Economics of Social Media Fake Accounts

Authors' names blinded for peer review

Abstract

The reports of fake social media accounts have caused increasing concerns about the economic and social viability of social media. But the shadow economy around social media fake accounts is still poorly understood, due to the lack of data, transparency, and reliable way of detecting fake accounts. This research uses game-theoretical analysis to understand what makes social media influencers buy fake accounts, how the existence of fake accounts impact consumers, advertisers, social media platforms, and the overall social welfare, further, how the platform's detection affects fake account purchasing behavior and whether the platform's detection can be socially optimal or not. The central contribution of this paper is the characterization of equilibrium scenarios. We find that in a pooling equilibrium, only the influencer with low content quality ("low type") buys fake accounts offensively while the high type one does not. However, in the "costly-separating" equilibrium, the purchasing behavior flips, i.e., only the influencer with high content quality buys fake accounts defensively while the low type does not. In addition, in the "costless separating" equilibrium, no influencer buys fake accounts. We also find that fake-account fighting strategies such as detection and consumer digital literacy education may sometimes exacerbate the fake account problem, which in turn results in a non-trivial impact of detection on social welfare, i.e., a platform's fake-account detection does not always improve social welfare. We find that the platform is not incentivized to implement the socially optimal detection level (i.e., it may under- or over-detect). Thus, we may not rely entirely on social media platforms to self-regulate their fake accounts. Finally, we extend our model to explain the case in which different types of influencers buy fake accounts simultaneously.

Key words: Influencer Economy, Fake Accounts, Social Media, Signaling, Game Theory

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: 0! It turned out that the microblog was infested with fake followers. This incidence is not alone in today's global social media market, already reaching 38 billion in 2019 and rising. Facebook shut down more than 5 billion fake accounts in 2019 and estimates that no less than five percent of the user counts are

fake.

By fake accounts, we mean social media accounts designed to impersonate real users with fake personal information (e.g. names and photos) and/or behaviors (e.g. follow, view, click, comment, and share). Fake accounts may be operated by computer programs, humans, or a combination. The rise of social media fake accounts is backed by a large underground economy for producing, selling, and buying fake accounts. On Oct 21, 2019, for example, the U.S. Federal Trade Commission (FTC) settled a lawsuit with Devumi, a company that made millions of dollars by manufacturing and selling fake accounts/services, on multiple platforms including Twitter, LinkedIn, and YouTube, to actors, athletes, musicians, and other high-prole individuals on social media who wanted to appear more popular and influential online.

Social-media fake accounts can cause a range of harms to individuals, firms, and society, including wasted advertising dollars and user attention, misleading users, undermining trust on social media, the spread of misinformation (Shao et al., 2018), and manipulation of public opinions (Hjouji et al., 2018). There is an urgent need among campaign managers, social media platforms, consumers, and policymakers to develop an understanding of and strategies for fighting social-media fake accounts.

The key enabler of social-media fake accounts is the rising of the "influencer economy," where social media influencers are paid by social media campaigns for product endorsements and placements among their followers. Because influencers' advertising/sponsorship revenue increases in their influence as measured by the number of followers, friends, views, likes, and comments, they have a strong economic incentive to buy fake followers and other influence indicators powered by fake accounts.

While it is intuitive that influencers are motivated to buy fake accounts, many other aspects of the fake-account ecosystem are unclear. For example, why would advertisers pay for fake accounts in the long run as they would realize, and discount, influencers who buy fake accounts, thus has a low conversion rate. A further question is that, knowing advertisers may penalize fake-account purchasers in the long run, what type of influencers are still incentivized to buy fake accounts or associated influence indicators. Furthermore, the welfare implications of fake accounts are far from clear. On one hand, fake accounts can obfuscate influence indicators (e.g. the number of followers) which consumers use to in their decision-making, thus may undermine consumer welfare. On the other hand, fake accounts may allow new influencers to amass a sizable audience more quickly, thus increase the competitiveness of the influencer market.

To address the aforementioned gaps of understanding, we build a game-theoretic model of fake social media accounts to answer the following specific questions: What is the equilibrium fake-account purchase behavior among social media influencers?

1. What is the social welfare impact of fake accounts?

2. How does fake account detection affect the equilibrium?

3. What is the social media platform's optimal level of fake-account detection? Is it motivated to implement a socially optimal detection level?

Answers to the above questions are of broad interest to campaign managers, social media platforms, consumers, and policymakers, as the concern about the prevalence of fake accounts and the associated problems (e.g., the role of fake accounts in misinformation campaigns) has grown rapidly in recent years. To our knowledge, most existing studies of fake accounts focus on examining fake accounts' activities (Stringhini et al., 2013) and developing detection techniques (Raturi, 2018; Yuan et al., 2019). A few studies investigate the political influence of social-media fake accounts on public opinions since the 2016 U.S. presidential election. So far the academic literature has offered little understanding of the economic implications of social media fake accounts on the influencer economy.

One may argue that social media platforms can use machine learning and other technologies to detect and remove fake accounts (or their fake activities such as clicks). But the truth is such technologies are far from reliable - fake account providers are constantly developing technologies and strategies to evade such detection. Furthermore, it is unclear whether social media platforms have the incentive to detect and remove all fake accounts – after all, they also get a share of advertising revenues generated by fake accounts.

We study the aforementioned questions in a game-theoretic model that captures important elements of the ecosystem, including the influencers, consumers, advertisers, and the platform. An important ingredient of the model is that a subset of consumers is "uninformed" and uses the number of followers as a signal to infer the quality of an influencer and guide their consumption decisions, in the spirit of Spence (1978)'s signaling model. The remaining "informed" consumers can observe the quality of the influencer and do not rely on such a heuristic. Advertisers care about the "real" consumers, but they too have to infer the quality of an influencer from the observed number of followers. The platform, which shares advertiser revenue with the influencer, can mount a detection effort that increases the cost of fake accounts, but also imposes a hassle on consumers for they may be inconvenienced by the detection efforts (e.g., a real user account may be wrongly flagged as a fake account). The platform can choose the detection intensity, defined as the percentage of fake accounts it aims to flag, given the limitation (e.g., precision) of the detection technology.

Our analysis suggests that there is a "pooling" equilibrium where an influencer with a low content quality (or "low type" influencer) purchases fake accounts to mimic a high-type influencer, who does not purchase fake accounts. Interestingly, there is also a "costly separating" equilibrium, where a high-type influencer purchases fake accounts to prevent a low-type influencer from mimicking, whereas a low-type influencer does not purchase fake accounts. Finally, there is a "costless separating" equilibrium where neither the low- nor high-type influencer buys any fake account. Interestingly, in the pooling equilibrium, the number of fake accounts bought by the low-type influencer increases with the platform's detection intensity and the proportion of informed consumers

on the platform, whereas in the costly separating equilibrium, the opposite is true for the number of fake accounts bought by the high-type influencer.

The findings suggest that (1) both high- and low-type influencers may purchase fake accounts under certain conditions; the low-type influencer does it offensively while the high-type, defensively; (2) fake-account fighting strategies such as detection and consumer digital literacy education may sometimes exacerbate the fake account problem.

A platform's fake-account detection does not always improve social welfare. In the pooling equilibrium, an increase in detection intensity makes the low-type influencer buy more fake accounts, thus, the overall impact on social welfare is negative. However, in the costly separating equilibrium, an increase in detection intensity makes the high type influencer buy fewer fake accounts, which increases social welfare. However, the social media platform is not incentivized to implement the socially optimal detection level. In fact, the platform may under- or over-detect. Given the platform's inefficiency in fake-account detection, it would be unwise to rely entirely on the social media platform to tackle the fake-account problem.

Our extended model further demonstrates that different types of influencers may buy fake accounts simultaneously. In an extension where there are three types of influencers (low-, medium-, and high-type), four equilibria arise. First, there is a fully separating equilibrium in which both high and medium types buy fake accounts but the low type doesn't. Then, there is a fully pooling equilibrium in which the medium and low type influencers buy fake accounts but the high type doesn't. Interestingly, there is a third, hybrid equilibrium where the high-type influencer is separated from the pack, but the medium and low types pool together. A fourth equilibrium is also a hybrid one where the high and medium type influencers pool together but the low type is separated. All of the three types of influencers buy fake accounts in the two hybrid equilibria. The number of different types of influencers who may buy fake accounts in the equilibrium, together with the simultaneity of different types of influencers buying at the same time, may explain why fake accounts are so prevalent in social media.

2 Related Literature

Our study contributes to the literature about the economics of online frauds that is a broad topic including a few streams, e.g., deceptive advertising, fake sales, misinformation, and click fraud. As we analyze a game in the context of Influencer Economy enabled by social media platforms, our model differs in a few dimensions.

First, in terms of the influencer's fake-account purchasing strategy, our paper has a closer connection to the analytical work in the literature of deceptive advertising as well as fake sales that usually study a game in which the sellers compete for a buyer by pricing decisions and deceptive tactics such as false advertising, fake purchases/reviews and so on. The extant studies mainly focus on

only one type of equilibrium outcome, for instance, Piccolo, Tedeschi, and Ursino (2018) study the sellers' deceptive advertising strategy and characterize a class of pooling equilibria where the L-type sellers deceive a Bayes-rational buyer as their central contribution, which is consistent with the finding by Mayzlin (2006) that firms with inferior products are more likely to lie. Whereas the most interesting findings of another paper (Corts, 2013) in the same stream are separating equilibria. Similarly, this game is also studied in the context of fake sales. Chen and Papanastasiou (2021)study a social learning process in which the seller manipulates the buyers' beliefs with a fake purchase. They also preclude the class of separating equilibria, i.e., a H-type seller (the good seller in their paper) never cheats using a fake purchase. Different from the sellers in the traditional e-commerce context, the influencers in our model compete for consumers (or followers) by signaling their popularity using fake accounts instead of a pricing decision. In addition, the advertisers as the other type of influencers' buyers compete for the influencer's social network as an advertising slot whose price is decided under an auction mechanism. As the price is not used by the influencer for signaling function, so the characterization of pooling and separating in our model is different from those in the extant studies. As a result, we have interesting findings from both costly separating and pooling equilibria, which is that the H-type influencer defensively buys fake accounts in the costly separating while the L-type one offensively buys fake accounts in the pooling equilibrium. In addition, we also find a costless separating equilibrium in which neither type buys fake accounts.

Second, our paper is related to another stream of work that focusing the online platform's anti-manipulation strategy on misinformation. As a bridge between consumers and information producers, the platform can either help consumers learn the true quality of products by information disclosure (Che and Hörner, 2018; Papanastasiou, Bimpikis, and Savva, 2018; Pennycook et al., 2020) or penalize the information producers for their manipulative behaviors (Corts, 2014; Papanastasiou, 2020). In particular, the existing studies also find that a more intensive anti-manipulation strategy can lead to a higher level of manipulation (Chen and Papanastasiou, 2021; Papanastasiou, 2020). By contrast, in our model, we only consider a basic anti-fake-account effort, e.g., the fake-account detection and removal, as the platform's decision. Different from the unilateral effect of information disclosure on consumers or penalty on the sellers, the platform's anti-fake-account effort can affect the consumers' and influencers' decisions simultaneously as the imperfect technology unavoidably misclassifies the real and fake accounts. The penalty on the influencer can be reflected by their cost inflation due to anti-fake-account detection. As a result, the consequence of the platform's strategy on the manipulation level is complicated, and we indeed find that the L-type influencer buys more fake accounts while the H-type one can buy either more or fewer fake accounts under certain conditions with the platform's anti-fake-account effort.

Third, our paper also has a connection to the work by Wilbur and Zhu (2009) that focuses on the advertisers' decisions, e.g., bid and budget, in the search advertising keywords auction to study how the search engines' revenues are affected by click fraud. They find that the search engines' revenues don't change under full information of click fraud but may increase or decrease with click

fraud under different levels of competitiveness among advertisers. In their work, the search engine is not a strategic player, they just suggest the search engines leverage neutral third parties to audit the click fraud considering the misclassification problem of the search engines' own detection. By contrast, in our paper, we endogenize the platform's anti-fake-account effort, and reveal how it affects the number of fake accounts, which in turn affects the advertisers' decisions and finally the platform's own revenue. We find that the platform can be incentivized to tackle the fake-account problem under certain conditions.

Finally, to the best of our knowledge, our paper is the first analytical one to study the social media fake account problem by modeling the platform, the influencer (can also be considered as a seller), consumers and advertisers as strategic players in one model, which enables us to analyze social welfare in a broader domain and produce a systematic understanding of the social media fake accounts problem, e.g., we can compare the platform's optimal anti-fake-account decision with the socially optimal one, and we find that the platform is only incentivized to implement an efficient anti-fake-account effort under a specific condition while it could exert an over or under effort in most cases.

3 The Model

We model the ecosystem for fake accounts to consist of four types of players: a social media platform, an influencer, a unit mass of consumers, and m advertisers. The influencer produces social media content and uses the platform to distribute it to consumers. The platform provides a mechanism for consumers to follow and consume the influencer's content, such as by enabling following, subscription, or two-way friendships on the platform. The advertiser leverages the influencer's network to promote his product through sponsored posts or product placements. The advertiser pays the platform to advertise and the platform shares the advertisement revenue with the influencer.

The influencer produces one unit of social media content. The quality of the content $\dot{q}$ is a random draw from two levels, q_H and q_L ($q_H > q_L$), with probabilities p and $1-p$, respectively. We call an influencer with q_H (q_L) content quality an H-type (L-type) influencer. We normalize the influencer's cost of production to zero.

Consumers must follow the influencer to receive her content. A consumer i's valuation for content is $\theta_i q$, where θ_i is consumer i's taste parameter, and is uniformly distributed on $[0,1]$. A consumer incurs a cost c for following the influencer. We interpret c as the opportunity cost of time.

Consumers are differently informed about the influencer's type at the time of their following decisions. We assume that a proportion γ of consumers are *informed*, that is, they know the true content quality q. Such consumers have accumulated knowledge about the influencer and the ability to judge the quality of the influencer's content. The remaining $1-\gamma$ proportion of consumers are *uninformed* – they do not know the true q, but know the distribution of q and can use the influ-

encer's popularity to update their belief about q in a Bayesian manner.[1] Many empirical studies have confirmed that social media consumers use popularity indicators to guide their consumption decisions. The parameter γ represents the level of social media literacy among the consumers. We assume consumers' informativeness level is independent of their taste θ for content quality.

A consumer additionally incurs a nuisance cost c_d from the platform's anti-fake-account effort (more details later).

Based on the aforementioned assumptions, a consumer i's utility is given by

$$u_i = \begin{cases} \theta_i q - c - c_d, & \text{if i follows the influencer} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

We note that for uninformed consumers, q is a random variable; they must make a following decision based on their expected utility. We assume that consumers are risk-neutral.

Advertisers derive value from advertising to real consumers, e.g. from their purchases, brand awareness, and email-sign-ups. Although a fake account may also mimic real users in generating impressions, clicks, and other metrics, they do not purchase from the advertiser or contribute to the advertiser's bottom line. We assume that the advertiser j derives a value v_j from advertising to a real consumer and, the valuations $\{v_j\}$ are independently drawn from a distribution with a cumulative distribution function G and support $[\underline{v}, \overline{v}]$.

The advertiser incurs a unit cost, normalized to 1, for advertising to a fake follower. This is because the advertiser must spend resources to track and follow up with fake accounts as they fake impressions or clicks on advertisement like real consumers but do not generate any return. In sum, an advertiser's valuation of advertising through the influencer is

$$V(v) = vn_r - n_f \tag{3.2}$$

where n_r is the number of real consumers reached and n_f is the number of fake accounts reached.

We assume that the advertiser cannot tell whether the influencer's followers are real users. Because advertisers are often uncertain about the influencer's quality, we assume that the advertiser does not know the influencer's quality either, but she knows the distribution of the influencer's quality and can update his belief after the influencer's popularity.[2]

[1] Though we restrict our model to the number of followers as the popularity indicator, our insights are generalizable to related popularity indicators such as the number of likes, the number of forwards, and the number of comments. One reason is that because fake accounts increasingly behave like real consumers, buying fake accounts also means buying fake likes, etc. Moreover, other fake popularity indicators are inevitably backed by fake accounts.

[2] We acknowledge that there should be also a proportion of advertisers are informed as the consumers. However, as there is a considerable amount of ad spending wasted (around $23 billion) annually due to online advertising fraud that keeps being complained about in the digital marketing industry, we argue that most of the advertisers are more or less uninformed, at least not perfectly informed, i.e., they are not able to exactly distinguish the H-type influencer from the L-type one and tell the number of fake accounts in the influencer's followers. Like the example mentioned

We assume that the influencer has only one advertising slot, and the advertisers compete for the slot via a sealed-bid second-price auction organized by the platform. Such an auction format for advertising has been used in the literature and is a special case of the popular generalized second price auction used in online advertising (Liu and Viswanathan, 2014). Specifically, the advertiser with the highest bid for the slot wins and pays the second highest bid. We make a simplifying assumption that the lowest valuation $\underline{v}$ is high enough such that the lowest-valuation advertiser is still willing to participate in the auction (Technical details are introduced in the proofs of the following Lemmas).

The auction revenue goes to the platform who then shares it with the influencer. We assume for each dollar of the advertising revenue, the platform shares λ with the influencer and keeps the remaining $1-\lambda$.[3] This assumption is consistent with the observation that social media platforms share revenue with their top influencers. We normalize the platform's cost of operation to zero. Thus, the platform's expected payoff is given by

$$\pi_{plat} = (1-\lambda)\prod(m) \tag{3.3}$$

where $\prod(m) = \int_{\underline{v}}^{\overline{v}} \left(n_r[v - \frac{1-G(v)}{g(v)}] - n_f\right)(\frac{dG(v)^m}{dv})dv$, see proofs in the Appendix.

The influencer's main decision is whether to purchase fake accounts. Let x denote the number of fake accounts purchased by the influencer. The unit price of fake social media accounts is $c_f c_f$. The influencer's expected payoff is given by

$$\pi_{inf} = \lambda\prod(m) - c_f x \tag{3.4}$$

The platform's main decision is its anti-fake-account effort $d(d \geq 0)$. The anti-fake-account effort includes fake-account detection and prevention. For example, the platform may use machine learning to detect abnormal behaviors of a user account. It may also deploy user verification technologies such as reCAPTCHA and two-factor authentication to make it harder to automate fake accounts. Here, $d = 0$ means the platform does nothing about the fake accounts. A higher effort d means, for example, more aggressive detection, more frequent scans, and/or stricter user verification.

No anti-fake-account technology is perfect. False positives are inevitable in fake account detection. While the intensified anti-fake-account effort can increase the unit cost of fake accounts, increased detection can also result in more legitimate accounts being misclassified, which adds to the nuisance costs of real consumers. Similarly, stricter user verification (e.g. reCAPTCHA) can also increase the hassle of legitimate users. Therefore, the maturity of the ani-fake-account technology can affect both the unit cost of fake accounts and consumers' nuisance costs. Specifically, we use a technology-

at the beginning, the advertiser would not expect a 0 conversion though she may realize the fake-account problem in advance, which indicates the advertiser's informedness is restricted.

[3]Or, we can say that social media platforms share their advertising revenue with the influencers to encourage the creation of quality content. It doesn't matter who is the leader in the revenue sharing process.

level parameter $\tau\,(1 > \tau \geq 0)$ to capture the maturity of the anti-fake-account technology. A higher technology level τ is associated with more mature and effective detection algorithms and prevention technologies. We let the consumer's nuisance cost be

$$c_d = \rho(1-\tau)d = \phi_1 d \tag{3.5}$$

where ρ is a constant and ϕ_1 is shorthand for $\rho\,(1-\tau)$. By this formulation, the nuisance cost is an increasing function of the anti-fake-account effort d and a decreasing function of the technology level τ.

We let the unit cost for fake accounts c_f be

$$c_f = \kappa + \frac{1}{1-\tau}d = \kappa + \phi_2 d \tag{3.6}$$

where κ is a constant and ϕ_2 is shorthand for $1/\,(1-\tau)$. By this formulation, the unit cost of fake accounts is an increasing function of the anti-fake-account effort d and the technology level τ.

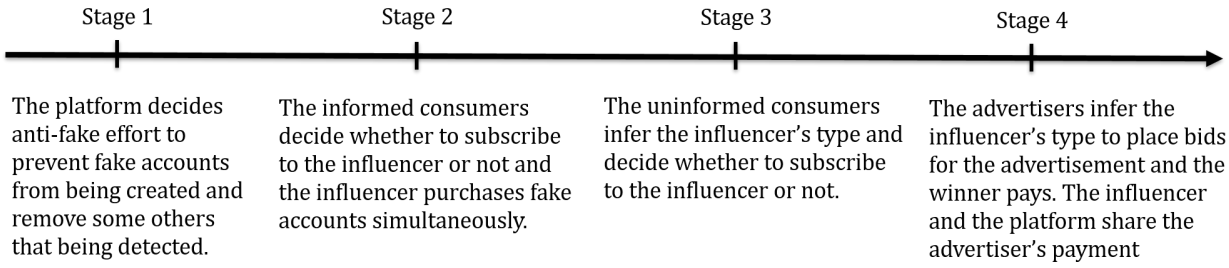


Figure 1: Game Timeline

The timeline of the game is as follows. At **time 1**, the platform decides its anti-fake-account effort, d. A non-zero anti-fake-account effort d would prevent some fake accounts from happening and remove some detected fake accounts. Nature draws the influencer's content quality q, the consumers' taste θ, and the advertisers' valuation v. At **time 2**, the informed consumers decide whether to follow the influencer. At the same time, the influencer decides the number of fake accounts x to buy. These fake accounts have survived detection and will not risk being removed in the future. After these decisions, the influencer has n_1 followers, which include n_0 informed consumers and x fake accounts. At **time 3**, uninformed consumers observe the displayed number of followers n_1 and decide whether to follow the influencer. After the uninformed consumers' decision, the influencer has n_2 followers. At **time 4**, the advertisers observe the displayed number of followers n_2 and place bids for the advertisement slot according to their valuations. Then the winning advertiser's advertisement is displayed to the influencer's followers along with the influencer's social media content. The winning advertiser pays the platform, who then shares the revenue with the influencer.

In practice, the key stakeholders are likely to make their decisions on an iterative basis: consumers may arrive and make their following decisions at different times; fake account detection (and re-

Notation	*Interpretation*
d	The platform's anti-fake-account effort
x	The influencer's decision variable, the number of fake accounts purchased
γ	The proportion of informed consumers, which also represents social media literacy level.
λ	The revenue sharing parameter between the influencer and the platform.
m	the number of advertisers in the auction
v	Value of advertising to each real consumer in the influencer's network.
G, $\underline{v}, \overline{v}$	The valuations of advertising to each real consumer are independently draws from a cumulative distribution function of G with a support, $[\underline{v}, \overline{v}]$
$\mu(\underline{v}, \overline{v})$	The average revenue that a real consumer can make.
n_1, n_2	The displayed number of followers by the end of different stages.
q_H, q_L	The content quality of H- and L-type influencer, respectively.
p	The probability of drawing the H-type influencer.
c, c_d	A consumer's cost of following and consume the influencer's content.
$p(H\|n)$	The conditional probability of an influencer to be H-type given the observed consumer count is n
u_i	Consumer i's utility function.
π_a, π_{inf}, π_{plat}	The expected payoff of the winning advertiser, the influencer, and the platform, respectively.
W	Social welfare.
$\theta \in [0,1], F$	Consumers' taste and its cumulative distribution function
$\tau \in [0,1)$	The platform's anti-fake-account technology level.
η_1, η_2, η_3	The conditions for selecting the unique equilibrium outcome.

Table 1: Notations

moval) and purchasing of fake accounts may also happen iteratively. Our model simplifies such a continuous process and uses a sequence of activities to best capture the environment for each decision. Specifically, the rationales for our choice of decision sequence is as follows:

- First, because uninformed consumers rely on the influencer's popularity to infer her quality, it is natural for them to wait for the popularity signal to materialize before determining whether to follow the influencer. In contrast, informed consumers already know the true quality and thus do not need to wait. That is why we assume informed consumers make their following decisions before uninformed ones.

- Second, one of the benefits of purchasing fake accounts is to convince uninformed consumers

to follow the influencer. Therefore, the influencer needs to purchase fake accounts before uninformed consumers make their decisions. Although we assume the informed consumers and the influencer move simultaneously, the model remains the same if the two decisions occur sequentially since they are independent of each other.

- Third, any survived fake account must have passed the platform's detection. That is why we assume the platform's detection decision occurs before the influencer's fake-account purchase decision. This decision order is most natural also because the cost of fake accounts depends on the platform's anti-fake-account effort.

- Finally, 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. Having the advertiser move after the uninformed consumers reflects this observation.

4 Equilibrium Analysis

4.1 The Displayed Number of Followers at Different Times

According to a consumer i's utility given in the main text, there are $n_0^H = \gamma[1 - F(\frac{c+\phi_1 d}{q_H})]$ ($n_0^L = \gamma[1 - F(\frac{C+\phi_1 d}{q_L})]$) informed consumers following the H-type (L-type) influencer. As $q_H > q_L$, we can have $n_0^H > n_0^L$ which means H-type influencer has more informed consumers than the L-type one. At **time** 2, the influencer buys x fake accounts. Therefore, the influencer's displayed number of followers will be $n_1 = n_0 + x$, where $n_0 \in \{n_0^H, n_0^L\}$. At **time 3,** an uninformed consumer uses inferred content quality through the signal of perceived popularity to form their utility

$$u_i = \begin{cases} \theta_i[p(H|n_1)q_H + (1 - p(H|n_1))q_L] - c - \phi_1 d, & if\ i\ follows\ the\ influencer \\ 0, & otherwise \end{cases} \tag{4.1}$$

Where $p(H|n_1)$ is the conditional probability of the influencer to be H-type given the displayed number of followers observed by the uninformed consumer is n_1. Then there are $\Delta n = (1-\gamma)[1 - F(\frac{c+\phi_1 d}{p(H|n_1)q_H + (1-p(H|n_1))q_L})]$ uninformed consumers added into the influencer's follower base. By the end of **time 3**, the influencer has n_2 followers, where $n_2 = n_1 + \Delta n = n_0 + x + \Delta n$.

4.2 Payoff Functions

According to the payoff functions of the advertiser, the influencer and the platform above, we are deriving them in a detailed format in this section. To simplify the analysis, we assume, without loss of generality, that there are two advertisers whose valuation of a real consumer, v, is uniformly distributed between $\underline{v}$ and $\overline{v}$, thus, $m = 2, G(x) = \frac{x-\underline{v}}{\overline{v}-\underline{v}}$ in the following analysis.

As the equilibrium payoff of the advertiser (v) is $\pi_a(v) = n_r \int_{\underline{v}}^{v} G(x)^{m-1} dx$, the simplified format

is given by

$$\pi_a(v) = \frac{n_r}{2(\overline{v}-\underline{v})}(v-\underline{v})^2 \quad , \tag{4.2}$$

Where $n_r = p(H|n_2)(n_0^H + \Delta n_H) + (1-p(H|n_2))(n_0^L + \Delta n_L)$, and$p(H|n_2)$ is the conditional probability of the influencer to be H-type given the observed number of followers by the advertiser is n_2.

As the expected revenue function for the auctioneer is $\int_{\underline{v}}^{\overline{v}} \left(n_r[v - \frac{1-G(v)}{g(v)}] - n_f\right)(\frac{dG(v)^m}{dv})dv$, the simplified format is given by

$$\prod = \mu(\overline{v}, \underline{v})n_r - n_f \tag{4.3}$$

Where $\mu(\overline{v}, \underline{v}) = \frac{\overline{v}+2\underline{v}}{3}$ is the average revenue that a real consumer can make.

Therefore, the platform's and the influencer's expected payoff functions are given by

$$\pi_{plat} = (1-\lambda)\prod = (1-\lambda)[\mu(\overline{v}, \underline{v})n_r - n_f] \tag{4.4}$$

$$\pi_{inf} = \lambda\prod - (\kappa + \phi_2 d)x = \lambda[\mu(\overline{v}, \underline{v})n_r - n_f] - (\kappa + \phi_2 d)x \tag{4.5}$$

where n_r, $\mu(\overline{v}, \underline{v})$, and$p(H|n_2)$ are the same as above, and $n_f = n_2 - n_r$.

4.3 Perfect Bayesian Equilibrium

In this study, we analyze pure strategy Perfect Bayesian Equilibrium (PBE). The displayed number of consumers by the end of **time 2** serves as a signal of the influencer's type to the uninformed consumers who update their belief using Bayes' rule. If the displayed number of consumers by the end of **time 2** is the same for the two types of influencers, the PBE is pooling equilibrium, otherwise, it's separating equilibrium. As a result, we might see that the two types of influencers buy different numbers of fake accounts in the pooling equilibrium. In particular, the relationship of the displayed number of consumers between the two types of influencers keeps stable from the end of **time** 2 to the end of **time** 3, i.e., the displayed number of consumers by the end of **time** 3 can also serve as a signal of the influencer's type to the advertisers for updating their beliefs about the number of real consumers in the influencer's account. The definition of pooling and separating equilibrium remains the same from either the uninformed consumers' or the advertisers' perspectives. As the out-of-equilibrium beliefs are arbitrarily defined in the signaling game, there will be multiple equilibria. In this study, we adopt the *lexicographically maximum sequential equilibrium* (LMSE), which is proposed by Mailath, Okuno-Fujiwara, and Postlewaite (1993), to conduct equilibria refinement. This LMSE refinement method has already been used in many management studies (Guo and Jiang, 2016; Jiang et al., 2016; Schmidt et al., 2015). The LMSE will select the unique and most profitable equilibrium outcome from the perspective of the type who has the most incentive to reveal her/his type. In our study, the H-type influencer has an incentive to reveal her/his true type while the

L-type influencer is trying to mimic the H-type one. Thus, a PBE is LMSE if it is most profitable for the H-type influencer. If the most profitable PBE is not unique for the H-type influencer, it also needs to be most profitable for the L-type influencer. In the following sections, we'll show how we determine the unique equilibrium outcome by LMSE.

Pooling Equilibrium

As mentioned above, in the pooling equilibrium, H-type and L-type influencers have the same displayed number of consumers by the end of **time 2** and **3**. Solving the signaling game along with the LMSE refinement (Mailath et al. 1993), we get the unique pooling equilibrium for the two types of influencers' fake accounts purchasing behavior, which is given in Lemma 1.

Lemma 1. *There exists a unique pooling equilibrium in which the L-type influencer purchases fake accounts while the H-type influencer doesn't under certain conditions. The numbers of fake accounts purchased by the two types of influencers are*

$$\begin{cases} x^*_{H,pool} = 0 \\ x^*_{L,pool} = \gamma \frac{q_H - q_L}{q_H q_L}(c + \phi_1 d) \end{cases} \tag{4.6}$$

and the condition is

$$[\lambda\mu(\overline{v}, \underline{v})p + \lambda p - \lambda - \kappa - \phi_2 d]\gamma(\frac{1}{q_L} - \frac{1}{q_H}) + \lambda\mu(\overline{v}, \underline{v})(1-\gamma)(\frac{1}{q_L} - \frac{1}{pq_H + (1-p)q_L}) \geq 0 \tag{4.7}$$

where $\mu(\overline{v}, \underline{v}) = \frac{\overline{v}+2\underline{v}}{3}$,*which also applies in the following Lemmas.*

Proof. See proofs for Lemma 1, the following Lemmas, and propositions in the Appendix. Q.E.D

Separating Equilibrium

Similarly, in the separating equilibrium, H-type and L-type influencers have a different displayed number of consumers by the end of **time** 2 and 3. In particular, there is a special case. As we know, the two types of influencers have a different number of informed, also called incumbent, consumers. If neither of them buys fake accounts, their display counts of consumers by the end of **time** 2 and 3 will be still different. If the equilibrium for this case exists, it should be also classified into the separating equilibrium according to our definition. To distinguish this special separating equilibrium from the general separating one in which at least one type of influencer will buy fake accounts, we call the general case *costly separating* equilibrium as there is an expense of purchasing fake accounts, while the special case *costless separating* equilibrium in which the growth of consumers is totally organic. We are not the first to propose the terms "costly" vs "costless", in fact, Guo, Xiao, and Zhang (2017) also use the two terms in their paper. The unique costly separating equilibrium for the influencers' fake accounts purchasing behavior is given in Lemma 2,

and the condition for the costless separating equilibrium is given in Lemma 3.

Lemma 2. *There exists a unique costly separating equilibrium in which the H-type influencer purchases fake accounts while the L-type one doesn't under certain conditions. The numbers of fake accounts purchased by the two types of influencers are, respectively:*

$$\begin{cases} x^*_{H,sep} = \frac{\lambda\mu(\overline{v},\underline{v})-\gamma(\kappa+\phi_2 d)}{\lambda+\kappa+\phi_2 d}\frac{q_H-q_L}{q_H q_L}(c+\phi_1 d) \\ x^*_{L,sep} = 0 \end{cases} \tag{4.8}$$

and the condition is

$$\lambda\mu(\overline{v},\underline{v}) \geq \gamma(\kappa+\phi_2 d) \tag{4.9}$$

Lemma 3. *when $\lambda\mu(\overline{v},\underline{v}) < \gamma(\kappa+\phi_2 d)$, there also exists a unique costless separating equilibrium in which neither type of the influencers purchases fake accounts.*

As widely discussed in the deceptive advertising literature, the deceptive behaviors are all from the L-type sellers (Chen and Papanastasiou, 2021; Piccolo, Tedeschi, and Ursino, 2018). Therefore, intuitively, we would think that the cheating behaviors with fake accounts are from the L-type influencer here. However, according to our model, we find that the cheating behaviors are not limited to the L-type influencer, the H-type one also has an incentive to buy fake accounts under certain conditions. The fake accounts are used by the H-type influencer to prevent the L-type one from mimicking her/him.

4.4 Equilibrium Characterization

From Lemma 1 and Lemma 2, we can find that the fake accounts purchasing behavior for the two types of influencers flip between the pooling and costly separating equilibria. To figure out what conditions drive the flipping pattern for the different types of influencers, we still use the LMSE method to select a unique equilibrium outcome from the pooling, costly, and costless equilibria proposed in Lemma 1, 2, and 3. The unique equilibrium outcome along with conditions is given in Proposition 1.

Proposition 1. *There exists a parameter space to determine the unique equilibrium outcome for the influencer's fake accounts purchasing behavior under different conditions. (1) If $\gamma \leq \eta_1$,$\lambda \geq \eta_2$, and $\mu(\overline{v},\underline{v}) > \eta_3$, the unique equilibrium is a pooling equilibrium as characterized in Lemma 1. (2) If $\gamma \leq \eta_1$ and $\mu(\overline{v},\underline{v}) \leq \eta_3$, the unique equilibrium is a costly separating equilibrium as characterized in Lemma 2. (3) if $\gamma > \eta_1$, the unique equilibrium is a costless separating equilibrium as characterized in Lemma 3. where γ is the proportion of informed consumers, λ is the influencer's revenue-sharing proportion from the platform, $\mu(\overline{v},\underline{v})$ is an advertiser's average valuation of a real consumer, and η_1, η_2, η_3 are functions of other parameters in the three conditions, respectively*

$$\begin{cases} \eta_1 = \frac{\lambda\mu(\overline{v},\underline{v})}{\kappa+\phi_2 d} \\ \eta_2 = \frac{(\kappa+\phi_2 d)\gamma(\frac{1}{q_L}-\frac{1}{q_H})}{[\mu(\overline{v},\underline{v})p+p-1]\gamma(\frac{1}{q_L}-\frac{1}{q_H})+\mu(\overline{v},\underline{v})(1-\gamma)(\frac{1}{q_L}-\frac{1}{pq_H+(1-p)q_L})} \\ \eta_3 = \frac{(\kappa+\phi_2 d)\gamma\frac{q_H-q_L}{q_H q_L}+\lambda(1-p)\gamma\frac{q_H-q_L}{q_H q_L}}{\lambda[\gamma(\frac{1}{pq_H+(1-p)q_L}-\frac{p}{q_H}-\frac{1-p}{q_L})+(\frac{1}{q_L}-\frac{1}{pq_H+(1-p)q_L})]} \end{cases} \tag{4.10}$$

We can interpret **Proposition 1** as 1) when the social media literacy is lower than a certain level, the influencer will buy fake accounts. In particular, when the advertiser's average valuation of a real consumer is higher, also the influencer can have a higher revenue sharing proportion, the L-type influencer is more likely to buy fake accounts while the H-type doesn't buy. 2) However, if the advertiser's average valuation of a real consumer is lower, the H-type influencer is more likely to buy fake accounts while the L-type doesn't buy regardless of revenue sharing proportion. 3) If we would like to prevent influencers from buying fake accounts, the social media literacy level must be higher than a certain level. Note it is not always guaranteed as the certain level could be greater than 1 that is the maximum social media literacy level.

We illustrate the equilibrium outcome in Figure 2 by a set of numeric examples. In each panel of Figure 2, we first fix the platform's anti-fake-account effort as an exogenous variable, then iterate two parameters with managerial implications, the probability of an influencer to be H-type and the platform's anti-fake-account technology level, over their range by fixing other parameters as constants. From the three panels in Figure 2, we can see that the boundaries splitting the equilibrium outcome space shift with the anti-fake-account effort. If the platform doesn't exert an anti-fake-account effort in panel (a), there is only one boundary to split the space with costly separating and pooling equilibrium regions. As the anti-fake-account effort increases in panel (b) and (c), there are two patterns: one pattern is that a second boundary emerges to create the costless separating equilibrium region in addition to the previous two regions, the other pattern is that it is easier for the H-type influencer to separate her/him from the L-type one through buying fake accounts while harder for L-type influencer to pool together with the H-type one using fake accounts.

We also find that there are two regions in which the equilibrium doesn't change over the anti-fake-account effort. One region is the bottom-left of Figure 2 where the probability of an influencer to be H-type is small (e.g. smaller than 0.17) and the anti-fake-account technology level is not sufficiently accurate (e.g. lower than 0.5), the *costly separating* equilibrium remains regardless of the anti-fake-account effort. The other region is the bottom-right of Figure 2 where the probability of an influencer to be H-type is large (e.g. larger than 0.5) and the anti-fake-account technology level is less accurate (e.g. still lower than 0.5), the pooling equilibrium remains. The equilibrium in other areas switches either between pooling and costly separating equilibrium or among pooling, costly separating and costless separating one under different conditions.

As we know, the probability of an influencer to be H-type can be also interpreted as the proportion

of H-type influencers on the platform, thus, the intuitions of the above findings from Figure 2 are, on one hand, if the H-type influencer finds she/he is in a minority group, she/he is more eager to separate her-/himself from the majority L-type ones because her/his content quality will be undervalued by the uninformed consumers, so that the H-type influencer has an incentive to exploit the fake accounts when the platform doesn't exert an anti-fake-account effort or the anti-fake-account technology level is not high enough. On the other hand, the uninformed consumers' expectation of content quality increases with the proportion of H-type influencers on the platform. When the proportion of H-type influencer is higher than a certain level, the L-type influencer will find it is profitable to mimic the H-type one as her/his content quality will be overvalued by the uninformed consumers, so that the L-type influencer will buy fake accounts to pool together with the H-type one when the platform doesn't exert an anti-fake-account effort or the anti-fake-account technology level is not high enough. However, for both H- and L- type influencers, the cost of fake accounts increases with anti-fake-account effort and technology level. If the benefit of cheating behaviors can not cover their cost, they will stop buying fake accounts, which is the costless separating region in Figure 2.

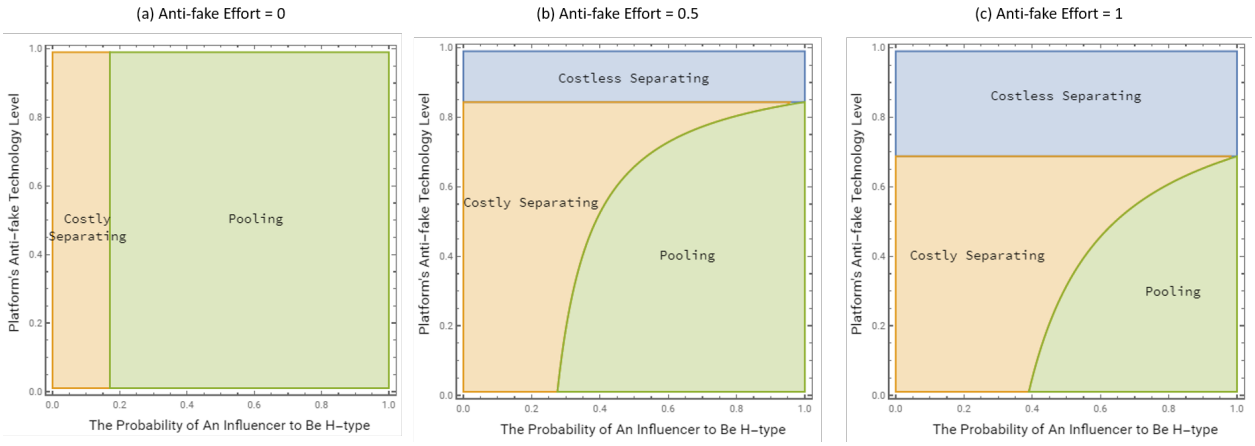


Figure 2: Illustration of the Equilibrium Outcome

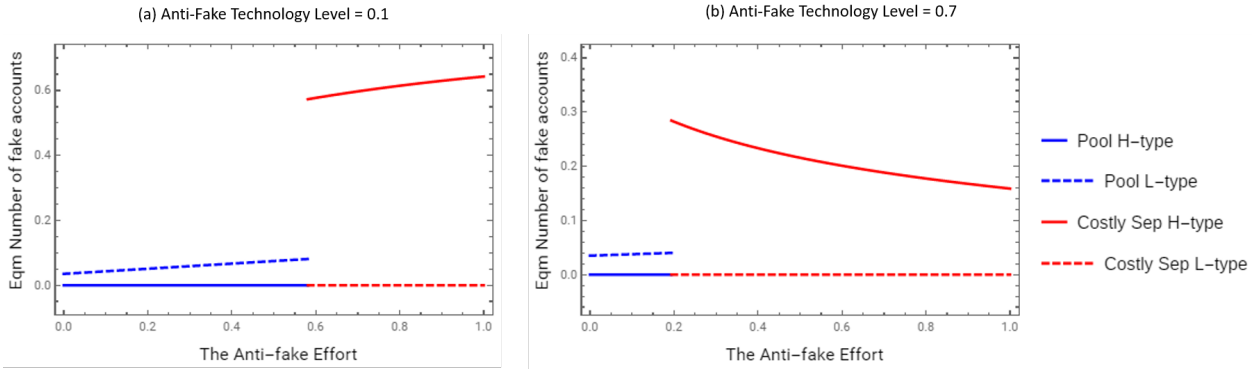
Note: $\underline{v} = 10, \overline{v} = 20, \lambda = 0.15, \gamma = 0.5, q_H = 20, q_L = 10, C = 5, \kappa = 0.8, \rho = 2.5$.

5 Equilibrium Properties of the Influencer's Fake-Account Purchasing Strategy

Given a fixed parameter space, we first examine how the pattern of the influencer's fake accounts purchasing strategy changes over the platform's anti-fake-account effort. The number of fake accounts purchased by the two types of influencers in the pooling and costly separating equilibrium is characterized in Lemma 1 and 2. By taking derivative of $x^*_{H,pool}$ and $x^*_{L,sep}$ with respect to anti-fake-account effort d, we have $\frac{\partial x^*_{L,pool}}{\partial d} > 0$ but $\frac{\partial x^*_{H,sep}}{\partial d}$ can be either > 0 or < 0 (see the details in Appendix, also for the following derivative analysis). Thus, the L-type influencer in the pooling

equilibrium buys more fake accounts with anti-fake-account effort while the number of fake accounts bought by the H-type influencer can either increase or decrease under certain conditions. From the proof of Lemma 1 and 2, we can find that, on one hand, in the pooling equilibrium, the number of informed consumers following the L-type influencer decreases with the anti-fake-account effort at a faster speed than that for the H-type one (i.e. $\frac{1}{q_L} > \frac{1}{q_H}$). Therefore, the L-type influencer needs to buy more fake accounts to catch up with the H-type one with the same displayed number of followers, i.e., to pool together with the H-type one. On the other hand, in the costly separating equilibrium, when the anti-fake-account technology level is lower, the number of informed consumers will be wrongly removed at a faster speed with the anti-fake-account effort. However, the cost of buying fake accounts increases at a slower speed. Thus, the H-type influencer prefers to buy more fake accounts at a lower cost to obtain a leading displayed number of followers, further to ensure the separating effect. As the anti-fake-account technology level increases, the number of mistakenly removed informed consumers increases in slower speed with the anti-fake-account effort but the cost of fake accounts increases at a faster speed, thus, the H-type one prefers to leverage the advantage of extant informed consumers with fewer bought fake accounts to obtain the desired displayed number of followers to separate themselves.

We populate the parameters to illustrate the relationship between the influencer's fake accounts purchasing strategy and the anti-fake-account effort in Figure 3. In Figure 3 (a), the two types of influencers buy more fake accounts with the anti-fake-account effort in both pooling and costly separating equilibrium. However, in Figure 3 (b), the two types of influencers have an opposite pattern of fake accounts purchasing strategy. The pattern of the L-type influencer's strategy in the pooling equilibrium remains but that of H-type one's strategy in the costly separating equilibrium flips because the anti-fake-account technology has increased.

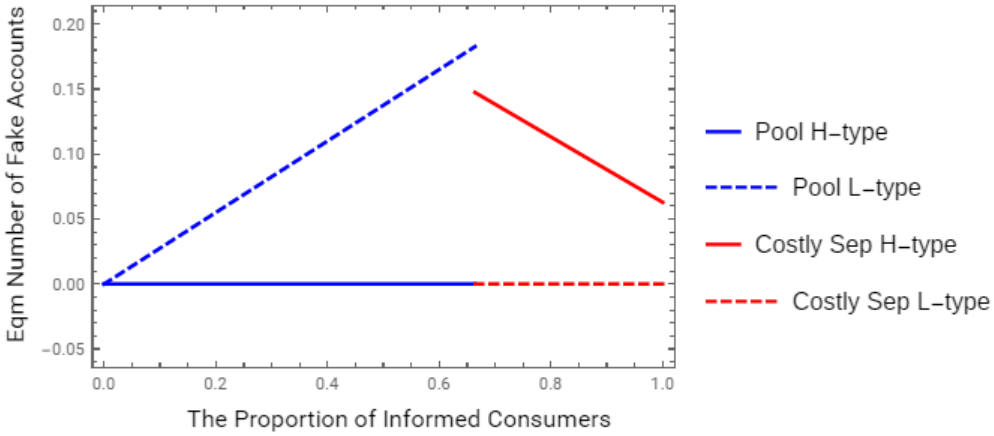


Note: $\underline{v} = 15, \overline{v} = 20, \lambda = 0.7, \gamma = 0.7, q_H = 20, q_L = 10, C = 1, p = 0.1, \kappa = 0.8, \rho = 2.5.$

Figure 3: Impact of anti-fake-account Effort on the Influencer's Fake Accounts Purchasing Strategy

As we stated above, the proportion of informed consumers γ also represents the consumers' social media literacy level. To investigate how this social media literacy level impacts the influencer's fake accounts purchasing strategy, we take the derivative of $x^*_{H,pool}$ and $x^*_{L,sep}$ with respect to the proportion of informed consumers γ, we have $\frac{\partial x^*_{L,pool}}{\partial \gamma} > 0$ and $\frac{\partial x^*_{H,sep}}{\partial \gamma} < 0$. In other words,

in the pooling (costly separating) equilibrium, the L-type(H-type) influencer buys more (fewer) fake accounts with the social media literacy level. The intuition behind is that, given the anti-fake-account effort and other parameters, as social media literacy level increases, more and more uninformed consumers will switch to be informed ones, which in turn follow the H-type influencer. As a result, the L-type influencer needs to buy more fake accounts to catch up with the H-type one follower base while the H-type influencer needs fewer fake accounts to separate themselves. The numeric illustration for this impact is shown in Figure 4.



Note: $\underline{v} = 10, \overline{v} = 20, \lambda = 0.15, q_H = 20, q_L = 10, C = 5, p = 0.5, \kappa = 0.8, \rho = 2.5, d = 0.4, \tau = 0.5.$

Figure 4: Impact of Social Media Literacy Level on the Influencer's Fake Accounts Purchasing Strategy

Above all, the equilibrium properties of the influencer's fake accounts purchasing strategy is given in Proposition 2.

Proposition 2. *(a) The platform's anti-fake-account effort can make the H-type influencer buys either fewer or more fake accounts in the costly separating equilibrium while the L-type influencer simply buys more in the pooling equilibrium. (b) The social media literacy education could either mitigate or exacerbate the fake-account problem as the H-type influencer buys more fake accounts in the costly separating equilibrium but the L-type influencer buys more in the pooling equilibrium.*

The above analysis on the properties of the equilibrium number of fake accounts purchased by the influencer suggests that (1) both H- and L-type influencers could buy fake accounts under certain conditions; the L-type influencer does it offensively while the H-type, defensively; (2) fake-account fighting strategies such as platform's detection and consumer digital literacy education does not necessarily help to solve the fake-account problem.

6 Social Welfare Analysis

In this study, social welfare is defined as the sum of expected payoffs for all players including the influencer, the platform, the advertiser, and the consumers. The social welfare in the pooling, costly

separating, and costless separating equilibrium is given by W_p^*, W_s^*, W_{ls}^* respectively as follows.

$$W_{\{p,s,ls\}}^* = pW_H^{*\{p,s,ls\}} + (1-p)W_L^{*\{p,s,ls\}} \tag{6.1}$$

$$\begin{cases} W_H^{*\{p,s,ls\}} = \pi_{inf,H}^{*\{p,s,ls\}} + \pi_{plat,H}^{*\{p,s,ls\}} + \pi_{a,H}^{*\{p,s,ls\}} + \Pi_{c,H}^{*\{p,s,ls\}} \\ W_L^{*\{p,s,ls\}} = \pi_{inf,L}^{*\{p,s,ls\}} + \pi_{plat,L}^{*\{p,s,ls\}} + \pi_{a,L}^{*\{p,s,ls\}} + \Pi_{c,L}^{*\{p,s,ls\}} \end{cases} \tag{6.2}$$

The above $\pi_{inf,H}^{*\{p,s,ls\}}, \pi_{plat,H}^{*\{p,s,ls\}}, \pi_{a,H}^{*\{p,s,ls\}}, \Pi_{c,H}^{*\{p,s,ls\}}, \pi_{inf,L}^{*\{p,s,ls\}}, \pi_{plat,L}^{*\{p,s,ls\}}, \pi_{a,L}^{*\{p,s,ls\}}, \Pi_{c,L}^{*\{p,s,ls\}}$ are introduced in the proofs of Lemma 1, 2, and 3 in the appendix, where π_a^*=$2E[\pi_a^*(v)] = 2\int_{\underline{v}}^{\overline{v}} \pi_a^*(v)g(v)dv = \frac{(\overline{v}-\underline{v})}{3}n_r^*$ is the total expected payoffs for all advertisers.

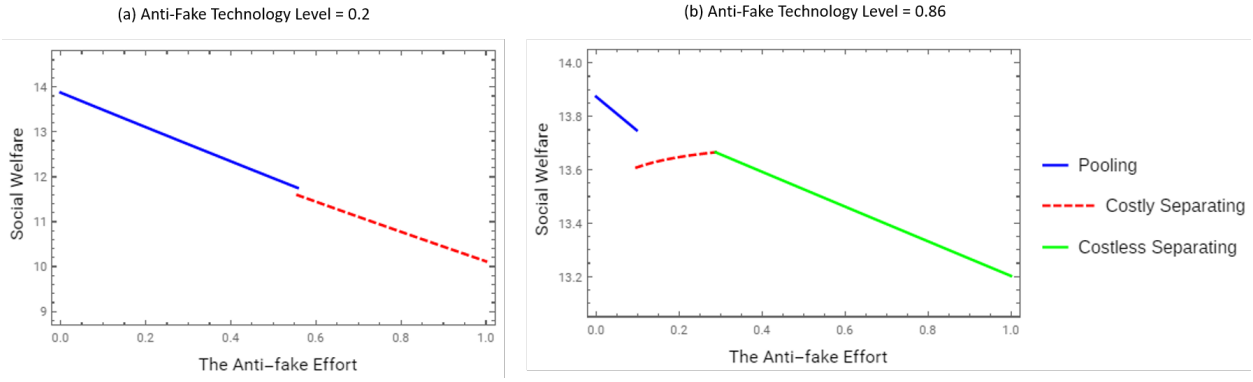
To understand the social welfare impact of fake accounts, we first examine how the platform's anti-fake-account effort d affects social welfare. Further, given the anti-fake-account effort, we also conduct comparative statics analysis to investigate how social welfare changes over the exogenous parameters of interest, the anti-fake-account technology level τ and the proportion of informed consumers γ, as what we did in Section 5.

6.1 Impact of Platform's Anti-fake-account Effort on Social Welfare

Taking derivative of W_p^*, W_s^*, and W_{ls}^* with respect to anti-fake-account effort d, we have $\frac{\partial W_p^*}{\partial d} < 0$ and $\frac{W_{ls}^*}{\partial d} < 0$. Thus, the social welfare decreases with the anti-fake-account effort in the pooling and costless separating equilibrium. However, the sign of $\frac{\partial W_s^*}{\partial d}$ is not deterministic. When the anti-fake-account technology level τ is lower, $\frac{\partial W_s^*}{\partial d} < 0$, but if τ is higher, $\frac{\partial W_s^*}{\partial d}$ could be either ¡ 0 or ¿ 0. It means that in the costly separating equilibrium, the social welfare could either increase or decrease with the anti-fake-account effort when the anti-fake-account technology level is high but just decreases with anti-fake-account effort when the anti-fake-account technology level is low. From the numeric examples in Figure 5 (a) and (b), we can see the opposite relationship between social welfare and the anti-fake-account effort in the costly separating equilibrium under different levels of anti-fake-account technology.

The intuition behind this pattern is that, first, as shown in formulas of advertiser's equilibrium payoff and the expected revenue of the influencer and the platform, we know that the welfare loss results from two parts: 1) the decrease in the number of real consumers, 2) the increase in the number of fake accounts. Then, from a consumer's utility function, we can understand that the platform's anti-fake-account effort harms the consumers' utility, which in turn decreases the number of real consumers. Thus, the platform's anti-fake-account effort results in welfare loss from the real consumers' part. According to Figure 2, we also learn that the platform's anti-fake-account effort makes the L-type influencer buy more fake accounts in pooling equilibrium so that the anti-fake-account effort results in the welfare loss from the fake accounts part in the pooling equilibrium. Overall, the platform's anti-fake-account effort harms the social welfare in the pooling equilibrium as well as in the costless separating equilibrium. However, Figure 2 also tells us that H-type

influencer buys either more or fewer fake accounts with the anti-fake-account effort in the costly separating equilibrium, which in turn either decreases or increases social welfare. As there could be two opposite forces of anti-fake-account effort on social welfare in the costly separating equilibrium, thus, the overall impact depends, i.e., the platform's anti-fake-account effort could either increase or decrease social welfare as we see from the red dashed lines in Figure 5.



Note: $\underline{v} = 10, \overline{v} = 20, \lambda = 0.15, \gamma = 0.7, q_H = 20, q_L = 10, C = 5, p = 0.5, \kappa = 0.8, \rho = 2.5$

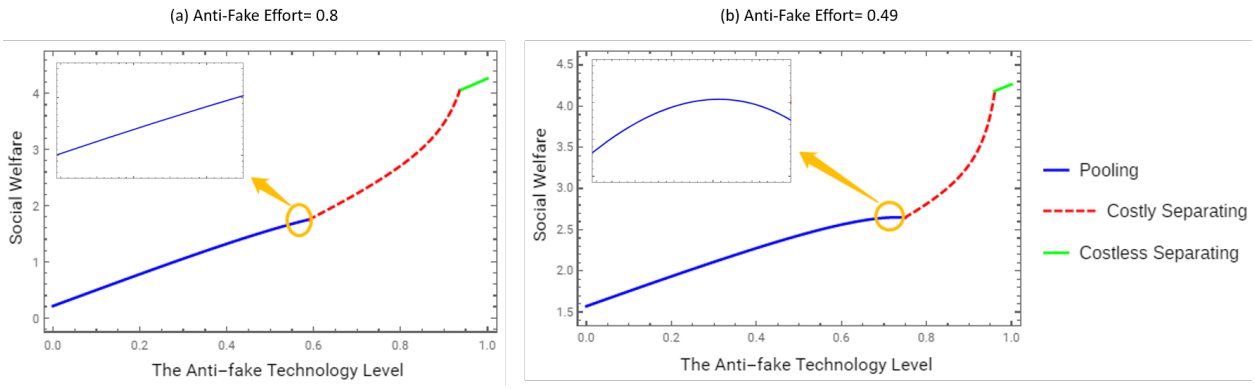
Figure 5: Impact of anti-fake-account Effort on Social Welfare

6.2 Impact of Platform's Anti-fake-account Technology Level on Social Welfare

Taking derivative of W_p^*, W_s^*, and W_{ls}^* with respect to the anti-fake-account technology level τ, we have $\frac{\partial W_s^*}{\partial \tau} > 0$ and $\frac{W_{ls}^*}{\partial \tau} > 0$. Thus, the social welfare increases with the anti-fake-account technology level in the costly and costless separating equilibrium. However, when the anti-fake-account technology level τ is lower, $\frac{\partial W_p^*}{\partial \tau} > 0$, but if τ is higher, $\frac{\partial W_s^*}{\partial d}$ could be either ¡ 0 or ¿ 0. Thus, in the pooling equilibrium, the social welfare could either increases or decreases with the anti-fake-account technology level. We populate two sets of parameters under two anti-fake-account effort, respectively. The pattern in Figure 6 (a) indicates that social welfare increases with anti-fake-account technology level in pooling, costly separating, and costless separating equilibria. However, in Figure 6 (b), we can find that social welfare could decrease with anti-fake-account technology level under certain conditions in the pooling equilibrium. This finding suggests that increasing the anti-fake-account technology level does not always improve social welfare.

6.3 Impact of the Proportion of Informed Consumers on Social Welfare

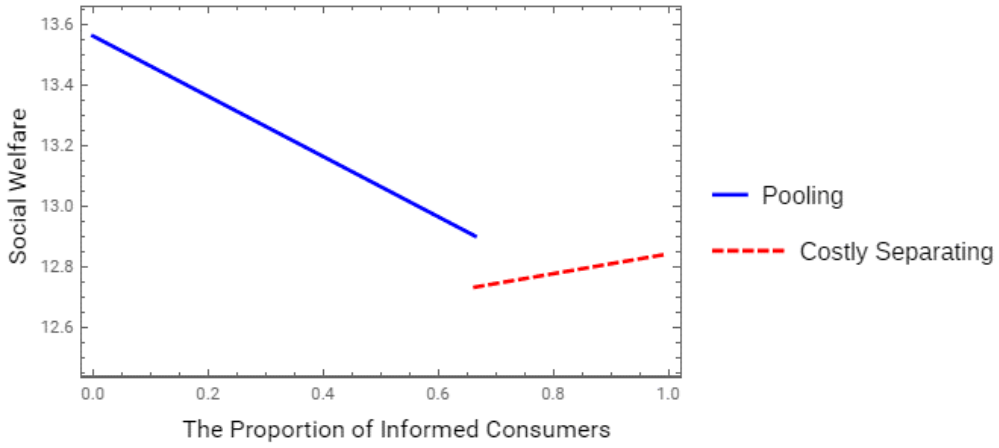
Taking derivative of W_p^*, W_s^*, and W_{ls}^* with respect to the proportion of informed consumers γ, we have $\frac{\partial W_p^*}{\partial \tau} < 0$, $\frac{\partial W_s^*}{\partial \gamma} > 0$ and $\frac{W_{ls}^*}{\partial \tau} = 0$. Thus, social welfare decreases (increases) with the proportion of informed consumers, i.e, the social media literacy level, in pooling (costly separating) equilibrium. In the costless separating equilibrium, social welfare doesn't change over the proportion of informed consumers. The intuition behind is that the higher proportion of informed consumers makes the L-type influencer buy more fake accounts in the pooling equilibrium as we state in **Proposition 2**, which further results in welfare loss. However, this pattern flips in the costly separating equilibrium.



Note: $\underline{v} = 10, \overline{v} = 20, \lambda = 0.9, \gamma = 0.9, q_H = 20, q_L = 10, C = 9, p = 0.24, \kappa = 0.8, \rho = 2.5.$

Figure 6: Impact of anti-fake-account Technology Level on Social Welfare

Overall, social media literacy education doesn't necessarily improve social welfare.



Note: $\underline{v} = 10, \overline{v} = 20, \lambda = 0.15, q_H = 20, q_L = 10, C = 5, p = 0.5, \kappa = 0.8, \rho = 2.5, d = 0.4, \tau = 0.5.$

Figure 7: Impact of Social Media Literacy Level on Social Welfare

Combining the above social welfare analysis, the formal conclusion is given in Proposition 3.

Proposition 3. *(a) The platform's anti-fake-account effort does not always improve social welfare. In particular, when the H-type influencer buys fake accounts defensively and the anti-fake-account technology level is high, the anti-fake-account effort improves social welfare, but in other cases, the detection harms social welfare. (b) Social welfare does not always increase with the anti-fake-account technology level. (c) Consumer's social media literacy education improves social welfare when H-type influencer buys fake accounts defensively but harms social welfare when L-type influencer buys fake accounts offensively.*

7 The Platform's Optimal Anti-fake-account Strategy

To understand whether we can rely entirely on the platform's anti-fake-account effort to tackle the fake-account problem, we analyze the platform's optimal anti-fake-account strategy to examine whether the platform is incentivized to implement the socially optimal anti-fake-account effort. The platform's profit in pooling, costly, and costless equilibrium is given by $\pi^{*p}_{plat}, \pi^{*s}_{plat}, \pi^{*ls}_{plat}$ respectively as described in the Appendix.

$$\pi^{*\{p,s,ls\}}_{plat} = p\pi^{*\{p,s,ls\}}_{plat,H} + (1-p)\pi^{*\{p,s,ls\}}_{plat,L} \tag{7.1}$$

As we learned from Proposition 1 and Figure 2, at a given point in the parameter space, the platform's anti-fake-account effort can affect the type of equilibrium. However, there is a parameter space in which a unique type of equilibrium remains regardless of the anti-fake-account effort, e.g., the bottom left (a costly separating equilibrium) or the bottom right (a pooling equilibrium) of Figure 2. For the parameter space where a unique type of equilibrium can hold, the platform only needs to decide the optimal anti-fake-account effort to maximize its profit in the corresponding equilibrium. For the parameter space where either of two, e.g., pooling and costly separating equilibrium, or any of the three equilibria can hold, the platform will globally select the optimal anti-fake-account effort to maximize its profit.

According to whether a particular type of equilibrium can hold or not, we divide the parameter space into the following regions: r_1= (only pooling equilibrium holds), r_2= (only costly separating equilibrium holds), r_3= (only costless separating equilibrium holds), r_4= (either costly or costless separating equilibrium holds), r_5= (either pooling or costly separating equilibrium holds), r_6= (either pooling or costless separating equilibrium holds), r_7= (any of pooling, costly and costless separating equilibrium could hold). In r_1, r_2 and r_3, the platform's profit dominates as long as the corresponding profit is greater than 0. In each of the remaining 4 regions, we need to investigate whether the platform's profit dominates or not, i.e., whether the platform's profit in one type of equilibrium is always higher than that in the other type(s). If the platform's profit in one type of equilibrium dominates in one region, then the optimal anti-fake-account effort in the corresponding equilibrium is its globally optimal strategy. While if no domination exists in a region, we may split the region further till the domination appears, which could be an endless process. The technical details about the domination in each region are discussed in the Appendix.

To maximize the platform's profit in any of the three equilibria, there is also a number of constraints to be satisfied that ensures the corresponding equilibrium holds. Thus, to find the platform's optimal anti-fake-account effort in one type of equilibrium as well as the corresponding socially optimal anti-fake-account effort is a constrained optimization problem.

The constrained optimization problem for the platform in the pooling equilibrium is

$$\begin{aligned} \max \quad & \pi^{*p}_{plat}(d) \\ \text{s.t.} \quad & \gamma \leq \eta_1;\ \lambda \geq \eta_2;\ \mu(\overline{v}, \underline{v}) > \eta_3 \end{aligned} \tag{7.2}$$

where η_1, η_2, η_3 are defined in Proposition 1. When we replace the above $\pi^{*p}_{plat}(d)$ with W^*_p but keep the constraints, we can also explore the socially optimal anti-fake-account effort.

Considering the complexity of parameter space, we employ a heuristic way to explore the solution. From region r_1, r_5, r_6, r_7, we iterate different combinations of the parameters with boundaries, e.g., γ, λ, τ, p, for those unbounded parameters, e.g, $q, c, \overline{v}$, we populate them with a set of numbers. We find that the platform, as well as the social planer's optimal anti-fake-account effort is always 0 in the pooling equilibrium as shown in Figure 8 (a). As we know, the platform's profit in the pooling equilibrium dominates in region $r_1$and r_6 (details in Appendix), thus, the optimal anti-fake-account effort is also the platform's globally optimal strategy. Although the parameter space is not fully covered, we can still produce some insights such as the platform could be totally inactive to fight fake-account problems but without harming social welfare. As a result, the L-type influencer buys fake accounts to pool with the H-type one.

Next, we look at the constrained optimization problem in the costly separating equilibrium, which is

$$\begin{aligned} \max \quad & \pi^{*s}_{plat}(d) \qquad\qquad (or \quad W^*_s) \\ \text{s.t.} \quad & \gamma \leq \eta_1;\ \mu(\overline{v}, \underline{v}) \leq \eta_3 \end{aligned} \tag{7.3}$$

We populate a set of parameters from region r_2 where the platform's profit in the separating equilibrium dominates, thus, at the given set of parameters, the optimal solution in the costly separating equilibrium is also the platform's globally optimal anti-fake-account effort. As shown in Figure 8 (a), the platform's optimal anti-fake-account effort could be either under or over the socially optimal anti-fake-account strategy. In addition, the two optimalities can be both 0 at the same time. The finding suggests that the platform could harm social welfare by exerting an over or under anti-fake-account effort, but it could also don't exert anti-fake-account effort at all. As a result, the H-type influencer buys fake accounts to separate themselves.

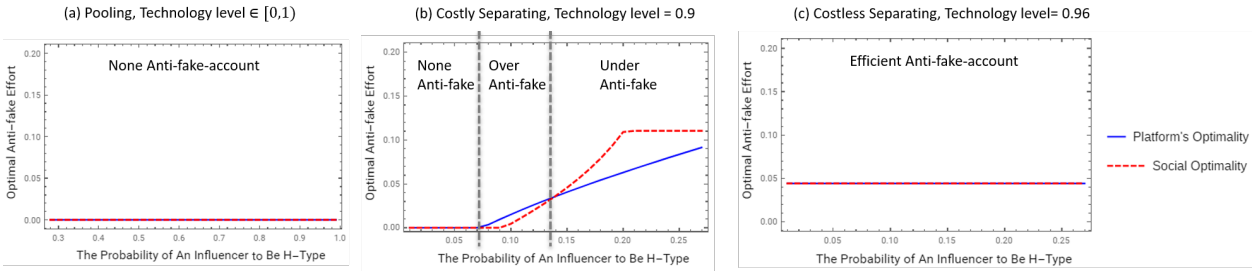
Finally, we solve the constrained optimization problem in the costless separating equilibrium as follows.

$$\begin{aligned} \max \quad & \pi^{*ls}_{plat}(d) \quad (or \quad W^*_{ls}) \\ \text{s.t.} \quad & \gamma > \eta_1 \end{aligned} \tag{7.4}$$

In region r_4 where both costly and costless separating equilibrium can hold, we iterate different sets of the parameters at which the platform's profit in the costless separating equilibrium dominates. Then we find that both the platform and the social planner can exert the same non-zero anti-fake-account effort as shown in Figure 8 (c). This finding suggests that under certain conditions, the platform could exert an efficient anti-fake-account effort to prevent both types of influencer from

buying fake accounts.

Above all, the platform could exert an under, over or efficient anti-fake-account effort under different conditions, therefore, we may not rely entirely on the social media platform to tackle the fake-account problem.



Note: $\underline{v} = 10, \overline{v} = 20, \lambda = 0.15, \gamma = 0.7, q_H = 20, q_L = 10, C = 5, \kappa = 0.8, \rho = 2.5.$

Figure 8: Anti-fake-account Effort: Platform's Optimality V.S. Socially Optimality

8 Model Extension

According to our observations from the real world, both H-type and L-type influencers could buy fake accounts at the same time on one platform. However, the prediction from our model demonstrates that when one type of influencer buys fake accounts, the other type doesn't buy. The reason why there is a gap between the model prediction and the real examples could be due to the number of influencer types is too limited in our model. To simplify the analysis, we only consider two types of influencers which could be too extreme to capture the real cases. Strictly speaking, the number of influencer types should be continuous. To judge whether an influencer is a H-type or $L-$type, the rule is relative rather than absolute, i.e., an influencer can be either a relatively high type one when compared with the extremely low type influencers or a relatively low type one when compared with the top influencers. Considering this situation, we extend our model by considering three types of influencers, low (L), medium (M), and high (H)-type.

By solving the signaling game with the three types of influencers, four equilibria arise which are shown in Table 2. See the proofs in Appendix.

Equilibrium	*High Type*	*Medium Type*	*Low Type*
Fully Separating	Buys	Buys	Doesn't buy
Fully Pooling	Doesn't buy	Buys	Buys
High type separated, but medium- and low type pool together	Buys	Buys	Buys
High- and medium type pool together, but the low-type separated	Buys	Buys	Buys

Table 2: Equilibria of Fake Accounts Purchasing Strategy for Three Types of Influencers

Similar to what we see in the game with two types, a fully separating with three types refers to

the case where the uninformed consumers and the advertisers see three different displayed number of consumers for each type, and a fully pooling means the displayed number for the three types of influencers are exactly the same to the uninformed consumers and the advertisers. In particular, there are two hybrid equilibria we don't see in the main model. 1) H-type influencer is separated by a different displayed number to the uninformed consumers and advertisers, but the M- and L-type pool together by showing the same displayed number. 2) The other one is that the H- and M-type pool together to the uninformed consumers and advertisers, but the L-type is separated.

From the results in Table 2, we can see in the fully separating equilibrium, both H-and M-types buy fake accounts but the L-type doesn't. Then, in the fully pooling equilibrium, the M- and L-type influencers buy fake accounts but the H-type doesn't. Interestingly, in the two hybrid equilibria, all of the three types of influencers buy fake accounts. This extended model demonstrates that different types of influencers may buy fake accounts at the same time, further, the concern about the above gap between the model prediction and the real world observations could have been addressed.

9 Discussion and Conclusion

In this paper, we study the mechanism why influencers buy fake accounts, how different types of influencers make fake accounts purchasing decisions under different conditions. We also explore how the amount of fake accounts in the platform changes over the key parameters with managerial implications. Then we examine the social welfare impact of fake accounts. Finally, we discuss whether the platform is incentivized to implement socially optimal anti-fake-account effort or not. We use the signaling game framework to analyze the influencers' strategy to influence consumers' consumption decisions. Our model tries to provide insights into the logic of fake accounts purchasing behavior, welfare, and policy implications.

The central contribution of the findings in this paper is the characterization of equilibrium outcomes. Under certain conditions, only the influencer with a low content quality buys fake accounts offensively to mimic the H-type one and mislead the uninformed consumers. However, the H-type influencer may also buy fake accounts defensively under certain conditions to separate them from the L-type one. In addition, it is also possible that no influencer buys fake accounts, usually when the platform conducts intensive anti-fake-account effort with a high technology level.

We also find that the platform's anti-fake-account effort can make L-type influencer buy more fake accounts in the pooling equilibrium while also make the H-type influencer buy fewer fake accounts in the costly separating equilibrium. As a result, a platform's anti-fake-account effort does not always improve social welfare.

For the platform's optimal anti-fake-account strategy, we may intuitively think that the platform will under detect the fake accounts when compared with the socially optimal one. However, we find that the platform can under, over, or efficiently exert an anti-fake-account effort. Thus, it would

be unwise to rely on the social media platform to fight the fake-account problem.

To explain why fake accounts are so prevalent on the social media platforms in the real world, we extend our model from two types of influencers to three types, then from hybrid equilibria arising in this extension, we further demonstrate that different types of influencers may buy fake accounts simultaneously, which is consistent with the real cases.

Finally, as we assume that a proportion of consumers (e.g., uninformed consumers) naively make subscription decisions according to their perceived popularity of the influencer, this is the main reason motivating the influencer to buy fake accounts because those consumers can generate advertising revenue. But we acknowledge that there could exist other motivations that are not incorporated into our model. In addition, we only consider the misleading effect of fake accounts on consumers. However, fake accounts may raise a harmful impact on consumers' user experiences like spamming, misinformation, privacy threats, and so on, which could further affect the consumers' consumption decisions. Limited by the complexity of the model, we don't consider such a polluting effect of fake accounts on consumers' utility. Those could be future work as an extension of our model.

References

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. DOI: 10.1093/qje/qjx044.Advance.

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. DOI: 10.2139/ssrn.3456139.

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. ISSN: 00222186. DOI: 10.1086/668835.

– (2014). "Finite optimal penalties for false advertising". In: *Journal of Industrial Economics* 62.4, pp. 661–681. ISSN: 14676451. DOI: 10.1111/joie.12064.

Guo, Xiaomeng and Baojun Jiang (2016). "Signaling Through Price and Quality to Consumers with Fairness Concerns". In: *Journal of Marketing Research* 53.6, pp. 988–1000.

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. DOI: 10.2139/ssrn.3039862.

Hjouji, Zakaria el et al. (2018). "The Impact of Bots on Opinions in Social Networks". URL: http://arxiv.org/abs/1810.12398.

Jiang, Baojun et al. (2016). "To Share or Not to Share: Demand Forecast Sharing in a Distribution Channel". In: *Marketing Science* 35.5, pp. 800–809.

Liu, De and Siva Viswanathan (2014). "Information asymmetry and hybrid advertising". In: *Journal of Marketing Research* 51.5, pp. 609–624. ISSN: 00222437. DOI: 10.1509/jmr.13.0074.

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. ISSN: 07322399. DOI: 10.1287/mksc.1050.0137.

Papanastasiou, Yiangos (2020). "Fake news propagation and detection: A sequential model". In: *Management Science* 66.5, pp. 1826–1846. ISSN: 15265501. DOI: 10.1287/mnsc.2019.3295.

Papanastasiou, Yiangos, Kostas Bimpikis, and Nicos Savva (2018). "Crowdsourcing Exploration". In: *Management Science* 64.4, pp. 1727–1746. ISSN: 15265501. DOI: 10.1287/mnsc.2016.2697.

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. ISSN: 15265501. DOI: 10.1287/mnsc.2019.3478.

Piccolo, Salvatore, Piero Tedeschi, and Giovanni Ursino (2018). "Deceptive advertising with rational buyers". In: *Management Science* 64.3, pp. 1291–1310. ISSN: 15265501. DOI: 10.1287/mnsc.2016.2665.

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.

Schmidt, William et al. (2015). "Signaling to Partially Informed Investors in the Newsvendor Model". In: *Production and Operations Management* 25.3, pp. 383–401.

Shao, Chengcheng et al. (2018). "The spread of low-credibility content by social bots". In: *Nature communications* 9.1, pp. 1–9. ISSN: 20411723. DOI: 10.1038/s41467-018-06930-7. arXiv: arXiv:1707.07592v1.

Spence, Michael (1978). "Job market signaling". In: *Uncertainty in economics*, pp. 281–306.

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.

Wilbur, Kenneth C. and Yi Zhu (2009). "Click fraud". In: *Marketing Science* 28.2, pp. 293–308. ISSN: 07322399. DOI: 10.1287/mksc.1080.0397.

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. ISBN: 9781450367479. DOI: 10.1145/3319535.3363198.

10 Appendix

Equilibrium Profit for the Advertiser and The Expected Revenue for the Influencer and the Platform

The equilibrium bid function is $b^*(v)$, and the indirect utility function of an advertiser (v) is

$$U^*(v) := U(b^*, v) = \varphi(b^*)V(v) - P(b^*) \tag{10.1}$$

where $\varphi(b^*)$ is the advertiser's winning probability and $P(b^*)$ is her expected payment.$V(v) = vn_r - n_f$ is the advertiser's valuation of the advertisement, where n_r and n_f are the number of real consumers and fake accounts respectively.We assume $V(v) \geq 0$ so that all advertisers are willing to participate the auction. Note that

$$\frac{dU^*(v)}{dv} = \frac{\partial U(b^*, v)}{\partial v} \mid_{b^*=V} + \frac{\partial U(v)}{\partial b^*} \frac{\partial b^*}{\partial v} \mid_{b^*=V} \tag{10.2}$$

As we know, the equilibrium bid function in the second price auction $b^* = V$ is the advertiser's optimal strategy, so $\frac{\partial U(v)}{\partial b^*} \mid_{b^*=V} = 0$ by applying the first order condition. Thus, $\frac{\partial U(v)}{\partial b^*} \frac{\partial b^*}{\partial v} \mid_{b^*=V} = 0$. Then we have

$$\frac{dU^*(v)}{dv} = \frac{\partial U(b^*, v)}{\partial v} \mid_{b^*=V} = \varphi(V)n_r = n_r G(v)^{m-1} \tag{10.3}$$

The advertiser with $\underline{v}$ is indifferent with advertising or not, so $U^*(\underline{v}) = 0$. Thus, the advertiser's (v) equilibrium expected payoff is

$$U^*(v) = n_r \int_{\underline{v}}^{v} G(x)^{m-1} dx \tag{10.4}$$

The expected revenue for the auctioneer (the platform and the influencer) $\prod(m)$ is total expected payment from all advertisers

$$\begin{aligned} \prod(m) &= mE[P^*(v)] = m \int_{\underline{v}}^{\overline{v}} P^*(v)g(v)dv \\ &= \int_{\underline{v}}^{\overline{v}} V(v)(\frac{dG(v)^m}{dv})dv - mn_r \int_{\underline{v}}^{\overline{v}} (\int_{\underline{v}}^{v} G(x)^{m-1}dx)(\frac{dG(v)}{dv})dv \\ &= \int_{\underline{v}}^{\overline{v}} \left(n_r[v - \frac{1 - G(v)}{g(v)}] - n_f \right) (\frac{dG(v)^m}{dv})dv \end{aligned} \tag{10.5}$$

Proof of Lemma 1.

By the end of **time 2**, the H-type (L-type) influencer has $n_1^H = n_0^H + x_H$ ($n_1^L = n_0^L + x_L$) followers. According to the definition of pooling equilibrium in this study, we have $n_1^H = n_1^L = n_1^{pool}$. Then at **time 3**, we identify one belief system for an uninformed consumer: if $n_1 < n_1^{pool}$, $p(H|n_1) = 0$; if $n_1 \geq n_1^{pool}$, $p(H|n_1) = p$.

To achieve a Perfect Bayesian Equilibrium (PBE), we must satisfy the Individual Rationality (IR) and Incentive Compatibility (IC) constraints for all players. It is easy to see that if the IR and IC constraints for the influencer can be met, those constraints can also be met for other players. Therefore, we only need to consider the IR and IC conditions for the influencer,i.e., there exists a PBE if and only if

$$\begin{cases} max_{n_1^H}\pi^p_{inf,H}(n_1^H \geq n_1^{pool}) \geq max_{n_1^H}\pi^p_{inf,H}(n_1^H < n_1^{pool}); \\ max_{n_1^L}\pi^p_{inf,L}(n_1^L \geq n_1^{pool}) \geq max_{n_1^L}\pi^p_{inf,L}(n_1^L < n_1^{pool}); \\ max_{n_1^H}\pi^p_{inf,H}(n_1^H \geq n_1^{pool}) \geq 0; \\ max_{n_1^L}\pi^p_{inf,L}(n_1^L \geq n_1^{pool}) \geq 0; \\ n_1^{pool} \geq n_0^H. \end{cases} \tag{10.6}$$

Solving the constraints inequations, we identify a range that the pooling equilibrium strategy n^1_{pool} must belong, which is

$$n_0^H \leq n_1^{pool} \leq \frac{1}{\lambda+\kappa+\phi_2 d}\{\lambda\mu(\overline{v},\underline{v})[p(n_0^H-n_0^L)+(\Delta n_p-\Delta n_L)]+\lambda[pn_0^H+(1-p)n_0^L)]+(\kappa+\phi_2 d)n_0^L\} \tag{10.7}$$

with the condition

$$\frac{1}{\lambda+\kappa+\phi_2 d}\{\lambda\mu(\overline{v},\underline{v})[p(n_0^H-n_0^L)+(\Delta n_p-\Delta n_L)]+\lambda[pn_0^H+(1-p)n_0^L)]+(\kappa+\phi_2 d)n_0^L\} \geq n_0^H \tag{10.8}$$

To ensure that all advertisers will participate as we assume, we also have $\underline{v} \geq v_0 = \frac{n_f}{n_r} = \frac{n_1^{pool}-(pn_0^H+(1-p)n_0^L)}{pn_0^H+(1-p)n_0^L+\Delta n_p}$.

We use $n_1^{pool*} \in [n_1^\alpha, n_1^\beta]$ to denote the H-type influencer's most profitable (unique) pooling equilibrium, i.e., $n_1^{pool*} = argmax_{n_1^{pool}}\lambda\{\mu(\overline{v},\underline{v})[pn_0^H+(1-p)n_0^L+\Delta n_p]-[n_1^{pool}-(pn_0^H+(1-p)n_0^L)]\}-(\kappa+\phi_2 d)(n_1^{pool}-n_0^H)$. And this n_1^{pool*} is also the L-type influencer's most profitable pooling equilibrium, i.e. $n_1^{H*} = n_1^{L*} = n_1^{pool*}$ It is easy to see that

$$n_1^{pool*} = n_1^\alpha = n_0^H \tag{10.9}$$

Thus, we know that the number of fake accounts purchased by the the two types of influencers in the most profitable pooling equilibrium are

$$\begin{cases} x^*_{H,pool} = n_1^{H*} - n_0^H = n_1^{pool*} - n_0^H = 0 \\ x^*_{L,pool} = n_1^{L*} - n_0^L = n_1^{pool*} - n_0^L = n_0^H - n_0^L \end{cases} \tag{10.10}$$

Proof of Lemma 2.

We have $n_1^H \neq n_1^L$ by the end of **time** 2, where $n_1^H = n_0^H + x_H$ and $n_1^L = n_0^L + x_L$. Then at **time 3**, we identify one belief system for an uninformed consumer:if $n_1 < n_1^{sep}$, $p(H|n_1) = 0$; if $n_1 \geq n_1^{sep}$, $p(H|n_1) = 1$. The following steps to solve the signaling game is similar to proof of **Lemma 1**.

Proof of Proposition 1.

We use *Lexicographically Maximum Sequential Equilibrium* (LMSE) method to select the unique equilibrium outcome from the most profitable pooling equilibrium, the most profitable costly separating, and costless separating equilibrium in Lemma 1,2,3 for the influencer's strategy. If both types of influencer can achieve the highest profit from the most profitable costly separating equilibrium than the other two, the unique equilibrium outcome that survives the LMSE is the most profitable costly separating equilibrium, while if both types of influencer can achieve higher profit from the other two equilibria, the unique LMSE outcome will be the corresponding one.

First, when the equilibrium is costless separating, we only need to compare the H-type influencer's profit between the pooling and costless equilibrium as the influencer's profit in costly separating equilibrium is 0.

By simplication, we can get $\pi_{inf,H}^{*ls} - \pi_{inf,H}^{*p} > 0$. In addition, we also have $\pi_{inf,L}^{*ls} - \pi_{inf,L}^{*p} = \frac{\lambda\mu(\overline{v},\underline{v})}{\kappa+\phi_2 d}(\pi_{inf,H}^{*ls} - \pi_{inf,H}^{*p}) > 0$. Therefore, the most profitable equilibrium is costless equilibrium when $\lambda\mu(\overline{v},\underline{v}) - \gamma(\kappa + \phi_2 d) < 0$.

Then, we compare the H-type influencer's profit between the pooling and costly separating equilibrium.

$\pi_{inf,H}^{*s} - \pi_{inf,H}^{*p} = \lambda\mu(\overline{v},\underline{v})(1 - \frac{c+\phi_1 d}{q_L}) + (\kappa+\phi_2 d)\gamma\frac{q_H - q_L}{q_H q_L}(c+\phi_1 d) - \lambda\{\mu(\overline{v},\underline{v})[\gamma(c+\phi_1 d)(\frac{1}{pq_H+(1-p)q_L} - \frac{p}{q_H} - \frac{1-p}{q_L}) + (1 - \frac{c+\phi_1 d}{pq_H+(1-p)q_L})] - (1-p)\gamma\frac{q_H-q_L}{q_H q_L}(c+\phi_1 d)\}$. In addition, the results of the comparison between the L-type influencer's profit between the pooling and the costly separating equilibrium should be consistent with that for the H-type influencer as

$$\pi_{inf,H}^{*s} - \pi_{inf,L}^{*s} = \pi_{inf,H}^{*p} - \pi_{inf,L}^{*p} = (\kappa + \phi_2 d)\frac{q_H - q_L}{q_H q_L}(c + \phi_1 d) \tag{10.11}$$

Thus, when $\pi_{inf,H}^{*s} - \pi_{inf,H}^{*p} \geq 0$, we also have $\pi_{inf,L}^{*s} - \pi_{inf,L}^{*p} \geq 0$, by incorporating the condition for the existence of costly separating equilibrium, $\lambda u(\overline{v},\underline{v}) - \gamma(\kappa + \phi_2 d) \geq 0$,then the unique LMSE outcome is the most profitable costly separating equilibrium.

And when $\pi_{inf,H}^{*s} - \pi_{inf,H}^{*p} < 0$, along with the condition for existence of pooling equilibrium, $[\lambda\mu(\overline{v},\underline{v})p + \lambda p - \lambda - \kappa - \phi_2 d]\gamma(\frac{1}{q_L} - \frac{1}{q_H}) + \lambda\mu(\overline{v},\underline{v})(1-\gamma)(\frac{1}{q_L} - \frac{1}{pq_H+(1-p)q_L}) \geq 0$, and the nonexistence condition of costless separating equilibrium, $\lambda\mu(\overline{v},\underline{v}) - \gamma(\kappa + \phi_2 d) \geq 0$ then the unique LMSE outcome is most profitable pooling equilibrium. The reason for incorporating the nonexistence condition of costless separating equilibrium is that pooling equilibrium is dominated by the costless separating equilibrium when it exists.

Details of Taking Derivatives of Variables

1) Taking Derivative of The Equilibrium Number of Fake Accounts with Resepect to anti-fake-account Effort

$\frac{\partial x^*_{L,pool}}{\partial d} = \gamma \frac{q_H - q_L}{q_H q_L}\phi_1 = \gamma \frac{q_H - q_L}{q_H q_L}\rho(1-\tau)$, as $q_H > q_L, \rho > 0, \gamma > 0$, and $1 > \tau \geq 0$, it is easy to see that $\frac{\partial x^*_{L,pool}}{\partial d} > 0$. Thus, $x^*_{L,pool}$ is monotonically increasing with d.

The analytical formula of $\frac{\partial x^*_{H,sep}}{\partial d}$ is complicated so that we are not able to tell its sign, i.e., whether it is greater than 0 or not. Thus, we employ an empirical method to discuss the possibilities. We iterate all combinations in the parameter spaces to find the maximum value of $\frac{\partial x^*_{H,sep}}{\partial d}$. If all found maximum values of $\frac{\partial x^*_{H,sep}}{\partial d}$ are negative, we can conclude $\frac{\partial x^*_{H,sep}}{\partial d} < 0$, i.e., $x^*_{H,sep}$ is monotonically decreasing with d. However, as long as we find at least two cases in which max value of $\frac{\partial x^*_{H,sep}}{\partial d}$ is positive and negative respectively, we can say that $\frac{\partial x^*_{H,sep}}{\partial d}$ can be either > 0 or < 0, i.e., $x^*_{H,sep}$ is not a monotonic function of d. By the empirical way, we know that the max value of $\frac{\partial x^*_{H,sep}}{\partial d}$ is positive (negative) when the anti-fake-account technology level is lower (higher). Thus, we can conclude that $x^*_{H,sep}$ is not a monotonic function of d. The following steps of taking derivatives are based on either way.

Proofs of Equilibria with Three Types of Influencer

The steps of solving signaling games are similar to that for the two-type case. Here we just specify the belief system and the results accordingly.

1) Fully Separating.

For fully separating, we have $n_1^H \neq n_1^M$, $n_1^H \neq n_1^L$, and $n_1^M \neq n_1^L$ at **time** 2, where $n_1^H = n_0^H + x_H$, $n_1^M = n_0^M + x_M$ and $n_1^L = n_0^L + x_L$. Then at **time 3**, we identify one belief system for an uninformed consumer:if $n_1 < n_1^{sepM}$, $p(H|n_1) = 0, p(M|n_1) = 0, p(L|n_1) = 1$; if $n_1^{sepM} \leq n_1 < n_1^{sepH}$, $p(H|n_1) = 0, p(M|n_1) = 1, p(L|n_1) = 0$; if $n_1 \geq n_1^{sepH}$, $p(H|n_1) = 1, p(M|n_1) = 0, p(L|n_1) = 0$.

Then, the numbers of fake accounts purchased by the three types of influencers in fully separating equilibrium are

$$\begin{cases} x^*_{H,sep} = \frac{\lambda\mu - \gamma(\kappa + \phi_2 d)}{\lambda + \kappa + \phi_2 d} \frac{q_H - q_L}{q_H q_L}(c + \phi_1 d) \\ x^*_{M,sep} = \frac{\lambda\mu - \gamma(\kappa + \phi_2 d)}{\lambda + \kappa + \phi_2 d} \frac{q_M - q_L}{q_M q_L}(c + \phi_1 d) \\ x^*_{L,sep} = 0 \end{cases} \tag{10.12}$$

2) *H*- type Separating, *M*- and *L*- types Pooling.

In this hybrid case, we have $n_1^H \neq n_1^M = n_1^L$ at **time 2**, where $n_1^H = n_0^H + x_H$, $n_1^M = n_0^M + x_M$ and $n_1^L = n_0^L + x_L$. Then at **time 3**, we identify one belief system for an uninformed consumer:if $n_1 > n_1^{pool}$, $p(H|n_1) = 1, p(M|n_1) = 0, p(L|n_1) = 0$; if $n_1 \leq n_1^{pool}$, $p(H|n_1) = 0, p(M|n_1) = p_M, p(L|n_1) = 1 - p_H - p_M$.

Then, the numbers of purchased fake accounts by the three types of influencers in this hybrid equilibrium are

$$\begin{cases} x^*_{H,sep} > n_1^{pool*} - n_0^H = \frac{\lambda(\mu+1)[n_0^H-(p_M n_0^M+p_L n_0^L)]+\lambda\mu(\Delta n_H-\Delta n_p)}{\lambda+\kappa+\phi_2 d} - (n_0^H - n_0^M) \\ x^*_{M,pool} = n_1^{pool*} - n_0^M = \frac{\lambda(\mu+1)[n_0^H-(p_M n_0^M+p_L n_0^L)]+\lambda\mu(\Delta n_H-\Delta n_p)}{\lambda+\kappa+\phi_2 d} \\ x^*_{L,pool} = \frac{\lambda(\mu+1)[n_0^H-(p_M n_0^M+p_L n_0^L)]+\lambda\mu(\Delta n_H-\Delta n_p)}{\lambda+\kappa+\phi_2 d} + (n_0^M - n_0^L) \end{cases} \tag{10.13}$$

3) H- and M- types pooling, L- type separating.

In this hybrid case, we have $n_1^H = n_1^M \neq n_1^L$ at **time 2**, where $n_1^H = n_0^H + x_H$, $n_1^M = n_0^M + x_M$ and $n_1^L = n_0^L + x_L$. Then at **time 3**, we identify one belief system for an uninformed consumer:if $n_1 < n_1^{pool}$, $p(H|n_1) = 0, p(M|n_1) = 0, p(L|n_1) = 1$; if $n_1 \geq n_1^{pool}$, $p(H|n_1) = p_H, p(M|n_1) = p_M, p(L|n_1) = 0$.

Then, the numbers of purchased fake accounts by the three types of influencers in this hybrid equilibrium are

$$\begin{cases} x^*_{H,pool} = n_1^{pool*} - n_0^H = \frac{\lambda\mu(p_H n_0^H+p_M n_0^M-n_0^L+\Delta n_p-\Delta n_L)+\lambda(p_H n_0^H+p_M n_0^M)+(\kappa+\phi_2 d)n_0^L}{\lambda+\kappa+\phi_2 d} - n_0^H \\ x^*_{M,pool} = n_1^{pool*} - n_0^M = \frac{\lambda\mu(p_H n_0^H+p_M n_0^M-n_0^L+\Delta n_p-\Delta n_L)+\lambda(p_H n_0^H+p_M n_0^M)+(\kappa+\phi_2 d)n_0^L}{\lambda+\kappa+\phi_2 d} - n_0^M \\ x^*_{L,sep} < n_1^{pool*} - n_0^L = \frac{\lambda\mu(p_H n_0^H+p_M n_0^M-n_0^L+\Delta n_p-\Delta n_L)+\lambda(p_H n_0^H+p_M n_0^M)+(\kappa+\phi_2 d)n_0^L}{\lambda+\kappa+\phi_2 d} - n_0^L \end{cases} \tag{10.14}$$

4) Fully Pooling.

In fully pooling, we have $n_1^H = n_1^M = n_1^L$ at **time 2**, where $n_1^H = n_0^H + x_H$, $n_1^M = n_0^M + x_M$ and $n_1^L = n_0^L + x_L$. Then at **time 3**, we identify one belief system for an uninformed consumer:if $n_1 < n_1^{pool}$, $p(H|n_1) = 0, p(M|n_1) = 0, p(L|n_1) = 1$; if $n_1 \geq n_1^{pool}$, $p(H|n_1) = p_H, p(M|n_1) = p_M, p(L|n_1) = 1 - p_H - p_M$.

Then, the numbers of purchased fake accounts by the three types of influencers in this hybrid equilibrium are

$$\begin{cases} x^*_{H,pool} = n_1^{pool*} - n_0^H = 0 \\ x^*_{M,pool} = n_1^{pool*} - n_0^M = n_0^H - n_0^M > 0 \\ x^*_{L,pool} = n_1^{pool*} - n_0^L = n_0^H - n_0^L > 0 \end{cases} \tag{10.15}$$