = 0 \\ x_M^* = \left(n_{in}^H - n_{in}^M\right) = l \left(\frac{c}{q_M} - \frac{c}{q_H}\right) \\ x_L^* = 0 \end{cases} \quad (59)$$

□