Journal of Machine Learning Research 21 (2020) 1-26 Submitted 6/17; Revised 7/19; Published 02/2

A model of fake data in data-driven analysis[*][†]

Xiaofan Li LI.X@UTEXAS.EDU
Andrew B. Whinston ABW@UTS.CC.UTEXAS.EDU
McCombs School of Business
The University of Texas at Austin
2110 Speedway
Austin, TX 78705, USA

Editor: Amos Storkey

Abstract

Data-driven analysis has been increasingly used in various decision making processes. With more sources, including reviews, news, and pictures, can now be used for data analysis, the authenticity of data sources is in doubt. While previous literature attempted to detect fake data piece by piece, in the current work, we try to capture the fake data sender's strategic behavior to detect the fake data source. Specifically, we model the tension between a data receiver who makes data-driven decisions and a fake data sender who benefits from misleading the receiver. We propose a potentially infinite horizon continuous time game-theoretic model with asymmetric information to capture the fact that the receiver does not initially know the existence of fake data and learns about it during the course of the game. We use point processes to model the data traffic, where each piece of data can occur at any discrete moment in a continuous time flow. We fully solve the model and employ numerical examples to illustrate the players' strategies and payoffs for insights. Specifically, our results show that maintaining some suspicion about the data sources and understanding that the sender can be strategic are very helpful to the data receiver. In addition, based on our model, we propose a methodology of detecting fake data that is complementary to the previous studies on this topic, which suggested various approaches on analyzing the data piece by piece. We show that after analyzing each piece of data, understanding a source by looking at the its whole history of pushing data can be helpful.

Keywords: data-driven analysis, fake data, game theory, point process

1. Introduction

With the growth of computational power and availability of data, we observe an increasing trend toward the adoption of data-driven approaches in various decision making processes such as quality of experience optimization (Jiang et al. (2017)) and media streaming (Zhang et al. (2005); Ganjam et al. (2015)). Compared with traditional techniques, data-driven approaches have shown potential advantages including efficiency, flexibility, robustness (Zhang et al. (2005)), and real-time agility (Jiang et al. (2017)).

[*]. This paper was presented at 2018 AEA Annual meeting in Philadelphia, in the session of Economics of News and Information. We are thankful for the comments of the participants and the chair of the session, Matthew Gentzkow. We take responsibility for all mistakes.

[†]. Declarations of interest: none

With the development of language, image, and video processing techniques, sources including reviews, photos, advertisements, and news can now be used as data for analysis (e.g.,Umbaugh (1997); Shi et al. (2016)). Recently, however, the authenticity of these sources has been a topical and controversial subject (Malbon (2013); Allcott and Gentzkow (2017)). Fake data senders can be incentivized by the benefit generated from misleading data receivers (Mayzlin et al. (2014); Luca and Zervas (2016)). Recent developments in adversarial machine learning also provide theories on and algorithms to produce fake data (e.g., Goodfellow et al. (2014); Springenberg (2015)). Without taking the possibilities of fake data into account, receivers who make decisions using data-driven approaches can suffer significantly.

There have been increasing numbers of studies on detecting fake data, especially fake news because of its threat to the whole of society. Most studies focus on two categories of features for each piece of news: its linguistic features and its network features. A news item's linguistic features include its lexical features, as in the "bag of words" approach, which identifies the usage of single significant words (e.g., Ott et al. (2013)), syntactic features (e.g., Oraby et al. (2017)), and semantic features (e.g., Rubin et al. (2016)). Network features apply to posts on social media such as Facebook and Twitter, to capture the posts' social features, such as the number of likes and re-tweets (e.g., Shu et al. (2018)). With such features, the studies run various classifiers, including Support Vector Machines (e.g., Zhang et al. (2012)) and Naïve Bayesian models (e.g, Oraby et al. (2017)), to classify each piece of news as fake or not. There are also studies that extract features and make classifications using deep neural networks (e.g., Ruchansky et al. (2017)). Besides directly proposing an algorithm of detecting fake data, there have also been discussions around how a platform, such as Facebook, should plan the detection efficiently from an optimal control perspective (Kim et al. (2018)).

To our knowledge, all the studies in this stream focus on identifying whether each single piece of data is fake or not, as the examples mentioned above. However, in many practical cases, whether a source rather than a specific piece of data is trustworthy is in the receiver's interest to decide. Identifying sources is different from identifying pieces of data, since trustworthy sources can make honest mistakes while misleading sources can also stream reliable data, such as the weather, to gain trust. In addition, even if one is interested in determining whether a specific piece of data is fake, knowing the history of its source can help. For example, a piece of news can be factually true but misleading because it is biased in terms of wording or selectively neglect some facts (Mullainathan and Shleifer (2002)). In such cases, one cannot conclude that the news is fake, even theoretically, by just looking at the piece itself. However, one can form an idea about it by studying the history of its source to see whether the source has a pattern of being systematically biased. Therefore, we propose a method to detect whether each source is trustworthy or not.

Analogous to the study of Antonelli et al. (2006), who propose a model of skin elasticity to detect whether a series of fingerprints are from fake fingers, we propose a game theoretic model to illustrate the behavior of a fake data sender (referred to as "sender" and "he" hereafter) who inserts fake data into a source and a data receiver (referred to as "receiver" and "she" hereafter) who decides whether to depend on this source to make data-driven decisions. We have a potentially infinite horizon continuous time model where the two agents mentioned above have asymmetric information such that the receiver does not initially know whether there is a fake data sender or not at the beginning and learns this information during the game. The receiver observes a stream of data from the source, which is a mixture of a stream of fake data from the sender and a stream of ordinary data from nature.[1] To model the real-time features of data-driven analysis, we assume that the data are observed piece by piece discretely but can be at any moment on the continuous time line (Kim et al. (2018)). Additionally, in our model, both the sender and the receiver are making decisions dynamically at every moment, contingent on what they have observed in the game. Specifically, the receiver decides whether to depend on this data stream, which, from the receiver's point of view, has some possibility of containing fake data. The sender decides the intensity of the fake data in the stream while facing a trade-off between an immediate gain from misleading the receiver and potential loss in the future due to the receiver's increasing suspicion.

To model the distinction between fake and ordinary data, we assume that each piece of data is a random variable and that the distribution of pieces of fake data is different from that of ordinary data. Following the spirit of generative models, from the perspective of the data generators, including the sender who generates fake data and nature who generates ordinary data, each piece of data is a draw from its distribution. The distribution of the fake data can either be manually specified by the sender or automatically generated with generative models developed following Goodfellow et al. (2014). From the receiver's perspective, she does not know the distributions of the data and can only use some degenerated distributions, based on her technology for detecting fake data, to approximate the generators' distributions. This technology can be any machine learning algorithm that characterizes each piece of data. The degenerated distributions of fake and ordinary data are then the distributions of their feature vectors generated from the algorithm. For example, if the receiver directly uses a fake data detections algorithm to detect each piece of data, she will obtain a binary observation on each piece of data: each piece is detected as either fake or true. Then, in this case, the receiver's degenerated distribution for fake data is a binary distribution, corresponding to her belief of the accuracy of the detection algorithm, that is, Pr(Detected as fake|fake) and Pr(Detected as true|fake). Similarly her degenerated distribution for ordinary data is also binary, characterized by Pr(Detected as fake|ordinary) and Pr(Detected as true|ordinary). Therefore, our model complements to the fake data detection literature mentioned above. If the receiver instead looks at some lower-level features of each piece of data, such as the linguistic and network features of news, she obtains higher-dimensional event spaces for her degenerated distributions. In our theoretical analysis in Sections 2 and 3, we allow all such possibilities by making no assumptions on the event spaces of the degenerated distributions. In the numerical illustrations in Sections 4 to 6, we assume the degenerated distributions to be binary, which is a practical method for the receiver and also graphically provides the most insights.

In the case in which a piece of data has strictly positive likelihoods in both the fake and ordinary data generation distributions, it is theoretically impossible for any fake data detection algorithm that is based on this single piece of data to be perfectly accurate. The biased news example above is such a case. Practically, even without this theoretical gap, no

[1] In our analysis, we consider the stream of ordinary data exogenous. However, even if whether including ordinary data can be decided by the sender, it is of the sender's interest to include such data in the stream to gain the receiver's trust.

detection algorithm can reach perfect accuracy. Therefore, the degenerated distributions will not be trivial. Our method fills this theoretical gap as well as overcomes the practical limitations by modeling the sender's behavior to incorporate information from his history.

Although Goodfellow et al. (2014) argues that, theoretically, one can generate data that follows the same distribution as ordinary data, this is not of interest to the sender, since such fake data will be useful instead of misleading to the receiver. However, although the sender wants the distributions of fake and ordinary data to be different from his perspective, he would like the degenerated distributions to be as close as possible from the receiver's perspective, to make the fake data harder for the receiver to detect.

We first assume that the degenerated distributions of both fake and ordinary data are invariant throughout the game and known by the receiver beforehand. A practical interpretation is that the receiver knows the distributions from previous experience, such as training. Later, we relax this assumption to analyze the case in which the receiver has inaccurate information about the distributions, which can happen if the sender is using a new technology to generate fake data while the receiver is still using the old training samples. In both cases, after each time the receiver observes a piece of data, she learns whether there is a sender through Bayesian updates of her belief of whether there is a sender. Since such learning is based on the degenerated distributions, the generators' distributions are actually irrelevant to the game. Therefore, in the remainder of the paper, distributions refer to the degenerated ones unless otherwise specified.

We prove the existence and uniqueness of a Markov Perfect Equilibrium (MPE) (Maskin and Tirole (2001)), which is a Perfect Bayesian Equilibrium where the players' strategies are Markovian. In our model, the state variable is the receiver's belief of whether a sender of fake data exists, since the receiver's payoff depends on this state variable while the sender has perfect information about this variable. We then use numerical examples to illustrate the players' strategies and payoffs to gain insights. On the sender's side, we show that the sender's optimal strategy is to be careful and not too aggressive to prevent the receiver from abandoning this source. In addition, we compare the sender's payoffs when the receiver anticipates the distribution of fake data correctly and incorrectly to show the sender's benefit from improving his technology for generating fake data. On the receiver's side, we show that, whether the receiver knows the distributions or not, the most important thing for her is to have at least some initial suspicion about the source. We also show the receiver's benefit from having a better understanding of the distributions, although, in practice, the receiver needs to balance this benefit with the cost of learning the distributions from new samples.

Apart from the data-driven and adversarial machine learning literature mentioned above, closely related literature also includes research on communication and deception games. In communication theories, the information the sender provides to the receiver can be non-strategic truth, strategically selected truth or fake information (e.g., Kamenica and Gentzkow (2011); Anderson and Smith (2013)). Compared with other models in this literature, we generalize them by proposing a dynamic model where the cost of sending fake information comes from the increasing suspicion of the receiver. Literature on deception games (e.g., Anderson and Smith (2013)) models how the receiver updates her belief based on her observations, and that the sender can deceive due to stochasticity. We generalize the model to fit our data-driven analysis settings, where each piece of information is observed discretely and can follow arbitrary distributions.

2. Model Development

To model the data traffic and the interaction between the sender and the receiver, we propose a game-theoretic model based on point processes. In this section, we first provide a brief overview of point processes and then specify the players' information and payoff structures in our model.

2.1. Point Processes

A point process is a type of stochastic process that models the occurrence of events as a series of random points in time or geographic space (Xu et al. (2014)). For example, in the context of this study, the receiver's observation of a piece of data can be modeled as a point occurring along the time line. We can describe such a point process by $N(t)$, which is an increasing non-negative integer-valued counting process such that $N(t_2) - N(t_1)$ is the total number of points that occurred within the time interval $(t_1, t_2]$. Most point processes can be fully characterized by the *conditional intensity function* defined as follows (Daley and Vere-Jones (2007)):

$$\lambda(t|\mathcal{H}_t) = \lim_{\Delta_t \to 0} \frac{\Pr(N(t+\Delta t) - N(t) > 0|\mathcal{H}_t)}{\Delta t} \tag{1}$$

where $\mathcal{H}_t$ is the *history* up to time instant t, which includes all the information before t. The intensity measures the probability of instantaneous point occurrence given the previous history. Specifically, given the history $\mathcal{H}_t$, the probability of a point occurring within $(t, t+\Delta t]$ is $\lambda(t|\mathcal{H}_t)\Delta t$.

It is worth noting that the commonly used Poisson processes can be seen as a special kind of point processes where the intensity $\lambda(t|\mathcal{H}_t)$ is independent of the history $\mathcal{H}_t$. If the intensity $\lambda(t|\mathcal{H}_t)$ is constant over the whole process, then the point process reduces to a homogeneous Poisson process, and, if the intensity $\lambda(t|\mathcal{H}_t)$ is not constant but can be a deterministic function of time t and independent of the history, then the point process reduces to a nonhomogeneous Poisson process.

2.2. Players' Information Structure

In the game we are modeling, there are two players, the sender and the receiver, and there is information asymmetry in the game: the sender has perfect information about the receiver whereas the receiver does not know whether there is a sender or not. Figure 1 is an illustration of their information structure, which we will discuss in detail in this section.

In the game, the receiver observes incoming data traffic, that is a mixture of a stream of fake data from the sender and a stream of ordinary data. To model the data traffic, we need to model two characteristics of each piece of data: the timing and the content.

We use two point processes to model the time of occurrence of each piece of fake data and ordinary data. Specifically, we use $N_a(t)$ to denote the stochastic process that counts the number of pieces of fake data up to time t and, similarly, use $N_0(t)$ to denote the count

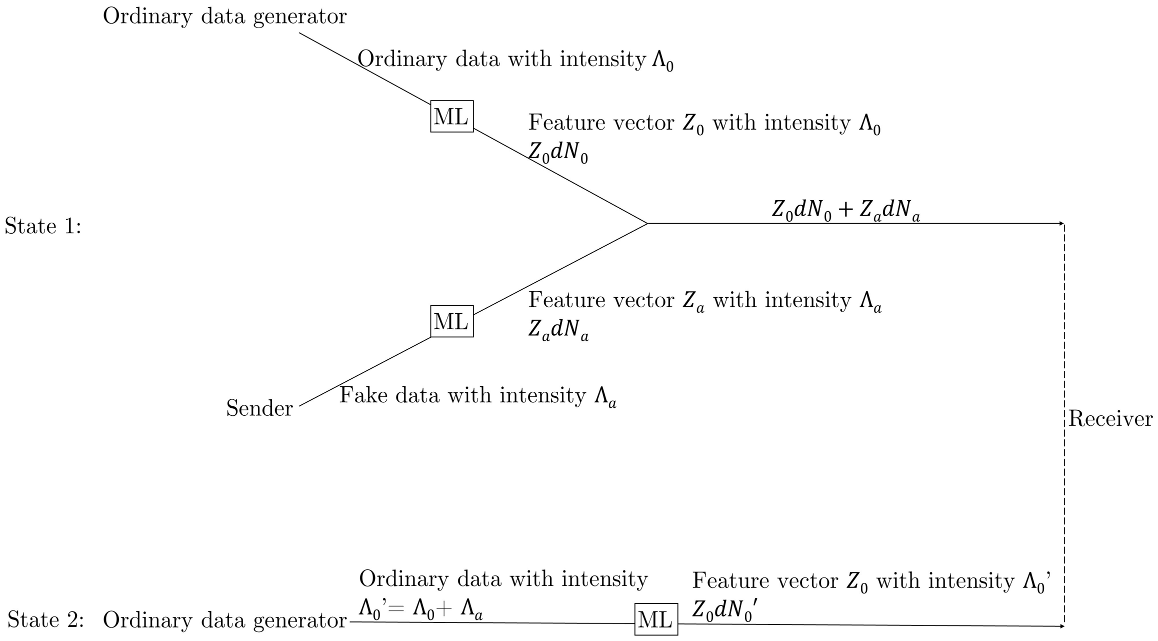
Ordinary data generator
Ordinary data with intensity Λ_0
ML
Feature vector Z_0 with intensity Λ_0
$Z_0 dN_0$
State 1:
$Z_0 dN_0 + Z_a dN_a$
ML
Feature vector Z_a with intensity Λ_a
$Z_a dN_a$
Sender
Fake data with intensity Λ_a
Receiver
Ordinary data with intensity
$\Lambda_0' = \Lambda_0 + \Lambda_a$
Feature vector Z_0 with intensity Λ_0'
$Z_0 dN_0'$
State 2: Ordinary data generator
ML

Figure 1: An illustration of the game

of ordinary data. Following (1), denote $\Lambda_0(t|\mathcal{H}_t)$ and $\Lambda_a(t|\mathcal{H}_t)$ as the conditional intensity functions of N_0 and N_a, respectively. We assume that $\Lambda_a(t|\mathcal{H}_t)$, the intensity of fake data, is decided by the sender, whereas $\Lambda_0(t|\mathcal{H}_t)$, the intensity of ordinary data, is exogenous and constant for the entire process, which implies that N_0 is a Poisson process. For simplicity of exposition, until further specification, Λ_a is used to denote the t dependent function, $\Lambda_a(t|\mathcal{H}_t)$, and Λ_0 is used to denote the constant that $\Lambda_0(t|\mathcal{H}_t)$ equals to.

For the content, we assume that the receiver uses a machine learning algorithm (denoted as ML in Figure 1) to characterize each piece of data. This algorithm transforms a piece of data into a feature vector. We use two random variables to model the feature vector of each piece of fake data and ordinary data, denoted as $Z_0, Z_a : \Omega \to \mathbb{R}$. The event space Ω is determined by the algorithm: it is binary when the algorithm is directly a piece-by-piece fake data detection algorithm and could comprise of tens of dimensions or even more when the algorithm summarizes the data's lower level features. In the present model, we do not go into the receiver's decision model for the machine learning algorithm, and thus assume Ω to be exogenous. Denote the distributions of Z_0, Z_a to be P_0, P_a respectively. The smaller the difference between P_0 and P_a, the better the technology the sender has, because it is harder for the receiver to see the difference of a piece of ordinary data and a piece of fake data. We assume P_0 and P_a to be invariant throughout the game. We first consider P_0 and P_a to be exogenous and then numerically compare results for different P_a's for insights. To summarize, the receiver observes the incoming data traffic as a stochastic process Y subject to

$$dY = Z_0 dN_0 + Z_a dN_a \tag{2}$$

However, given this observation of Y, the receiver does not know whether there is a sender or not. Formally speaking, there are two states of the world: in State 1, there is a sender and Y comes from (2). In State 2, there is no sender and Y is subject to $dY = Z_0 dN_0'$, where N_0' is a point process whose intensity Λ_0' satisfies $\Lambda_0' = \Lambda_0 + \Lambda_a$. At each time t, the receiver has a belief $q(t) \in [0,1]$ that the world is in State 1 and a belief $1 - q(t)$ that the world is in State 2. The belief q is Bayesian updated through the observation of Y as follows[2]:

$$dq = \frac{q(1-q)(\frac{\Lambda_a P_a(dY) + \Lambda_0 P_0(dY)}{\Lambda_0 + \Lambda_a} - P_0(dY))}{q \cdot \frac{\Lambda_a P_a(dY) + \Lambda_0 P_0(dY)}{\Lambda_0 + \Lambda_a} + (1-q) P_0(dY)} \tag{3}$$

When proposing (3), we are making three assumptions. First, the receiver is learning whether there is a sender from and only from the content of the data, which means that the timing of data is not informative to the receiver. We assume that $\Lambda_0' = \Lambda_0 + \Lambda_a$ in State 2 to make sure that the timing of data in both states is expected by the receiver to be the same, which makes the timing does not include information of the existence of the sender. Second, we assume that the receiver knows P_0 and P_a beforehand, which could come from training. In the latter part of our paper, we relax this assumption to analyze and compare the case in which the receiver does not have accurate information about the distributions. Third, we assume that the receiver correctly anticipates Λ_0 and Λ_a, conditional on the world in State 1, which is a mandatory assumption for Nash equilibrium. We will discuss our solution concept in detail in Section 3.

The information is asymmetric in the game. Although the receiver does not know whether the world is in State 1 or State 2, the sender knows about his own existence, meaning that the world is in State 1. Besides the timing and content of the fake data he sends, we assume that the sender also knows the timing and content of ordinary data, which is natural. We also assume that the sender knows the receiver's initial belief, $q(t_0)$. Then from (3), the sender knows the receiver's belief, $q(t)$, at any moment t.

2.3. Players' Payoff Structure

With the information structure above, we next model the decision variables and payoff structure for both players.

First, it is worth noting that the receiver decides whether to use the focal data source for data-driven analysis instead of picking and using the pieces of data from this source that appears authentic. With the random variable characterization of both ordinary and fake data, using data selectively will significantly undermine the effect of the data-driven analysis, because substantial selection bias would be introduced. For example, even when the receiver believes that most of the ordinary reviews should be positive and the fake ones are more likely to be negative, ignoring all the negative reviews will make the analysis pointless.

Thus, the receiver needs to decide whether to abandon this data source, fully depend on it, or partly depend on it. Denote $p(t) \in [0,1]$ as the receiver's dependence on the focal data source at time t, which is her decision variable. Normalize the receiver's payoff from

[2] For a proof, see the Appendix

outside options, such as using other data sources or abandoning data-driven analysis, as zero and her gain from depending on a piece of ordinary data as a unit. On the other hand, we assume that the receiver suffers a loss L when she depends on a piece of fake data. With the belief process $q(t)$ that the world is in State 1, by dynamically choosing $p(t)$ based on the history before t, the receiver maximizes her expected payoff, *i.e.*,

$$\begin{aligned} \max_{p_{t=t_0}^{\infty}} \quad & E[\int_{t_0}^{\infty} p((1-q)(\Lambda_0+\Lambda_a)+q(\Lambda_0-L\Lambda_a))dt] \\ \text{s.t.} \quad & p\in[0,1] \quad \forall t\in[t_0,\infty) \end{aligned} \tag{4}$$

where p, q, Λ_a are functions of t.

The sender's goal is to mislead the receiver. Therefore, his payoff depends on how much fake data is consumed by the receiver and how much the receiver depends on such fake data. By dynamically choosing the fake data intensity $\Lambda_a(t)$, the sender maximizes his expected payoff, *i.e.*,

$$\begin{aligned} \max_{\Lambda_{a}{}_{t=t_0}^{\infty}} \quad & E[\int_{t_0}^{\infty} e^{-r(t-t_0)}p\Lambda_a dt] \\ \text{s.t.} \quad & \Lambda_a\in[0,c] \quad \forall t\in[t_0,\infty) \end{aligned} \tag{5}$$

where r is the time discount factor and p, Λ_a are functions of t. We did not assume a time discount factor for the receiver, and we explain why in the equilibrium analysis. We assume that the capacity of the sender is c, that is, $\Lambda_a(t)\in[0,c]$ for all t.

3. Equilibrium Analysis

In the present research, we focus on the MPE of the game, which is a Perfect Bayesian equilibrium where the players' strategies are Markovian. In our model, the state variable is the receiver's belief of whether a sender of fake data exists, since the receiver's payoff depends on this state variable while the sender has perfect information about this variable. In other words, we focus on the equilibrium where, at each time t, both players' strategies depend on t only through the receiver's current belief $q(t)$, that is, $p(t)$ and $\Lambda_a(t)$, can be written as deterministic functions of $q(t)$. In the remainder of the paper, p and Λ_a denotes $p(q(t))$ and $\Lambda_a(q(t))$ unless further specification.

To simplify the notation, define function $g(\cdot; q, \Lambda_a)$ as

$$g(\cdot;q,\Lambda_a)=\frac{q(1-q)(\frac{\Lambda_a P_a(\cdot)+\Lambda_0 P_0(\cdot)}{\Lambda_0+\Lambda_a}-P_0(\cdot))}{q\cdot\frac{\Lambda_a P_a(\cdot)+\Lambda_0 P_0(\cdot)}{\Lambda_0+\Lambda_a}+(1-q)P_0(\cdot)}$$

Then, equation (3) can be rewritten as

$$dq=g(dY;q,\Lambda_a)$$

Give the properties of point processes, we know that, in any time period dt, the likelihood of $dN_0+dN_a\geq 2$ is $O(dt^2)$. Therefore, with likelihood $1-O(dt^2)$, dN_0 and dN_a are either 0 or 1 and are not both 1. Therefore,

$$dq=g(Z_0;q,\Lambda_a)dN_0+g(Z_a;q,\Lambda_a)dN_a \tag{6}$$

Now, we start analyzing the equilibrium strategies of the players.

Because both players' strategies depend only on q and equation (6) suggests that the receiver's strategy p will not influence the evolution of q, the receiver's current decision will not have any impact in the future. Therefore, the receiver's dynamic optimization problem is equivalent to optimization at each static time point t. This is why we did not assume a time discount factor for the receiver, since it does not affect her strategy. With this argument, the receiver's optimal strategy p for problem (4) can thus be characterized by

$$\begin{aligned} p \in \arg\max_{p} \quad & p((1-q)(\Lambda_0+\Lambda_a)+q(\Lambda_0-L\Lambda_a)) \\ \text{s.t.} \quad & p \in [0,1] \end{aligned} \tag{7}$$

for each $q \in [0,1]$, where Λ_a is a function of q.

For the sender's dynamic optimization problem (5), we define a value function V and derive the associated Hamilton-Jacobi-Bellman (HJB) equation, whose solution suggests the sender's optimal strategy.

Definition 1 *The value function V is defined as the maximum of the expected value of the sender's payoff from state $q \in [0,1]$ to the final state,* i.e.,

$$\begin{aligned} V(q) \doteq \max_{\Lambda_a(q(t))_{t=t_0}^{\infty}} \quad & E[\int_{t_0}^{\infty} e^{-r(t-t_0)}p(q(t))\Lambda_a(q(t))dt|q(t_0)=q] \\ \text{s.t.} \quad & \Lambda_a(q(t)) \in [0,c] \quad \forall t \in [t_0,\infty) \end{aligned} \tag{8}$$

We apply Bellman's principle of optimality (Bellman and Kalaba (1965)) to the above definition and obtain that

$$V(q(t_0)) = \max_{\Lambda_a(q(t_0))} E[e^{-r(t_0+dt-t_0)}V(q(t_0+dt)) + e^{-r(t_0-t_0)}p(q(t_0))\Lambda_a(q(t_0))dt] \tag{9}$$

With $e^{-r(t_0+dt-t_0)} = e^{-rdt} = 1-rdt$ and $e^{-r(t_0-t_0)} = 1$, (9) leads to

$$V(q(t_0)) = \max_{\Lambda_a(q(t_0))} E[(1-rdt)V(q(t_0+dt)) + p(q(t_0))\Lambda_a(q(t_0))dt] \tag{10}$$

Due to the Markov property of the state variable $q(t)$, (10) can be reorganized as

$$E[rdtV(q(t_0+dt))] = \max_{\Lambda_a(q(t_0))} \{E[V(q(t_0+dt)) - V(q(t_0))] + p(q(t_0))\Lambda_a(q(t_0))dt\} \tag{11}$$

With $E[V(q(t_0+dt))] = V(q(t_0+dt))$ and $q(t_0) = q$ by definition (8), differentiate both sides of (11) with respect to t and omit the higher order infinitesimal on the left hand side, we obtain the HJB equation:

$$rV(q) = \max_{\Lambda_a(q)} \{E[\frac{dV(q)}{dt}|_{t=t_0}] + p(q)\Lambda_a(q)\}, \tag{12}$$

with $V(1) = 0$ as the the terminal condition. This terminal condition comes from the following argument: When $q(t_0) = 1$, $dq = 0$ because of equation (3). Therefore the receiver holds the belief of $q(t) = 1$ for all t, meaning that the receiver is always sure that there is a fake data sender. In this case, it is natural to assume that the sender cannot mislead the receiver and therefore gets zero payoff.

For simplicity of exposition, we use V, p and Λ_a to denote $V(q), p(q)$ and $\Lambda_a(q)$. We also use V' to denote $V'(q)$, which is the value of $\frac{dV}{dq}$ at point q. With equation (6), we obtain

$$\begin{aligned} E[\frac{dV(q)}{dt}|_{t=t_0}] &= E[\frac{dV}{dq}|_q \cdot \frac{dq}{dt}|_{t=t_0}] \\ &= V'E[g(Z_0; q, \Lambda_a)\frac{dN_0}{dt}|_{t=t_0} + g(Z_a; q, \Lambda_a)\frac{dN_a}{dt}|_{t=t_0}] \\ &= V' \cdot (E[g(Z_0; q, \Lambda_a)]\Lambda_0 + E[g(Z_a; q, \Lambda_a)]\Lambda_a) \end{aligned} \tag{13}$$

Take (13) into (12), we have

$$rV = \max_{\Lambda_a} p\Lambda_a + V' \cdot (E[g(Z_0; q, \Lambda_a)]\Lambda_0 + E[g(Z_a; q, \Lambda_a)]\Lambda_a), \tag{14}$$

which can be separated into two conditions:

$$\begin{aligned} &\Lambda_a \in \arg\max_{\Lambda_a} p\Lambda_a + V' \cdot (E[g(Z_0; q, \Lambda_a)]\Lambda_0 + E[g(Z_a; q, \Lambda_a)]\Lambda_a) \\ &\qquad \text{s.t.} \quad \Lambda_a \in [0, c] \end{aligned} \tag{15}$$

and

$$rV = p\Lambda_a + V' \cdot (E[g(Z_0; q, \Lambda_a)]\Lambda_0 + E[g(Z_a; q, \Lambda_a)]\Lambda_a), \tag{16}$$

where $V(1) = 0$.

Conditions (15) and (16) fully characterize the sender's optimal strategy. Thus, in all, a MPE is a 3-tuple (p, V, Λ_a), where each entry is a function of q, such that conditions (7),(15),(16) are satisfied.

Next, we discuss the solution to these conditions in the following proposition and its proof.

Theorem 2 *There exists a unique MPE.*

Proof We first provide a 3-tuple (p, V, Λ_a) such that (7), (15), (16) are satisfied, then show its uniqueness.

First consider condition (7). When the receiver's expected payoff, $(1 - q)(\Lambda_0 + \Lambda_a) + q(\Lambda_0 - L\Lambda_a)$, is positive, the receiver's only optimal strategy is $p = 1$; when it is negative, her only optimal strategy is $p = 0$; when it equals to 0, her optimal strategy is any $p \in [0, 1]$. Therefore, condition (7) is equivalent to

$$p \begin{cases} = 1, \text{when} \quad (1 + L)q - 1 \le 0 \quad \text{or} \quad \Lambda_a < \frac{\Lambda_0}{(1+L)q-1} \\ \in [0, 1], \text{when} \quad (1 + L)q - 1 > 0 \quad \text{and} \quad \Lambda_a = \frac{\Lambda_0}{(1+L)q-1} \\ = 0, \text{when} \quad (1 + L)q - 1 > 0 \quad \text{and} \quad \Lambda_a > \frac{\Lambda_0}{(1+L)q-1} \end{cases} \tag{17}$$

It is easy to verify that a pair of (p, Λ_a) that satisfy the following conditions (18) and (19) will suffice both (17) and the condition that $\Lambda_a \in [0, c]$ for all $q \in [0, 1]$.

$$p \begin{cases} = 1, \text{when } q \le \frac{c+\Lambda_0}{c(1+L)} \\ \in (0, 1), \text{when } q > \frac{c+\Lambda_0}{c(1+L)} \end{cases} \tag{18}$$

$$\Lambda_a = \begin{cases} c, \text{when } q \le \frac{c+\Lambda_0}{c(1+L)} \\ \frac{\Lambda_0}{(1+L)q-1}, \text{when } q > \frac{c+\Lambda_0}{c(1+L)} \end{cases} \tag{19}$$

Next, we try to find a pair of (p, V), with Λ_a fully characterized by (19), that can suffice (15), (16) and (18). Consider the first order condition of (15). When Λ_a is not at the border, i.e., $q > \frac{c+\Lambda_0}{c(1+L)}$, we have

$$p + V' \cdot \frac{d(V'(E[g(Z_0; q, \Lambda_a)]\Lambda_0 + E[g(Z_a; q, \Lambda_a)]\Lambda_a))}{d\Lambda_a} = 0 \tag{20}$$

Taking (19) into (20), we have

$$p + V' \cdot \int_\Omega a(x)dx = 0, \tag{21}$$

where

$$a(x) = \frac{(P_0(x) - P_a(x))(-1+q)(P_a^2(x)q + P_0^2(x)(-1+q+Lq)^2 + P_0(x)P_a(x)(-1+q+2Lq))}{(LP_0(x) + P_a(x))^2 q}$$

Taking (19), (21) into (16), we have

$$rV = V'\Lambda_0 \cdot \int_\Omega s(x)dx, \tag{22}$$

where

$$s(x) = \frac{(P_0(x) - P_a(x))(-1+q)(P_a(x) + P_0(x)(-1+q+Lq))}{(LP_0(x) + P_a(x))(-1+q+Lq)} - \frac{a(x)}{(1+L)q-1}$$

Solving (22), we have

$$V = Ce^{F(q)}, \tag{23}$$

where $F(\cdot)$ is an antiderivative of $\frac{r}{\Lambda_0 \cdot \int_\Omega s(x)dx}$ and C is any constant.

Taking (23) into (18),(21), we have

$$p = \begin{cases} 1, \text{when } q \le \frac{c+\Lambda_0}{c(1+L)} \\ -\frac{r}{\Lambda_0 \cdot \int_\Omega s(x)dx} \cdot Ce^{F(q)} \cdot \int_\Omega a(x)dx, \text{when } q > \frac{c+\Lambda_0}{c(1+L)}, \end{cases} \tag{24}$$

where C is selected such that p is continuous at $q = \frac{c+\Lambda_0}{c(1+L)}$. With this C, (23) and (24) fully characterize V and p.

Following the derivations above, the 3-tuple (p, V, Λ_a) characterized by (23), (24) and (19) suffice the conditions (7), (15) and (16).

Next, we show the uniqueness by exploiting the zero-sum feature of the game. Here, zero-sum does not mean that the aggregated payoff of the players is literally zero, but that the sender benefits from misleading the receiver, which provides the receiver disutility. This feature leads to the fact that any pair of strategies of the players leads to a Pareto efficient allocation, where neither player can be strictly better off while leaving the other player not worse off.

If the MPE of the game is not unique, there will exist two different 3-tuple, (p, V, Λ_a) and (p^*, V^*, Λ_a^*). Consider a given initial belief q. Since p is the receiver's best response to Λ_a, she does weakly better at (p, Λ_a) than at (p^*, Λ_a), thus the sender does weakly worse at (p, Λ_a) than at (p^*, Λ_a). Again, because Λ_a^* is the sender's best response to p^*, he does weakly better at (p^*, Λ_a^*) than at (p^*, Λ_a). Combining these two arguments, the sender does weakly better at (p^*, Λ_a^*) than at (p, Λ_a), i.e., $V^*(q) \geq V(q)$. However, our starting point is arbitrary, $V^*(q) \leq V(q)$ can therefore be derived with the same logic. Therefore, $V^*(q) = V(q)$. Since these arguments hold for any initial belief q, we obtain $V^* = V$. Conditions (16) and (20) show that (p, Λ_a) are uniquely determined by V and its derivative. Therefore, $(p, V, \Lambda_a) = (p^*, V^*, \Lambda_a^*)$, contradicting the assumption at the beginning.

■

4. An Illustrative Example

In this section, to illustrate and analyze the equilibrium strategies, we use an example where both distributions of Z_0 and Z_a, P_0 and P_a, are binary distributions. This is the case when the receiver uses the results of a fake data detection algorithm as her only observations. Other than this practical interpretation and computational convenience, another advantage of choosing binary distributions is that they make it easier to measure the efficiency of the sender's technology for producing fake data. Specifically, assume that there are two states in the event space of Z_0 and Z_a, Ω, denoted State M and State N. Z_0 is realized as State M with probability p_0 and as State N with probability $1 - p_0$; Z_a is realized as State M with probability p_a and as State N with probability $1 - p_a$. The efficiency of the sender's technology can be measured by the difference between P_0 and P_a, which, in this case, reduces to $|p_0 - p_a|$. When $|p_0 - p_a| = 0$, the sender's technology is flawless and the receiver has no way to distinguish between ordinary and fake data, which makes our research trivial. Thus, we study the case where $|p_0 - p_a| \neq 0$.

Our focus in this section is to illustrate both players' strategies and payoffs and understand how the sender's technology will influence them. First, we are interested in how the game evolves with time and we obtain the following results when both players are playing their equilibrium strategies.

Theorem 3 *When $0 < q < 1$, $E[dq/dt] > 0$. Therefore, there are only two stable states: $q = 0$ and $q = 1$. If the receiver's initial belief $q_0 > 0$, $q \to 1$ when $t \to \infty$.*

Proof See the Appendix. ■

From Theorem 3, we observe that if the receiver's initial belief is strictly between 0 and 1, then this game is like a Ponzi Scheme: At the end of the game, the receiver learns that there is a fake data sender; however, she lost value through the learning process, which can be seen as the cost of learning, and the sender also gains through the game.

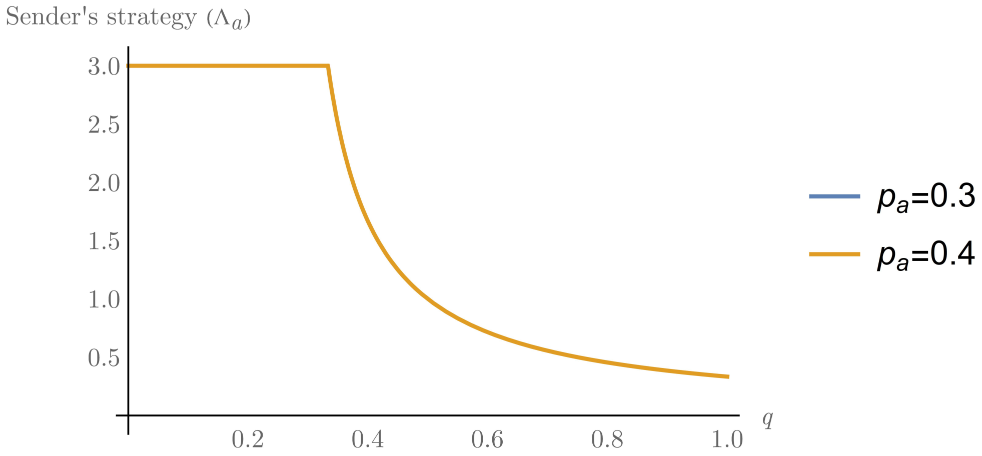

Figure 2: Comparison of the sender's strategies

To visualize both players' strategies and payoffs, we use a numerical example. Because we are interested in the influence of the sender's technology, we set parameters including L, c, r, Λ_0, p_0 as fixed and analyze the change of p_a. Specifically, we set $L = 3, c = 3, r = 0.1, \Lambda_0 = 1, p_0 = 0.1$, and compare the strategies and payoffs between cases where $p_a = 0.3$ and $p_a = 0.4$.

Figure 2 depicts the sender's strategies in the two cases. In both cases, the sender fully employs his capability when the receiver's belief is below a threshold, and his attacking frequency starts declining after a threshold with the increase in the belief. The intuition is that, when the receiver is more suspicious of the data source, the sender needs to be more cautious to prevent the receiver from abandoning the source. The sender's strategies fully overlap in the two cases, suggesting that the strategies remain the same for different technologies. This is because given the receiver's belief, her optimization problem (7) is independent of the sender's technology, which makes the intensity the sender requires to keep the receiver using the source independent of the sender's technology.

Figure 3 depicts the receiver's strategies. With a lower belief, implying greater trust in the data source, the receiver will fully depend on the source. As this belief increases, the receiver will rely less on the source. The dependence converges to 0 as the belief converges to 1. It is worth noting that the receiver's strategies *converge* to 0 continuously instead of jumping to 0 after a threshold because, in the equilibrium, the sender also carefully manages his intensity to ensure that he is not being too aggressive, which provides the receiver incentive to keep using the data source even if she has a significant belief that there exists a fake data sender. Because in such cases, the sender's attacking intensity is low enough to make the receiver still expect a benefit from the source. Another observation from Figure 3 is that, when the sender has better technology ($p_a = 0.3$), the receiver will be depend less on the data source, given the same belief.

Figure 4 depicts the sender's value function $V(q)$, which can be interpreted as the sender's expected payoff throughout the game when the receiver's initial belief $q(t_0) = q$. It is straightforward that V is monotonically decreasing with q because the more trust the receiver has in the beginning, the more room there is for the sender to extract profit.

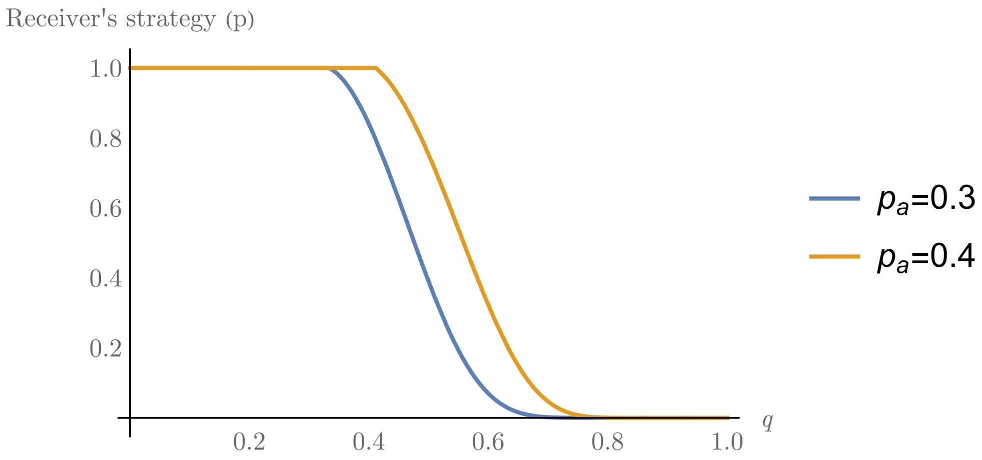

Figure 3: Comparison of the receiver's strategies

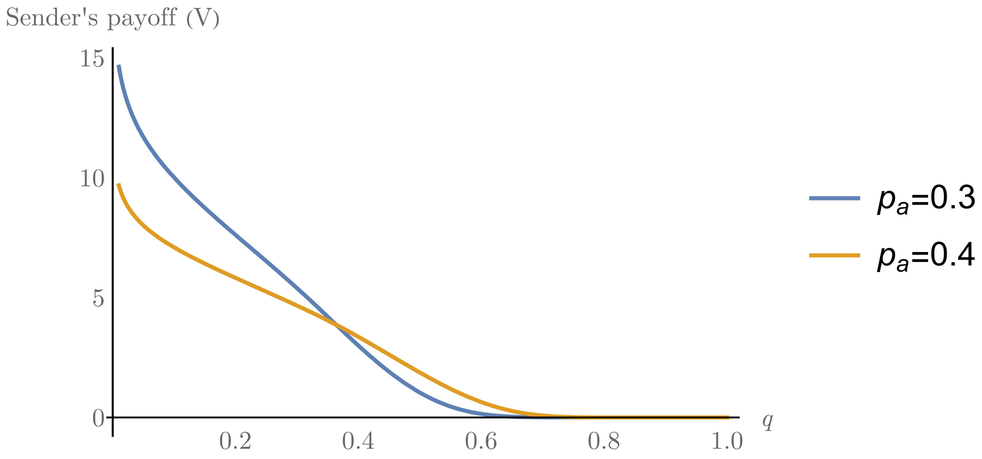

Figure 4: Comparison of the sender's payoffs

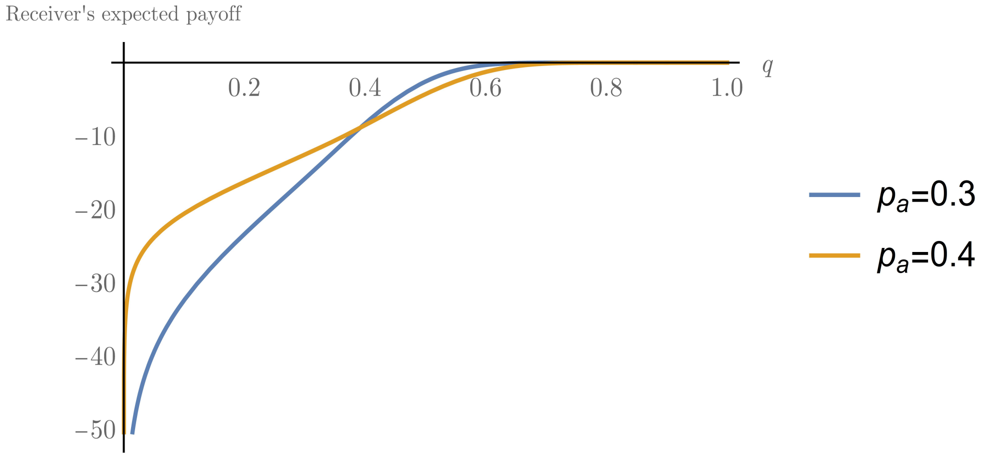

Figure 5: Receiver's expected payoffs

Comparing the two cases in which the sender's technology differs, we observe that, if the receiver's initial belief is low, the sender is better off with superior technology; meanwhile, if the initial belief of the receiver is high, the sender is better off with inferior technology. The intuition here is that having superior technology has two effects: First, the difference between the distributions of ordinary data and fake data is smaller, and, therefore, the receiver will update her belief more slowly. Second, as shown in Figure 3, the receiver will be more cautious and depend on the data source less, especially when her belief is high. When the initial belief of the receiver is lower, there is a longer period in terms of belief when the receiver's strategies—given the sender's different technologies—do not diverge, such that the first effect dominates the second effect. On the other hand, when the receiver's initial belief is relatively high, the second effect will dominate the first. This contrast between these two effects leads to the comparison of sender payoff, shown in Figure 4.

With Theorem 3, the receiver's expected payoff (4) can be calculated (for details, see the Appendix) and is shown in Figure 5.

When there is a fake data sender, it is no surprise that the receiver needs to incur more costs of learning the existence of the sender if she starts with a lower initial belief, and this cost converges to infinity as the initial belief approaches 0. Therefore, having an initial belief that is strictly greater than 0 can significantly reduce this cost. On the other hand, from Figure 3, we see that the receiver's strategy holds the same when her belief is under a threshold, implying that, when there is no fake data sender, having a reasonably low initial belief that is different from 0 will not hurt the receiver at all. Taking both possibilities into account, it is suggested that, when dealing with data, one should have some degree of suspicion (a reasonably low, but strictly positive, initial belief) in case there is a fake data sender.

5. Off-equilibrium Analysis

In the previous section, we illustrated equilibrium strategies and payoffs, assuming that the receiver knows accurately the distributions from training. However, in practice, this assumption could be too restrictive, especially when the sender is also keep improving his technology, such that the old training samples cannot work perfectly. In this section, we relax this assumption and examine a case in which the receiver does not perfectly anticipate the sender's technology, to gain more insights.

In this case, following the same arguments as before, we continue to assume Z_0 and Z_a to be binary distributions, and maintain the previous notations p_0 and p_a to model the distributions of Z_0 and Z_a, respectively. However, in this case, the receiver anticipates the sender's technology, p_a, to be p'_a, which is not equal to p_a. Therefore, instead of using the equilibrium strategy where the technology is p_a, the receiver uses the equilibrium strategy where the technology is p'_a, which is suboptimal in this case. Additionally, the receiver updates belief q based on her incorrect anticipation, meaning that, in function $g(\cdot)$, the parameter p_a is substituted by p'_a and Λ_a is substituted by the receiver's anticipation of the sender's attacking intensity, which is the sender's equilibrium strategy when his technology is p_a. Assume that the sender knows how the receiver estimates his technology and therefore updates his strategy to be optimal given the receiver's strategy. In other words, in this case, equation (7) does not hold because the receiver is not optimizing, whereas conditions (15) and (16) hold given a certain p, which is the receiver's equilibrium strategy with technology p'_a. Following the derivations in the proof of Theorem 2, V and Λ_a can be solved with (15) and (16).

We say that the receiver *underestimates* the sender's technology if $p_0 < p_a < p'_a$ or $p'_a < p_a < p_0$. The larger $(p_0 - p'_a)/(p_0 - p_a)$ is, the *more underestimated* we say the sender is. Practically, this underestimation is likely to be due to the sender's improvement in technology.

Analogous to Theorem 3, we are still interested in how the game evolves in time. However, the result is different for the following property.

Theorem 4 *Assume that $Lc > \Lambda_0$. If the sender is underestimated, then there exists a unique $q_e \in (0,1)$, such that:*
i) when $0 < q < q_e$, $E[dq/dt] > 0$
ii) when $q = q_e$, $E[dq/dt] = 0$
iii) when $q_e < q < 1$, $E[dq/dt] < 0$
Therefore, besides $q = 0$ and $q = 1$, q_e is another stable state. If $0 < q(t_0) < 1$, as $t \to \infty$, q will fluctuates around q_e and the sender's intensity will fluctuate around $\Lambda_a(q_e)$. Specifically, $\Lambda_a(q_e)$ satisfies:
i) $\Lambda_a(q_e) = \frac{p_0 - p'_a}{Lp_0 + p'_a - p_a - Lp_a}\Lambda_0$, if $\frac{p_0 - p'_a}{p_0 - p_a} < \frac{c}{c+\Lambda_0}(1+L)$.
ii) $\Lambda_a(q_e) = c$, if $\frac{p_0 - p'_a}{p_0 - p_a} \geq \frac{c}{c+\Lambda_0}(1+L)$

Proof See the Appendix. ■

The assumption at the beginning is trivial: when $Lc < \Lambda_0$, there is actually no game because the sender is not capable enough to hurt the receiver as much as her gain, so the receiver will always be fully dependent on the data source and still benefit even when the data are polluted.

From Theorem 4, as $t \to \infty$, q does not converge to 1, therefore the receiver's expected payoff at each time does not converge to 0, and, so, we cannot use equation (4) to evaluate her expected payoff throughout the game. Instead, we use the intensity of fake data in the stream in the long term to evaluate the receiver's payoff. If the receiver's initial belief is $q(t_0) = 1$, she will abandon the source at all times; and, if the receiver's initial belief is $q(t_0) = 0$, she will be attacked with the sender's capacity, c, at all times. Other than those two trivial cases, in the long term, the fake data intensity will fluctuate around $\Lambda_a(q_e)$. Since $(p_0 - p'_a)/(p_0 - p_a)$ characterizes how the sender is underestimated, we see that as long as the receiver does not underestimate the sender too much, the receiver will end up with $\Lambda_a(q_e) = \frac{p_0 - p'_a}{Lp_0 + p'_a - p_a - Lp_a}$, which is easily shown to be smaller than c. Actually, in practical cases such as the following numerical example, this difference is significant. Therefore, following the same arguments as in the equilibrium case, the receiver is suggested to have a low but non-zero initial belief because this will not hurt the receiver at all if there is no sender but will benefit the receiver significantly if there is a sender. In addition, we obtain the following result:

Theorem 5 *If the receiver has an initial belief* $0 < q(t_0) < 1$ *and underestimates the sender's technology, she will receive more fake data in the long run if the sender is more underestimated. Formally,* $\Lambda_a(q_e)$ *is an increasing function with respect to* $(p_0 - p'_a)/(p_0 - p_a)$.

Proof See the Appendix. ■

This result shows that it is beneficial for the receiver to have a better understanding of the sender's technology, even with some upfront cost, because it will benefit her payoff in the long term.

In this section, we still use a numerical example to show the effects of underestimating the sender's technology. Specifically, we keep assuming $L = 3, c = 3, r = 0.1, \Lambda_0 = 1, p_0 = 0.1$ while setting $p_a = 0.3$ and $p'_a = 0.4$.

Figure 6 depicts the sender's strategy when the receiver underestimates his technology. It is no surprise that, under the same belief, the sender is attacking with a higher intensity than in the equilibrium case, because he benefits more from the receiver's greater dependence on the data source than in the equilibrium strategy (as shown in Figure 3).

Figure 7 depicts the comparison between the sender's payoff in the off-equilibrium case in which $p_a = 0.3$ and $p'_a = 0.4$ and in those two equilibrium cases where $p_a = 0.3, 0.4$. First, the sender is strictly better off than the case in which equilibrium strategies are implemented as $p_a = 0.4$. The intuition is straightforward that the receiver is using the same strategy in these two cases while the sender has superior technology. The comparison between the off-equilibrium case and the $p_a = 0.3$ equilibrium case is more complicated. In this comparison, the sender's technologies are the same, so the difference is due to the receiver's underestimation of the sender's technology.

This underestimation has two effects: First, as shown in Figure 2, the receiver has greater dependence when her belief is high. Second, the Bayes's Rule suggests that, for each piece of suspicious data that is realized as State M, the receiver will update her belief

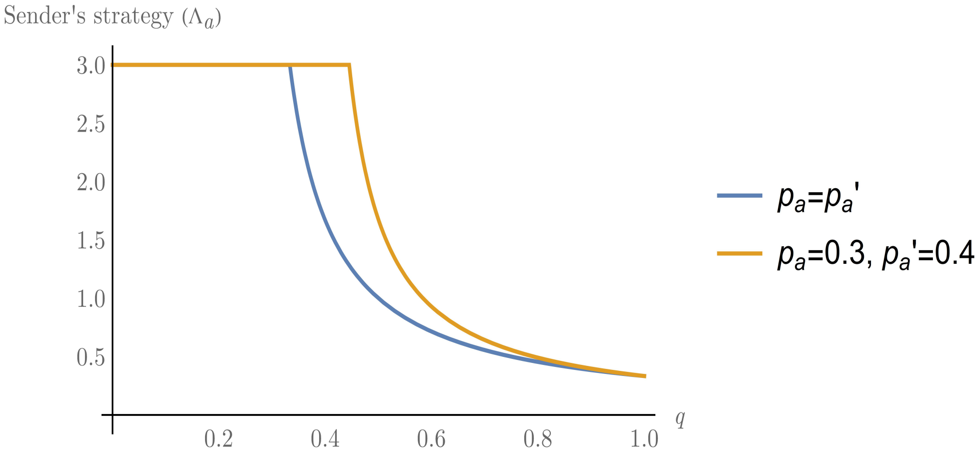

Figure 6: Comparison of the sender's equilibrium strategy and strategy when the receiver underestimates the sender's technology

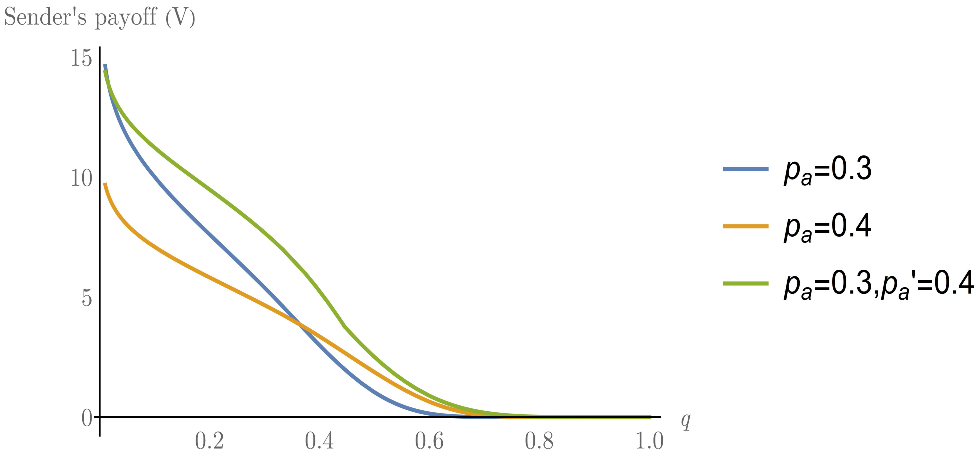

Figure 7: Comparison of the sender's payoffs when the receiver underestimates the sender's technology

more toward a sender existing when she underestimates the sender's technology. Therefore, when the receiver's belief is relatively low, such that she is fully dependent on the source, this underestimation will make her belief update more quickly and, thus, reduce the sender's payoff. Aggregating these two effects leads to what we observe in Figure 6: if the receiver's initial belief is high, the sender is better off with this underestimation; however, when the initial belief gets lower, this difference shrinks.

Because improving technology to generate fake data is costly for the sender, this result provides insight for the sender in deciding how much effort to exert on sharpening his technology. In the cases where the receiver's initial belief is either low or the belief is high, improving the technology will not give the sender a big boost in payoff; however, when the belief is moderate, the sender might want to spend more effort on improving his technology.

6. Simulation

In this section, we illustrate the fact that one cannot be perfectly accurate in detecting fake data by looking at each piece of data separately when the distributions of true and fake data have the same support, meaning that the data have strictly positive likelihoods in the distributions of both the true and fake data . We then use simulation data to show that our method addresses this problem.

6.1. Detection of the Data Piece by Piece

Denote the feature vector of the piece of data of interest as x. Assume that it has strictly positive likelihoods in the distributions of both the true and fake data, i.e., $P_a(x) > 0$ and $P_0(x) > 0$. Denote the prior belief that this piece of data is fake as p. Then the posterior belief after analyzing this piece of data will be $\frac{pP_a(x)}{pP_a(x)+(1-p)P_0(x)}$, which is not surprisingly strictly between 0 and 1, meaning that such an algorithm cannot be perfectly accurate in determining whether this piece of data is fake or not.

6.2. Detection of the Source

In our model, we illustrated an approach to how the receiver updates her belief about the source, based on her understanding of the sender's strategic behavior. As Theorem 3 suggests, with this approach, if a sender of fake data exists, then the receiver's belief of this converges to 1 with time. However, if the receiver is naïve in terms of not recognizing that the sender has a consistent strategy, this belief will not converge to 1, meaning that the receiver will not be able to detect the fake source with certainty. We use simulation data to show this difference by comparing a strategic receiver following our approach and a naïve receiver. Both types of receiver apply the same machine learning algorithm to obtain the realization x of each piece of data. However, the strategic receiver will update her belief based on her understanding of the sender's strategy, which is to be more careful as the receiver becomes suspicious, as shown in Figure 2. The naïve receiver will ignore this fact that the sender can be strategic and thus update her belief by assuming the sender is always at his full capacity.

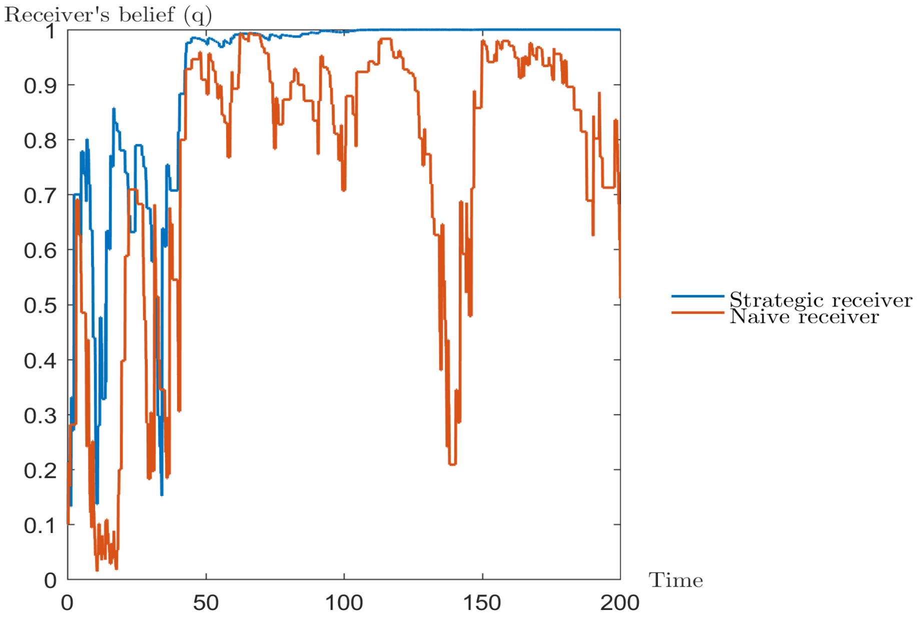

Figure 8: Comparison of the belief process of a strategic receiver and a naïve receiver

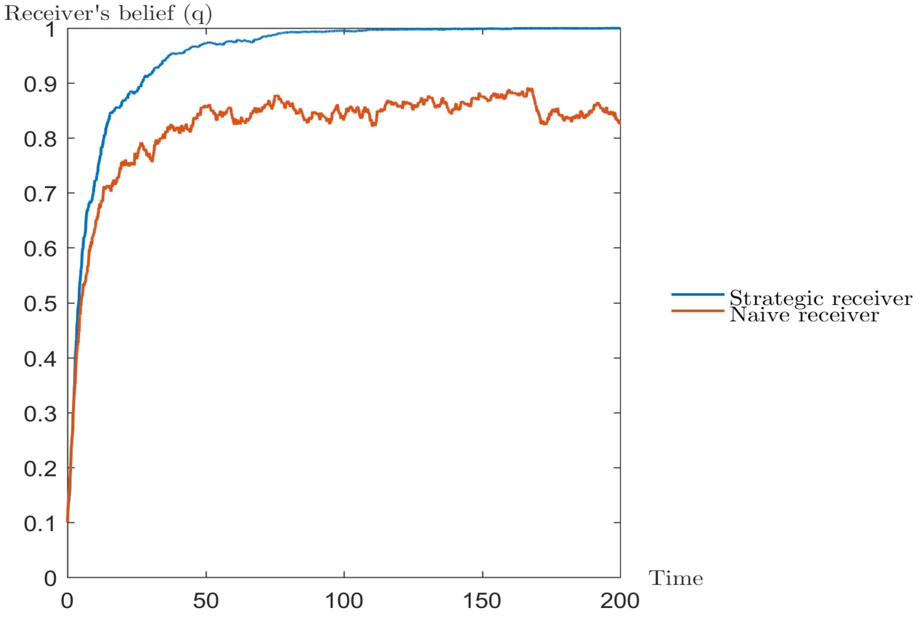

Figure 9: Comparison of the mean of 100 samples of the belief process of a strategic receiver and a naïve receiver

To generate simulation data, we use our model and set the parameters $L = 3, c = 3, r = 0.1, \Lambda_0 = 1, q(t_0) = 0.1, p_0 = 0.1, p_a = 0.4$. We look at the how the belief of both types' receivers' is updated in the first 200 units of time. An example is shown in Figure 8.

We can see that the naïve receiver's belief fluctuates significantly while the strategic receiver's belief converges to 1, which is the ground truth. This is because that when the sender becomes careful, the naïve receiver will start to think the sender is innocent whereas the strategic receiver will not, because the latter knows this is part of the sender's strategy.

We generate the process above 100 times. Figure 9 shows the mean of the belief processes. Even when the fluctuation is smoothed out by averaging, we can see that a naïve receiver's belief will not converge to 1.

These results show that, in the case in which the distributions of true and fake data have the same support, so that the sender can potentially apply such a strategy we modeled, it is important to recognize this strategy and be a strategic receiver, besides detecting each piece of data separately.

7. Conclusions and Future Directions

In the present research, we propose a game-theoretic model based on point processes to model data traffic that potentially contains fake data and show insights for both the sender and the receiver of fake data. Specifically, we show that, for the receiver, having some initial degree of suspicion of the data source and understanding that the sender can be strategic are very helpful, even when the receiver does not accurately anticipate the fake data sender's technological capabilities. In our observation, current data-driven research is ignorant of the potential existence of fake data, possibly because their data sources are reasonably reliable. However, with more and more sources—such as reviews, news and photos—being analyzed as data these days and people relying increasingly heavily on data, there could be more fake data senders in the future because they can see the profitability. Therefore, in light of the results of this research, we suggest remaining suspicious of the source when making data-driven decisions. Additionally, from the sender's strategy we find from our model, we propose an approach of detecting fake data source, which complements the piece-by-piece fake data detection algorithms in previous studies.

Discussions remain open in many aspects around this issue. First, the timing of the fake data can be more strategic. For example, the sender could decide to push a large amount of fake data at once to influence some significant decisions by the receiver, and the receiver could also learn about the existence of a sender from the timing of the data she receives. Second, practically speaking, the receiver could decide the event space of each piece of data, Ω, by choosing different machine learning algorithms to check them. An Ω with a higher dimension means a higher cost for the receiver but also allows the receiver to gain a better understanding on whether there is fake data. How to balance this trade-off is also an important issue for the receiver. Third, the fake data sender could improve his technology for generating fake data through the game and the receiver could also gain a better understanding of the characteristics of fake data through observation. It will be interesting to incorporate these possibilities to see how both sides should exert effort on dynamically improving themselves.

Appendix

Proof of Equation 3

Proof

$$\begin{aligned} dq &= q(t+dt) - q(t) \\ &= Pr[\text{State } 1|dY] - q \\ &= \frac{Pr[dY|\text{State } 1]Pr[\text{State } 1]}{Pr[dY|\text{State } 1]Pr[\text{State } 1] + Pr[dY|\text{State } 2]Pr[\text{State } 2]} - q \\ &= \frac{(\Lambda_a P_a(dY) + \Lambda_0 P_0(dY))q}{(\Lambda_a P_a(dY) + \Lambda_0 P_0(dY))q + (\Lambda_a + \Lambda_0)P_0(dY)(1-q)} - q \\ &= \frac{(\Lambda_a P_a(dY) + \Lambda_0 P_0(dY))(q-q^2) - (\Lambda_a + \Lambda_0)P_0(dY)(1-q)q}{(\Lambda_a P_a(dY) + \Lambda_0 P_0(dY))q + (\Lambda_a + \Lambda_0)P_0(dY)(1-q)} \\ &= \frac{q(1-q)(\frac{\Lambda_a P_a(dY) + \Lambda_0 P_0(dY)}{\Lambda_0 + \Lambda_a} - P_0(dY))}{q \cdot \frac{\Lambda_a P_a(dY) + \Lambda_0 P_0(dY)}{\Lambda_0 + \Lambda_a} + (1-q)P_0(dY)} \end{aligned}$$

■

Proof of Theorem 3

Proof First, for any $q \in [0,1]$, we have

$$\begin{aligned} E[dq/dt] &= E[(g(Z_0; q, \Lambda_a)dN_0 + g(Z_a; q, \Lambda_a)dN_a)/dt] \\ &= \Lambda_0 E[g(Z_0; q, \Lambda_a)] + \Lambda_a E[g(Z_a; q, \Lambda_a)] \\ &= \Lambda_0(p_0 \cdot g(M; q, \Lambda_a) + (1-p_0) \cdot g(N; q, \Lambda_a)) + \Lambda_a(p_a \cdot g(M; q, \Lambda_a) + (1-p_a) \cdot g(N; q, \Lambda_a)) \end{aligned} \tag{25}$$

Calculating $g(M; q, \Lambda_a)$, we have

$$\begin{aligned} g(M; q, \Lambda_a) &= \frac{q(1-q)(\frac{\Lambda_a P_a(M) + \Lambda_0 P_0(M)}{\Lambda_0 + \Lambda_a} - P_0(M))}{q \cdot \frac{\Lambda_a P_a(M) + \Lambda_0 P_0(M)}{\Lambda_0 + \Lambda_a} + (1-q)P_0(M)} \\ &= \frac{q(1-q)(\frac{\Lambda_a p_a + \Lambda_0 p_0}{\Lambda_0 + \Lambda_a} - p_0)}{q \cdot \frac{\Lambda_a p_a + \Lambda_0 p_0}{\Lambda_0 + \Lambda_a} + (1-q)p_0} \end{aligned} \tag{26}$$

Similarly, we have

$$g(N; q, \Lambda_a) = \frac{q(1-q)(\frac{\Lambda_a(1-p_a) + \Lambda_0(1-p_0)}{\Lambda_0 + \Lambda_a} - (1-p_0))}{q \cdot \frac{\Lambda_a(1-p_a) + \Lambda_0(1-p_0)}{\Lambda_0 + \Lambda_a} + (1-q)(1-p_0)} \tag{27}$$

Taking (26), (27) into (25), after simplification, we have

$$E[dq/dt] = \frac{(p_0 - p_a)^2(1-q)^2 q \Lambda_a^2(\Lambda_0 + \Lambda_a)}{((1-p_0)\Lambda_0 + (1 - p_0 + p_0 q - p_a q)\Lambda_a)(p_a q \Lambda_0 + p_0(\Lambda_0 + \Lambda_a - q\Lambda_a))} \tag{28}$$

On the right hand side, the only term that is not straightforwardly larger than 0 is $1-p_0+p_0q-p_aq$. If $p_0>p_a$, $1-p_0+p_0q-p_aq \geq 1-p_0>0$; if $p_0<p_a$, $1-p_0+p_0q-p_aq \geq 1-p_0+p_0-p_a=1-p_a>0$. Therefore, we have $E[dq/dt]>0$, when $0<q<1$.

■

Calculation of the receiver's payoff in the equilibrium case

With Theorem 2, we know that if $0<q(t_0)<1$,

$$\begin{aligned} E[\int_0^\infty p((1-q)(\Lambda_0+\Lambda_a)+q(\Lambda_0-L\Lambda_a))dt] &= E[\int_{q(t_0)}^1 p((1-q)(\Lambda_0+\Lambda_a)+q(\Lambda_0-L\Lambda_a))\frac{dq}{dt}dt] \\ &= E[\frac{dq}{dt}]\int_{q(t_0)}^1 p((1-q)(\Lambda_0+\Lambda_a)+q(\Lambda_0-L\Lambda_a))dt \end{aligned}$$

So, with (28), the receiver's expected payoff can be rewritten as a deterministic form, which then can be numerically calculated.

Proof of Theorem 4

Proof Same as the proof of Theorem 3, we still have the equation (25). However, in this case, the receiver is updating her belief based on her wrong anticipation, meaning that in function $g(\cdot)$, p_a is substituted by p'_a, and Λ_a follows (19). Then follow the same steps as in the proof of Theorem 3 and after simplification, we have

$$E[dq/dt]=\begin{cases} \frac{c(p_0-p'_a)(1-q)q((p_0-p_a)\Lambda_0\Lambda_a+c(-p_a\Lambda_a+p'_aq(\Lambda_0+\Lambda_a)-p_0(q(\Lambda_0+\Lambda_a)-\Lambda_a)))}{(c(p_0(1-q)+p'_aq)+p_0\Lambda_0)(c(1-p_0(1-q)-p'_aq)+\Lambda_0-p_0\Lambda_0)}, \text{when } q\leq\frac{c+\Lambda_0}{c(1+L)} \\ -\frac{(p_0-p'_a)(1-q)((p_0-p'_a)\Lambda_0-(Lp_0+p'_a-p_a-Lp_a)\Lambda_a)}{(1-p'_a+L(1-p_0))(Lp_0+p'_a)}, \text{when } q>\frac{c+\Lambda_0}{c(1+L)} \end{cases}$$

Because when $q\leq\frac{c+\Lambda_0}{c(1+L)}$, $\Lambda_a=c$,

$$E[dq/dt]=\begin{cases} \frac{c^2(p_0-p'_a)(1-q)q(p_0-p_a-p_0q+p'_aq)(c+\Lambda_0)}{(c(p_0-p_0q+p_aq)+p_0\Lambda_0)(c(1-p_0+p_0q-p'_aq)+(1-p_0)\Lambda_0)}, \text{when } q\leq\frac{c+\Lambda_0}{c(1+L)} \\ -\frac{(p_0-p'_a)(1-q)((p_0-p'_a)\Lambda_0-(Lp_0+p'_a-p_a-Lp_a)\Lambda_a)}{(1-p'_a+L(1-p_0))(Lp_0+p'_a)}, \text{when } q>\frac{c+\Lambda_0}{c(1+L)} \end{cases} \tag{29}$$

Simplifying (22), we have that: when $q\leq\frac{c+\Lambda_0}{c(1+L)}$, $E[dq/dt]$ has the same symbol as $\frac{p_0-p_a}{p_0-p'_a}-q$; when $q>\frac{c+\Lambda_0}{c(1+L)}$, $E[dq/dt]$ has the same symbol as $\frac{Lp_0+p'_a-p_a-Lp_a}{p_0-p'_a}\Lambda_a-\Lambda_0$.

When $\frac{p_0-p'_a}{p_0-p_a}<\frac{c}{c+\Lambda_0}(1+L)$, it's easy to see that $\frac{p_0-p_a}{p_0-p'_a}-q>0$ when $q\leq\frac{c+\Lambda_0}{c(1+L)}$ and it's also easy to show that $\frac{Lp_0+p'_a-p_a-Lp_a}{p_0-p'_a}>0$. Therefore, if there exists q_e such that $\Lambda_a(q_e)=\frac{p_0-p'_a}{Lp_0+p'_a-p_a-Lp_a}\Lambda_0$, it will be a stable state. From condition (15), we have $\Lambda_a(0)=c$, $\Lambda_a(1)=\frac{1}{L}\Lambda_0$ and Λ_a is strictly monotonously decreasing when $\Lambda_a<c$. Because $\frac{1}{L}\Lambda_0<\frac{p_0-p'_a}{Lp_0+p'_a-p_a-Lp_a}\Lambda_0<c$, there exists a unique q_e such that $\Lambda_a(q_e)=\frac{p_0-p'_a}{Lp_0+p'_a-p_a-Lp_a}\Lambda_0$.

When $\frac{p_0-p'_a}{p_0-p_a}>\frac{c}{c+\Lambda_0}(1+L)$, we have $\frac{Lp_0+p'_a-p_a-Lp_a}{p_0-p'_a}\Lambda_a-\Lambda_0<0$, therefore $q_e=\frac{p_0-p_a}{p_0-p'_a}$ is the only stable state other than 0 and 1 and $\Lambda(q_e)=c$.

When $\frac{p_0-p'_a}{p_0-p_a}=\frac{c}{c+\Lambda_0}(1+L)$, $q_e=\frac{p_0-p_a}{p_0-p'_a}=\frac{c+\Lambda_0}{c(1+L)}$ and $\Lambda(q_e)=c$. ■

Proof of Theorem 5

Proof First, it's obvious that when $\frac{p_0-p'_a}{p_0-p_a} \geq \frac{c}{c+\Lambda_0}(1+L)$, $\Lambda_a(q_e)$ equals to c and is larger than the $\Lambda_a(q_e)$ when $\frac{p_0-p'_a}{p_0-p_a} < \frac{c}{c+\Lambda_0}(1+L)$.

Then, when $\frac{p_0-p'_a}{p_0-p_a} < \frac{c}{c+\Lambda_0}(1+L)$, $\Lambda_a(q_e) = \frac{p_0-p'_a}{Lp_0+p'_a-p_a-Lp_a}\Lambda_0 = \frac{1}{(L+1)\frac{p_0-p_a}{p_0-p'_a}-1}\Lambda_0$, which is increasing with respect to $\frac{p_0-p'_a}{p_0-p_a}$.

Therefore, $\Lambda_a(q_e)$ is an increasing function with respect to $\frac{p_0-p'_a}{p_0-p_a}$. ■

References

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

Axel Anderson and Lones Smith. Dynamic deception. *The American Economic Review*, 103(7):2811–2847, 2013.

Athos Antonelli, Raffaele Cappelli, Dario Maio, and Davide Maltoni. Fake finger detection by skin distortion analysis. *IEEE Transactions on Information Forensics and Security*, 1 (3):360–373, 2006.

Richard Bellman and Robert E Kalaba. *Dynamic Programming and Modern Control Theory*, volume 81. Citeseer, 1965.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media, 2007.

Aditya Ganjam, Faisal Siddiqui, Jibin Zhan, Xi Liu, Ion Stoica, Junchen Jiang, Vyas Sekar, and Hui Zhang. C3: Internet-scale control plane for video quality optimization. In *USENIX Symposium on Networked Systems Design and Implementation*, pages 131–144, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Junchen Jiang, Shijie Sun, Vyas Sekar, and Hui Zhang. Pytheas: Enabling data-driven quality of experience optimization using group-based exploration-exploitation. In *USENIX Symposium on Networked Systems Design and Implementation*, 2017.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *The American Economic Review*, 101(6):2590–2615, 2011.

Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM, 2018.

Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016.

Justin Malbon. Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, 36(2):139–157, 2013.

Eric Maskin and Jean Tirole. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219, 2001.

Dina Mayzlin, Yaniv Dover, and Judith Chevalier. Promotional reviews: An empirical investigation of online review manipulation. *The American Economic Review*, 104(8): 2421–2455, 2014.

Sendhil Mullainathan and Andrei Shleifer. Media bias. Technical report, National Bureau of Economic Research, 2002.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. *arXiv preprint arXiv:1709.05295*, 2017.

Niels Ott, Ramon Ziai, Michael Hahn, and Detmar Meurers. Comet: Integrating different levels of linguistic modeling for meaning assessment. In *Joint Conference on Lexical and Computational Semantics, Volume 2: International Workshop on Semantic Evaluation*, volume 2, pages 608–616, 2013.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.

Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM, 2017.

Zhan Shi, Gene Moo Lee, and Andrew B Whinston. Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS quarterly*, 40(4), 2016.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

Scott E Umbaugh. *Computer Vision and Image Processing: A Practical Approach Using CViptools with Cdrom*. Prentice Hall PTR, 1997.

Lizhen Xu, Jason A Duan, and Andrew Whinston. Path to purchase: A mutually exciting point process model for online advertising and conversion. *Management Science*, 60(6): 1392–1412, 2014.

Hu Zhang, Zhuohua Fan, Jiaheng Zheng, and Quanming Liu. An improving deception detection method in computer-mediated communication. *Journal of Networks*, 7(11): 1811, 2012.

Xinyan Zhang, Jiangchuan Liu, Bo Li, and Y-SP Yum. Coolstreaming/donet: A data-driven overlay network for peer-to-peer live media streaming. In *Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 3, pages 2102–2111. IEEE, 2005.