REPUTATION GAMES AND POLITICAL ECONOMY

Cheng Sun

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

ECONOMICS

Adviser: Wolfgang Pesendorfer

June 2015

Abstract

This dissertation studies the applications of reputation games in social media and finance as well as decision games in political economy. Chapter 1 develops a reputation game in which a biased but informed expert makes a statement to attract audiences. The biased expert has an ideological incentive to distort his information as well as having a reputation concern. The expert knows that his expertise may vary in different topics, while the audiences cannot identify such differences. The biased expert is more likely to announce his favorite message when he knows less about it. Moreover, the biased expert is less willing to lie when the audiences have better outside options, and such improvements in outside options may benefit both the expert and the audiences.

Chapter 2 studies a credit rating game with a credit rating agency(CRA), an issuer and an investor. The privately informed and biased CRA provides a rating on the issuer's project, and the investor decides to purchase the project or not according to the report. As long as the CRA obtains a contract, he will inflate the rating. When the default risk is high, the CRA tells the truth. Moreover, he is more likely to tell the truth when the issuers private benefit is larger. When the default risk is low, the CRA sends a good rating. He is more likely to inflate the rating if the issuer has a higher private benefit.

Chapter 3 presents a model in secessions and nationalism, with a special emphasis on the role of civil war. In our model, a disagreement on secession between the central government and the minority group leads to disastrous military conflicts. As a result, the tremendous potential cost of the war distorts the political choice of the minority group, and helps the central government to exploit them both economically and politically. Several key ingredients, such as population, per capita income and perceived winning chance of the civil war, play an essential role in the decision making

process of the minority group. I also conduct an empirical test of this model, which supports the major findings stated above.

Acknowledgements

I would like to express my deepest gratitude to my advisor Wolfgang Pesendorfer for his advice, patience, encouragement and invaluable insights throughout all these years. It has been an honor and a privilege to be his student, and a wonderful experience to grow under his academic guidance.

I also want to express my sincere gratitude to Stephen Morris for his tremendous support, enlightenment and wisdom. It has been a great pleasure to learn from him, and I am very thankful for all his time, attention and input.

Moreover, I would like to thank Faruk Gul, Roland Benabou, Thomas Romer, Dilip Abreu, Sylvain Chassang and Juan Pablo Xandri for their valuable suggestions and helpful comments about my work. Thank you also to all my friends at the Economics Department for many useful conversations. Thank you to Jiangmin Xu, for his infinite patience and constant support. Thank you to all my non-econ friends, for making Princeton feel like home.

Special thanks to Ellen Graf for her extraordinary administrative support and contributions to the Microeconomic Theory Group. Thanks also go to Kathleen DeGennaro and Laura Hedden for their exceptional effort in administering the graduate program.

In addition, I would like to thank the seminar participants at Princeton University, Guanghua School of Management of Peking University, Tsinghua University School of Economics and Management, School of Business of Renmin University of China and Central University of Finance and Economics for help comments on my work.

Finally, thank you to my parents, to whom I owe everything.

To My Parents

Contents

Chapter 1

Reputation Game on Social Media

1.1 Introduction

Created in March 2006, Twitter immediately gained popularity, and has since become a global platform for public information transmission in real time. It enables users to send and read short 140-character messages instantly, which changes the way for people to distribute and discover information. As a public intellectual (or an expert), one could respond to any issue instantly by sending a Tweet, while the 140 words are only sufficient to show one's position not argument. The expert maintains his popularity by attracting as many followers as possible, and he cares about the size of followers in the current period as well as his reputation on the Twitter platform in general. An average user on Twitter decides to follow the expert or not before reading the message, and would benefit from receiving a correct prediction or at least some valuable information.

The aspect of the social network that I will focus on is how the expert and the audiences interact. The expert's popularity is evaluated by the number of his followers, and he would like to attract as many followers as possible. The audience's decision on following happens before receiving the message, so it is based on the

expert's reputation as well as his expected strategy. Moreover, the expert provides comments on a wide range of issues, on which his expertise may vary dramatically. Unfortunately, the audience may not be able to identify such differences, since many issues are not repetitive and their connections are unobservable. Consequently, the expert could take advantage of the information asymmetry and manipulate the message accordingly. However, the expert and the audience are not the only players involved in this information transmission process. If the audience are not following, they may use the tradition media as their information source, which is represented by outside options in this paper. This outside option measures how intense the competition between the social media and the traditional media is. When the outside option improves or weakens, both the expert and the audience would adjust their strategies correspondingly.

On such platforms like Twitter, a standard question would then be how a biased but informed expert attracts audiences. The private signal is noisy, and either an incorrect signal or a distorting message could mispredict the true state. An expert without self interest will honestly report his private information regardless of his reputation. However, a biased expert has to balance the benefit of stating his ideological preference against the potential reputation loss. Therefore, a message would not merely represent a prediction on the true state, but would also partially reveal whether the expert is honest or not. The reputation literature has long recognized the effect of reputation incentives and ideological preferences on the player's strategy, such as Sobel (1985), in which the expert builds a reputation by providing reliable information. My goal here is to investigate how the additional aspects of the social network platform (i.e.,Twitter) affect the strategies and the outcomes of the game.

I consider a reputation model in infinite horizon, in which a long-run player (i.e., the expert) sends messages to a sequence of short-run players (i.e., general Twitter users). This is a standard setting in the reputation game literature, and, for example,

Mailath and Samuelson (2001) also discussed a reputation game with long lived firms facing short term incentives. In each period, a continuum of short run players enter the game, and decide whether to follow the expert or not. The expert receives a noisy signal at the beginning of each period, and sends a corresponding message to the followers. The noisy signal is generated from a natural state, which is an independently and identically distributed random variable. If the audience decided to follow the expert, they will get a message on the true state of the world, which will be revealed at the end of that period. If they choose not to follow, they will reach out for other resources to learn the true state. In my model, I use a fixed outside option to measure the credibility of such a resource.

The general results establish the existence of a reputation equilibrium as well as its behavioral properties. Even though the biased expert may begin with a very high prior reputation, it will vanish in the long run. People may be uncertain about whether a rising star on Twitter has self interest in a certain issue. However, they are able to tell whether the one with a long track record is honest or not. This is a standard result in the reputation literature, for example, Benabou and Laroque (1992) conclude that the public will be able to tell whether the insider is trustworthy or not. I also show that the expert will announce his preferred position after receiving his favorable signals and may mix otherwise. The expert does have informative signals about the true state, thus he knows that it would be more costly to send his preferred message after receiving an unfavorable signal than a favorable one. This is a stronger conclusion than what is commonly found in the reputation literature. For example, Morris (2001) suggested that the expert may randomize after both types of signal. Such a stronger conclusion comes from the discrete decision rule of the followers as well as the binary message space for the expert. Moreover, the expert almost surely choose his preferred message for a high reputation, and randomize after an unfavorable signal if his reputation drops. When the expert has a good reputation, he will be

able to attract a significant size audience to follow regardless of the actual message. Therefore, his instantaneous benefit outweighs the reputation concern in this case. Another interesting observation would be that the expert almost surely pools with his preferred message for a very low reputation, since the reputation would not drop much after an incorrect prediction in this case.

The main contribution of my paper is introducing uncertainty in expertise to model the fact that the expert comments on a wide range of issues. The basic model provides an analytical tool for the reputation game on social networks, while I connect my model to social-network platforms like Twitter by implementing a few new elements. I begin with the observation that the sender's expertise may vary in different topics, while the followers cannot identify each of them. The audiences do know there is a wide distribution of the expertise, but their limited understanding prevent them from identifying each individual topic. As a result, the expert is more likely to lie after an unfavorable signal with low quality than one with high quality. On the one hand, the high quality signal improves the average credibility of his prediction, and then attracts more audiences after any types of signal. Moreover, the expert knows that distorting a high type signal is more likely to lead to a misprediction, which consequently hurts his reputation more. Therefore, he would truthfully report his private signal with high quality. On the other hand, lying is less costly after a low signal, since it is less likely to be an incorrect prediction. Thus, he would rather send his preferred message and exploit his reputation after a low type signal. I assume there is no learning process of the expertise and the sender cares about the size of the followers instead of the quality of the signal. On a social network, the expert responds to any random topics immediately, which are not necessarily repetitive. This random arrival of topics discourages the audience to learn the topics as well as the corresponding expertise. I did not include learning in the reputation updating process directly, but possible effects of it are considered in the paper.

The uncertainty on the expert's quality has been discussed in the reputation literature as well. Ottaviani and Sorensen (2006b) analyzed the case in which a privately informed expert is concerned about the perceived quality of his signal and credibly communicates only part of its information. In their paper, the expert with a unknown type provides advice on a single issue, while, in my paper, an expert offers comments on a variety of issues. Instead of the perceived quality, the expert in my model is concerned about his reputation only.

Another contribution of my model is to discuss how the expert and the audience respond to changes in outside options. I use the outside option to measure the alternative information source for each individual within the audience, which can be interpreted as tradition medias. The expert is more likely to tell the truth, when the audience have better outside options. However, the effect on the size of followers is ambiguous. Two cases of experts on social network platforms are extended here to explain its effect. The first case considers the condition in which the outside options are improved for the audience. The expert is more likely to tell the truth in general, while certain improvements could benefit both the expert and the followers. The second case introduces naive individuals within the audience, who will follow the expert regardless of the message. Consequently, the expert is less likely to tell the truth, and will almost surely chooses his preferred message under a very high or a very low reputation. To explain the first case, I begin with an example of a special correspondent in health care for Wall Street Journal (WSJ), whom only people with knowledge in health care or finance would consider to follow. The potential followers are better informed in this field than an average person, and could predict the true state correctly with a higher chance as well. Therefore, this correspondent cannot lie as frequently as facing the general public, since the size of his followers could drop to zero to force him to leave the social-network platform. However, the effect on the size of followers is ambiguous, which depends on the actual distribution of the audience's

outside options. This is consistent with the observation that the highly specialized social network stars are more likely to be objective, but not necessarily less popular. The second example is a commentator from Fox News. A significant number of the followers are loyal supporters. Even if he is known to be almost surely biased, this group of audiences will stay with him. Therefore, the instantaneous benefit of pooling with his ideological preference outweighs the reputation damage for both the low and the high reputation. In addition, this commentator is more likely to stick to his preference in general. These two cases show that the better the audiences' outside option is, the more likely the expert will truthfully reveal the signal.

The idea of outside option is not new in the reputation literature. For example, Ely and Valimaki (2003) also talks about outside options. In their model, the true state is not revealed, thus the long-lived agent may prefer a certain separating action, which actually hurts the short run player. Moreover, when the long run player's reputation incentive is sufficiently strong, short run agents would rather not participate. However, in my setting, the reputation updating is based on both the action and the true state, hence an incorrect prediction would hurt both the expert's reputation and the audience's payoff. Such a difference would keep the expert from extreme separating actions and would encourage the audiences to remain in the game.

In addition, my paper is offering silence as a choice to the expert. When a new event happens, the expert could either respond with a definitive statement, or simply remain silent. In this paper, I introduce an uninformative signal, and assume the honest person will not respond to a topic after this kind of signals. This assumption offers an alternative to the biased expert, which hurts the instantaneous payoff less. Therefore, the expert would use this alternative instead of his unfavorable message after the uninformative signal. Even after an unfavorable signal, the choice of silence would reduce the probability of an unfavorable message. However, since silence does not improve the reputation as much as sending an unfavorable message, silence can-

not totally replace this option. The role of silence has been a popular topic in the reputation literature as well as in the finance literature. Sharfstein and Stein (1990) was focusing on the uncertainty in expertise with two experts, who would use herding to share the blame. Ottaviani and Sorensen (2006a) also used herding in a multi-expert model, and the experts decide between herding and anti-herding to boost their reputations. My paper is focusing on the single expert case, and the choice of silence acts like a wedge between lying and truth telling after an unfavorable signal.

My paper is closely related to Benabou and Laroque (1992), Morris (2001) and Ely and Valimaki (2003). Benabou and Laroque (1992) describes a symmetric signal space, in which the insider provides manipulative recommendations to investors and makes profit from the fluctuation of stock prices. Similar to his model, the biased expert in my paper also benefits from the presence of a committed honest type. However, in my model, the biased expert has an asymmetric ideological preference, and the Discussion section will show that introducing a strategic honest type would not affect the main conclusion. Morris (2001) introduces two types of strategic player, one who shares the same preference as the decision maker, and one with an ideological preference. The reputation concern is built into the model by applying a higher weight to the future payoff. When the future is sufficiently important, no information is revealed in the first period. This is different from my paper, where the biased expert would almost surely tell the truth with a significant reputation concern. Instead of persuading the audiences of the true state of the world, the biased expert in my model wishes to broadcast his ideological preference to as many audiences as possible. Therefore, he would never be pooling with his unfavorable position. Ely and Valimaki (2003) analyzes the effect of outside option in a bad reputation model, and the reputation is updated according to the action only. The reputation updating rule in my model is based on both the expert's action and the revealed state of the world. This weaker assumption guarantees that the expert could have a positive number of

followers and stay in the game. The reputation game in this paper is similar to the reputation games initiated by Crawford and Sobel (1982) and Sobel (1985). Benabou and Laroque (1992) studies a reputation game with noisy signals and shows how good reputation helps the insider exploit profits from investors. Morris (2001) and Ely and Valimaki (2003) explain how the strategic good advisor is forced to lie to enhance his reputation. Ottaviani and Sorensen (2006a) and Ottaviani and Sorensen (2006b) extend Sobel's reputation game with multiple experts and uncertainty in expertise. More recent work such as Chen, Kartik, and Sobel (2008) and Chen (2011) examine the characteristics of the cheap-talk equilibria with a focus on the technical analysis. However, my paper is not exactly cheap talk, since the followers' action is based on the credibility of the message instead of their belief of the true state.

The remainder of this paper is organize as follows. In Section 1.2, I describe a general setup of the model and also define the reputation equilibrium. In Section 1.3, I analyze the basic model with a noisy signal. Section 1.4 considers a variation of the basic model with multiple topics as well as a continuum of topics. Section 1.5 is focusing on the effect of outside option and Section 1.6 introduces the choice of silence to the basic model. Finally Section 1.7 concludes.

1.2 The Model

Consider the following situation. There is a social network platform, where a long lived expert sends a message each period to predict the natural state of the world, and a continuum of short lived audience decides to follow the expert or not before reading the message. The expert has a private signal about the true state each period, which provides credibility to his message. The audience lack such expertise, and therefore consider to follow the expert for his prediction. If the expert is honest, he would

truthfully report his information. Otherwise, he may distort the message to fulfill his ideological preference.

The natural state of the world $w_t \in \{-1, 1\}$ at time t is an i.i.d random variable, and both states would happen with equal probability. The expert may observe a noisy informative signal $s_t \in \{-1, 1\}$ of type $\tau_t \in (0, 1)$ or an uninformative signal $s_t = 0$ of type $\tau_t = 0$, where the signal s_t and the type τ_t are both private information to the expert. In each period, τ_t is an i.i.d random variable with a distribution of $G(.)$, and s_t is an i.i.d random variable whose probability distribution depends on the true state w_t and the type τ_t. Let $P(s_t|w_t, \tau_t)$ denote the probability of getting a signal s_t given the true state w_t and type τ_t. Since the private information is valuable, the informative signal would predict the true state with the credibility better than flipping a coin. More specifically, $P(s_t = w_t|w_t, \tau_t) = \frac{1+\tau_t}{2} = 1 - P(s_t = -w_t|w_t, \tau_t)$. On the other hand, $s_t = 0$ means that the signal is uninformative, and the expert cannot learn anything about the state w_t. Thus, on average, the expert is able to correctly predict the state with a probability strictly greater than $\frac{1}{2}$.

The audience are uncertain about whether the expert is biased or not. With probability ρ_t, the expert honestly reports his signal. With probability $1 - \rho_t$, the expert is normal, meaning that he is biased with an ideological preference of $m_t = 1$. I use H and N to denote the honest and normal type respectively in this paper. The prior belief ρ_1 is given at the beginning of the game. The individual γ within the audience will decide to follow or not before receiving the message m_t. If he chooses to follow the expert, then he could only benefit from such an action when the message is the same as the true state. Otherwise, he will get an outside option of γ, which follows a distribution $F(.)$. After the followers receive m_t, the true state of the world w_t is publicly observed. At the beginning of the next period, period $t + 1$, another continuum of the audiences rationally updates their belief about the type of expert, according to the public history h^{t+1}. In period 1, the public history

includes the prior belief ρ_1 only. In the period t, the public history would be $h^t = \{\rho_1, m_1, w_1, ..., m_{t-1}, w_{t-1}\}$. The updated reputation at the beginning of period t could be written as $\rho_t = Pr(H|h^t) = Pr(H|\rho_{t-1}, m_{t-1}, w_{t-1})$, and ρ_t contains all the public information up to period t. Therefore, this paper will be focusing the Markovian equilibrium with respect to repuation ρ_t. To summarize, the timing of this game would be as below: the audiences decide to follow or not based on ρ_t, and then a new message m_t is sent by the expert after observing s_t and τ_t, and finally the true state w_t is revealed at the end of that period. Period $t+1$ repeats the process in period t.

In each period, an individual within the continuum of short lived audience is indexed by γ in $[\underline{\gamma}, \bar{\gamma}]$, with $\frac{1}{2} \leq \underline{\gamma} < \bar{\gamma} \leq 1$. The individual γ's utility depends on his decision of following, the expert's message, and the state of the world. If the individual γ chooses not to follow, he will get the outside option γ, which measures the credibility of his other possible information source on the true state of the world. It is the the opportunity cost of following the expert, or an alternative to the expert's message. Moreover, it is bounded below by $\frac{1}{2}$, since anybody can toss a coin and correctly predict the true state with 50% chance. If the audience does follow the expert, he would get 1 with $m_t = w_t$, $\frac{1}{2}$ with $m_t = 0$, and 0 otherwise. Define $\pi(\rho_t) = Pr(w_t = m_t|\rho_t) + \frac{1}{2}Pr(m_t = 0|\rho_t)$ as the credibility of the prediction, where π shows how the audience would project the expert's type and action upon reputation ρ_t. Let $D(\gamma, \pi)$ denote the decision for the audience γ facing credibility π. Clearly, the individual γ would choose to follow at $\pi(\rho_t) = \gamma$, and this helps to eliminate any uncertainty at this point. Then

$$D_(\gamma, \pi) = \begin{cases} 1 \text{ if } \pi(\rho_t) \geq \gamma \\ 0 \text{ otherwise.} \end{cases} \tag{1.2.1}$$

Thus, the audience γ's payoff can be written as below

$$u_\gamma(h^t) = \begin{cases} \gamma \text{ if } D_(\gamma, \pi) = 0, \\ \pi(\rho_t) \text{ otherwise.} \end{cases} \tag{1.2.2}$$

The total number of the audience is normalized be 1, and let ϕ denote the size of the followers at time t. Then, $\phi(\rho_t) = \int_{\underline{\gamma}}^{\bar{\gamma}} D(\gamma, \pi) dF(\gamma) = F(\pi(\rho_t))$, where $F(.)$ is the C.D.F of γ. For each individual γ, his decision depends on the shared belief on reputation ρ_t and the expert's corresponding strategy. Therefore, as far as someone with a better outside option $\gamma' > \gamma$ would follow, he would do the same thing. This equation on the size of the followers catches the action of the person, who is indifferent between following or not. To simplify the mathematical analysis, I will assume $\gamma \in U(\underline{\gamma}, \bar{\gamma})$ in the next few sections, where U denotes the uniform distribution. In the Discussion section, I will show that the main results of this paper would not change after removing the assumption on uniform distribution.

As mentioned above, the honest expert will play non-strategically, and report his signal truthfully. The normal expert's always wants to send his preferred position $m_t = 1$ independent of the state, but also wish to attract as many followers as possible. I use $M(s_t, m_t; \tau_t, \rho_t) \in [0, 1]$ to define the probability of the normal type announcing m_t after s_t of type τ_t with the updated reputation ρ_t, where ρ_t represents the history h^t. Therefore, his strategy could be described as $M(\tau_t, \rho_t) = \{M(s_t, m_t; \tau_t, \rho_t)\} \in [0, 1]^9$, where $s_t, m_t \in \{-1, 0, 1\}$. Let $\beta \in [\underline{\beta}, 1)$ denote the future discount value, and the total utility of the normal expert after signal s_t would be

$$U(M, \phi | h^t) = \sum_{j=t}^{\infty} \beta^{j-t} \sum_{m_j} M(s_j, m_j; \tau_j, \rho_j) u(w + m_j \phi(\rho_j)) \tag{1.2.3}$$

Here, w measures the endowment of the normal expert in each period, and the instantaneous payoff $u(.)$ is a monotonically increasing, twice differentiable, concave

function. The concavity of the instantaneous payoff shows that the expert is risk averse, while $m_t\phi(\rho_t)$ connects the ideological preference $m_t = 1$ with the size of followers.

A Markov strategy of the short run audience γ, maps his belief on the expert's reputation and projected action to a decision of following, that is $D(\gamma, \rho_t; M) : [\underline{\gamma}, \bar{\gamma}] * [0,1] * [0,1]^9 \to \{0,1\}$, The aggregate decision of the whole audience or the size of the followers is $\phi(\rho_t; M) = \int_{\underline{\gamma}}^{\bar{\gamma}} D(\gamma, \rho; M) dF(\gamma) = F(\pi(\rho_t; M)))$. A Markov strategy of the long run normal expert is function specify the probability of truth telling after each signal s_t, that is, $M(\tau_t, \rho_t) : T * [0,1] \to [0,1]^9$.

Definition 1.2.1. *A strategy profile (D, M) together with belief ρ_t is a reputation equilibrium if*

(1) it is a perfect Bayesian equilibrium,

(2) the utility of the normal type is increasing in reputation ρ_t over $[0,1]$.

Perfect Bayesian equilibrium is defined in Fudenberg and Tirole (1991). Monotonicity of the utility function captures the normal expert's concern on reputation. Similar monotonicity conditions in reputation setups have been discussed Benabou and Laroque (1992) as well as a more recent work by Lee and Liu (2013) . There exists Bayesian equilibrium which is not reputation equilibrium. For example, the expert could be babbling with $m_t = 1$ on an interval of reputation and then maximize his utility on other reputation values.

1.3 The Baseline Case

I begin with the benchmark case, in which there is an informative signal predicting the state correctly with probability $P(s_t = w_t | w_t) = p$, and incorrectly with probability $P(s_t = -w_t | w_t) = 1 - p$. The type τ_t is eliminated in this section to simplify notations. The honest type will truthfully report his signal, and would not announce $m_t = 0$.

Therefore, the discussion can be reduced to $s_t, m_t \in \{-1, 1\}$. Moreover, I assume that γ follows $U(\frac{1}{2}, 1)$, a uniform distribution between $\frac{1}{2}$ and 1. The honest type could predict the state with probability p, which is the upper bound for the credibility of the message. As a result, any follower with $\gamma > p$ will not consider to follow at all. On the other hand, the worst prediction would happen when ρ_t drops to 0 and the normal type sends $m_t = 1$. This condition will lead to credibility of $\frac{1}{2}$, the lower bound of γ with $F(\frac{1}{2}) = 0$. According to this assumption, for any strictly positive reputation ρ_t, the size of the followers will never shrink to 0.

1.3.1 Inference Within and Across Periods

Rational audiences use their prior on the expert's type ρ_t, and their expectation of the normal type's strategy $M(\rho_t)$ to infer the credibility π, as

$$\pi(\rho_t) = \rho_t p + (1 - \rho_t)\frac{[(2p-1)(M(1,1;\rho_t) + M(-1,-1;\rho_t)) + 2(1-p)]}{2} \tag{1.3.1}$$

π increases with both $M(1,1;\rho_t)$ and $M(-1,-1;\rho_t)$, since the audiences always benefit from the normal expert telling the truth. Moreover, $\frac{[(2p-1)(M(1,1;\rho_t)+M(-1,-1;\rho_t)+2(1-p)]}{2}$ is smaller than p, for $M(1,1;\rho_t), M(-1,-1;\rho_t) \in [0,1]$, which guarantees the audiences' payoff or the size of audiences always increases with the reputation ρ_t for any expected strategy $M(\rho_t)$.

According to the timing discussed above, the audiences update the expert's reputation ρ_{t+1} with the public history at the beginning of time $t+1$. Thus, ρ_{t+1} would be affected by both the message m_t and the revealed state w_t.

$$\rho_{t+1}(m_t, w_t; \rho_t) = \begin{cases} \frac{\rho_t}{\rho_t + (1-\rho_t)(M(m_t,m_t;\rho_t) + \frac{1-p}{p}(1 - M(-m_t,-m_t;\rho_t)))} & \text{if } m_t = w_t \\ \frac{\rho_t}{\rho_t + (1-\rho_t)(M(m_t,m_t;\rho_t) + \frac{p}{1-p}(1 - M(-m_t,-m_t;\rho_t)))} & \text{if } m_t = -w_t \end{cases} \tag{1.3.2}$$

After the state w_t becomes public information, the reputation will benefit from a correct prediction, or more precisely, $\rho_{t+1}(m_t, w_t = m_t) \geq \rho_{t+1}(m_t, w_t = -m_t)$. The equality only happens with the strategy $M(-m_t, -m_t; \rho_t) = 1$. When the expert has perfect signals with $p = 1$, his reputation will drop to $\rho_t = 0$ immediately after the first incorrect prediction. On the other hand, when $p \to \frac{1}{2}$, the updated reputation is almost the same after different w_t values. In general, the higher the quality of the signal p is, the more the expert's reputation will be hurt after a incorrect prediction. However, as far as the signal is noisy or p is strictly below 1, the reputation will never drop to 0 from a positive value.

Since ρ_t contains all the information in h^t, from a subjective point of view, ρ_t is a martingale,

$$E[\rho_{t+1}|h^t] = \sum_{m_t, w_t} \rho_{t+1}(m_t, w_t; \rho_t) * Pr(m_t, w_t; \rho_t) = \rho_t \tag{1.3.3}$$

1.3.2 Analysis of The Reputation Equilibrium

Let V_{s_t, m_t} denote the discounted expected sum of the normal type's utility after announcing m_t with signal s_t, and $W(\rho_t)$ denote the discounted expected sum at the beginning of period t. Then V_{s_t, m_t} for $(s_t, m_t) \in \{-1, 1\}^2$ are calculated as below

$$V_{s_t, m_t}(\rho_t) = u(w + m_t \phi(\rho_t)) + \beta \sum_{w_t} P(s_t|w_t) W(\rho_{t+1}(m_t, w_t; \rho_t)) \tag{1.3.4}$$

This paper will focus on reputation equilibrium, in which $W(\rho_t)$ is nondecreasing and continuous in ρ_t. Let C_+ denote the space of such functions on $[0, 1]$, endowed with the norm of uniform convergence.

Definition 1.3.1. *A reputation equilibrium of the dynamic game corresponds to the strategy of the normal type $M(.)$ as a function of his reputation: $M(s_t, m_t; .)$: $\{-1, 1\}^2 * [0, 1] \to [0, 1]^4$ and an associated nondecreasing value function $W(.)$:*

$[0,1] \to R$, *such that for all* ρ_t*:*

$$V_{s_t,s_t}(\rho_t) > V_{s_t,-s_t}(\rho_t) \text{ implies that } M(s_t, s_t; \rho_t) = 1, \tag{1.3.5}$$

$$V_{s_t,s_t}(\rho_t) < V_{s_t,-s_t}(\rho_t) \text{ implies that } M(s_t, s_t; \rho_t) = 0. \tag{1.3.6}$$

and

$$W(\rho_t) = \frac{1}{2} \sum_{s_t} \max\{V_{s_t,s_t}(\rho_t), V_{s_t,-s_t}(\rho_t)\} \tag{1.3.7}$$

where the function $V_{s_t,m_t}(\rho_t)$ *is defined from* W *by (3.4).*

This definition shows how the strategy M is defined by the discounted expected sum $V_(s_t, m_t)$ of announcing m_t after s_t. When the payoff is strictly better after a certain message, the expert will play a pure strategy. Otherwise, the expert will mix between two of them. Moreover, the discounted future payoff W is defined by allowing the expert to choose the optimal strategy after any signal, where the prior of each signal is $\frac{1}{2}$.

Proposition 1.3.1. *There is a unique Markovian reputation equilibrium of the dynamic game, with a unique continuous nondecreasing value function* W*.*

In order to prove the existence and the uniqueness of both the strategy M and the value W. I will first show that the strategy is uniquely defined for each value function W, and then explain how W is uniquely defined with such strategy M. Therefore, I can conclude the proposition above. The lemma below summarize the equilibrium strategy for any value function W.

Lemma 1.3.1. *The normal expert always announces* $m_t = 1$ *after signal* $s_t = 1$*. After* $s_t = -1$*, he either mixes between two messages or announces* $m_t = 1$*.*

I begin with characterizing the equilibrium strategy for each signal with a nondecreasing value function W. Here it begins with the strategy after $s_t = 1$. Announcing

$m_t = 1$ provides a higher instantaneous payoff, but hurts the future reputation, and consequently reduces the future payoff. Also, the expert is always better off with $m_t = s_t$, that is $V_{m_t,m_t} \geq V_{-m_t,m_t}$, given any $m_t \in \{-1, 1\}$. This characteristic leads to the result that the normal type will always announce $m_t = 1$ after $s_t = 1$, or $M(1, 1; \rho_t) = 1$. The argument is as following: If $M(1, 1; \rho_t) \neq 1$, then announcing $m_t = -1$ is at least as good as $m_t = 1$ after signal $s_t = 1$. Consequently, $V_{-1,1}(\rho_t) \leq V_{1,1}(\rho_t) \leq V_{1,-1}(\rho_t) < V_{-1,-1}(\rho_t)$ and the expert announces $m_t = -1$ after $s_t = -1$. This means the normal type will announce $m_t = -1$ more frequently than $m_t = 1$, or $m_t = 1$ actually improves the reputation of the normal type. Therefore, announcing $m_t = 1$ increases both instantaneous payoff and future reputation, or the normal type strictly prefers $m_t = 1$ to $m_t = -1$, contradiction. Another interesting finding would be that the normal type never plays the pure strategy of telling the truth after $s_t = -1$. Otherwise, the reputation will not change in period $t + 1$ after whatever the normal type announces. Then the normal type will announce $m_t = 1$ for sure, contradiction. Thus, $M(-1, -1; \rho_t) < 1$.

With $M(1, 1; \rho_t) = 1$, the credibility of the expert is updated as $\pi(\rho_t) = \rho_t p + (1 - \rho_t)\frac{1+(2p-1)M(-1,-1;\rho_t)}{2}$, which is strictly above $\frac{1}{2}$ for any $\rho_t > 0$ and $M(-1, -1; \rho_t) \in [0, 1]$. When $\gamma \in U[\frac{1}{2}, 1]$, the size of followers $\phi(\rho_t) = (2p - 1)(\rho_t + (1 - \rho_t)M(-1, -1; \rho_t))$ increases with $M(-1, -1; \rho_t)$ and ρ_t. The normal type's forecast is more likely to be accurate, when he tells the truth, and consequently would attract more followers. With $p < 1$, or the signal being noisy, the reputation can approach 0, but remains strictly positive. Therefore, with the assumption $\gamma \in [\frac{1}{2}, 1]$, the followers with $\gamma \to \frac{1}{2}$ will always follow the expert and the size of followers is positive.

Similar to the paper Benabou and Laroque (1992), I first solve the two-period game, taking the valuation W of future reputation as payoff in the second period. Then I use the restriction on W and the contract mapping theory to show the existence

and uniqueness . Define $F_{-1}(\rho_t)$ the net gain of announcing $m_t = 1$ instead of $m_t = -1$ after $s_t = -1$.

$$F_{-1}(\rho_t) = u(w + \phi_t) + \beta(pW(\rho_{t+1}(m_t = 1, w_t = -1)) + (1-p)W(\rho_{t+1}(m_t = 1, w_t = 1)))$$
$$- u(w - \phi_t) - \beta(pW(\rho_{t+1}(m_t = -1, w_t = -1)) + (1-p)W(\rho_{t+1}(m_t = -1, w_t = 1))) \tag{1.3.8}$$

As $\rho_t \to 1$, the updated reputation ρ_{t+1} does not change much from ρ_t. Then, the normal type prefers $m_t = 1$, since the reputation gain of $m_t = -1$ in future period is not enough to compensate the immediate loss. Or more precisely, the normal type would state his preferred position, regardless to their private signal. However, since $F_{-1}(\rho_t)$ drops with ρ_t with any $M(-1, -1; \rho_t)$ and $\beta > \underline{\beta} > 0$, it will fall to 0 at some reputation level. As a result, the normal type cannot stick to $m_t = 1$ any more. The normal type cannot switch to mimic the honest type either, as such a decision would have the reputation ρ_{t+1} remains the same as ρ_t, which makes announcing $m_t = 1$ a strictly better choice. Therefore, there exists a mixing strategy for lower reputation ρ_t. When ρ_t approaches 0, the normal type would still randomize between the two messages, but would announce $m_t = 1$ almost sure. For $\rho_t \to 0$, the normal type does not have much reputation concern, and would rather biased toward his ideological preference. The analysis above showed that for any nondecreasing utility function W, the equilibrium strategy M is uniquely defined. With the contraction mapping theorem, I can also show W is uniquely define. A more detailed proof can be found in the Appendix. In the following proposition, I assume that γ is uniformly distributed on the interval $[\frac{1}{2}, 1]$.

Proposition 1.3.2. *When $\beta \in (\underline{\beta}, 1)$ and $p \in (\frac{1}{2}, 1)$, there exists $\bar{\rho}$ that for $\rho_t \geq \bar{\rho}$ the normal type always sends $m_t = 1$. When $\rho_t < \bar{\rho}$, the normal type send $m_t = 1$ after $s_t = 1$ and randomize after $s_t = -1$. The probability of $m_t = -1$ is strictly positive and converges to 0 as ρ_t converges to 0.*

The first part of *Proposition 2* is discussed above already, and here I will demonstrate that $M(-1,-1;\rho_t) \to 0$ for $\rho_t \to 0$. According to the continuity of $F_{-1}(\rho_t)$, the solution to $F_{-1}(\rho_t) = 0$ is continuous as well. If $M(-1,-1;\rho_t) \to 0$ is not true, then it is either 0 or converge to a positive value. It cannot be 0, since the discount factor β is bounded below by a positive $\underline{\beta}$, and the reputation always matters. Assume that $M(-1,-1;\rho_t)$ would converge to a positive number ϵ. Then $\lim_{\rho_t \to 0} F_{-1} = u(w+(2p-1)\epsilon) - u(w-(2p-1)\epsilon) > 0$, contradiction. Therefore, the normal expert would rather state his preferred position $m_t = 1$, when the reputation is almost 0. The mixing strategy shows how the expert weight the incentive of future reputation against instantaneous payoff. For every low reputation ρ_t, announcing $m_t = -1$ would not benefit ρ_{t+1} much. As a result, the expert would rather enjoy the instantaneous benefit, and send $m_t = 1$ with a significantly high probability.

With $M(1,1;\rho_t) = 1$, the reputation will decrease over time after announcing $m_t = 1$ and increase after $m_t = -1$. Next I will check whether the honest type could rebuild his reputation in the long run. The expected reputation be the honest type would be

$$E[\rho_{t+1}|h^t, H] = \frac{1}{2} \sum_{w_t=-1,1} \left(\frac{\rho_t p^2}{Pr(m_t = w_t|w_t)} + \frac{\rho_t (1-p)^2}{1 - Pr(m_t = w_t|w_t)}\right) \geq \rho_t \qquad (1.3.9)$$

with strict inequality when ρ_t is neither 0 nor 1. This inequality does not depend on p value or $M(-1,-1;\rho_t)$ value. The only necessary condition is $Pr(m_t = w_t|w_t) < p$, which is supported by $\rho_t < 1$. Thus, we can conclude that the honest type's reputation will improve and is a strict sub-martingale. For any prior belief ρ_1, it will converges almost surely to some stationary variable on $[0,1]$. In Appendix, I will show that the reputation would converge to 1 in the long run. This result works in most of the reputation environment, for example, Benabou and Laroque (1992) also discussed this. $E[\rho_{t+1}|h^t, H] * \rho_t + E[\rho_{t+1}|h^t, N] * (1-\rho_t) = E[\rho_{t+1}|h^t] = \rho_t$. $E[\rho_{t+1}|h^t, H] \geq \rho_t$

leads to the result $E[\rho_{t+1}|h^t, N] \leq \rho_t$, also with strict inequality when ρ_t is different from 0 or 1. Therefore, ρ_t is a super-martingale from the normal type's point of view. Similar to Benabou and Laroque (1992), I can draw the following conclusion.

Lemma 1.3.2. *From any prior belief ρ_1, the equilibrium process $\{\rho_t\}_{t\in N}$ converges almost surely to 0 as t goes to infinity if the sender is normal.*

The normal type sends $m_t = 1$ after $s_t = 1$, and send $m_t = 1$ with positive probability after $s_t = -1$. His forecast would be less credible than the honest type, or his reputation is expected to drop gradually. Thus, in the long run, the audiences will be able to identify the type of the expert. All this argument is based on the fact that the game will be continued for infinite many period, which means that neither the expert nor the audiences will drop out from the game. To support such statement, we need the assumption $p < 1$. If $p = 1$, the reputation ρ_t and the size of followers drop to 0 immediately after an incorrect forecast. Consequently, the normal type may rather tells the truth and keep his reputation constant at the initial reputation ρ_1. The question remains would be whether $\min \gamma = \frac{1}{2}$ is necessary for the nonzero audience condition as well. This assumption will be removed in the later part of this paper, and I will explain how $\underline{\gamma}$ would affect the equilibrium strategy of the biased expert.

1.3.3 Comparative Statics

According to the definition above, the expected future valuation W also depends on the future discount factor β and the quality of the private signal p. I will discuss the effect of the discount factor β, which W increases with. Consider the cutoff point $\rho_t = \bar{\rho}$ for some β value. When β increases to $\beta + \epsilon$, the instantaneous payoff remains the same. However, the expected future payoff of announcing $m_t = 1$ does not grow as much as the case with $m_t = -1$. Thus, the expert should have switched to mixing

at such ρ_t already, or $\bar{\rho}$ increases with β value. With similar argument, the normal type would mix with a higher $M(-1,-1;\rho_t)$ value, when β grows. When $\beta \to 1$, the effect of the instantaneous on reputation diminishes, and the expert mainly cares about their reputation. A detailed proof is available in the appendix.

Lemma 1.3.3. *The value function W increases with the discount factor β. Moreover, the larger β is, the higher the cutoff point $\bar{\rho}$. For $\rho_t < \bar{\rho}$, the probability of truth telling $M(-1,-1;\rho_t)$ also grows with β.*
When $\beta \to 1$, $\bar{\rho} \to 1$ and $M(-1,-1;\rho_t) \to 1$ for $0 < \rho_t < \bar{\rho}$.

1.4 Uncertainty in The Quality of the Signal

In the previous section, there is only uncertainty in the expert's preference and the expert sends messages on a single issue. However, on social networks, the expert provides comments on a wide range of issues. Moreover, those issues may follow a random arrival, and are not necessarily repetitive. As a result, the audiences might be unable to tell the quality of the expert's private signal s_t. For various topics, the expert might be more familiar with one topic, and know less about another. The expert understands his own strength and weakness. The audiences know the distribution in the expert's quality on various topics, but are unable to identify his expertise in each topic. For example, a Macro economist can explain a specific monetary policy well, but be unable to predict the result of a certain budget negotiation in the Congress. Unfortunately, the public only know this person as a good economist and expect him to answer any question on U.S. Economy. But the lack of expertise could not stop his from participating the discussion, or expressing his own position. The expert may even take advantage of such information asymmetry to state his preferred position. As a columnist on New York Times, Paul Krugman participates discussions on every single popular issue, not even limited to Economic conversations. He wrote on a va-

riety of topics, such as income distribution, taxation and monetary policy, which an average reader can hardly differentiate from his field of international trade. Because of the information asymmetry on expertise, the average reader is still willing to listen to him on any economic issues. Therefore, the question becomes how the expert adjusts his strategy when the quality of the signal is uncertain to the followers.

I begin with the two-type case to show how the expert's strategy is affected by such uncertainty, and then extend the discussion to the continuous type case. The two type case provides insight into any finite topic case, while the continuous type offers an example of infinite topic cases. In this section, I assume no learning in expertise. However, if there are finite many type of topics within the infinite horizon, the topic could repeat in the future. The repetition will lead to learning of the type. If learning is allowed, then the game would be more similar to the baseline with a single type known to the public. For the continuous type case, it is possible to have a different topic in each period, without affecting the prior of the type or being repetitive.

1.4.1 Discrete Type of Signals

To capture this uncertainty, we now introduce two kinds of noisy signal, $\tau^h = 2(p + \epsilon) - 1$ type and $\tau^l = 2(p - \epsilon) - 1$ type. These two types are i.i.d and equally likely to happen in each period, or $Pr(\tau^h) = Pr(\tau^l) = \frac{1}{2}$. With a high type signal, $Pr(s_t = w_t | w_t, \tau^h) = p + \epsilon$; with a low type signal, $Pr(s_t = w_t | w_t, \tau^l) = p - \epsilon$. Since the quality of the signal is between $\frac{1}{2}$ and 1, ϵ is bounded above by $\max\{p - \frac{1}{2}, 1 - p\}]$. The expert knows whether he receives a high type or a low type, while the audiences only know the prior of $\frac{1}{2}$ at each period t. The timing is exactly the same as before. At the beginning of period t, the audiences decide to follow or not. Then the expert gets a noisy signal s_t of type τ_t, and sends out a message m_t accordingly. After the followers receives the message m_t, the true state w_t is revealed, and the public adjust their belief on the type of the expert. The payoff for the expert and the audiences

are the same as before. Since the normal expert knows the quality of the signal s_t, his strategy would depends on ρ_t, s_t, and the type τ_t. However, the public history only includes $\{m_t, w_t\}$ at the end each period t, so the audiences' decision or the size of followers still varies with ρ_t only.

In this section, I will compare the strategy of two types, and also check whether the main conclusions in the benchmark case remains the same. I define the normal type's strategy $M(\rho_t, \tau_t) = \{M(s_t, m_t; \rho_t, \tau_t)\}$ with $\rho_t \in (0, 1)$ and $\tau_t \in \{\tau^h, \tau^l\}$. $M(\rho_t, \tau_t) : [0, 1] * \{h, l\} \to [0, 1]^4$, where $M(s_t, m_t; \rho_t, \tau_t)$ is the probability of sending m_t after a signal s_t of type τ_t with reputation ρ_t.

The credibility is $\pi(\rho_t) = \frac{1}{2}(\pi(\rho_t, \tau^h) + \pi(\rho_t, \tau^l))$, which depends on the reputation ρ_t only. With nonzero reputation ρ_t, the credibility is strictly above $\frac{1}{2}$ and the size of the audiences will always be positive. Moreover, the credibility increases with $M(s_t, s_t; \rho_t, \tau^h)$, $M(s_t, s_t; \rho_t, \tau^l)$, and the reputation ρ_t. These characteristics are necessary for the existence of reputation equilibrium.

This reputation updating rule confirms that the reputation will drop after $m_t = 1$ and improve after $m_t = -1$. Also, a nonzero reputation ρ_t will not jump into a zero reputation in the next period. The question remains would be how the reputation will change in the long run for the honest type and normal type expert. $E(\rho_{t+1}|h^t) = \rho_t$, ρ_t is a martingale for any reputation updating rule, and this is a very general conclusion for the information transmission game. $E(\rho_{t+1}|h^t, H) \geq \rho_t$ and $E(\rho_{t+1}|h^t, N) \leq \rho_t$ are true for this modified version of model, with strict inequality for $\rho_t \neq 0, 1$. The reason is that the argument in the previous section only need one of the most basic characteristics, such as $Pr(m_t = w_t|w_t) \in [\frac{1}{2}, p]$. Repeat the process in proof of Lemma 2, I can conclude that the reputation for the honest type will converge to 1 in the long run, while the normal type's reputation will converge to 0.

Let $V_{\tau_t, s_t, m_t}(\rho_t)$ denote the discounted expected sum of the normal type's utility after announcing m_t with signal s_t of type τ_t given reputation ρ_t, and $W(\rho_t)$ denote

the discounted expected sum at the beginning of period t.

$$V_{\tau_t,s_t,m_t}(\rho_t) = u(w + m_t\phi(\rho_t)) + \beta \sum_{w_t} P(s_t|w_t,\tau_t) W(\rho_{t+1}(m_t, w_t; \rho_t)) \tag{1.4.1}$$

Lemma 1.4.1. *A reputation equilibrium of the dynamic game corresponds to the strategy of the normal type M as a function of his reputation and the quality of his signal, and an associated value function $W(\rho_t) : [0,1] \to R^+$, such that for all ρ_t:*

$$V_{\tau_t,s_t,s_t}(\rho_t) > V_{\tau_t,s_t,-s_t}(\rho_t) \text{ implies that } M(s_t, s_t; \rho_t, \tau_t) = 1, \tag{1.4.2}$$

$$V_{\tau_t,s_t,s_t}(\rho_t) < V_{\tau_t,s_t,-s_t}(\rho_t) \text{ implies that } M(s_t, s_t; \rho_t, \tau_t) = 0. \tag{1.4.3}$$

and

$$W(\rho_t) = \frac{1}{2} \sum_{s_t} \sum_{\tau_t} Pr(\tau_t) \max\{V_{\tau_t,s_t,s_t}(\rho_t), V_{\tau_t,s_t,-s_t}(\rho_t)\} \tag{1.4.4}$$

where function $V_{\tau_t,s_t,m_t}(\rho_t)$ is defined from M and W. Moreover, there exists a unique Markovian reputation equilibrium of this dynamic game, with a continuous nondecreasing value function W.

The normal type still announces $m_t = 1$ after $s_t = 1$, whether the expert has a signal of high quality or not. Here is a rough intuition. Announcing $m_t = 1$ always damages the expert's reputation, while $m_t = -1$ improves it. If the normal type will mix after a high quality signal of $s_t = 1$, he is indifferent between announcing 1 and -1 in such a case. However, if he is facing a low type signal of $s_t = 1$, the true state is less likely to be $w_t = 1$, or his continuation payoff for announcing $m_t = 1$ would be lower. On the other hand, the continuation payoff of $m_t = -1$ goes up. Therefore, the expert would strict prefers $m_t = -1$ after a low quality signal of $s_t = 1$. Similar arguments suggest that the normal expert would strictly prefers $m_t = -1$ after $s_t = -1$ for both the high and low type signal. Therefore, the normal expert on average announcing $m_t = -1$ more often, or $m_t = -1$ leads to lower reputation, contradiction. We repeat

the process, and could show the normal expert would not mix after a low type signal of $s_t = 1$ as well.

When reputation $\rho_t \to 1$, reputation gain of announcing $m_t = -1$ is not enough to compensate the instantaneous loss. Therefore, the normal type would state his preferred position $m_t = 1$, or $M(-1, -1; \rho_t, \tau^h) = M_t(-1 - 1; \rho_t, \tau^l) = 0$ for large ρ_t value. Compare the expected payoff for announcing $m_t = 1$ after $s_t = -1$ for both high type and low type signal, and $V_{h,-1,1} < V_{l,-1,1}$ for any ρ_t value. To explain this, the signal of high type predicts the true state better, or the expert understands that he will suffer more reputation damage by lying. Consequently, there exists a cutoff $\bar{\rho}^h$, such that when the reputation ρ_t falls below $\bar{\rho}^h$, the expert would start to mix between $m_t = 1$ and $m_t = -1$ after $s_t = -1$ with high type signal first.

$W(\rho_t)$ is uniquely defined according to the lemma above, and monotonically increasing in reputation. A detailed proof can be found in the Appendix. Next, I will consider how the strategy looks like when the normal type start to mix after at least one type of $s_t = -1$. A natural claim would be: the expert know he is reputation is more likely to drop after lying with a high quality $s_t = -1$, thus he is less willing to lie after a high type signal than a low type one. Actually, I can draw a general conclusion of $M(-1, -1; \rho_t, \tau^h) > M(-1, -1, ; \rho_t, \tau^l)$ for $\rho_t < \bar{\rho}^h$, and the scratch of the argument is as below. If this is not true, then there exists some ρ_t with $M(-1, -1; \rho_t, \tau^h) \leq M(-1, -1, ; \rho_t, \tau^l)$. That is to say the expert is more likely to tell the truth after a low quality $s_t = -1$ than a high quality $s_t = -1$. Then the reputation $\rho_{t+1}(-1, -1)$ is larger than or equal to $\rho_{t+1}(-1, 1)$, which leads to the result $V_{\tau^h,-1,-1}(\rho_t) \geq V_{\tau^l,-1,-1}(\rho_t)$. If the expert is indifferent between telling the truth or lying after a high quality signal of $s_t = -1$, or $V_{\tau^h,-1,-1}(\rho_t) = V_{\tau^h,-1,1}(\rho_t)$, then he would strictly prefer to announce $m_t = 1$ after a low quality signal of $s_t = -1$, with $V_{\tau^l,-1,-1}(\rho_t) \leq V_{\tau^h,-1,-1}(\rho_t) = V_{\tau^h,-1,1}(\rho_t) < V_{\tau^l,-1,1}(\rho_t)$. Similarly, if the expert is indifferent between telling the truth or lying after a low quality signal of $s_t = -1$, he

would strictly prefer to tell the truth after a high type $s_t = -1$. Both cases contradicts the claim of $M(-1,-1;\rho_t,\tau^h) \leq M(-1,-1;\rho_t,\tau^l)$.

Proposition 1.4.1. *When there are two types of signals τ^h and τ^l, there exists a reputation level $\bar{\rho}^h$ such that the expert will send $m_t = 1$ for $\rho_t > \bar{\rho}^h$. If $\rho_t < \bar{\rho}^h$, the expert mixes after $s_t = -1$ and tells the truth after $s_t = 1$. $M(-1,-1;\rho_t,\tau^h) > M(-1,-1;\rho_t,\tau^l)$ for any $\rho_t < \bar{\rho}^h$.*

Here, the quality of the high type $p+\epsilon$ and low type $p-\epsilon$ are symmetric to their mean value p, and both are equality likely to happen. The readers may wonder if this assumption is necessary for the conclusion above. Redefine the quality of high type as $p+(1-q)\epsilon$ and the low type as $p-q\epsilon$. The type τ_t is still i.i.d, but the two types are not necessarily equally likely to happen any more. The probability of facing a high type signal is q, while the probability for low type is $1-q$. As far as $(p+(1-q)\epsilon)Pr(\tau^h)+(p-q\epsilon)Pr(\tau^l) = p$, the expected credibility of the honest type would be the same p. Moreover, the ranking of the updated reputation after each possible realization would not change either. These are enough to support the statement above for the case with two possible qualities of the signals. Moreover, $M(-1,-1;\rho_t,\tau^h) > M(-1,-1;\rho_t,\tau^l)$ can be extended to finite many topics, as far as $\tau^h > \tau^l$. I can prove this by repeating the binary case on any pairs within a finite topic case.

1.4.2 Continuous Type of Signals

At the beginning of this section, I assume that there are only two types of signal, with high quality or low quality. The argument above suggests that the expert is more likely to tell the truth in the field which he is more familiar with. Instead of two discrete values, I now assume that the quality of the signal τ_t could be any value within $[\tau^l,\tau^h]$ and follows a uniform distribution. I define $\tau^h = 2(p+\epsilon)-1$ and $\tau^l = 2(p-\epsilon)-1$,

with the corresponding $P(s_t = w_t|w_t, \tau_t) \in U(p-\epsilon, p+\epsilon)$. The timing is exactly the same as other parts of this section, and, at the beginning of each period, an i.i.d type τ_t is randomly selected by the nature. The type of the signal at each period is private information for the expert, so the audiences' strategy and reputation updating are still based on the message and revealed state. For the normal expert, their strategy $M(s_t, m_t; \rho_t, \tau_t)$ is a function of both the type τ_t and the reputation ρ_t. Thus the normal type's strategy is defined as $M(s_t, m_t; \rho_t, \tau_t) : \{-1, 1\}^2 * [0, 1] * [\tau^l, \tau^h] \to [0, 1]$. I update the reputation according to the the strategy specified above

$$\rho_{t+1}(\rho_t, m_t, w_t) = \frac{\rho_t Pr(m_t|w_t, H)}{\rho_t Pr(m_t|w_t, H) + (1-\rho_t)\int_{\tau^l}^{\tau^h} Pr(m_t|w_t, \tau_t, N) dG(\tau_t)} \tag{1.4.5}$$

and the credibility of the expert would be $\pi_t = p\rho_t + (1-\rho_t)\int_{\tau^l}^{\tau^h} Pr(m_t = w_t|w_t, \tau_t, N) dG(\tau_t)$. Let $W(\rho_t)$ still denote the expected future payoff for the expert at the beginning of the period t, and $V_{\tilde{p}, s_t, m_t}(\rho_t)$ denote the expected payoff for announcing m_t after s_t of quality $\tilde{p}$. Then the normal type will announce m_t instead of $-m_t$, if and only if $V_{\tilde{p}, s_t, m_t}(\rho_t) \geq V_{\tilde{p}, s_t, -m_t}(\rho_t)$. Thus, W is defined as $W(\rho_t) = \int_{\tau^l}^{\tau^h} \frac{1}{2} \sum_{s_t} \max_{m_t} \{V_{\tau_t, s_t, m_t}(\rho_t)\} dG(\tau_t)$.

The expert will announce $m_t = 1$ after $s_t = 1$ on any topic. If this is not true, then there is a topic $\hat{\tau}$, that the expert is indifferent between sending $m_t = 1$ or $m_t = -1$. As a result, the expert will send $m_t = -1$ after $s_t = 1$ of any $\tau_t < \hat{\tau}$ and all types of $s_t = -1$. Then, $m_t = -1$ is more likely to be announced by the normal expert, or it will hurt the reputation of the expert, contradiction. Therefore, the expert will announce $m_t = 1$ after $s_t = 1$ as the benchmark case. With a similar analysis, the expert's strategy can be reduced to $M(s_t, m_t; \rho_t, \tau_t) \in \{0, 1\}$.

Now, I will look at the strategy for different ρ_t value, and describe the equilibrium for each reputation. Begin with $\rho_t \to 1$, the expert could state his preferred message $m_t = 1$ for any signal s_t of any quality $\tilde{p}$. When ρ_t keeps dropping, the expert may

tell the truth after $s_t = -1$ of type $2(p+\epsilon)-1$. Or , equivalently, there exists $\bar{\rho}$, that the normal type states his preferred position only with $\rho_t \geq \bar{\rho}$. When ρ_t falls below $\bar{\rho}$, the normal expert start to announce $m_t = -1$ after $s_t = -1$ of larger τ_t value.This is following the fact that $\rho_{t+1}(1,-1) < \rho_{t+1}(1,1) < \rho_t < \rho_{t+1}(-1,-1) < \rho_{t+1}(-1,1)$. More precisely, there exists a $\tau(\rho_t) \in [\tau^l, \tau^h]$ such that, the expert sends $s_t = -1$ for $\tau_t \geq \tau(\rho_t)$ and $s_t = 1$ otherwise. The expert knows that he is lying when he announces $m_t = 1$ after $s_t = -1$, and he also knows that the reputation would drop more after incorrectly distorting a signal of a higher quality. Therefore, he would rather distort the message after the low quality signal. Announcing $m_t = 1$ always damages the reputation, which also mean the expert announces it more frequently. Therefore, the expert will send $m_t = 1$ after $s_t = 1$, and $m_t = -1$ occasionally after $s_t = -1$. To summarize the strategy I just discussed, the strategy of the expert is defined as following.

Definition 1.4.1. *The expert is ideological on issue τ_t, if he chooses $m_t = 1$ for both signals; the expert is informative on issue τ_t if the message matches the signal.*

With the help of the definition above, a formal statement about the reputation equilibrium strategy is described as below.

Proposition 1.4.2. *There is a $\hat{\rho}$ such that for $\rho_t \geq \hat{\rho}$, the expert is ideological on every issue $\tau_t \in [\tau^l, \tau^h]$. For $\rho_t < \hat{\rho}$, there exists $\tau(\rho_t) \in (\tau^l, \tau^h)$, such that the expert is ideological on all issues $\tau_t < \tau(\rho_t)$ and informative o every issue $\tau_t \geq \tau(\rho_t)$.*

This conclusion, together with the two type case shows that the expert is more likely to tell the truth on the topic that he is more familiar with. The expert sacrifices his instantaneous payoff on high type signals by truthfully reporting the signal, and exploits his reputation on low type signals by babbling his ideological preference. The continuous case connects to the social networks in two ways. First, instead of two or finite many topics, the expert offers guidances on a wide range of issues. Second,

the no learning assumption is built upon the assumption of non-repetitive topics, and the continuum of topics offers sufficient topics for non-repetitive random draw in the infinite horizon game.

1.5 Quality of The Audiences' Outside Option

In this paper, I am using outside options to model the competition between the expert on a social network platform and traditional media platform. In the previous sections, I assume the outside option γ follows a uniform distribution $U[\frac{1}{2}, 1]$. Two interesting characters of this distribution will be commented here. First, $F(\frac{1}{2}) = 0$, or there will be no followers if the credibility drops to 0. If, instead, $F(\frac{1}{2})$ is strictly positive, then there exists a naive group within the audience, who will follow regardless to the expert's reputation and strategy. Second, $F(\pi_t)$ is positive as far as $\pi_t > \frac{1}{2}$. Therefore, once ρ_t is positive, no matter what strategy the normal type is choosing, the size of the followers will be positive. If I remove this assumption, then the size of the followers may drop to 0 for some reputation ρ_t and its corresponding strategy $M(\rho_t)$. To study the game with such assumption, I will add a restriction in which the expert would drop out from the game with $\phi_t = 0$ and receive a reservation payoff $\frac{u(w)}{1-\beta}$.

In the following parts, I will consider two cases, one with $F(\frac{1}{2}) > 0$ and one with $F(\frac{1}{2} + \delta) = 0$ for a positive δ. As mentioned in the introduction, the outside option measures the competition between the social network and traditional media platforms. $F(\frac{1}{2} + \delta) = 0$ implies an intense competition from traditional media platforms. Every individual γ is able to prediction the true state better than flipping the coin, but their outside option is not sufficient to totally drive the social network out of competitions. $F(\frac{1}{2}) > 0$ indicates that a weak competition from traditional media platforms, which helps the expert to further exploit the audience.

1.5.1 $F(\frac{1}{2}+\delta)=0$ for a positive δ.

I return to the baseline model with a single type topic, and the signal's credibility is fixed at p. To compare with the baseline case, I assume the marginal distribution stays the same as $f(\gamma)=2$ for $\gamma \in (\frac{1}{2}+\delta, 1)$, with $F(1)=1$ and $F(\frac{1}{2}+\delta)=0$. The question would be whether the expert would drop out of the game after his reputation hits some cutoff point, or he would remain in the game for any reputation ρ_t. Let $M(s_t, m_t; \rho_t, \delta)$ denote the probability of announcing m_t after s_t. The credibility of the expert would be the same as Section 3, where $\pi(\rho_t) = \rho_t p + (1-\rho_t)\frac{(2p-1)(M(1,1;\rho_t,\delta)+M(-1,-1;\rho_t,\delta))+2(1-p)}{2}$, and the corresponding size of followers would be $\phi(\rho_t) = \max\{0, (2p-1)(2\rho_t - 1 + (1-\rho_t)(M(1,1;\rho_t,\delta)+M(-1,-1;\rho_t,\delta)) - 2\delta\}$. Thus the size of followers is either 0, or exactly 2δ smaller than in the baseline case for any strategy M with ρ_t. And in this section, I will denote the baseline case as $\delta = 0$. Let $W(\rho_t;\delta)$ denote the expected future payoff at the beginning of the period t with reputation ρ_t as before. Then the payoff $V_{s_t,m_t}(\rho_t,\delta)$ for the norma expert announcing m_t to signal s_t would be

$$V_{s_t,m_t}(\rho_t,\delta) = \begin{cases} \frac{u(w)}{1-\beta} \text{ if } \pi(\rho_t) - 2\delta \leq 0, \\ u(w + m_t\phi(\rho_t)) + \beta \sum_{w_t} Pr(s_t|w_t) W(\rho_{t+1}(m_t, w_t)) \text{ otherwise.} \end{cases} \tag{1.5.1}$$

In this paper, I am focusing on the reputation equilibrium, in which $W(\rho_t;\delta)$ is monotonically nondecreasing with ρ_t and ϕ is nonnegative. Therefore, I could repeat the argument in Section 3, and claim that the normal expert will be announcing $m_t = 1$ after $s_t = 1$.

Next, I will show there exists an equilibrium, in which the expert will never quit the game for any reputation ρ_t. Let $F_{-1}(\rho_t,\delta) = V_{-1,1}(\rho_t,\delta) - V_{-1,-1}(\rho_t,\delta)$ be the net gain of announcing $m_t = 1$. There is a cutoff $\rho(\delta)$, below which

the expert will start to mix after signal $s_t = -1$. Then, consider the opportunity cost of announcing $m_t = 1$ for $\rho_t < \rho(\delta)$. $F_{-1}(\rho_t, \delta)$ is non decreasing in $M(-1,-1;\rho_t,\delta)$. To stay in the game, the size of the audiences must be positive, and define $M^*(-1,-1;\rho_t,\delta) = \max\{0, \frac{\frac{2\delta}{2p-1}-\rho_t}{1-\rho_t}\}$. $M^*(-1,-1;\rho_t,\delta)$ describes the strategy that the audiences with outside option $\frac{1}{2}+\delta$ is indifferent between following or not. Thus, as far as $M(-1,-1;\rho_t,\delta)$ is strictly above such a lower bound, the expert would be able to stay in the game. $F_{-1}(M^*(-1,-1;\rho_t,\delta);\rho_t,\delta) < 0 < F_{-1}(1;\rho_t,\delta)$, so there is a unique $M(-1,-1;\rho_t,\delta)$ for the mixing strategy equilibrium. When $\rho_t \to 0$, the equilibrium strategy would also approach the lower bound $\lim_{\rho_t \to 1} M^*(-1,-1;\rho_t,\delta) = \frac{2\delta}{2p-1}$, which is strictly positive. The existence and uniqueness of the value function W is guaranteed by repeating the argument in Section 3. The finding above could be summarized here.

Lemma 1.5.1. *When $F(\frac{1}{2}+\delta) = 0$, $M(1,1;\rho_t,\delta) = 1$ for all ρ_t and $M(-1,-1;\rho_t,\delta) \geq \max\{0, \frac{\frac{2\delta}{2p-1}-\rho_t}{1-\rho_t}\}$ for $\rho_t < \rho(\delta)$. When $\rho_t \to 0$, $M(-1,-1;\rho_t,\delta)$ converges to $\frac{2\delta}{2p-1}$.*

Next, I will compare the payoff and the strategy under this setting against the benchmark case. I will show that the expert will have a payoff bounded above by $W(\rho_t, 0)$ and the mixing strategy $M(-1,-1:\rho_t,\delta)$ bounded below by $M(-1,-1:\rho_t,0)$, where $W(\rho_t,0)$ and $M(-1,-1:\rho_t,0)$ are the corresponding values defined in section 3. To get this conclusion, I will start with the two period case, then use backward induction for any finite horizon case, and use the convergence to show that it will work with the infinite case. Here is the scratch of the proof. Let $W^s(\rho_t,\delta)$ denote the continuation payoff and $M^s(-1,-1:\rho_t,\delta)$ the mixing strategy, when there are s periods left. In the one period game, the expert will be announcing $m_1 = 1$ for sure, thus the payoff $W^1(\rho_1,\delta)$ is bounded above and $M^1(-1,-1:\rho_1,\delta)$ is bounded below by the bench mark case. Consider the t period game, and assume it is true for the $t-1$ period. For any reputation ρ_1 and its correspond strategy $M^t(-1,-1;\rho_1,0)$ in the benchmark case. Since $u(.)$ is concave and $m_1 = 1$ decreases the expert's

reputation, -2δ will hurt the payoff after the message $m_t = 1$ much more than the one after $m_t = -1$. Therefore, the expert would strict prefer to announce $m_t = -1$ with such a probability $M^t(-1, -1; \rho_1, 0)$, or $M^t(-1, -1; \rho_1, \delta) \geq M^t(-1, -1; \rho_1, 0)$. Next, consider the payoff or the size of audiences in each period. To keep the same size of audiences as in the benchmark case at the period $s \leq t$, the expert need to mix with a probability of at least $M^s(-1, -1; \rho_{t+1-s}, 0) + \frac{2\delta}{(2p-1)(1-\rho_{t+1-s})}$. However, $M^s(-1, -1; \rho_{t+1-s}, 0) + \frac{2\delta}{(2p-1)(1-\rho_{t+1-s})}$ is not sufficient for the equilibrium, and the expert will have a strategy lower than $M^t(-1, -1; \rho_1) + \frac{2\delta}{(2p-1)(1-\rho_1)}$ to make the expert indifferent. Therefore, the expert will have smaller payoff. Let t goes to ∞, and the monotonicity still work.

Proposition 1.5.1. *The expert is more likely to tell the truth and also get fewer followers, when the outside option is strictly better.*

In this example, I put a mass of 2δ on the point $\gamma = 1$, and the followers of any credibility would decrease by 2δ. This improvement leads to a loss in followers and a gain in the audience's welfare. For the tradition media, a smaller size of followers means more people are choosing them instead. Therefore, it is a pareto improvement for the audience and traditional media platforms.

Instead of putting a mass of 2δ on the point $\gamma = 1$, I place 2δ of mass on the point $\frac{1}{2} + \delta$. Consider $\lim_{\gamma \to \frac{1}{2}+\delta^-} F(\gamma) = 0$ and $F(\frac{1}{2} + \delta) = 2\delta$. Then the m.d.f. $f(\gamma) = 2$ for $\gamma \in (\frac{1}{2} + \delta, 1)$. Thus, the size of followers is either 0 or the same as in the baseline, for any credibility. Repeat the process in this section, i find that

Proposition 1.5.2. *There exists such an improvement on outside options, that the expert is more likely to tell the truth and will get more followers.*

1.5.2 $F(\frac{1}{2}) = 2\delta$ for a positive δ.

It is not intuitive to have any γ strictly below $\frac{1}{2}$, since any individual could flip a coin and predict the state of nature correctly with 50% chance. However, there would be some irrational audiences who will follow an expert regardless to any information. Let $F(\frac{1}{2}) = 2\delta$, and $f(\gamma) = 2$ for $\gamma \in (\frac{1}{2}, 1-\delta)$. Such an adjustment only adds weight to the point $\frac{1}{2}$ without affecting the marginal distribution of the audience. The interesting question would be how the existence of irrational audience will affect the strategy of the expert.

With a similar argument as Section 3, I can claim that the expert strictly prefers to announce $m_t = 1$ after signal $s_t = 1$. Therefore, I can reduce the problem to the strategy after signal $s_t = -1$ as before. Let $M^+(-1,-1;\rho_t,\delta)$ denote the probability of telling the truth after $s_t = -1$, and the corresponding size of audiences would be $\phi_t = (2p-1)(\rho_t + (1-\rho_t)(M^+(-1,-1;\rho_t,\delta)) + 2\delta$. To compare with the benchmark case, let $\delta = 0$ denote such condition. This distribution of followers adds 2δ units of audiences to the expert for any strategy choice. As a result, when the reputation drops to almost 0 and reputation incentive also approaches 0, the expert will prefer to announce $m_t = 1$. If I improve the reputation from $\rho_t = 0$, then there will be a cutoff where the expert is indifferent between announcing $m_t = 1$ and $m_t = -1$. For a low reputation, the reputation incentive is not enough to compensate the instantaneous loss, which is bounded below by $u(w+2\delta) - u(w-2\delta)$. Thus, the lower bound $\underline{\rho}^+$ exists, below which the expert will play the pooling equilibrium. For a higher reputation $\rho_t \to 1$, the expert also strictly prefers announcing 1, since announcing $m_t = 1$ would not hurt the reputation as much as it can benefit the instantaneous payoff. Consequently, there is a switching point $\bar{\rho}^+$, above which the expert would pool with message $m_t = 1$. To generalize such a conclusion, I have a statement as following.

Proposition 1.5.3. *There exists an interval* $(\underline{\rho}^+, \bar{\rho}^+) \subset (0,1)$*, such that the expert of a reputation* ρ_t *within would mix after* $s_t = -1$*. Otherwise, the expert would play the babbling strategy of sending* $m_t = 1$*.*

To prove the proposition above, solving the two period equilibrium with $W^+(\rho_t)$ as expected future payoff is sufficient. I can repeat the argument in the Appendix for Section 3, and show the monotonicity of W. With monotonicity, and the existence and uniqueness of W can be concluded according to the contract mapping theorem.

1.6 Silence as A Choice

In this section, I will introduce the uninformative signal $s_t = 0$. For the honest type, he has two possible actions: randomizing between $m_t = 1$ and $m_t = -1$ or remaining silent. I will compare the normal type's strategy with respect to this two conditions. At the beginning of each period, the expert either receive an informative signal $s_t \in \{-1, 1\}$ or an uninformative signal $s_t = 0$. I use type τ^l to denote uninformative signal, and τ^h to denote informative signal. Comparing with the two type signal case in the previous section, this is a special case for $p - \epsilon = \frac{1}{2}$. This is not exactly the same as the condition of $p - \epsilon$ strictly above $\frac{1}{2}$, as both true states are equally likely to happen after a s_t of $p - \epsilon = \frac{1}{2}$, and the strategy would not depend on s_t at all. In the $p - \epsilon > \frac{1}{2}$ case, the expert would strictly prefer to announce $m_t = 1$ after $s_t = 1$, as he knows $s_t = 1$ contains some valuable information. However, with $p - \epsilon = \frac{1}{2}$, the expert would just randomize after the uninformative signal.

In this section, I assume the informative signal will happen with probability q, and the uninformative signal will happen with probability $1 - q$. When the signal is informative, it could predict the state correctly with probability p. I will begin with the case that the honest type randomize between $m_t \in \{-1, 1\}$ equally after the uninformative signal $s_t = 0$, and then consider the case the honest type send $m_t = 0$

after $s_t = 0$. $m_t = 0$ means the honest type stay silent, and this section will show how the choice of silence will affect the strategy of the normal expert.

1.6.1 Expert Is Never Silent

In this part, the honest expert will randomize between $\{-1, 1\}$ after the uninformative signal, so the normal type will send a message in the space $\{-1, 1\}$ as well. This case is equivalent to the condition where $p - \epsilon \to \frac{1}{2}$. Let $M^1(s_t, m_t; \rho_t)$ denote the probability of sending m_t after signal s_t in this case. In the next part, I will use $M^2(s_t, m_t; \rho_t)$ to denote the strategy when the expert would keep silence after some signal. Since the signal spaces for the high type and the low type are different, s_t is sufficient to reveal the type of the signal as well. The timing is exactly the same as before, and thus the audience would make the decision based on the reputation ρ_t only, while the reputation ρ_{t+1} is updated according to the message m_t together with the revealed state w_t.

I claim that the normal type would send $m_t = 1$ after $s_t = 1$ as the standard case. Otherwise, the expert would strictly prefer sending $m_t = -1$ after $s_t = 0$ and $s_t = -1$, or $m_t = 1$ would actually improve the expert's reputation, contradiction. Also the strategy after signal $s_t = 0$ will not affect the updated reputation not the size of the followers, since any message have half chance to be correct after such signal. If the reputation is known to be $\rho_t = 1$, the credibility of the message m_t would be $Pr(m_t = w_t|H) = \bar{p} = qp + \frac{1}{2}(1 - q)$, which is denoted by $\bar{p}$ in this section. In general, the credibility for the expert with reputation ρ_t would be $\pi(\rho_t) = \frac{1}{2} + \frac{q(2p-1)}{2}(2\rho_t - 1 + (1 - \rho_t)(M^1(-1, -1; \rho_t) + M^1(1, 1; \rho_t)))$, which depends on the probability of truthfully reporting after an informative signal. Thus any action after an uninformative signal could affect the reputation only.

With very higher reputation ρ_t, the expert will announce $m_t = 1$ after any signal. When ρ_t drops below some cutoff point $\bar{\rho}^1$, the expert will star to mix after the

informative signal $s_t = -1$. Let $W^1(\rho_t)$ denote the expected future payoff at the beginning of period t with reputation t. Then the cutoff $\bar{\rho}^1$ is defined as solution ρ_t to the equation

$$u(w + q(2p-1)\rho_t) + \beta((1-p)W^1(\frac{\rho_t\bar{p}}{\rho_t\bar{p} + 1 - \rho_t}) + pW^1(\frac{\rho_t(1-\bar{p})}{\rho_t(1-\bar{p}) + 1 - \rho_t}))$$
$$- u(w - q(2p-1)\rho_t) - \beta W^1(1) = 0 \tag{1.6.1}$$

Since $\bar{p}$ is strictly smaller than p, the updated reputation looks like the audiences facing an expert with a lower quality. Moreover, the size of followers is also shrinking by a factor q. On average, the audiences would think the message is less informative with a smaller q value, and fewer of them would be willing to follow the expert.

When the reputation drops further down, the expert would start to mix after the informative signal $s_t = -1$ or even possibly the uninformative signal $s_t = 0$, while the expert would state $m_t = 1$ after the informative signal $s_t = 1$. Moreover, the expert is always more likely to announce $m_t = 1$ with a larger s_t. He would announce $m_t = 1$ with more than half chance after $s_t = 0$, which is equivalent to announcing $m_t = -1$ after $s_t = 1$ of low type with certainty. Since this is a special version of the two-type signal case, and the conclusion comes from the analysis of the previous section.

1.6.2 Expert Can Choose Silence

Now assume the honest type will be silent or send message $m_t = 0$ after the uninformative signal $s_t = 0$. Consequently, the normal type would choose among all three messages $m_t \in \{-1, 0, 1\}$. $M^2(s_t, m_t; \rho_t)$ still denotes the probability of announcing m_t after s_t for an expert with reputation ρ_t. This time, I adjust the credibility $\pi_t = Pr(w_t = m_t|\rho_t) + \frac{1}{2}Pr(m_t = 0)$, which means $m_t = 0$ does not generate disutility for the person with $\gamma = \frac{1}{2}$. Therefore, the updated credibility for the normal type would be $\pi(\rho_t) = \frac{1}{2} + \frac{q(2p-1)}{2}(2\rho_t - 1 + \frac{1-\rho_t}{2}(M^2(-1,-1;\rho_t) + M^2(1,1;\rho_t) +$

$2 - M^2(-1, 1; \rho_t) - M^2(1, -1; \rho_t)))$. Thus, if the expert use the same strategy in both cases, or $M^1(s_t, m_t; \rho_t) = M^2(s_t, m_t; \rho_t)$, the number of the followers would be the same. However, $m_t = 0$ would give the expert higher instantaneous payoff than $m_t = -1$. Thus, if a pure strategy of $m_t = 1$ is not sufficient, $m_t = 0$ would be the next choice before $m_t = -1$. Otherwise, $m_t = 0$ would immediately bring the reputation up to $\rho_{t+1} = 1$ with a higher instantaneous payoff then $m_t = -1$. It also means that any strategy in the no silence case cannot be supported as equilibrium any more, after silence is introduced.

With a very high reputation ρ_t, the expert will announce $m_t = 1$ after any signal. When ρ_t drops below some cutoff point $\bar{\rho}^1$, the expert will star to mix with message $m_t = 0$ after the informative signal $s_t = -1$. Let $W^2(\rho_t)$ denote the expected future payoff at the beginning of period t with reputation t. Then the cutoff $\bar{\rho}^2$ is defined as solution ρ_t to the equation

$$u(w + q(2p-1)\rho_t) + \beta((1-p)W^2(\frac{\rho_t qp}{\rho_t qp + 1 - \rho_t}) + pW^2(\frac{\rho_t q(1-p)}{\rho_t q(1-p) + 1 - \rho_t}))$$
$$- u(w - q(2p-1)\rho_t) - \beta W^2(1) = 0 \tag{1.6.2}$$

where $W^2(1) = W^1(1) = \frac{\beta}{1-\beta}u(w + q(2p-1))$. Since qp is strictly smaller than $\bar{p}$, and $q(1-p)$ is also strictly smaller than $1 - \bar{p}$, the expert's reputation would drop more comparing to the non-silence case. The immediate payoff for announcing $m_t = 1$ is the same, while the updated reputation is strictly smaller. However, the alternative $m_t = 0$ leads to the same reputation $\rho_{t+1} = 1$ as $m_t = -1$, but a higher instantaneous payoff. Then, the equilibrium strategy in the case without silence cannot be supported as an equilibrium any more, as the expert is facing a higher alternative everywhere. Therefore, the expert would stop playing pooling strategy of $m_t = 1$ with a higher cutoff.

The discussion above could be generalized for all ρ_t, when I compare the strategy $M^1(s_t, m_t; \rho_t)$ against $M^2(s_t, m_t; \rho_t)$. It is easy to see that $M^1(1, 1 : \rho_t) = M^2(1, 1; \rho_t) = 1$. If this is not true, then the expert is indifferent between $m_t = 1$ and $m_t = 0$ after $s_t = 1$, then the expert is even less likely to announce $m_t = 1$ after $s_t \neq= 1$, and then the normal expert announces $m_t = 1$ less than the honest type, contradiction. Second, I want to claim that $M^2(0, -1; \rho_t) = 0$. Otherwise, the expert may announce $m_t = -1$ after $s_t = 0$ as well as $s_t = 0$. This means the expert would strictly prefer $m_t = -1$ after $s_t = -1$, and $s_t = 0$ leads to a higher reputation than $s_t = -1$ contradiction. The comparisons are summarized as the proposition below.

Proposition 1.6.1. *After any signal s_t, if the expert would announce $m_t = -1$ with positive probability in the non-silence case, he would announce $m_t = 0$ with positive probability in the silence case. Moreover, he would not announce $m_t = -1$ after $s_t = 0$ if he could remain silent.*

1.7 Conclusion

I have analyzed the reputation game on social media with a model in which a biased but informed expert sends a message to attract audiences. A key feature of my model is that the biased expert takes advantage of his good reputation to attract followers. I added a few elements to the baseline model to explain how different aspects of a social-network platform, such as Twitter, would affect the expert's strategy. I first showed that if the audiences cannot identify his expertise in each topic, the expert is more likely to announce his favorite message when he knows less about it. This result suggests that the expert will be ideological on his less familiar issues, and exploit the good reputation that he has accumulated. I then found that when the audiences have better outside options, the expert is more likely to tell the truth, and both could be better off with certain improvements in outside options. The expert is competing

against the outside options for followers. Therefore, the more informative the outside option is, the more informative he has to be. Finally, I introduced silence as a choice and explained how silence acts as a wedge between telling the truth and lying. The expert would rather be silent than announce his unfavorable position, when the signal is not very informative.

In the construction of the model, I have made use of a number of simplifying assumptions. First of all, I assume the honest type plays nonstrategically, and the biased expert benefits from the good reputation. This assumption can be relaxed to allow the honest type to play strategically. If the honest type cares about sending a correct prediction as well as attracting followers, he would play politically correctly after a low reputation and truthfully report his signal after a high reputation. When the future is sufficiently important, both the types are very likely to play politically correctly and send uninformative messages. This is consistent with findings in Morris (2001) and Ely and Valimaki (2003), which also concluded that the reputation concern will lead to uninformative outcomes. Another assumption is on the distribution of the outside option. Since the utility function is concave and the composite of two concave functions is concave, the conclusion remains the same as long as the cumulative distribution is concave. Finally, I assume there is no learning when the audiences cannot identify different topics. If the audiences learn to identify the expertise, the game becomes the single topic case with the quality of the signal alternating in each period.

1.8 Appendix

1.8.1 The Baseline Case

The proof in this section is following the same fashion as Benabou and Laroque (1992). I will focus on the reputation equilibrium with the value function $W(\rho_t)$ of the expert, which is nondecreasing and continuous in his reputation ρ_t. C_+ denotes the space of such functions on $[0,1]$, endowed with the norm of uniform convergence.

Proof of Proposititon 1. F_{-1} was defined as the net gain of announcing $m_t = 1$ instead of $m_t = -1$, and it can be expanded as

$$\begin{aligned}
&F_{-1}(M(-1,-1;\rho_t):\rho_t,W) = u(w+(2p-1)(\rho_t+(1-\rho_t)M(-1,-1;\rho_t)))\\
&+\beta(pW(\frac{\rho_t}{1+(1-\rho_t)\frac{p}{1-p}(1-M(-1,-1;\rho_t))})+(1-p)W(\frac{\rho_t}{1+(1-\rho_t)\frac{1-p}{p}(1-M(-1,-1;\rho_t))}))\\
&-u(w-(2p-1)(\rho_t+(1-\rho_t)M(-1,-1;\rho_t)))-\beta(pW(\frac{\rho_t}{\rho_t+(1-\rho_t)M(-1,-1;\rho_t)}))
\end{aligned} \tag{1.8.1}$$

$u(.)$ and $W(.)$ are both strictly increasing in $M(-1,-1;\rho_t)$. Therefore, this net gain increases with $M(-1,-1;\rho_t)$. At $M(-1,-1;\rho_t) = 0$, $F_{-1}(0;\rho_t,W)$ is always positive. The question remains to be whether $F_{-1}(1;\rho_t,W)$ is positive as well. If this is true, then announcing $m_t = 1$ is strictly preferred. Otherwise, there exists a $M(-1,-1;\rho_t) \in (0,1)$, such that $F_{-1}(M(-1,-1;\rho_t);\rho_t,W) = 0$. Let $M^*(-1,-1;\rho_t)$ be the probability of announcing $m_t = -1$ after $s_t = -1$ in the equilibrium, then

$$M^*(-1,-1;\rho_t) = \begin{cases} 0 \text{ if } F_{-1}(1;\rho_t,W) \geq 0 \\ F_{-1}^{-1}(0;\rho_t,W) \text{ if } F_{-1}(1;\rho_t,W) < 0 \end{cases} \tag{1.8.2}$$

where $F_{-1}^{-1}(0;\rho_t,W)$ denote the solution to $F_{-1}(M(-1,-1;\rho_t);\rho_t,W) = 0$. Thus,

$$\begin{aligned}
&V_{-1,1}(M^*(-1,-1;\rho_t);\rho_t,W) = u(w+(2p-1)(\rho_t+(1-\rho_t)M^*(-1,-1;\rho_t)))+\\
&\beta(pW(\frac{\rho_t}{1+(1-\rho_t)\frac{p}{1-p}(1-M^*(-1,-1;\rho_t))})+(1-p)W(\frac{\rho_t}{1+(1-\rho_t)\frac{1-p}{p}(1-M^*(-1,-1;\rho_t))}))
\end{aligned} \tag{1.8.3}$$

Define the operator $T(\rho_t; W) = \frac{1}{2}V_{-1,1}(\rho_t, W) + \frac{1}{2}V_{1,1}(\rho_t, W)$.

First, I want to show that $V_{1,1}(\rho_t, W)$ and $V_{-1,1}(\rho_t, W)$ increases with ρ_t. Define $V_{-1,1}(\rho_t, W) = V_{-1,1}(M(-1,-1;\rho_t);\rho_t, W)\&V_{1,1}(\rho_t, W) = V_{1,1}(M(-1,-1;\rho_t);\rho_t, W)$, then $V_{-1,1}(\rho_t, W)$ increases with ρ_t. The argument would be as below. If $M(-1,-1;\rho_t)$ increases with ρ_t, this is true $V_{-1,1}(\rho_t, W)$. If $M(-1,-1;\rho_t)$ decreases with ρ_t, this is true by the definition of $V_{-1,1}(M(-1,-1;\rho_t);\rho_t, W) = V_{-1,-1}(M(-1,-1;\rho_t);\rho_t, W)$. Therefore $V_{-1,1}(\rho_t, W)$ will increase with both ρ_t and W. Redefine $\eta(\rho_t) = \rho_t + (1-\rho_t)M(-1,-1;\rho_t)$, and then $V_{-1,1}(M(-1,-1;\rho_t);\rho_t, W) = u(w + (2p-1)\eta(\rho_t)) + \beta(pW(\frac{(1-p)}{1-p\eta(\rho_t)/\rho_t}) + (1-p)W(\frac{p}{1-(1-p)\eta(\rho_t)/\rho_t}))$. If η_t decreases with ρ_t, then $\eta(\rho_t)/\rho_t$ also decreases with ρ_t, which leads $V_{-1,1}$ decreases with ρ_t, contradiction. Therefore, we can see that η_t increases with ρ_t, and $V_{-1,1}(M(-1,-1;\rho_t);\rho_t, W) = u(w + (2p-1)\eta(\rho_t)) + \beta(pW(\frac{(1-p)\rho_t}{1-p\eta(\rho_t)}) + (1-p)W(\frac{p\rho_t}{1-(1-p)\eta(\rho_t)}))$ also increases with ρ_t. Consequently, $T(\rho_t; W)$ increases in ρ_t for any W.

Second, prove that $T(\rho; W)$ is nondecreasing in (ρ, W). Consider $(\rho^1, W^1) \geq (\rho^2, W^2)$. If $M^1(-1,-1;\rho^1, W^1) \geq M^1(-1,-1;\rho^2, W^2)$ or $W^1 = W^2$, $T(\rho^1; W^2) \geq T(\rho^1; W^2)$ is true for sure. Look at the case, $M^1(-1,-1;\rho^1, W^1) < M^2(-1,-1;\rho^2, W^2)$ and $W^1 > W^2$. However, $F_{-1}(M(-1,-1;\rho_t) : \rho_t, W)$ decreases with W, and $0 = F_{-1}(M^1(-1,-1;\rho_t) : \rho_t, W^1) < F_{-1}(M^2(-1,-1;\rho_t) : \rho_t, W^2) = 0$, contradiction.

T maps C_+ continuously into itself, and is non decreasing in W. Moreover, $T(W + c) = T(W) + \beta c$ for any constant C. Therefore, by Blackwell's theorem, T is a contracting mapping, and since C_+ with sup norm is complete, it has a unique fixed point W. And the discussion above confirms that W is non decreasing in reputation.

Proof of Lemma 2. This is a standard result, and I will follow the steps of Benabou and Laroque (1992) to show that $\lim_{T\to\infty} E(\rho_T|\rho_t, H) = 1$. Define $f(\rho) = E(\rho_{t+1}|\rho_t = \rho, H)$, and $f(\rho) \geq \rho$ with strict inequality for $\rho \in (0,1)$. Therefore, $\{1-\rho_t\}_{t\in N}$ is a

super-martingale, and would converge to a nonnegative random variable $1-\rho_\infty$. Let μ_t denote the distribution of ρ_t, and $E(\rho_\infty)=\int_0^1 f(\rho)d\mu_\infty$. Then ρ_∞ would be 0 or 1 almost surely, otherwise, $E(\rho_\infty)>\int_0^1 f(\rho)d\mu_\infty$.

Consider the process $\{y_t=\frac{1}{\rho_t}\}_{t\in N}$, which is a super-martingale, and would converge to a finite variable almost surely. Thus, $\{\rho_t\}$ cannot converge to 0, or $\rho_\infty=1$. Following similar steps, $\lim_{T\to\infty}E(\rho_T|\rho_t,N)=0$.

Proof of Lemma 3. Let $W^s(\rho_t;\beta)$ denote the expected payoff with reputation ρ_t and discount factor β, when there are s periods remains. To prove this lemma, I claim that in any finite game with s periods left, the probability of telling the truth after the unfavorable signal $M^s(-1,-1;\rho_t,\beta)$ and the expected future payoff $W^s(\rho_t;\beta)$ increases with β.

I will begin with the two period case, and extend to the infinite horizon. When there is only one period left, the normal type will announce $m_t=1$ for sure, and $W^1(\rho_t;\beta)=u(w+(2p-1)\rho_t)$ increases in ρ_t. Let $F^s_{-1}(M^s(-1,-1;\rho_t,\beta))$ denote the opportunity cost of announcing $m_t=-1$ after signal $s_t=-1$ when there are s periods.

$$
\begin{aligned}
&F^s_{-1}(M^s(-1,-1;\rho_t,\beta);\rho_t,\beta)=u(w+(2p-1)(\rho_t+(1-\rho_t)M^s(-1,-1;\rho_t,\beta)))\\
&+\beta(pW^{s-1}(\frac{\rho_t}{1+(1-\rho_t)\frac{p}{1-p}(1-M^s(-1,-1;\rho_t,\beta))})+(1-p)W^{s-1}(\frac{\rho_t}{1+(1-\rho_t)\frac{1-p}{p}(1-M^s(-1,-1;\rho_t,\beta))}))\\
&-u(w-(2p-1)(\rho_t+(1-\rho_t)M^s(-1,-1;\rho_t,\beta)))-\beta(pW^{s-1}(\frac{\rho_t}{\rho_t+(1-\rho_t)M^s(-1,-1;\rho_t,\beta)}))
\end{aligned}
\tag{1.8.4}
$$

Now check the condition for $s=2$. $F^2_{-1}(M^2(-1,-1;\rho_t,\beta);\rho_t,\beta)$ increases with $M^2(-1,-1;\rho_t,\beta)$ and decreases with β. Therefore, the solution $M^2(-1,-1;\rho_t,\beta)$ to $F^2_{-1}(M^2(-1,-1;\rho_t,\beta);\rho_t,\beta)=0$ increases with β, and the equilibrium strategy

$$\tilde{M}^2(-1,-1;\rho_t,\beta)=\max\{0,(M^2(-1,-1;\rho_t,\beta)|F^2_{-1}(M^2(-1,-1;\rho_t,\beta);\rho_t,\beta)=0)\}.$$

When there are two periods remained, the payoff would be

$$
\begin{aligned}
&W^2(\rho_t;\beta)=u(w+(2p-1)(\rho_t+(1-\rho_t)\tilde{M}^2(-1,-1;\rho_t,\beta)))\\
&+\beta(\frac{1}{2}W^1(\frac{\rho_t}{1+(1-\rho_t)\frac{p}{1-p}(1-\tilde{M}^2(-1,-1;\rho_t,\beta))})+\frac{1}{2}W^1(\frac{\rho_t}{1+(1-\rho_t)\frac{1-p}{p}(1-\tilde{M}^2(-1,-1;\rho_t,\beta))}))
\end{aligned}
\tag{1.8.5}
$$

where $W^2(\rho_t;\beta)$ increases with both ρ_t and β.

Now, assume the statement is true for $T-1$ period case, and check the T case. It is clear that $F^T_{-1}(M^T(-1,-1;\rho_t,\beta);\rho_t,\beta)$ increases with $M^T(-1,-1;\rho_t,\beta)$ and decreases with β. The equilibrium strategy $\tilde{M}^T(-1,-1;\rho_t,\beta)$ will be $\max\{0,(M^T(-1,-1;\rho_t,\beta)|F^T_{-1}(M^T(-1,-1;\rho_t,\beta);\rho_t,\beta) = 0)\}$ increases with β. When there are T periods remained, the payoff would be

$$W^T(\rho_t;\beta) = u(w+(2p-1)(\rho_t+(1-\rho_t)\tilde{M}^T(-1,-1;\rho_t,\beta)))$$
$$+\beta(\frac{1}{2}W^{T-1}(\frac{\rho_t}{1+(1-\rho_t)\frac{p}{1-p}(1-\tilde{M}^T(-1,-1;\rho_t,\beta))})+\frac{1}{2}W^{T-1}(\frac{\rho_t}{1+(1-\rho_t)\frac{1-p}{p}(1-\tilde{M}^T(-1,-1;\rho_t,\beta))}))$$
(1.8.6)

where $W^T(\rho_t;\beta)$ increases with both ρ_t and β.

1.8.2 Uncertainty in The Quality of the Signal

Here is the updated reputation for the expert, after history $\{\rho_t, m_t, w_t\}$.

$$\rho_{t+1} = \begin{cases} \frac{\rho_t}{1+\frac{1-\rho_t}{2p}[(1-p-\epsilon)(1-M(-1,-1;\rho_t,\tau^h))+(1-p+\epsilon)(1-M(-1,-1;\rho_t,\tau^l))]} \text{ if } m_t=1,\ w_t=1 \\ \frac{\rho_t}{1+\frac{1-\rho_t}{2(1-p)}[(p+\epsilon)(1-M(-1,-1;\rho_t,\tau^h)+(p-\epsilon)(1-M(-1,-1;\rho_t,\tau^l))]} \text{ if } m_t=1,\ w_t=-1 \\ \frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2p}[(p+\epsilon)M(-1,-1;\rho_t,\tau^h)+(p-\epsilon)M_t(-1,-1;\rho_t,\tau^l)]} \text{ if } m_t=-1,\ w_t=-1 \\ \frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2(1-p)}[(1-p-\epsilon)M(-1,-1;\rho_t,\tau^h)+(1-p+\epsilon)M(-1,-1;\rho_t,\tau^l)]} \text{ if } m_t=-1,\ w_t=1 \end{cases}$$
(1.8.7)

This reputation updating rule shows that the reputation after announcing $m_t=1$ increases with the probability of truthfully reporting after $s_t=-1$, and the reputation after $m_t=-1$ decreases with it. Moreover, the ranking among the updated reputation would be $\rho_{t+1}(1,-1) < \rho_{t+1}(1,1) < \rho_t < \rho_{t+1}(1,-1), \rho_{t+1}(1,-1)$. The relationship between $\rho_{t+1}(1,-1)$ and $\rho_{t+1}(-1,-1)$ depends on whether $M(-1,-1;\rho_t,\tau^h)$ or $M(-1,-1;\rho_t,\tau^l)$ is larger.

The expected size of audience would be

$$\phi(\rho_t) = (2p-1)\rho_t + \frac{\rho_t}{2}((2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h) + (2(p-\epsilon)-1)M(-1,-1;\rho_t,\tau^l) \tag{1.8.8}$$

$\phi(\rho_t)$ increases with both $M(-1,-1;\rho_t,\tau^h)$ and $M(-1,-1;\rho_t,\tau^l)$, since the credibility improved with the probability of telling the truth.

The corresponding payoff $V_{\tau_t,s_t,m_t}(M;\rho_t)$ for announcing m_t after $s_t = -1$ would be

$$\begin{aligned}
V_{\tau^h,-1,1}(\rho_t) &= u(w + (2p-1)\rho_t + \frac{1-\rho_t}{2}((2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h) + (2(p-\epsilon)-1)M(-1,-1;\rho_t,\tau^l)) \\
&+ \beta((p+\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2(1-p)}((p+\epsilon)(1-M(-1,-1;\rho_t,\tau^h)) + (p-\epsilon)(1-M(-1,-1;\rho_t,\tau^l)))}) \\
&+ (1-p-\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2p}((1-p-\epsilon)(1-M(-1,-1;\rho_t,\tau^h)) + (1-p+\epsilon)(1-M(-1,-1;\rho_t,\tau^l)))})) \qquad (1.8.9) \\
V_{\tau^h,-1,-1}(\rho_t) &= u(w - (2p-1)\rho_t - \frac{1-\rho_t}{2}((2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h) + (2(p-\epsilon)-1)M(-1,-1;\rho_t,\tau^l)) \\
&+ \beta((p+\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2p}((p+\epsilon)M(-1,-1;\rho_t,\tau^h) + (p-\epsilon)M(-1,-1;\rho_t,\tau^l))}) \\
&+ (1-p-\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2(1-p)}((1-p-\epsilon)M(-1,-1;\rho_t,\tau^h) + (1-p+\epsilon)M(-1,-1;\rho_t,\tau^l))})) \qquad (1.8.10) \\
V_{\tau^l,-1,1}(\rho_t) &= u(w + (2p-1)\rho_t + \frac{1-\rho_t}{2}((2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h) + (2(p-\epsilon)-1)M(-1,-1;\rho_t,\tau^l)) \\
&+ \beta((p-\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2(1-p)}((p+\epsilon)(1-M(-1,-1;\rho_t,\tau^h)) + (p-\epsilon)(1-M(-1,-1;\rho_t,\tau^l)))}) \\
&+ (1-p+\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2p}((1-p-\epsilon)(1-M(-1,-1;\rho_t,\tau^h)) + (1-p+\epsilon)(1-M(-1,-1;\rho_t,\tau^l)))})) \qquad (1.8.11) \\
V_{\tau^l,-1,-1}(\rho_t) &= u(w - (2p-1)\rho_t - \frac{1-\rho_t}{2}((2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h) + (2(p-\epsilon)-1)M(-1,-1;\rho_t,\tau^l)) \\
&+ \beta((p-\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2p}((p+\epsilon)M(-1,-1;\rho_t,\tau^h) + (p-\epsilon)M(-1,-1;\rho_t,\tau^l))}) \\
&+ (1-p+\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2(1-p)}((1-p-\epsilon)M(-1,-1;\rho_t,\tau^h) + (1-p+\epsilon)M(-1,-1;\rho_t,\tau^l))})) \qquad (1.8.12)
\end{aligned}$$

Proof of Lemma 4. I begin with the case $\rho_t \to 1$. Since the expert reputation would not drop much after either message, announcing $m_t = 1$ is preferred after both

signals. Then the payoff after signal $s_t = -1$ would be

$$V_{\tau^h,-1,1}(\rho_t) = u(w + (2p-1)\rho_t) + \beta((p+\epsilon)W(\frac{\rho_t}{1+\frac{p}{1-p}(1-\rho_t)}) + (1-p-\epsilon)W(\frac{\rho_t}{1+\frac{1-p}{p}(1-\rho_t)})) \tag{1.8.13}$$

$$V_{\tau^h,-1,-1}(\rho_t) = u(w - (2p-1)\rho_t) + \beta W(1) \tag{1.8.14}$$

$$V_{\tau^l,-1,1}(\rho_t) = u(w + (2p-1)\rho_t) + \beta((p-\epsilon)W(\frac{\rho_t}{1+\frac{p}{1-p}(1-\rho_t)}) + (1-p+\epsilon)W(\frac{\rho_t}{1+\frac{1-p}{p}(1-\rho_t)})) \tag{1.8.15}$$

$$V_{\tau^l,-1,-1}(\rho_t) = u(w - (2p-1)\rho_t) + \beta W(1) \tag{1.8.16}$$

This shows that $V_{\tau^h,-1,1}(\rho_t) - V_{\tau^h,-1,-1}(\rho_t) < V_{\tau^l,-1,1}(\rho_t) - V_{\tau^l,-1,-1}(\rho_t)$, or the marginal cost of announcing $m_t = 1$ is higher for the τ^h type. Therefore, when both marginal costs decrease with ρ_t, there exists a $\bar{\rho}^h$ for each W, such that $0 = V_{\tau^h,-1,1}(\rho_t) - V_{\tau^h,-1,-1}(\rho_t) < V_{\tau^l,-1,1}(\rho_t) - V_{\tau^l,-1,-1}(\rho_t)$. The expert starts to mix after $s_t = -1$ of the high type, when ρ_t drops below $\bar{\rho}^h$. For the updated reputation $\rho_{t=1}$, the order would be $\rho_{t+1}(1,-1) < \rho_{t+1}(1,1) < \rho_t < \rho_{t+1}(-1,-1) < \rho_{t+1}(-1,1)$. As a result, the order for the payoff would be $V_{\tau^h,-1,1}(\rho_t) < V_{\tau^l,-1,1}(\rho_t)$ and $V_{\tau^h,-1,-1}(\rho_t) < V_{\tau^l,-1,-1}(\rho_t)$. The payoff can be written as below

$$\begin{aligned}
&V_{\tau^h,-1,1}(\rho_t) = u(w + (2p-1)\rho_t + \frac{1-\rho_t}{2}(2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h))+ \\
&\beta((p+\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2(1-p)}(2p-(p+\epsilon)M(-1,-1;\rho_t,\tau^h))})+ \\
&(1-p-\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2p}(2(1-p)-(1-p-\epsilon)M(-1,-1;\rho_t,\tau^h))}))
\end{aligned} \tag{1.8.17}$$

$$\begin{aligned}
&V_{\tau^h,-1,-1}(\rho_t) = u(w - (2p-1)\rho_t - \frac{1-\rho_t}{2}(2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h))+ \\
&\beta((p+\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2p}(p+\epsilon)M(-1,-1;\rho_t,\tau^h)}) + (1-p-\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2(1-p)}(1-p-\epsilon)M(-1,-1;\rho_t,\tau^h)}))
\end{aligned} \tag{1.8.18}$$

$$\begin{aligned}
&V_{\tau^l,-1,1}(\rho_t) = u(w + (2p-1)\rho_t + \frac{1-\rho_t}{2}(2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h))+ \\
&\beta((p-\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2(1-p)}(2p-(p+\epsilon)M(-1,-1;\rho_t,\tau^h))})+ \\
&(1-p+\epsilon)W(\frac{\rho_t}{1+\frac{1-\rho_t}{2p}(2(1-p)-(1-p-\epsilon)M(-1,-1;\rho_t,\tau^h))}))
\end{aligned} \tag{1.8.19}$$

$$\begin{aligned}
&V_{\tau^l,-1,-1}(\rho_t) = u(w - (2p-1)\rho_t - \frac{1-\rho_t}{2}(2(p+\epsilon)-1)M(-1,-1;\rho_t,\tau^h))+ \\
&\beta((p-\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2p}(p+\epsilon)M(-1,-1;\rho_t,\tau^h)}) + (1-p+\epsilon)W(\frac{\rho_t}{\rho_t+\frac{1-\rho_t}{2(1-p)}(1-p-\epsilon)M(-1,-1;\rho_t,\tau^h)}))
\end{aligned} \tag{1.8.20}$$

$M(-1,-1;\rho_t,\tau^h)$ is uniquely define by the equation $F_{\tau^h,-1}(\rho_t) = V_{\tau^h,-1,1}(\rho_t) - V_{\tau^h,-1,-1}(\rho_t) = 0$. With such $M(-1,-1;\rho_t,\tau^h)$, if $F_{\tau^l,-1}(\rho_t) = V_{\tau^l,-1,1}(\rho_t) - V_{\tau^l,-1,-1}(\rho_t) > 0$, the expert would announce $m_t = 1$ after the low type signal $s_t = 1$;

otherwise, the expert would mix after $s_t = -1$ of both type. Then the payoffs follow above, and $\{M(-1,-1;\rho_t,\tau_t)\}_{\tau_t=h,l}$ is uniquely defined by the two equation system $\{F_{\tau_t,-1}(\rho_t) = V_{\tau_t,-1,1}(M;\rho_t,W) - V_{\tau_t,-1,-1}(M;\rho_t,W) = 0\}_{\tau_t=h,l}$. Therefore, the analysis above shows that for each ρ_t, there is a vector $\{M(-1,-1;\rho_t,\tau_t)\}_{\tau_t=h,l}$ uniquely defined as lemma 4.

Define the operator $T(\rho_t;W) = \frac{1}{4}V_{\tau^h,-1,1}(\rho_t,W)+\frac{1}{4}V_{\tau^h,1,1}(\rho_t,W)+\frac{1}{4}V_{\tau^l,-1,1}(\rho_t,W)+\frac{1}{4}V_{\tau^l,1,1}(\rho_t,W)$. I will check the monotonicity for the value function $T(\rho_t;W)$. There are three cases: (1) $M(-1,-1;\rho_t,\tau^h) = M(-1,-1;\rho_t,\tau^l) = 0$; (2) $M(-1,-1;\rho_t,\tau^h) > M(-1,-1;\rho_t,\tau^l) = 0$; (3) $M(-1,-1;\rho_t,\tau^h) > M(-1,-1;\rho_t,\tau^l) > 0$. For case (1), $T(\rho_t;W) = u(w+(2p-1)\rho_t)+\beta(\frac{1}{2}W(\frac{\rho_t}{1+\frac{p}{1-p}(1-\rho_t)})+\frac{1}{2}W(\frac{\rho_t}{1+\frac{1-p}{p}(1-\rho_t)}))$, which obviously increases in ρ_t. Repeat the argument in lemma 2 on $M(-1,-1;\rho_t) > 0$, Case (2) can be proved as well. Consider case (3), with respect to two variables $M(-1,-1;\rho_t,\tau^h)$ and $\delta M(-1,-1;\rho_t) = M(-1,-1;\rho_t,\tau^h)-M(-1,-1;\rho_t,\tau^l)$. These two variables would move together with ρ_t. Consequently, by considering the vector $\{M(-1,-1;\rho_t,\tau^h),\delta M(-1,-1;\rho_t)\}$ and repeating the argument on lemma 2, we can also conclude that $T(\rho_t;W)$ increases with ρ_t.

Second, prove that $T(\rho;W)$ is nondecreasing in (ρ,W).

Consider $(\rho^1,W^1) \geq (\rho^2,W^2)$. If $(M(-1,-1;\rho^1,W^1,h),\delta M(-1,-1;\rho^1,W^1)) \geq (M(-1,-1;\rho^2,W^2,\tau^h),\delta M(-1,-1;\rho^2,W^2))$ or $W^1 = W^2$, $T(\rho^1;W^2) \geq T(\rho^1;W^2)$ is true for sure. Then look at the case, $(M(-1,-1;\rho^1,W^1,\tau^h),\delta M(-1,-1;\rho^1,W^1)) < (M(-1,-1;\rho^2,W^2,\tau^h),\delta M(-1,-1;\rho^2,W^2))$ and $W^1 > W^2$. However, $F_{-1,\tau^h}(\rho_t)$ decreases with W, and $0 = F_{-1,\tau^h}(M_t^1(-1,-1) : \rho_t,W^1) < F_{-1,\tau^h}(M_t^2(-1,-1) : \rho_t,W^2) = 0$, contradiction.

T maps C_+ continuously into itself, and is non decreasing in W. Moreover, $T(W+c) = T(W) + \beta c$ for any constant C. Therefore, by Blackwell's theorem, T is a contracting mapping, and since C_+ with sup norm is complete, it has a unique fixed point W. And the discussion above confirms that W is non decreasing in reputation.

Proof of Proposition 4. I have showed in this paper, that the expert will announce $m_t = 1$ after $s_t = 1$ of any type. I will check and verify this claim in a two period setting, with W as the nondecreasing utility function for the section period. Then, I will use the equilibrium to prove monotonicity, existence, and uniqueness of the value function W. The way to define the cutoff for pooling equilibrium $\bar{\rho}$ is exactly the same as the signal issue or double issue case, and the condition of $\rho_t < \bar{\rho}$ will be considered as below.

For $\rho_t < \bar{\rho}$, I claim that the expert would tell the truth for signals of high quality and lie after signals of low quality. If this is not true, then there exists some point $\hat{\tau}$ and δ, that the expert would tell the truth for $\tau_t \in [\hat{\tau} - \delta, \hat{\tau})$ and lie for $\tau_t \in (\hat{\tau}, \hat{\tau} + \delta]$. Then the expert would be tell the truth for $\tau_t \in [\tau^l, \hat{\tau})$ and lie for $\tau_t \in (\hat{\tau}, \tau^h]$. Then, the implied equilibrium would be the expert tells the truth below $\hat{\tau}$ while lying above $\hat{\tau}$, and the updated reputation for different history would follow such relationship $\rho_{t+1}(1,-1) < \rho_{t+1}(1,1) < \rho_t < \rho_{t+1}(-1,1) < \rho_{t+1}(-1,-1)$. Consequently, $W(\rho_{t+1}(-1,-1)) - W(\rho_{t+1}(1,-1)) > W(\rho_{t+1}(-1,1)) - W(\rho_{t+1}(1,1))$. Then

$$\frac{d(V_{\tau_t,-1,1}(\hat{\tau};\rho_t,W) - V_{\tau_t,-1,1}(\hat{\tau};\rho_t,W))}{d\tau_t} =$$
$$[W(\rho_{t+1}(-1,1)) - W(\rho_{t+1}(1,1))] - [W(\rho_{t+1}(-1,-1)) - W(\rho_{t+1}(1,-1))] < 0 \tag{1.8.21}$$

This equation means that the expert would tell the truth for $\tilde{p} > \hat{p}$,contradiction.

The credibility of the expert would be

$$\pi(\rho_t) = \frac{1}{2} + \frac{1}{2}((2p-1)\rho_t + (1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \tau_t dF(\tau_t)) \tag{1.8.22}$$

which decreases with $\hat{\tau}$. The updated reputation in the continuous case would be

$$\rho_{t+1} = \begin{cases} \frac{p\rho_t}{p+(1-\rho_t)\int_{\tau^l}^{\hat{\tau}} \frac{1-\tau_t}{2} dF(\tau_t)} \text{ if } m_t = 1,\ w_t = 1 \\ \frac{(1-p)\rho_t}{(1-p)+(1-\rho_t)\int_{\tau^l}^{\hat{\tau}} \frac{1-\tau_t}{2} dF(\tau_t)} \text{ if } m_t = 1,\ w_t = -1 \\ \frac{p\rho_t}{p\rho_t+(1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \frac{1+\tau_t}{2} dF(\tau_t)} \text{ if } m_t = -1,\ w_t = -1 \\ \frac{(1-p)\rho_t}{(1-p)\rho_t+(1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \frac{1+\tau_t}{2} dF(\tau_t)} \text{ if } m_t = -1,\ w_t = 1 \end{cases} \tag{1.8.23}$$

This updating rule leads to the ranking of $\rho_{t+1}(1,-1) < \rho_{t+1}(1,1) < \rho_t < \rho_{t+1}(-1,-1) < \rho_{t+1}(-1,1)$. The reputation after announcing $m_t = 1$ decreases with $\hat{\tau}$, while the reputation after announcing $m_t = -1$ increases with $\hat{\tau}$. The payoff for type $\hat{\tau}$ of announcing $m_t = 1$ would be

$$\begin{aligned} &V_{\hat{\tau},-1,1}(\hat{\tau};\rho_t,W) = u(w+(2p-1)\rho_t+(1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \tau_t dF(\tau_t)) \\ &+\beta(\frac{1-\hat{\tau}}{2} W(\frac{p\rho_t}{p+(1-\rho_t)\int_{\tau^l}^{\hat{\tau}} \frac{1-\tau_t}{2} dF(\tau_t)}) + \frac{1+\hat{\tau}}{2} W(\frac{(1-p)\rho_t}{(1-p)+(1-\rho_t)\int_{\tau^l}^{\hat{\tau}} \frac{1+\tau_t}{2} dF(\tau_t)})) \end{aligned} \tag{1.8.24}$$

$$\begin{aligned} &V_{\hat{p},-1,1}(\hat{p};\rho_t,W) = u(w-(2p-1)\rho_t-(1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \tau_t dF(\tau_t)) \\ &+\beta(\frac{1-\hat{\tau}}{2} W(\frac{(1-p)\rho_t}{(1-p)\rho_t+(1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \frac{1-\tau_t}{2} dF(\tau_t)}) + \frac{1+\hat{\tau}}{2} W(\frac{p\rho_t}{p\rho_t+(1-\rho_t)\int_{\hat{\tau}}^{\tau^h} \frac{1+\tau_t}{2} dF(\tau_t)})) \end{aligned} \tag{1.8.25}$$

The marginal cost of lying after the unfavorable signal $V_{\hat{\tau},-1,1}(\hat{\tau};\rho_t,W) - V_{\hat{\tau},-1,1}(\hat{p};\rho_t,W)$ is monotonically decreasing in $\hat{\tau}$. Therefore, a $\hat{\tau}$ value is uniquely decided by $V_{\hat{\tau},-1,1}(\hat{\tau};\rho_t,W) - V_{\hat{\tau},-1,1}(\hat{\tau};\rho_t,W) = 0$. Repeat the process, a non decreasing value function W exists, and is uniquely defined.

Chapter 2

The Credit Rating Game with Self Interested Issuers

2.1 Introduction

In the past few year, credit rating agencies (CRAs) have been blamed heavily for the subprime mortgage crisis. People believe that CRAs were too lax in the ratings of some structured products, and the AAA ratings for a large proportion of subprime residential mortgage-backed securities did support such over-rating arguments. Becker and Milbourn (2011) provides some empirical evidence on the over rating problem, while Skreta and Veldkamp (2009) provides some theory foundations. However, this conflict of interest between CRAs and investors is generated by the business model of CRAs, in which their principal source of revenue comes from the issuers not the investors. The issuer is willing to pay the CRA only if they can benefit from a rating. Therefore, the issuer's self-interest will affect the strategy of the CRAs.

One interesting fact about the issuers is that they have private benefit or loss after the failure of a project, and such private benefit or loss cannot be shared by the investor. There is a large literature on CRAs discussing the over-rating intermediaries,

initiated by Lizzeri (1999), and many elements related to the CRA's strategy have been covered. However, surprisingly, the role of the issuer's private benefit has not attracted much attention at all. In security markets and business operations, they do play a significant role in the issuer's decision making process. Consider the tax credit from a failed investment, the Research and Development spillover effect from a unsuccessful new medication, or the reputation loss after a collapsing affiliated regional factory. All these examples show that the failure of an investment could lead to positive or negative effects to the issuer, and consequently would affect the issuer's choice of approaching the CRA or not.

Another key factor about the issuer is the probability of default, or the quality of the project. The classical literature on reputation games with ideological preferences can be classified into two types. One type is the game with perfect private signals, in which an incorrect prediction will lead to an immediate reputation drop, such as Mailath and Samuelson (2001). The other type is the game with noisy signals, where the reputation will not be affected much if it is very high or very low, such as Benabou and Laroque (1992). In the credit rating games, the project could be a risk-free investment, or a risky one with a high premium. For example, the corporate bond of a well-established blue-chip company is almost as safe as a bank deposit, and a high yield junk bond is named after its high default risk. However, even a junk bond might have a significant probability of success. Therefore, there is asymmetry in the information quality, in which the blue-chip bond leads to a perfect signal and the junk bond means a noisy signal. When a new bond is issued in the market, the investor's belief is a randomization of both. Then, the question becomes which effect will dominate, the perfect signaling game or the noisy signaling game. My answer is that it depends on the probability of the default of the bad type.

My model builds on the standard reputation game literature like Benabou and Laroque (1992) as well as its more recent applications, such as Mathis, McAndrews,

and Rochet (2009). I consider a financial market in a two-period horizon with three players: the CRA, the issuer and the investor. The CRA is long lived while the issuer and the investor are short lived. In each period, the new issuer wants to raise cash for a new project. The project is equally likely to be good or bad. The good project is risk-free,while the bad project will default with a positive probability. The project's quality is unknown to the issuer and the investor, but the CRA can observe it. At the beginning of each period, the issuer will approach the CRA, and the CRA will propose a fee and offer a rating report according to his private signal. The issuer decides to purchase or not based on the fee and the report. After reading the rating, the investor chooses to invest or not and how much he pays for this project. If the investor refuses to purchase the security, he will get some reservation benefit of having cash on hand. At the end of the period, all three players can observe whether the project defaults or not. I assume the good project should always be financed, but no project will get the investment without the CRA. This creates a role for the CRA who has private information about the project and communicates a rating to the market. The CRA can be one of the two types: the honest type or the profit-maximizing type. His reputation is measured by the probability of being honest, and his strategy depends on his reputation.

The main contribution of my paper is introducing the CRA's private benefit. In my model, the investor will get a promised return after a success, while the issuer collects some private benefit (loss) after a default. When the private benefit is negative, the promised return is used to compensate both the issuer and the investor's reservation value. When the private benefit is larger than the promised return, the CRA would rather have a default than a success. However, the effect of the private benefit on the CRA's strategy does not depend on its value. Instead, the effect depends on the quality of the project. When the project is very unlikely to default, the credit rating game behaves like the noisy signaling game. Then the CRA is more likely to tell the

truth when the private benefit decreases. When the project is very likely to default, the credit rating game is similar to the perfect signaling game and the CRA is more likely to tell the truth when the private benefit increases.

Another contribution of the paper is showing that the CRA is more likely to give a good rating as long as he has a contract. This is consistent with the empirical evidence of CRA's over-rating actions as well as the theoretical prediction on the strategy of a biased sender's distorting behavior. However, the mechanism is different from those of the classical reputation games. In classical reputation games, such as Benabou and Laroque (1992), the assumption that the payoff increases with reputation is necessary for this conclusion. Based on this assumption and the sender's ideological preference, the sender is more likely to distort the message toward his biased position. However, in my model, the payoff may decrease with the CRA's reputation given some private benefit value, and the conclusion still remains true. The fact that only bad projects default is sufficient for this result. Therefore, by introducing the issuer's private benefit, I can show that fewer restrictions are required for the over-rating result.

This paper belongs to the literature in reputation games, and is closely related to Benabou and Laroque (1992), Mathis, McAndrews, and Rochet (2009) and Bolton, Freixas, and Shapiro (2012). Benabou and Laroque (1992) shows how the insider provides manipulative recommendations to investors and makes profit from the fluctuation of stock prices. Similar to his model, the biased CRA in my paper also benefit from his private information as well as the existence of a commitment type. A more recent application of the reputation game in credit rating agencies (CRAs) is Mathis, McAndrews, and Rochet (2009). Mathis, McAndrews, and Rochet (2009) uses an infinite horizon game to model the CRA's behavior, and is focusing on the effect of the CRA's other revenue sources. They find that as long as the revenue from the non-rating operation is sufficient, the CRA always tells the truth. Both papers use the monotonic relationship between the reputation and the sender's payoff to build

up their main results, while my model relieves this assumption. Bolton, Freixas, and Shapiro (2012) does not use a commitment type to leverage the benefit, and instead, they discussed the competition among multiple CRAs. However, they did consider the effect of the quality of an investment as I do in this paper, but their conclusion was focusing on a different direction.

This paper is also related to the literature in financial intermediaries. Lizzeri (1999) considers a single period game, and discusses what privately informed parties will reveal to uninformed parties. His focus is on the strategic manipulation of information by the certification intermediaries. His result is the foundation of the binary state in my model. Even if the original information is continuous, the monopoly intermediary chooses to reveal only whether quality is above some minimal standard. The intermediaries only provide a binary result, and that is why I choose the binary environment for the market. Kuhner (2001) and Faure-Grimaud, Peyrache, and Quesada (2009) also discuss the role of financial intermediaries in the one period framework. Faure-Grimaud, Peyrache, and Quesada (2009) identifies the optimal contract between a rating agency and a firm, and they analyze how the ownership contracts affect the optimal solution. They show that the CRA will fully disclose information at the equilibrium. Kuhner (2001) claims that the CRAs are more credible if their ratings cannot become self-fulfilling from an ex-post point of view.

The reputation building process of this paper is inherited from the reputation game literature. This literature is initiated by Crawford and Sobel (1982) and Sobel (1985), who study the sender-receiver game in the finite repeated game with perfect monitoring. Then Benabou and Laroque (1992) and Sharfstein and Stein (1990) study the application of the reputation game in finance. Sharfstein and Stein (1990) is focusing on the uncertainty in expertise with two experts in the financial market, who would use herding to share the blame. Benabou and Laroque (1992) studies a reputation game with noisy signals and shows how good reputation helps the insider

exploit profits from investors. Morris (2001) and Ely and Valimaki (2003) explain how the strategic good advisor is forced to lie to enhance his reputation. Ottaviani and Sorensen (2006a) and Ottaviani and Sorensen (2006b) extend Sobel's reputation game with multiple experts and uncertainty in expertise. Ottaviani and Sorensen (2006c) is specifically focusing on the application on the financial forecasting. This paper develops the theory of reputation cheap talk and shows how the forecasters endeavor to convince the market that they are well informed.

The remainder of this paper is organize as follows. In Section 2.2, I describe a general setup of the model and also define the reputation equilibrium. In Section 2.3, I analyze the basic model with a positive private benefit for the issuer. Section 2.4 discusses the role of the private benefit in the CRA's decision making process. Finally Section 2.5 concludes.

2.2 The Model

Consider a two period game with three types of risk-neutral agents: issuers, CRAs, and investors. In each period, a cashless firm (issuer) wants to issue a security to finance an investment. The quality of investment is unknown, and the investment characterized by its probability of default. Good projects always succeed. Bad projects may succeed with probability λ, or fail with probability $1-\lambda$. A successful project would yield R to the investor and 0 to the issuer. This means that the issuer would hand all the benefit to the investor. A failed project yields 0 benefit to the investor, but a positive return of r to the issuer. When the project fails, the issuer can decline any request on return from the investor. At the same time, the issuer would enjoy tax benefit and spill-over effect from the R&D expense on this project. I use r to measure this private benefit to the issuer. Either the investor or the issuer can collect the return if and only if the investment actually happens. If the investor refused to

purchase the security, every party ends up with 0. The investment is equally likely to be good or bad. This uncertainty of the investment creates a role for the credit rating agency (CRA), which can perfectly observe the quality of the project and communicate a rating to the investor. I normalize the cost of this private signal to be 0.

The CRA is a long-run player with a discount factor β, and the payoff would be the discounted sum of stage game payoffs. Issuers and investors are short-run players, who care about the current payoff only. In the market, a CRA would have businesses with a different issuer in each period, and the CRA's reputation would affect the actual contract between the issuer and the CRA. There are infinitely many investors in the market, and only their aggregate decision matters. Consequently, I use a short-live investor to model the aggregate behavior of the buy side.

The CRA can be one of the two types: a honest type or a normal type. The honest type always tells the truth, and the normal type maximizes the continuation payoff. I use H and N to denote these two types respectively. The issuer and the investor are uncertain about the CRA's type. I use p_t to denote the probability that the CRA is the honest type and p_1 is the prior reputation. The issuer and the investor observe the same public history, whether the project is financed or not and whether it is successful or not. They share the common belief, and will update their belief on the reputation p_t accordingly.

The timing of the game in the period t is as follows. The CRA posts their fee ϕ_t at the beginning of the period, at which a rating can be purchased. If an issuer approaches the CRA, the CRA will obtain the signal $s_t \in \{G, B\}$ and produce a credit report $m_t \in \{g, b\}$ accordingly. After observing the report m_t, the issuer decides whether to purchase and distribute the report or not. According to his expected return, the investor will post a price θ_t to the issuer. The cost of the investment is also normalize to be 0, and θ_t is a spread for the issuer. The investor has a reservation

value of 1. Thus, the price θ_t measures the net gain beyond this reservation value, and it depends on the reputation p_t and the actual report m_t. If the expected return is below the reservation value, there will be no investment at all. Consequently, no rating is necessary for this project. At the end of this period, all three agencies will observe the true state $w_t \in \{s, f\}$, if the security has been successfully issued. If it is not issued, the state $w_t = \varnothing$, or no information about the true state will be disclosed. I use w_t to denote whether the investment succeeds or fails. Both the issuer and the investor will update their belief on the reputation according to the public history H_t. At the beginning of the game the public history includes only the prior belief $\{p_1\}$. In period $t+1$, it also includes the information $\{m_t, w_t\}$. Let H_t denote the history up to period t, and then $H_{t+1} = \{p_1, ..., m_t, w_t\}$.

The issuer benefits from obtaining a good rating on the project as well as hiring a CRA with a high reputation of being honest, since the investor's willingness to pay depends on both. Therefore, the contract between the issuer and the CRA will reflect both incentives. In a multiple-period game, reputation costs create an incentive for the CRA to tell the truth, since the short-run issuer and the short-run investor can learn the type from the public history. However, the issuer would not be willing to honor the contract without a good rating at all, and only the security with a good rating can survive in the market. As a result, the CRA has to balance between these two incentives for an optimal strategy.

The investor chooses to purchase one unit of the investment or not. There is a reservation utility for the investor, which is normalized to be 1. The investor could hold his money in cash. Therefore, a larger return is necessary for the commitment of putting lots of money in this investment vehicle. I have the following assumptions on the return of the investment:

Assumption 1. $\frac{\lambda+1}{2}R < 1$.

Assumption 2. $R > 1$.

The first assumption shows that there will be no investment if no information is revealed. It also implies that no investment would happen without the CRA. The high probability of the investment being bad creates a role for the CRA, who could increase the market efficiency. Consequently, the CRA profits from this information provision process. If there is no rating on the investment, or $m_t = \varnothing$, it also conveys information to the investor. However, the investment would not happen at all under such a condition. To simplify the game, I let $m_t = \{g, b\}$, and claim that there is no investment after $m_t = b$. Therefore, there will be no w_t revealed after $m_t = b$. The second assumption says that an investor is always willing to buy a good investment. When the project's probability of being good is above a certain cutoff, the investment will happen.

Define $\pi(p_t) = Pr(w_t = s|m_t = g, p_t)$, the probability of success after a good rating given the reputation p_t. In the later part, I use p_t instead of H_t to represent the history, since p_t contains all necessary information. Then, in each period, the investor posts the price θ_t according to their expected net return, which can be written as

$$\theta(\pi(p_t)) = \begin{cases} R\pi(p_t) - 1 \text{ if } R\pi(p_t) - 1 \geq 0 \\ 0 \text{ otherwise.} \end{cases} \tag{2.2.1}$$

This equation shows that θ measures the investor's maximum willingness to pay. If the expected return is lower than the reservation value, no investment will happen at all.

The CRA will be able to collect all the profit, and ϕ makes the issuer indifferent between hiring the CRA or not. The instantaneous payoff for the CRA after sending

a good message would be

$$\phi(\pi(p_t)) = \begin{cases} R\pi(p_t) + r(1-\pi(p_t)) - 1 \text{ if } R\pi(p_t) - 1 \geq 0 \\ 0 \text{ otherwise.} \end{cases} \tag{2.2.2}$$

As mentioned above, the honest type will play non-strategically, and report his signal truthfully. The normal type always wants to send a good report to collect the fee, but he also wishes to get a higher fee from an investment. I use $\alpha_t(s_t; p_t) \in [0,1]$ to define the probability of the normal type providing a good report after a signal s_t at the time t. The CRA's strategy could be described as $\alpha_t(p_t) = (\alpha_t(g; p_t), \alpha_t(b; p_t)) \in [0,1]^2$. Let β denote the discount factor, and the inter-temporal payoff for the CRA after signals $\{s_1, s_2\}$ would be

$$U(\alpha, p_1 | s_1, s_2) = \alpha_1(s_1; p_1)\phi(\pi(p_1)) + \beta\alpha_2(s_2; p_2)\phi(\pi(p_2)) \tag{2.2.3}$$

where p_2 denote the reputation updated after $\{p_1, m_1, w_1\}$.

The payoff function implies that the CRA will extract all the benefit, both from the investor and issuer. It also shows that the CRA has reputation concern, since the payoff in the second period depends on the public history or his updated reputation from the first period. But how the reputation affects the payoff depends on the relationship between R and r. If the return R to the investor in a successful project is larger than the private benefit r in a failed project for the issuer, than the effect of R dominates in the CRA's payoff. In that case, the payoff will be 0 before $R\pi(p_t)$ reaches the reservation value 1, and increasing with $\pi(p_t)$ afterward. Overall, the payoff is increasing with π. However, when the private benefit r is more than the promised return R, the payoff is decreasing with π after an actual investment. This would eliminate the monotonicity, and the standard reputation concern in the information transmission game would not function as usual.

At the end of the first period, all three players would observe one of the three possible outcomes: Success($w_1 = s$) when a good project is finance, or a bad projects is financed and gets lucky; Failure($w_t = f$) when a bad project is finance and gets unlucky; or No finance($w_t = \varnothing$). With the strategy $\alpha_1(s_1, p_1)$ defined above, the posterior reputation p_2 in the second period would be

$$p_2(m_1 = g, w_1 = s; p_1) = \frac{p_1}{p_1 + (1 - p_1)(\alpha_1(g; p_1) + \alpha_1(b; p_1)\lambda)} \tag{2.2.4}$$

$$p_2(m_1 = g, w_1 = f; p_1) = 0 \tag{2.2.5}$$

$$p_2(m_1 = b, w_1 = \varnothing; p_1) = \frac{p_1}{p_1 + (1 - p_1)(2 - \alpha_1(g; p_1) - \alpha_1(b; p_1))} \tag{2.2.6}$$

This reputation updating rule shows that a failure after financing will lead to 0 reputation. However, a success after financing or no financing may lead to a higher or lower reputation, and it depends on the actual strategy implemented. With the strategy α_t, the reputation p_t and the reputation updating rule, I can define the equilibrium as follows:

Definition 2.2.1. *A strategy $\{\alpha_t\}$ together with the reputation $\{p_t\}$ is the equilibrium of this credit rating game if*

(1) it is a perfect Bayesian equilibrium,

(2) $\alpha_1(p_1)$ maximized the inter-temporal profit of the CRA,

(3) $\alpha_2(p_2)$ maximized the profit at period 2.

In an equilibrium of this credit rating game, a normal CRA maximizes profits, investor's and issuer's expectations are correct, and they update their beliefs rationally. In what follows, I will refer such a strategy $\{\alpha_t\}$ as the equilibrium and a payoff at the equilibrium is uniquely determined at each reputation p_t. In the next section, I will focus on the baseline case, where the private benefit r is strictly smaller than the return R.

2.3 The Baseline Case

I begin with the benchmark case, in which the promised return R is larger than the private benefit r. As discussed in the previous section, the benefit from promised return R will dominate in calculating the CRA's payoff. In the market, after a failed project, the issuer could at least recover some lost from tax benefit, which would lead to a nonzero r value. However, the tax benefit and other spillover effect may not pass the total benefit from a successful project. Therefore, $r < R$ would be a more common case in a credit rating game.

2.3.1 Inference Within and Across Periods

The rational issuer and investor would use their prior in the CRA's type p_t, and their expectation of the normal type's strategy $\alpha_t(s_t; p_t)$, to infer the probability of success after a good rating

$$\pi_t(p_t) = \frac{p_t + (1-p_t)(\alpha_t(g;p_t) + \alpha_t(b;p_t)\lambda)}{p_t + (1-p_t)(\alpha_t(g;p_t) + \alpha_t(b;p_t))} \tag{2.3.1}$$

The equation above shows that π_t increases with the reputation p_t and the probability of truth telling after good signal $\alpha_t(g;p_t)$, but decrease with $\alpha_t(b;p_t)$. The fact that π_t increases in p_t together and $R > r$ create the reputation incentive for the CRA. Since lying would hurt the instantaneous payoff, π_t is decreasing with $\alpha_t(b;p_t)$.

2.3.2 Analysis of The Reputation Equilibrium

In this section, I will use backward induction to solve the two period rating game. In the second period, the CRA does not have reputation concern, and sending a good rating is a dominant strategy. There will be a lower bond for $\bar{p}_2$, and there will be no investment when p_2 falls below this boundary. Here is a lemma describing the strategy and payoff in the second period.

Lemma 2.3.1. *The normal CRA always announce $m_2 = g$ in the second period. Define $\bar{p}_2 = 1 - \frac{R-1}{1-R\lambda}$*

(1) When $p_2 < \bar{p}_2$, there is no investment in the second period and the payoff is 0.

(2) When $p_2 \geq \bar{p}_2$, the investor will purchase the project and CRA's instantaneous payoff is $\frac{(1-p_2)(1-\lambda)}{2-p_2} r + \frac{p_2+(1-p_2)(1+\lambda)}{2-p_2} R - 1$.

According to the lemma above, the payoff is discontinuous at the cutoff point $1 - \frac{R-1}{1-R\lambda}$, and $\phi(p_2)$ equals to $\frac{1}{R}$ at this point. Once the reputation p_2 passes this cutoff, the payoff will jump from 0 to a positive value immediately. This cutoff defines the reputation where the investor will start to take the project. The rational investor knows that the normal CRA's rating does not contain any information at all. So they only value the probability that the CRA is honest. The reservation value for the investor leads to the discontinuity in utility, while the monotonicity comes from the classical reputation game settings. Such discontinuity creates incentive for the CRA to keep the reputation stay above this cutoff.

Let V_{s_1,m_1} denote the discounted expected sum of the normal type's utility after announcing m_1 with signal s_1. Then V_{s_1,m_1} for $(s_1, m_1) \in \{g, b\}^2$ can calculated as below

$$V_{g,g}(p_1) = \phi(\frac{p_1 + (1-p_1)(\alpha_1(g,p_1) + \lambda\alpha_1(g,p_1))}{p_1 + (1-p_1)(\alpha_1(g,p_1) + \alpha_1(g,p_1))}) + \beta\phi(\pi_2(\frac{p_1}{p_1 + (1-p_1)(\alpha_1(g;p_1) + \alpha_1(b;p_1)\lambda)})) \tag{2.3.2}$$

$$V_{b,g}(p_1) = \phi(\frac{p_1 + (1-p_1)(\alpha_1(g,p_1) + \lambda\alpha_1(g,p_1))}{p_1 + (1-p_1)(\alpha_1(g,p_1) + \alpha_1(g,p_1))}) + \lambda\beta\phi(\pi_2(\frac{p_1}{p_1 + (1-p_1)(\alpha_1(g;p_1) + \alpha_1(b;p_1)\lambda)})) \tag{2.3.3}$$

$$V_{s_1,b}(p_1) = \beta\phi(\pi_2(\frac{p_1}{p_1 + (1-p_1)(2 - \alpha_1(g;p_1) - \alpha_1(b;p_1))})) \tag{2.3.4}$$

The payoff monotonically increases with the reputation, and therefore, the CRA has reputation concern. Or more specifically, the CRA will have better reputation after providing a bad rating. This leads to the result $\alpha_1(g,p_1) + \alpha_1(g,p_1) \geq 1$, or on average the CRA is more likely to give good ratings. Here is the argument. If this is not true, than $\alpha_1(g,p_1) + \alpha_1(g,p_1) < 1$ and $\alpha_1(g,p_1) + \lambda\alpha_1(g,p_1) < 1$. Consequently, the reputation would drop after a bad rating, and go up after a good rating with

a success. It implies that a good rating is strictly preferred after a good signal, or $\alpha_1(g, p_1) = 1$, contradiction.

Next, I will discuss whether the CRA could mix after both signal, nor just one of them. I begin with the condition that the CRA may randomize after both signal, and claim that the CRA will get 0 payoff in the second period after a good rating. With probability $1-\lambda$, sending a good rating after a bad signal will lead to 0 reputation and 0 utility. It means that the utility of sending a good message after a good signal is at least as good as sending a good message after a bad signal. However, no true state is revealed after a bad rating, so the utility after a bad message is not state dependent. In order to randomize after both signals, the utility after a good rating cannot be state depend as well. Therefore, the second period payoff after a good rating in the first period must be 0. If the updated reputation after a successful financing is below the cutoff in the second period, then the CRA will get 0 payoff in the second period. Thus, the CRA will have the same payoff after the good message regardless to the true state revealed to the public. Consequently, the CRA may mix after both signals. This result is driven by two assumptions. First, a default after a good rating leads to 0 utility in the second period. Second, the reservation value creates a cutoff in reputation, and the second period utility is 0 once the updated reputation is below the cutoff.

Now, consider the case that the CRA randomizes after at most one of the signals. If the updated reputation after a successful finance is above the cutoff, $V_{g,g}(p_1) > V_{b,g}(p_1)$ and the CRA cannot randomize after two signal at the same time. If the CRA randomize after the good signal, he would strictly prefer send bad rating after bad signals. On average, he is more likely to send ratings than good ratings contradiction. Thus, the expert would always give a good rating after a good signal.

According to the definition of V_{s_1,m_1}, the one-stage deviation principle can be written as

(1) If $V_{s_t,g}(p_1) > V_{s_t,b}(p_1)$, $\alpha_1(s_1, p_1) = 1$ is part of the equilibrium.

(2) If $V_{s_t,g}(p_1) < V_{s_t,b}(p_1)$, $\alpha_1(s_1, p_1) = 0$ is part of the equilibrium.

(3) If $V_{s_t,g}(p_1) = V_{s_t,b}(p_1)$, $\alpha_1(s_1, p_1) \in [0, 1]$ is part of the equilibrium.

The analysis above defines the one-stage deviation principle in the equilibrium. Together with the Bayesian reputation updating rule and the payoff maximizing property, the equilibrium payoff is well defined. However, the equilibrium strategy may not be unique when the payoff is 0, or no investment happens at all. The lemma below will discuss the 0 payoff case.

Proposition 2.3.1. *There exists a value p^*, such that if $p_1 < p^*$, the CRA's payoff is 0.*

Here is the sketch of the proof. I begin with a p_1 very close to 0, and will show that the payoff is 0. Case 1: If updated reputation after a bad message is above $\bar{p}_2$, then $\alpha_1(g, p_1)$ and $\alpha_1(a, p_1)$ are very close to 1. It means that the normal type is very unlikely to give a bad rating. As a result, the updated reputation after a successful financing will be below $\bar{p}_2$ for sure, and the payoff in the first period is also 0. This payoff means that the CRA can be better off by sending a bad rating, contradiction. Case 2: If the updated reputation after a bad message is below $\bar{p}_2$, but the payoff after good rating is positive. Then the CRA would strictly prefer to send a good message, or $\alpha_1(g, p_1) = \alpha_1(a, p_1) = 1$. This strategy leads to a updated reputation of 1 after a bad rating, contradiction. This argument is true for small p_1 values. Once the prior reputation p_1 goes above some cutoff, the payoff will be positive. A more detailed proof is presented in the Appendix.

The explanation shows that a positive payoff cannot be supported when the reputation falls below p^*. A CRA with a low reputation will not be able to help the issuer to find an investor at all. For the low reputation with 0 payoff, the equilibrium strategy is not unique. However, for the nonzero payoff, the strict monotonicity above the discontinuity point will lead to unique strategy for the maximum payoff.

Next I will discuss the equilibrium for the CRA with a high reputation. In this paper, the CRA can perfectly observe the signal, but the signal does not perfectly reveal the true state of the world. In a classical perfect signaling game, the reputation concern would push the sender to tell the truth. However, on the imperfect signaling game, sending preferred message is the optimal choice with a higher reputation. This leads to a interesting question, whether this credit rating game more similar to the perfect signaling game or the imperfect signaling game. In this paper, β measures how much the future or the reputation matters, while λ measure how imperfect the bad signal is. Moreover, a good rating with a failing investment immediate ruins the reputation, which resembles the effect of a perfect signaling game.

Proposition 2.3.2. *For every $\lambda > 1 - \frac{1}{\beta}$, there exists a p^{**}, such that the CRA gives good ratings after both signals for $p_1 \geq p^{**}$.*

Here is the sketch of the proof. When the prior reputation p_1 is almost 1, the updated reputation in the second period will be almost the same as in the first period. Then the IC of giving a good rating after a bad signal become $(1 + \lambda\beta)(R - 1) \geq \beta(R - 1)$. When λ is sufficiently large, the current benefit of giving a good rating dominates. Therefore, the CRA will always give good ratings for high reputations. The IC after the bad signal will be

$$\phi(\frac{p_1 + (1 - p_1)(1 + \lambda)}{p_1 + 2(1 - p_1)}) + \lambda\beta\phi(\pi_2(\frac{p_1}{p_1 + (1 - p_1)(1 + \lambda)})) \geq \beta(R - 1) \tag{2.3.5}$$

The cutoff p^{**} will be the point where the IC above binds.

When β is low or λ is large, the result is similar to the classical reputation game with imperfect signaling. A small β means the CRA cares less about the future, while a large λ value shows the signal is more noisy. Similar to those games with noisy signals, when the reputation falls below this cutoff, the CRA may mix after at least one of the two signals. But on average, the CRA is more likely to give the CRA

a good rating than he should. The empirical literatures have shown that the CRAs overate the securities in general.

In a reputation game with noisy signals, the property above remains the same with any β value. However, in this credit rating game, the reputation will drop to 0 with probability $1-\lambda$ after a misreported good rating. When λ is small enough, the instantaneous benefit cannot compensate the damage of a 0 reputation. The lemma below summarizes this result.

Proposition 2.3.3. *For every $\lambda < 1 - \frac{1}{\beta}$, there exists a p^{***}, such that the CRA tells the truth for $p_1 \geq p^{***}$.*

When λ is small, the current benefit of telling the truth dominates. The IC after the bad signal will be

$$\beta(1-\lambda)\phi(\frac{p_1 + (1-p_1)(1+\lambda)}{2-p_1}) \geq R - 1 \tag{2.3.6}$$

The LHS of the inequality measures the damage of the misreporting after a bad signal, while the RHS is the benefit of the lying in the first period. If $1-\lambda$ is large enough, the effect of perfect signaling will dominate, and the CRA will tell the truth with a high prior reputation. A more extreme case would be comparing the benefit of truth telling at the cutoff of no investment to the instantaneous payoff of lying. Then the equation (3.6) becomes $\beta(1-\lambda)\frac{r(R-1)}{R} \geq R-1$, or $\beta > \frac{R}{r(1-\lambda)}$. The cutoff for purchasing the investment depends on the promised return R and the quality of the signal $1-\lambda$. Both variables affect the expected return for the investor. However, r also affects the CRA's benefit. Given everything else fixed, the CRA's payoff increases with r as well. At the switching point, the CRA's reputation incentive is $\beta(1-\lambda)\frac{r(R-1)}{R}$, which strictly increases with the private benefit r. When this is above the incentive for lying, the CRA will hit the reputation switching point before hitting the strategy switching point. The truth telling strategy cannot be supported,

since the reputation is too low to attract second period investment. Therefore, the CRA will start to overate the issuer, either hit the IC or the switching reputation. But the discussion before shows that the all good rating equilibrium does not exist in this case.

To summarize the discussion on equilibrium strategy and payoff above, I have the following proposition.

Proposition 2.3.4. *In the two period credit rating game, there exists a value p^*, such that if $p_1 < p^*$, the CRA's payoff is* 0.
(1) For every $\lambda > 1 - \frac{1}{\beta}$, there exists a p^{**}, such that the CRA gives good ratings after both signals for $p_1 \geq p^{**}$. The CRA randomizes after at least one of the signals in the first period.
(2) For every $\lambda < 1 - \frac{1}{\beta}$, there exists a p^{***}, such that the CRA tells the truth for $p_1 \geq p^{***}$. The CRA randomizes after at least one of the signals in the first period.
(3) $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$.
(4) $\alpha_2(s_2, p_2) = 1$ for any p_2.

2.3.3 Comparative Statics

In this part, I will consider how different variables in this model would affect the CRA's strategy. I begin with the discount factor and the quality of the bad signal $1 - \lambda$. The CRA is more likely to tell the truth with a larger β, since β measures how much the CRA cares about the future. I use $1 - \lambda$ to describe how precise the signal is. Therefore, the larger $1 - \lambda$ is, the more lying will hurt the CRA's payoff and the more likely the CRA will tell the truth. This result is consistent with the conclusion in Proposition 4.

The other two variables in this game are the promised return R and the private return r. I can rewrite the utility function as $\phi(\pi_t) = R - 1 - (1 - \pi_t)(R - r)$. This utility function implies that the private return r and the promised return R will affect

the equilibrium in an opposite way. The difference in returns between two state $R-r$ will affect the payoff through the coefficient β and $1-\lambda$. Following the same argument after Proposition 2 and Proposition 3, when $\lambda > 1 - \frac{1}{\beta}$, the CRA is more likely to tell the truth when R increases or r decreases. A detailed proof can be found in the Appendix.

The lemma below concludes all the comparative statistics discussed above.

Proposition 2.3.5. *The larger β or the smaller λ is, the more likely the CRA will tell the truth after a bad signal. The effect of R and r on the CRA's equilibrium strategy depends on the discount factor β and the quality of the signal $1-\lambda$.*
(1) For every $\lambda > 1 - \frac{1}{\beta}$, the CRA is more likely to tell the truth after the bad signal when R increases or r decreases.
(2) For every $\lambda < 1 - \frac{1}{\beta}$, the CRA is more likely to tell the truth after the bad signal when R decreases or r increases.
(3) For $\lambda = 1 - \frac{1}{\beta}$, the CRA's equilibrium strategy would not change with R or r.

2.4 The Role of Private Benefits

In the previous section, I solved the baseline model with the assumption $r \in (0, R)$. The equilibrium strategy has been described according to the value of the discount factor. In this section, I will show how r's value will affect the equilibrium strategy, with a negative private benefit ($r < 0$) or a significantly large private benefit $r > R$. First, I will begin with the negative private benefit.

2.4.1 The Negative Private Benefit $r < 0$

In the baseline case, I analyze the equilibrium strategy with a positive private benefit $r \in (0, R)$. A positive r value means that the issuer is able to collect some private benefit after the project's failure, such as tax credits or spillover effects from Research

and Development. However, the failure of this investment could also have negative effects on the issuer's other operations. To model this negative effect, I will assume $r < 0$ in the following discussion.

The net return to the investor will be the same as the baseline case, but the CRA's fee is defined in a slightly different manner.

$$\phi(\pi(p_t)) = \begin{cases} R\pi(p_t) + r(1-\pi(p_t)) - 1 \text{ if } R\pi(p_t) + r(1-\pi(p_t)) - 1 \geq 0 \\ 0 \text{ otherwise.} \end{cases} \tag{2.4.1}$$

The promised return R is used to compensate for both the reservation value of the investor and the private loss of the issuer. Because of this private loss, some projects with a positive net return cannot get financed. The CRA will get a contract from the issuer unless the sum of the issuer and the investor's return is negative. This helps to explain why some companies are more likely to get financed after their independence from their original holding companies. When they are affiliated with the holding company, their failure will affect the operation of the holding company's other businesses as well. Therefore, even if the affiliation itself can benefit from a project, the holding company may not have sufficient incentives to obtain a good rating and get the project financed. In my model, this correlation is denoted by a negative r value. After the affiliation gains independence, the negative r value immediately turns to 0. Given the same coefficients (β, λ, R), the investment is more likely to get finance then. In general, the project is more likely to get financed when the private benefit or promised return is larger.

Proposition 2.4.1. *There exists a value p^*, such that if $p_1 < p^*$, the project will not be financed. The cutoff p^* decreases with R and r.*

The expected net gain $R\pi(p_t) - 1$ is split between the CRA and the issuer to compensate for the issuer's loss in the failed case. This setting eliminates the dis-

continuity in the $r > 0$ case, but the CRA's payoff still increases with π_t. Thus, the reputation concern of the expert remains the same. We can repeat the argument in Section 3, and all the qualitative results remain true.

2.4.2 The Private Benefit Larger than The Promised Return

Now, I assume that the private benefit is larger than the promised return. This could happen when the spillover effect from the current project is much larger than the immediate benefit that it could generate. One example is the pharmaceutical industry. The failure of one drug's clinical trial is not the end of the world. Instead, it may help to develop the next new project, which could be even more profitable.

Similar to the baseline case, the cutoff point for the investment depends on the promised return R and the reservation value 1. The private benefit only affects the CRA's payoff, not the investor's decision. Therefore, the CRA with a low reputation cannot get the contract as before, while the CRA with a high reputation will act according to the value of the discount factor. However, the reputation incentive is reversed now, since $\phi(\pi_t) = r-1-(r-R)\pi_t$ implies that ϕ decreases in π_t. Therefore, the payoff in the first period achieves maximum at the cutoff, where the investor is indifferent between investing or not. Moreover, it will be minimized at $\pi_t = 1$ with a value $R-1$. In the second period, sending a good report is still a dominant strategy. With these properties, I revisit the IC of the truth telling equilibrium given $\beta > \frac{1}{1-\lambda}$ and the IC of giving good ratings given $\beta < \frac{1}{1-\lambda}$

$$\beta(1-\lambda)\phi(\frac{p_1+(1-p_1)(1+\lambda)}{2-p_1}) \geq R-1 \tag{2.4.2}$$

$$\phi(\frac{p_1+(1-p_1)(1+\lambda)}{p_1+2(1-p_1)}) + \lambda\beta\phi(\pi_2(\frac{p_1}{p_1+(1-p_1)(1+\lambda)})) \geq \beta(R-1) \tag{2.4.3}$$

The LHS of both inequalities will decrease with π_t right now. Thus, any strategy work for the reputation $p_1 = 1$ will work for $p_1 < 1$ as long as the investment is happening.

The IC is not working once the investment stops with the updated reputation p_2. This cutoff helps to define $\frac{p^{**}}{p^{**}+(1-p^{**})(1+\lambda)} = \bar{p}_2$ and $p^{***} = \bar{p}_2$, and the CRA will randomize when the reputation is below such a cutoff.

Since the CRA's payoff decreases with π_t, the interesting question becomes whether the over-rating statement $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$ is still true. In the following discussion, I will show that the CRA will still inflate the rating in two different cases.

I begin with the $\beta > \frac{1}{1-\lambda}$ case first. If $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$ is not true, $\alpha_1(g, p_1) + \alpha_1(b, p_1) < 1$. Consider the prior reputation falls below p^{***}. The updated reputation p_2 after a bad rating is smaller than $\bar{p}_2$, which means the payoff is 0. Hence the CRA either strictly prefers to announce a good rating and get a positive payoff, or get 0 payoff. Giving the assumption $\alpha_1(g, p_1) + \alpha_1(b, p_1) < 1$, the payoff is 0. Therefore, the CRA's strategy will follow $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$ or he gets a 0 payoff.

Then consider the $\beta < \frac{1}{1-\lambda}$ case, and I also want to show either the payoff is 0 or $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$. If neither is true, then we can assume $\alpha_1(g, p_1) + \alpha_1(b, p_1) < 1$ and the payoff is positive. If $\alpha_1(g, p_1) + \alpha_1(b, p_1) < 1$, the CRA randomizes after both signals and the payoff after a good rating is the same after both signals. It means that the second-period payoff after a good rating is 0. However, $\alpha_1(g, p_1) + \alpha_1(b, p_1) < 1$ means the reputation after the good rating is better than after the bad rating in the second period. Therefore, if the payoff after the good rating is 0, the payoff after the bad rating is 0 as well. I assume the payoff is positive, which means the CRA strictly prefers to send a good rating and the good rating leads to a positive first period payoff, resulting in a contradiction.

The analysis above shows that, even if the reputation concern changes after $r > R$, all the main conclusion remains the same. It implies that the over-rating statement $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$ is working for any r value, even if there is an inverse reputation

concern. The inverse reputation concern helps to support the pure strategy when the reputation drops. With the reservation value, the pure strategy will stop at a certain cutoff. If there is no reservation value, then the same strategy will work for any prior reputation. The main proposition is modified as below.

Proposition 2.4.2. *In the two period credit rating game, for every $r > R$, there exists a value p^*, such that if $p_1 < p^*$, the CRA's payoff is* 0.
(1) For every $\lambda > 1 - \frac{1}{\beta}$, there exists a $p^{**} = \frac{(1+\lambda)\bar{p}_2}{1-\bar{p}_2(1+\lambda)\bar{p}_2}$, such that the CRA gives good ratings after both signals for $p_1 \geq p^{**}$. The CRA randomizes after at least one of the signals in the first period.
(2) For every $\lambda < 1 - \frac{1}{\beta}$, there exists a $p^{***} = \bar{p}_2$, such that the CRA tells the truth for $p_1 \geq p^{***}$. The CRA randomizes after at least one of the signals in the first period.
(3) $\alpha_1(g, p_1) + \alpha_1(b, p_1) \geq 1$.
(4) $\alpha_2(s_2, p_2) = 1$ for any p_2.

2.5 Conclusion

I have analyzed the reputation game with a model in which a privately informed CRA sells a credit rating report to an issuer for profit. A key feature of my model is that the CRA's signal is perfect on failed projects, but noisy on successful projects. This information asymmetry leads to different equilibrium strategies for different default risk levels. The main contribution of the paper is that I introduce the issuer's private benefit after the default, and explain how this private benefit would affect the CRA's strategy. When the project is very unlikely to default, the credit rating game behaves like the noisy signaling game. Then the CRA is more likely to tell the truth when the private benefit decreases. When the project is very likely to default, the credit rating game is similar to the perfect signaling game and the CRA is more likely to tell the truth when the private benefit increases. Another contribution is showing that the

CRA is more likely to give a good rating as long as he has a contract. I have proved that it is not necessary to assume the payoff will increase with the reputation.

2.6 Appendix

2.6.1 The Baseline Case

Proof of Proposition 1: I begin with a p_1 sufficiently small, and will show that the payoff is 0. Let $p_1 = \epsilon$ and $\epsilon \to 0$. The probability of success after a good rating in the first period and the updated reputation in the second period would be

$$\pi_1(p_1) = \frac{p_t + (1-p_t)(\alpha_t(g;p_t) + \alpha_t(b;p_t)\lambda)}{p_t + (1-p_t)(\alpha_t(g;p_t) + \alpha_t(b;p_t))} \tag{2.6.1}$$

$$p_2(m_1 = g, w_1 = s; p_1) = \frac{p_1}{p_1 + (1-p_1)(\alpha_1(g;p_1) + \alpha_1(b;p_1)\lambda)} \tag{2.6.2}$$

$$p_2(m_1 = b, w_1 = \varnothing; p_1) = \frac{p_1}{p_1 + (1-p_1)(2 - \alpha_1(g;p_1) - \alpha_1(b;p_1))} \tag{2.6.3}$$

Case 1: If updated reputation after a bad message is above $\bar{p}_2$, then $\alpha_1(g,p_1)+\alpha_1(a,p_1)$ is bounded below by $1 + \frac{\bar{p}_2 - \epsilon}{\bar{p}_2(1-\epsilon)}$. This approaches 2 when $\epsilon \to 0$. As a result, the updated reputation after a successful financing will be below $\frac{\epsilon}{\epsilon + (1-\epsilon)\lambda(1+\frac{\bar{p}_2-\epsilon}{\bar{p}_2(1-\epsilon)})}$. This is also almost 0, when $\epsilon \to 0$, and below $\bar{p}_2$ if ϵ is sufficient small. The probability of success is bounded above by $\frac{\epsilon+(1-\epsilon)(1+\lambda)}{\epsilon+2(1-\epsilon)}$, which is below $\frac{1}{R}$ with a very small ϵ. Then the payoff is 0 after good rating and positive after a bad rating. This payoff means that the CRA can be better off by sending a bad rating, contradiction to $\alpha_1(g,p_1) + \alpha_1(a,p_1) > 1$. Case 2: If the updated reputation after a bad message is below $\bar{p}_2$, but the payoff after good rating is positive. Then the CRA would strictly prefer to send a good message, or $\alpha_1(g,p_1) = \alpha_1(a,p_1) = 1$. This strategy leads to a updated reputation of 1 after a bad rating, contradiction. Define $p^* = \max(p_1 | \pi_1(p_1) < \frac{1}{R} \& p_2(m_1, w_1; p_1) < \bar{p}_2)$.

Then I will show once the payoff is positive for some reputation p_1', it is always positive for $p_1 > p_1'$. If the payoff is 0 in the second period after a good rating, then

the IC becomes

$$\phi(\frac{p_1+(1-p_1)(\alpha_1(g,p_1)+\lambda\alpha_1(g,p_1))}{p_1+(1-p_1)(\alpha_1(g,p_1)+\alpha_1(g,p_1))})=\beta\phi(\pi_2(\frac{p_1}{p_1+(1-p_1)(2-\alpha_1(g;p_1)-\alpha_1(b;p_1))})) \tag{2.6.4}$$

Suppose there is an equilibrium strategy α_1 for the reputation p_1'. When the reputation increases to p_1, both $LHS(\alpha_1,p_1')$ and $RHS(\alpha_1,p_1')$ will increase. To get the new equilibrium, I will rebalance the strategy, and the payoff is larger than $\min(LHS(\alpha_1,p_1'),RHS(\alpha_1,p_1'))$, which is greater than the equilibrium payoff at p_1.

Proof of Proposition 5: The comparative statics on discount factor β and project risk λ come directly from the definition of the payoff and the reputation updating rule. Here I will focus on the comparative static on the promised return R and private benefit r.

I can rewrite the utility function as $\phi(\pi_t)=R-1-(1-\pi_t)(R-r)$. This utility function implies that the private return r and the promised return R will affect the equilibrium in an opposite way.

The cutoff for 0 payoff is defined by $\frac{1}{R}$. Therefore, the larger R is, the less like the CRA will get 0 payoff. However, for the 0 payoff, the equilibrium strategy is not unique, so the comparative statics is for the none zero payoff only. There are two cases, one with 0 payoff in the second period after a good rating, and one with positive payoff.

Case 1: the payoff is 0 in the second period, then the IC is

$$R-1-(1-\pi_1)(R-r)=\beta(R-1-(1-\pi_2)(R-r)) \tag{2.6.5}$$

Since $\pi_1<\pi_2$, this will only happen with $\lambda>1-\frac{1}{\beta}$. I can write $\frac{\partial(LHS-RHS)}{\partial R}=\pi_1-\beta\pi_2<0$ and $\frac{\partial(LHS-RHS)}{\partial r}=1-\beta-(\pi_1-\beta\pi_2)>0$. When R increases with r decreases, $LHS<RHS$ and the CRA is more likely to send a bad rating.

Case 2: the payoff is positive in the second period after a good rating, then the CRA is sending good ratings after good signals. The IC after the bad signal is

$$R-1-(1-\pi_1)(R-r) = \beta(1-\lambda)(R-1)-\beta((1-\pi_2(m_1=b))-\lambda(1-\pi_2(m_1=g,w_1=g)))(R-r) \tag{2.6.6}$$

I can rewrite this equation as

$$(R-1)(1-\beta(1-\lambda)) = ((1-\pi_1)-\beta((1-\pi_2(b))-\lambda(1-\pi_2(g,g))))(R-r) \tag{2.6.7}$$

Given $R-1>0$ and $R-r>0$, the comparative statics depend on the value of $1-\beta(1-\lambda)$. If this is positive, the CRA is more likely to tell the truth when R increases or r decreases. If it is negative, the CRA is less likely to tell the truth when R increases or r decreases.

Chapter 3

Fiscal Policy, Ethnicity and Secession

3.1 Introduction

Over the second half of the twentieth century, the number of independent nations almost tripled, from 74 to 193. 25 of these new countries were created in the 1960s in Africa after the abolition of colonial rule. Another wave of border changes was brought by the breakups of former Soviet Union and Czechoslovakia in the 1990s. Other than these two major sequences of events, there are various types of separatist movement closely related to the establishment of new nations all over the globe. However, not all attempts or threats of secession by the minority regions succeeded. Even with large-scale violent conflicts and deaths of more than 1000 per year, some separatists still failed to secede. This may lead us to wonder why such movements or even a consequent civil war would ever happen in the first place.

In order to answer these questions, my paper presents a model on secession and nationalism, with a special emphasis on the role of public goods. In my model, a disagreement on secession between the central government and the minority group

leads to a disastrous military conflicts. As a result, the tremendous potential cost of the war distores the political choice of the minority group, and consequently hurts the minority group both economically and politically. I conduct an empirical test of this model and find that, per capita income and perceived winning chance of the civil war play the most important role in the decision making process of the minority group. In addition, I find that population and cultural difference would also affect the probability of a civil war, but their impact is smaller compared to that of income and winning chance.

My main contribution is to explain the motivation of secession crises and internal conflicts. I find that ideological preference and fiscal transfer in policy do matter when the minority regions try to declare independence. I build my argument on two assumptions. The first assumption is that the minority may not value the public service from the central government as much as the tax revenue they sent. Therefore, residents from the minority region would rather provide their own public service, such as education, health care and infrastructure. The second assumption is that the majority group and minority group hold different views on nationalistic policy about language, religious belief, and cultural tradition. The residents in the minority region prefer to implement a public policy that favors their own ethnic group and respects their heritage.

My other contribution is to show the provision of more local public good induces an economic incentive for political integration. Under federation, the minority share the cost of the public good, and the marginal benefit of their expense on public good can be higher in comparison to the situation of gaining independence. And with tax revenue transferred from the minority region, the majority region is strictly better off than supporting the public service alone. Moreover, as the majority of the federation, the agents can implement their preferred nationalistic policy without accommodating others. This agrees with the popular category of explanation led by Buchanan and

Faith (1987), which pointed out that the central government may not have sufficient revenue to finance the production of public goods. In a related work, Gradstein (2004) showed that the majority region might need tax revenue or fiscal transfer from the minority region to finance public goods, and the majority region can improve their bargaining positions even with the option to secede.

Moreover, I conduct an empirical test focusing on the proxies of the variables in the theory section. I extend the analyses in Fearon and Laitin (2003) for another 10 years with two more variables, the GINI coefficients and the religious tension, to measure income heterogeneity and cultural difference. I also coded the civil war differntly to highlight the importance of both economic and political incentive of internal ethnic conflict. Most of the independent variables on my test have the same sign as the Fearon and Laitin (2003) or Collier and Hoeffler (2004), but my theory section provide a different argument for such result.

There is a larger theory literature explaining the motivation of secession crises and internal conflicts. Collier and Hoeffler (2004) raised the theory of greed and grievance, which suggests that economic opportunities and severe grievance were the main factors fueling insurgencies. Another strand of literature offers similar explanations. For instance, after observing the secessionist movements within the former socialist economies, Berkowitz (1997) pointed out that the resource-rich peripheral regions, such as Chechens of Russia, were not willing to make net payment to the fiscal federations and declared independence. Relatively poorer peripheral regions, such as Slovakia in the former Czechoslovak Republic, may prefer private goods to public goods and enjoy a higher welfare in secession. Fearon and Laitin (2008) showed that in Afghanistan the Taliban survived not by terrorist attack or suicide bombing, but through providing their local supporters with public education, health service, and protection of smuggling to Pakistan. There are also papers focusing on explaining why would central governments work hard to prevent secession. Fearon and Laitin

(2011) argued that, in 31 out of 103 ethnic civil wars between 1945 and 2008, the violence was between members of a regional ethnic group and recent migrants from other parts of the country. The local ethnic group considered themselves to be the indigenous sons-of-soil, while the migrants of the dominant group took for granted that they could come in search of land or job within the border of their own country. In Olofsgard (2003), the increasing return to scale in production created economic incentive against separation. In that case, the majority group might even be willing to accommodate political policy to avoid separation. However, this theory fails to explain the observations that the large countries are more likely to be involved in civil war.

Other than the theory literatures discussing the economic and political incentive of secession, there is also a large group of empirical papers which tried to predict the risk of civil war breakout. They cannot observe people's view by running surveys or polls on secession. Thus, insurgencies appear as a secondary evidence of the internal conflict between ethnic groups under the same federation. Unable to find proxies for grievances and opportunities, Collier and Hoeffler (2004) used corresponding political and economic variables to estimate the risk of civil war breakout. To reveal the relationship between ethnicity and civil war, Fearon and Laitin (2003) added factors of ethinic and religious characteristics with the standard economic variables to predict the risk for civil war. Both found GDP per capita, population and being a natural resource exporter are highly significant. Collier and Hoeffler (2004) also observed that secondary schooling and GDP growth reduced conflict risk, as these two variables are related to the income forgone by enlist as a rebel. Fearon and Laitin (2003) pointed out that the new or unstable countries are more likely to be exposed to civil wars. Both literatures dropped the hypothesis that the ethinic or religious characteristics explains the civil war.

This paper is structured as follows. Section 3.2 introduces the setup of the basic model and the socially optimal solution. Section 3.3 solves the model for secession with civil war, and discusses the nationalistic policy in a framework of homogeneous income. In Section 3.4, we extend the discussion by introducing heterogeneity in regional income distributions. In Section 3.5, some empirical results of civil war are reviewed to assess the analytical conclusion. Section 3.6 summarizes the major findings of the paper and briefly discusses its limitation.

3.2 The Model

Consider two regions, indexed $k = A, B$, which form a federation at the beginning. Region A has a population $\frac{1}{1+d}N$, and the measure of individuals in region B is $\frac{d}{1+d}N$, where $0 < d < 1$, and the total population is N. I assume that the population of each region is fixed. When two regions stay together, only region A has access to the production technology of a local public good, and the production of the public good in region A entails spillover effects to residents in region B. Moreover, only residents of region A are the direct recipients of this public good. This effect induces an economic incentive for political integration. With the probability of secession, however, individuals with a low preference for the public good would rather require a referendum on secession. Gradstein (2004) used a similar setting to argue that the option to secede would distort the political choices made by individual regions to improve their bargaining position. In the case of secession, both regions get access to the production technology, provide their own public good, and cannot benefit from the spillover from each other any more. The cost of producing g unit of public good per capita is

$$c(g) = \frac{g^2}{2}$$

and the total cost is $\frac{N}{1+d}\frac{g^2}{2}$ under unification. This production function has also been used in Gradstein (2004), and enables the derivation of closed form solutions.

Let y_i^k denote the income of person i in region k, where y_i^k is exogenous. The mean incomes of region A and region B are y^A and y^B, respectively. The corresponding incomes of region A and B are y_m^A and y_m^B. The public good is financed by the tax revenue available to the central government located in region A. Private consumption is supported by the after tax income of the residents.

Other than the private good and the local public good, individuals also care about the nationalistic policy in ideology. In my model, people in the same region share the same ideology preference. Each region carries their own ethnic, religious, and linguistic tradition. I assume the preferred policy and implemented policy are both in a single-dimensional space. Let x_k denote the preferred position of region k, x denote the implemented national policy in the case of unity, and x^k denote the policy of region k in the case of secession. Similar to Olofsgard (2003), if the country stays together, I define the utility of the nationalistic policy as a decreasing function of the metric distance between the most preferred policy of the respective group x_k, and the actual national policy, x, according to $-(x_k - x)^2$. And the utility of the public good would be $-(x^k - x_k)^2$, if they separate successfully.

If two regions stay together, the utility of individual i in region A is

$$-(x - x_A)^2 + \alpha g + y_i^A(1 - \tau)$$

and the utility of individual j in region B is

$$-(x - x_B)^2 + \beta g + y_j^B(1 - \tau)$$

with the budget constraint

$$\frac{N}{1+d}\frac{g^2}{2} \leq \frac{N}{1+d}\tau y^A + \frac{dN}{1+d}\tau y^B$$
$$0 \leq \tau < 1$$

Here, α and β denote the parameters reflecting public preferences in region A and region B respectively. The residents in region A are the direct recipients of this local provided public good g, and the residents of region B consume the spillover effect only. For example, the government may sponsor some health care program that benefit the patients mostly from a specific ethnic group, or whose treatment is against the belief of some religion. Another example is a public schooling system whose primary language is some group's native language, and thus the direct recipients of such budget are the native speakers of this language. I use this differentiation between two groups to model the greed and grievance theory in Collier and Hoeffler (2004). The benefit to the periphery is not as much as to the central region, or more specifically $\alpha > \beta$. Let τ specify the tax rate of the whole country. To avoid unnecessary complications, we assume that there exists a solution to the welfare maximization problem if and only if $\tau < 1$.

If the regions separate, the utility of individual i in region k is

$$-(x^k - x_k)^2 + \alpha g + y_i^k(1-\tau^k)$$

with the budget constraint

$$\frac{g^2}{2} \leq y^k\tau^k$$
$$0 \leq \tau^k < 1$$

Since both will produce their own public good locally, we assume the marginal value of one unit public good is α for each region. Under unification, the total marginal value of the public good is $\alpha + \beta$. Thus, the federation is more socially efficient comparing to this condition. Moreover, region B has to bear the cost alone as an independent country, which may provide the economic rationale for them to form a federation with region A to share the cost.

3.2.1 Social Optimal Solution

I will show the social optimal solution and decentralized solution to this model as the benchmark case. Allocations in this economy will specify the amount of public good g produced, and the tax rate implemented. The social optimal problem is to maximize the aggregate welfare of the federation with two regions,

$$\max -(x - x_A)^2 + \alpha g + y^A(1 - \tau) + d(-(x - x_B)^2 + \beta g + y^B(1 - \tau))$$

such that,

$$\frac{g^2}{2} \leq y^A\tau + dy^B\tau$$
$$0 \leq \tau < 1$$

We can solve the social optimal nationalistic policy and the provision of local public good as,

$$x = \frac{x_A + dx_B}{1 + d}$$
$$g = \alpha + d\beta \qquad \text{if } (\alpha + d\beta) \leq \sqrt{2(y_A + dy_B)}$$

We now consider the decentralized problem, where region A determines the amount of the public good produced, and region B determines the tax revenue that they are willing to transfer to region A. We assume both decisions are made by the representatives elected by their own region. As a result, residents in region B would not be willing to give any tax revenue to region A, and the local public good is financed by the tax revenue from region A only. Thus for resident i in region A, the problem becomes,

$$\max -(x^A - x_A)^2 + \alpha g + y^A(1 - \tau^A)$$

with the budget constraint

$$\frac{g^2}{2} \leq y^A \tau^A$$
$$0 \leq \tau < 1$$

Then we have

$$x^A = x_A$$
$$g = \alpha \qquad \text{if } \frac{1}{2}\alpha^2 < y^A$$

Comparison with the welfare optimal solution reveals that the public good provided under decentralization is too low. Also, region A will not be better off staying with region B if this policy is implemented. This result echoes the findings of Oates (1972), whereby inter-regional spillovers are not internalized through decentralized decision making. However, a marginal increase in the provision of public good financed by the tax revenue from region B can benefit both regions. Therefore, both regions paying for the locally provided public good is a Pareto improvement relative to the decentralized outcome, and it creates an economic incentive of unity for both regions.

3.3 Bargaining under Federation with Homogeneous Income

To examine how civil war can lead to exploitation of minority by the majority, a simple bargaining game of the nationalistic fiscal and ideological policy is introduced in this section. In the game, I assume that free secession can take place only if the majority region agree to it. In reality, there are very few constitutions that explicitly guarantee the right of free secession, and people can easily find evidence from the long term bleeding conflicts in Balkan or Caucasus. To further diminish the individual incentive of the minority region to support secession, I design a punishment phase for the minority region in the case where they lose the war. However, to limit exploitation of minority by the majority, I give some bargaining power to the minority region in negotiating fiscal policy and the provision of public good. The equilibrium of this game will be compared to the social optimal problem and the decentralized problem in the previous section.

The model of bargaining is a sequential game, and has the following timing. First, there are two simultaneous referenda, one in each region, where residents vote for or against secession. Free separation is implemented if that alternative gains a majority of the votes in region A. Second, the individuals vote again, but this time on the representatives of their own regions. If both regions agree to stay together in stage one, then the representative from region A will make an offer of the national tax rate, the public goods and the nationalistic ideological policy. If the representative from region B accepts, then they will implement this nationalistic policy. Otherwise, region B refuse to pay the tax, but they can still enjoy the spillover of any local public good produced by region A accompanied by the implemented national ideological policy. When region A votes for secession in the first stage, they will separate peacefully, and both regions make their own policy decisions afterwards. If region B is the only one

that vote for secession, there will be a civil war between the two regions, and region B can gain independence with probability p. The civil war will destroy part of the output in both regions. If region B suceeds, two regions will make their own policy decisions. If region B loses the war, they have to pay for the damage in region A first, and then accept any nationalistic policy chosen by region A.

In this section, I will discuss the baseline case, where, in each region, the individuals have homogeneous income level, and there is no factor mobility across regions. The result will be contrasted to the case with heterogeneous income for each region, analyzed in the next section. We also assume that region A, as the majority of the whole country, lacks the ability to credibly commit to an accommodating nationalistic policy to avoid secession from region B.

3.3.1 Fiscal Policy under Unification

First, consider the tax rate and the public good provision if two regions stay together. If region B refuses to pay for the local public good, the problem would be the same as the decentralized scenario, and $g = \alpha$ will be produced in region A. The bargaining proposal made by region A's representative will maximize his utility, while guaranteeing region B's utility of not paying tax, i.e.:

$$\max -(x - x_A)^2 + \alpha g + y^A(1 - \tau)$$

such that

$$\frac{g^2}{2} \leq y^A\tau + dy^B\tau$$

$$-(x_B - x)^2 + \beta g + y^B(1 - \tau) \geq -(x_B - x_A)^2 + \beta\alpha + y^B$$

$$\tau < 1$$

The solution is

$$x = x_A$$
$$g = \alpha \frac{y^A + dy^B}{y^A} \qquad \text{if } \frac{1}{2}\alpha^2 < \frac{y^A}{y^A + dy^B} y^A \text{ and } \frac{\beta}{\alpha} > \frac{y^A + dy^B}{2dy^A}$$

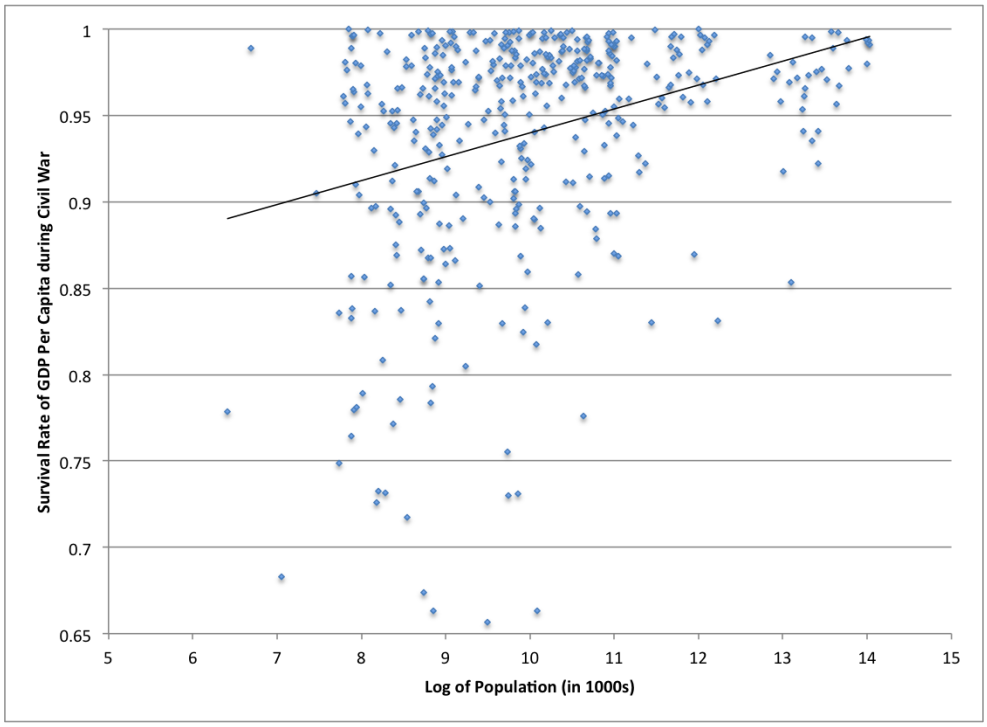
Comparing this with the decentralized outcome, I observe residents in region A attains a higher welfare without hurting region B. Also people from region B are strictly better off by paying the tax if $\frac{\beta}{\alpha} > \frac{y^A + dy^B}{2dy^A}$. Thus, the willingness of region B's residents to finance the public good is an increasing function of the spillover effect. In other words, the larger $\frac{\beta}{\alpha}$ is, the more they are willing to join the federation and finance the public good.

3.3.2 Fiscal Policy under Secession

In my model, free secession is not allowed, and central government will prevent secession through military means. After the civil war, residents in both regions have only $w(N)$ of their productivity left, where $w(N)$ is a monotonic increasing function of the federation's total population N. Since most of the conflicts will take place on the border area between regions, it can hardly incur cost to the vast majority living elsewhere. Thus, the larger total population is, the less likely each individual could be hurt during a war. Figure 1 shows the trend of survival rate with different population size during a civil war, based on the data used by Fearon and Laitin (2003). The figure echoes my assumption that the survival rate will increase with the size of the population during the war.

With probability p, region B will gain independence, and make their own policy decision after the civil war. Moreover, since they would produce the local public good themselves, the marginal utility would be α rather than the marginal spillover effect

Figure 3.3.1: Survival Rate of GDP Per Capita, 1945-1999



β. Thus, the problem for region B is

$$\max -(x^B - x_B)^2 + \alpha g + w(N)y^B(1-\tau)$$

with the budget constraint

$$\frac{g^2}{2} \leq w(N)y^B\tau$$

The solution is

$$\begin{aligned} x^B &= x_B \\ g &= \alpha \qquad\qquad \text{if } \frac{1}{2}\alpha^2 < w(N)y^B \end{aligned}$$

Therefore, if the damage of war is not huge or $\frac{1}{2}\alpha^2 < w(N)y^B$, region B will be able to provide the same level of public good as region A in the decentralized case. Similarly, region A will produce $g = \alpha$ units of local public good, if $\frac{1}{2}\alpha^2 < w(N)y^A$.

By implementing their preferred policy in ideology, people in region B benefit from secession. However, to the social planner's view, it is not efficient for both regions to produce their own local public good. Considering the damage of the war, the residents from region A are worse off than allowing for free secession. Thus, if region B can win the civil war with certainty, region A are not willing to stop the war with military power.

With probability $1-p$, region B will lose the civil war and pay for the damage in region A. At the end of this section, I will justify the choice of punishment phase rather than no punishment. If the civil war is disastrous, residents in region B may not have any private consumption left after compensating region A. Hence, when $dw(N)y^B \leq (1-w(N))y^A$ or $w(N) \leq \frac{y^A}{y^A+dy^B}$, people in region A would collect everything from region B as the punishment. Now, region A makes its decision upon the maximization problem

$$\max -(x-x_A)^2 + \alpha g + w(N)(y^A + dy^B)(1-\tau)$$

with the budget constraint

$$\frac{g^2}{2} \leq w(N)(y^A + dy^B)\tau$$
$$\tau < 1$$

which yields the solution:

$$x = x_A$$
$$g = \alpha \qquad \text{if } \frac{1}{2}\alpha^2 < w(N)(y^A + dy^B)$$

When $\tau < 1$, the public good provision is identical to the decentralized case with initial per capita producticity of $w(N)(y^A + dy^B) < y^A$, and the utility of residents in

region A is less than in the decentralized case. To summarize, with $w(N) \leq \frac{y^A}{y^A+dy^B}$, the civil war will always mean utility loss to region A, regardless of the winner. Thus, if region A expect region B to vote for secession with certainty and the civil war would be catastrophic, region A should vote for secession as well or allow for free secession. The punishment phase exists only if it can effectively decrease the incentive of region B to secede. One related example would be Czechoslovakia, whose breakup was peacefully resolved by parliament after growing nationalistic tensions.

When region A vote against secession, region B will vote for secession if their expected utility can be improved after declaring independence,

$$p(\frac{1}{2}\alpha^2 + w(m)y^B) + (1-p)(-(x_A - x_B)^2 + \alpha\beta) \geq -(x_A - x_B)^2 + \alpha\beta\frac{y^A + dy^B}{y^A} + y^B(1 - \frac{y^A + dy^B}{y^A}\frac{\alpha^2}{2y^A})$$

The left hand side is the expected utility of region B after declaring independence, and the right hand side is the utility under unification. p signified the opportunity of winning, $w(N)$ captures the economic cost, and $\|x_A - x_B\|$ reflects the political incentive of initiating a civil war. To simplify this relationship, we get

$$p(\frac{1}{2}\alpha^2 + w(m)y^B + (x_A - x_B)^2 - \alpha\beta) \geq \alpha\beta\frac{dy^B}{y^A} + y^B(1 - \frac{y^A + dy^B}{y^A}\frac{\alpha^2}{2y^A})$$

When p, $w(N)$, and $\|x_A - x_B\|$ are larger, the relationship is more likely to hold. Oppositely, fixing $\frac{y^A}{y^B}$, the inequality is less likely to hold with a larger y^B or $\frac{\beta}{\alpha}$. Separation means a political gain for region B, when they win the civil war. Thus, an increase in cultural difference, measured as an increase in $\|x_A - x_B\|$, will increase the political motives for separation of region B. Once p increases, region B is more likely to win the war, or more likely to benefit from the political gain, and thus they are more willing to declare independence. Similarly, a larger $w(N)$ means smaller cost of civil war, and this consequently provides region B with economic incentive to declare war. $\frac{\beta}{\alpha}$ describes the economic gain of region B sharing the expense of public good with region A. Therefore, a larger ratio of marginal benefit creates economic

incentive for region B to stay with region A. Also, keeping $\frac{y^B}{y^A}$ constant, the higher y^B is, the less likely region B will vote for secession. Because $\tau < 1$, residents in region B can have private consumption. With the higher income, the damage of the war results in a greater lost to their private consumption, and hence they prefer staying with region A under unity.

If the loss from a war is mild, civilians in regions B would have positive private income left after punishment, and resident in region A can recover all the cost during the war. As a result, when $dw(N)y^B > (1-w(N))y^A$ or $w(N) > \frac{y^A}{y^A+dy^B}$, the problem of individuals in region A is

$$\max -(x - x_A)^2 + \alpha g + y^A(1 - \tau)$$

with the budget constraint

$$\frac{g^2}{2} \leq w(N)(y^A + dy^B)\tau$$

Hence, the solution is

$$\begin{aligned} &x = x_A \\ &g = \frac{w(N)(y^A + dy^B)}{y^A}\alpha \qquad\qquad \text{if } \frac{1}{2}\alpha^2 < \frac{y^A}{w(N)(y^A + dy^B)}y^A \end{aligned}$$

In this case, $g = \frac{w(N)(y^A+dy^B)}{y^A}\alpha > \alpha$, so the local public good will be provided at a more efficient level comparing to the decentralized scenario, and region A is better off than allowing for free secession. Therefore, this national fiscal policy creates an economic incentive for region A to keep region B together with military enforcement. With a high probability of winning the civil war, and a low potential lost from the civil war, region A will implement this punishment phase to intensify the stability of the federation. Region B will vote for secession, if they can gain independence

and will be better off. Relegating the analytical details to the appendix, I obtain a similar result to the previous case. When p, $w(N)$ and $\|x_A - x_B\|$ are larger or per capita income and $\frac{\beta}{\alpha}$ are smaller, the relationship is more likely to hold. In contrast to the low $w(N)$ case, region A's residents now can benefit from winning the war comparing to allowing for free secession. This rise in their expected welfare could reinforce their decision to fight against any secession effort. On the other hand, individuals from region B now have the private consumption after losing the war, but the overall war damage to their welfare will still increase with productivity. Thus, the conclusion is that a high average income will reduce region B's inclination to opt for independence. Even in the case of a catastrophic war, both parties involved may not expected a significant mortality at the early stage. Rather, they did make the decision of declaring war based on a high perceived $w(N)$, similar to this scenario.

The findings of the two cases above are summarized in Proposition 1 below. In this paper, I define the stable equilibrium as the unique Nash Equilibrium, where both regions cannot be better off by deviation, $\tau < 1$ is satisfied, and under unification the minority group B is strictly better off by paying tax.

Proposition 3.3.1. *(1) The stable equilibrium of nationalistic policy exists if* $\frac{\alpha^2}{2} < \min\{w(N)y^A, w(N)y^B, \frac{{y^A}^2}{w(N)(y^A+dy^B)}\}$, *and* $\frac{\beta}{\alpha} > \frac{y^A+dy^B}{2dy^A}$.
(2)The minority is more likely to declare independence, if p, $w(N)$ *and* $\|x_A - x_B\|$ *are larger or per capita income and* $\frac{\beta}{\alpha}$ *are smaller.*

An increase in the average productivity, or region B's relative marginal benefit of public good incurs a higher utility loss during the war, and consequently decreases the economic incentive to separate. A rise in the perceived winning probability or the survival rate of civil war increases the expected post-war welfare of residents from region B, and then encourages more of them to support secession. A larger cultural distance increases the political incentives for separation.

3.3.3 Further Discussion

As promised earlier, here I discuss the effectiveness of the punishment phase. After winning the civil war, region A has the alternative of returning to the pre-war political mechanism. In the case, A's welfare maximization problem becomes

$$\max -(x - x_A)^2 + \alpha g + w(N)y^A(1 - \tau)$$

with the budget constraint

$$\frac{g^2}{2} \leq w(N)(y^A + dy^B)\tau$$

which gives the following solution:

$$\begin{aligned} &x = x_A \\ &g = \frac{(y^A + dy^B)}{y^A}\alpha \qquad\qquad \text{if } \frac{1}{2}\alpha^2 < \frac{w(N){y^A}^2}{(y^A + dy^B)} \end{aligned}$$

The public good offering without punishment is much higher than the punishment phase. Although an ideal altruistic would favor this option, the voters care about their own welfare only in my model. To persuade them to support the other alternative, a comparison in utility is more convincing. Moreover, the sufficient condition for a stable equilibrium is most stringent in this non-punishment state, which is not desirable as stable equilibria are preferred in my model. The utility of residents in region A would be

$$\frac{1}{2}\alpha^2\frac{y^A + dy^B}{y^A} + w(N)y^A = \frac{1}{2}\alpha^2 + w(N)y^A + \frac{1}{2}\alpha^2\frac{dy^B}{y^A}$$

Given the existence of a stable equilibrium, this is always lower than the utility in punishment phase regardless of the value of $w(N)$. Additionally, without punishment

the agents in region B would achieve a higher welfare once they lose the civil war, which would strengthen their support for secession. Both arguments reveal that the punishment phase is a better option for region A. We list below the utility of region A and region B's residents for reference.

Table 3.3.1: Comparison of Utilities in Punishment and Non-punishment Phases

	Region A	Region B
Without Punishment	$\frac{1}{2}\alpha^2\frac{y^A+dy^B}{y^A}+w(N)y^A$	$-(x_A-x_B)^2+\alpha\beta\frac{y^A+dy^B}{y^A}+w(N)y^B-\frac{(y^A+dy^B)y^B}{2y^{A2}}$
if $w(N)\leq\frac{y^A}{y^A+dy^B}$ With Punishment	$\frac{1}{2}\alpha^2+w(N)(y^A+dy^B)$	$-(x_A-x_B)^2+\alpha\beta$
if $w(N)>\frac{y^A}{y^A+dy^B}$ With Punishment	$\frac{1}{2}\alpha^2\frac{w(N)(y^A+dy^B)}{y^A}+y^A$	$-(x_A-x_B)^2+\alpha\beta\frac{w(N)(y^A+dy^B)}{y^A}$ $+(w(N)y^B-\frac{(1-w(N))y^A}{d})(1-\frac{w(N)(y^A+dy^B)\alpha^2}{2y^{A2}})$

$\tau<1$ is another preliminary assumption that we would like to justify here. When $\tau=1$, any agents would be indifferent between working or not, and the output is nonzero only if they work. With a significant proportion of non-working citizens, the budget constraint and the original fiscal plan would not be supportable. Thus, even if we did not assume $\tau<1$, the supposed equilibrium with $\tau=1$ cannot be sustained. According to the data from the world bank, government revenue in a typical year ranges from below 10% of GDP to around 50%. Within this range, the tax revenue still increases with the tax rate, which supports the setup of my budget constraint.

People may wonder why we still assume that region A could take all the region B's output survived after winning the war in the punishment phase. We can indeed assume part of region B refuse to work or pay for war damage, but this would not affect B's utility function or the sufficient condition for $\tau<1$. Even if region A cannot collect any compensation from region B, they will still produce α units of local public good to maximize their social welfare. With the presupposition that region A can produce α in the case of losing the civil war, this condition is guaranteed. As a result, I can still assume that region A can use the law enforcement to collect their compensation after the civil war.

3.4 Bargaining under Federation with Heterogeneous Income

In the previous section, I assumed that individuals have homogeneous income and ideology preferences if they are from the same region. But in the data, there is a wide distribution in income levels within the same country or district. For a high income people, the ideological gain of declaring independence is much smaller than their private consumption lost in the case of civil war. Thus, some conclusions may change under the situation of heterogeneous income. Also, the income distribution is seldom a symmetric one, and the median income is normally below the mean income.

Let y_R^A denote the income of the representative in region A, and y_R^B denote the income of the representative in region B. Under federation, if region B refuses to pay for the local public good, the problem becomes

$$\max -(x - x_A)^2 + \alpha g + y_R^A(1 - \tau^A)$$

with the budget constraint

$$\frac{g^2}{2} \leq y^A \tau^A$$

Therefore, $g = \frac{y^A}{y_R^A}\alpha$ if $\frac{1}{2}\alpha^2 \frac{y^A}{y_R^A} < y_R^A$, and $x = x_A$. The bargaining proposal made by region A's representative will maximize his utility, while guaranteeing region B's utility of not paying tax, i.e. the proposal would be,

$$\max -(x - x_A)^2 + \alpha g + y_R^A(1 - \tau)$$

such that

$$\frac{g^2}{2} \leq y^A\tau + dy^B\tau$$
$$-(x_B - x)^2 + \beta g + y_R^B(1-\tau) \geq -(x_B - x_A)^2 + \beta\alpha\frac{y^A}{y_R^A} + y_R^B$$

The solution is

$$x = x_A$$
$$g = \alpha\frac{y^A + dy^B}{y_R^A} \qquad \text{if } \frac{1}{2}\alpha^2 < \frac{y_R^A}{y_A + dy_B}y_R^A \text{ and } \frac{\beta}{\alpha} > \frac{y^A + dy^B}{2dy_R^A}\frac{y_R^B}{y^B}$$

The utility of person i in region A with income y_i^A is

$$u_i(y_R^A) = \alpha^2\frac{y^A + dy^B}{y_R^A} + y_i^A - y_i^A\frac{\alpha^2(y^A + dy^B)}{2y_R^{A2}}$$

Hence, the most preferred representative for i would be the y_R^A, such that

$$\max_{y_R^A} u_i(y_R^A) = \alpha^2\frac{y^A + dy^B}{y_R^A} + y_i^A - y_i^A\frac{\alpha^2(y^A + dy^B)}{2y_R^{A2}}$$

Solving this maximization problem, we obtain $y_R^A = y_i^A$, and we observe that $u_i(y_R^A)$ will decrease with $\|y_R^A - y_i^A\|$. Thus the Condorcet winner of the game will be the one with the median income y_m^A, and the implemented public good of $\alpha\frac{y^A+dy^B}{y^A}\frac{y^A}{y_m^A}$. The difference between this result and the homogeneous one is the factor $\frac{y^A}{y_m^A}$. Similar to this case, after a civil war, the elected representative of region A and region B would be the median voter as well, regardless to whoever the winner of the civil war is. Moreover, the public good offered is amplified by $\frac{y^k}{y_m^k}$, a ratio due to the distortion in the symmetric income distribution. With a quadratic per capita cost function of the public good, the tax rate would rise even faster, and hit the 100% boundary sooner.

We list the current public good delivery, tax rate, and necessary condition for a stable equilibrium of each condition below.

Table 3.4.1: Comparison between Homogeneous and Heterogeneous Income

	Homogeneous Income	Heterogeneous Income
Under Unification		
Public good	$\frac{y^A+dy^B}{y^A}\alpha$	$\frac{y^A+dy^B}{y_m^A}\alpha$
Tax rate	$\frac{1}{2}\alpha^2\frac{(y^A+dy^B)}{{y^A}^2}$	$\frac{1}{2}\alpha^2\frac{(y^A+dy^B)}{{y_m^A}^2}$
Necessary conditions for stable equilibria	$\frac{1}{2}\alpha^2\frac{(y^A+dy^B)}{{y^A}^2}<1$	$\frac{1}{2}\alpha^2\frac{(y^A+dy^B)}{{y_m^A}^2}<1$
	$\frac{\beta}{\alpha}>\frac{(y^A+dy^B)}{2dy^A}$	$\frac{\beta}{\alpha}>\frac{(y^A+dy^B)}{2dy_m^A}\frac{y_m^B}{y^B}$
With Civil War Onset		
If region B wins		
Public good	α	$\frac{y^B}{y_m^B}\alpha$
Tax rate	$\frac{\alpha^2}{2w(N)y^B}$	$\frac{1}{2}\alpha^2\frac{y^B}{w(N){y_m^B}^2}$
Necessary conditions for stable equilibria	$\frac{1}{2w(N)}\alpha^2<\min\{y^A,y^B\}$	$\frac{1}{2w(N)}\alpha^2<\min\{\frac{{y_m^A}^2}{y^A},\frac{{y_m^B}^2}{y^B}\}$
If region B loses and $w(N)\leq\frac{y^A}{y^A+dy^B}$		
Public good	α	$\frac{y^A}{y_m^A}\alpha$
Tax rate	$\frac{\alpha^2}{2w(N)(y^A+dy^B)}$	$\frac{\alpha^2}{2w(N)(y^A+dy^B)}\frac{(y^A)^2}{(y_m^A)^2}$
Necessary conditions for stable equilibria	$\frac{\alpha^2}{2w(N)(y^A+dy^B)}<1$	$\frac{\alpha^2}{2w(N)(y^A+dy^B)}\frac{(y^A)^2}{(y_m^A)^2}<1$
If region B loses and $w(N)>\frac{y^A}{y^A+dy^B}$		
Public good	$\frac{w(N)(y^A+dy^B)}{y^A}\alpha$	$\frac{w(N)(y^A+dy^B)}{y_m^A}\alpha$
Tax rate	$\frac{1}{2}\alpha^2\frac{w(N)(y^A+dy^B)}{{y^A}^2}$	$\frac{1}{2}\alpha^2\frac{w(N)(y^A+dy^B)}{{y_m^A}^2}$
Necessary conditions for stable equilibria	$\frac{1}{2}\alpha^2\frac{w(N)(y^A+dy^B)}{{y^A}^2}<1$	$\frac{1}{2}\alpha^2\frac{w(N)(y^A+dy^B)}{{y_m^A}^2}<1$

If the income distribution is symmetric, or the median and mean income always coincide, then heterogeneity will not affect any conclusions from the homogeneous case. However, with a median income below the mean income level, some conclusions from the previous section would change. Comparing to the symmetric case, both the federation and independent regions will provide more public good, as $\{\frac{y^A}{y_m^A},\frac{y^B}{y_m^B}\}>1$.[1] Low income individuals natually prefer a bigger government to a smaller one, so the both the tax rate and public good provision will rise together. Subsequently, the preferred tax rate may surpass the 100% limit for a stable equilibrium sooner, or, mathematically speaking, $\frac{1}{2}\alpha^2<\frac{y_R^A}{y_A+dy_B}y_R^A$ is less likely to hold with a lower $\frac{y_m^A}{y^A}$. Nevertheless, the cutting off point, where region B exhausts all the private income to

[1] According to the data set available from the world bank, for most of the countries, the ratio of median and mean per capita income is between 0.5 and 1.

pay for the punishment, remains the same, since the total transfer only depends on the mean income level.

Given $w(N) \leq \frac{y^A}{y^A+dy^B}$, individual i in region B with income y_i^B will be in favor of separation if

$$p(\frac{1}{2}\alpha^2 + w(N)y_i^B) + (1-p)(-(x_A - x_B)^2 + \alpha\beta\frac{y^A}{y_m^A}) \geq -(x_A - x_B)^2 + \alpha\beta\frac{y^A + dy^B}{y_m^A} + y_i^B(1 - \frac{y^A + dy^B}{y_m^A}\frac{\alpha^2}{2y_m^A})$$

It is straightforward to verify that the difference between the LHS and the RHS of the above is either always increasing or always decreasing in y_i^B. When it is increasing, all agents in region B with income y_i^B above median income y_m^B are in favor of separation whenever the latter prefers separation, and all agents with income below the median income are in favor of unification whenever the median income agent is in favor of unification. When it is decreasing in i's income, the reverse is true. Thus, it suffices to determine region B's decision upon the preference of the median voter, which can be developed as

$$p(\frac{1}{2}\alpha^2 + w(N)y_m^B) + (1-p)(-(x_A - x_B)^2 + \alpha\beta\frac{y^A}{y_m^A}) \geq -(x_A - x_B)^2 + \alpha\beta\frac{y^A + dy^B}{y_m^A} + y_m^B(1 - \frac{y^A + dy^B}{y_m^A}\frac{\alpha^2}{2y_m^A})$$

An identical argument for $w(N) \leq \frac{y^A}{y^A+dy^B}$ implies that the median income voter is decisive in region B, and they will vote for secession if

$$\begin{aligned}
&p(\frac{1}{2}\alpha^2\frac{y^B}{y_m^B} + w(N)y_m^B) + \\
&(1-p)(-(x_A - x_B)^2 + \alpha\beta\frac{w(N)(y^A + dy^B)}{y_m^A} + \frac{y_m^B}{y^B}(w(N)y^B - \frac{(1-w(N))y^A}{d})(1 - \frac{w(N)(y^A + dy^B)\alpha^2}{2{y_m^A}^2})) \\
&\geq -(x_A - x_B)^2 + \alpha\beta\frac{y^A + dy^B}{y_m^A} + y_m^B(1 - \frac{y^A + dy^B}{y_m^A}\frac{\alpha^2}{2y_m^A})
\end{aligned}$$

With all the necessary conditions for a stable equilibrium in nationalistic policy, the major conclusion for the last section remains the same after following a similar argument. When p, $w(N)$ and $\|x_A - x_B\|$ are larger or per capita income and $\frac{\beta}{\alpha}$ smaller, the relationship is more likely to hold. As discussed earlier in this section, here the sufficient conditions for a stable nationalistic policy are stricter in comparison to the homogeneous case. On the contrary, there is no monotonic relationship between $\frac{y_m^A}{y^A}$ ($\frac{y_m^B}{y^B}$) and region B's decision of announcing independence. Fixing the mean income

of both regions, a lower median income would diminish the economic cost of the civil war, but a corresponding higher tax and public good provision may intensify the benefit of unification. Without a numerical assumption of the key ingredients in those inequalities, it is not clear which effect will dominate. One possible explanation would be that some quadratic function of $\frac{y_m^A}{y^A}$ ($\frac{y_m^B}{y^B}$) may be related to the decision making process. This implication will be tested in the empirical section of this paper.

To conclude the outcomes of this section, we have the following proposition

Proposition 3.4.1. *(1) The stable equilibrium of nationalistic policy exists if* $\frac{\alpha^2}{2} < \min\{w(N)\frac{{y_m^A}^2}{y^A}, w(N)\frac{{y_m^B}^2}{y^B}, \frac{{y_m^A}^2}{w(N)(y^A+dy^B)}\}$, *and* $\frac{\beta}{\alpha} > \frac{(y^A+dy^B)}{2dy_m^A}\frac{y_m^B}{y^B}$.
(2) When $\frac{y^A+dy^B}{1+d}$, *or* ($\frac{\beta}{\alpha}$) *increases, the minority is less willing to announce independence.*
(3)When p, $w(N)$ *or* $\|x_A - x_B\|$ *increases, the minority is more willing to announce independence.*

The proposition implies that there is a smaller parameter rang to obtain a stable equilibrium under income heterogeneity. It also shows that A rise in the average productivity ($\frac{y^A+dy^B}{1+d}$), or region B's relative marginal benefit of public good ($\frac{\beta}{\alpha}$) decreases the economic incentive to separate by leading to a higher utility lost during the war. A higher perceived winning probability or survival chance after civil war increases the expected post-war welfare of residents from region B, and consequently encourage more of them to support secession. A larger gap in cultural difference increases the political incentives for separation. There is no systematic monotonicity between the $\frac{y_m^A}{y^A}$ ($\frac{y_m^B}{y^B}$) and region B's decision of announcing independence.

3.5 Empirical result

The model outlined in this paper emphasizes some of the challenging issues raised by the disintegration of countries, with particular focus on the role of nationalistic

fiscal and ideological policy. It stresses the importance of both economic and cultural incentives in the referenda of secession. More specifically, the discussion in the previous sections entails a number of prediction for the policy outcomes. In brief, when p, $w(N)$, and $\|x_A - x_B\|$ are larger, the relationship is more likely to hold. Oppositely, fixing $\frac{y^A}{y^B}$, the inequality is less likely to hold with a larger y^B or $\frac{\beta}{\alpha}$. An increase in cultural difference, measured as an increase in $\|x_A - x_B\|$, will increase the political motives for separation of region B. Higher p or $w(N)$ means smaller cost of civil war, and this consequently provides region B with economic incentive to declare war. $\frac{\beta}{\alpha}$ describes the economic gain of region B contributing to the expense of public good with region A. Also, keeping $\frac{y^B}{y^A}$ constant, a higher y^B means a greater loss to their private consumption after war.

3.5.1 Define the determinants of civil war onset

Countries with a large population, or low average productivity tend to face a high risk of civil war breakout. To measure these effects directly, we use two variables: log of population and GDP per capita. These two variables are extracted from the World Bank and Penn World Table. Fearon and Laitin (2003) believed that the per capita income is a proxy for a state's overall financial , administrative, police and military capabilities. Therefore, a higher income should be associated with a lower risk of civil war onset. They also included population as one of the determinants of civil war onset, as a larger population makes it more challenging to keep tabs on who is doing what at the local level which increase the potential recruits to an insurgency.

The relative marginal benefit of public good $\frac{\beta}{\alpha}$ cannot be estimated directly, but the non-contiguous regions of any country naturally will not get as much spillover from the central government as the contiguous regions. Moreover, the central government may provide some public service targeting the local residents, which is very costly for

the residents from a contiguous region to obtain. Thus, I will use the dummy variable - non-contiguous state - as an indicator of low $\frac{\beta}{\alpha}$.

The perceived probability p of winning the civil war for the minority group is another variable which cannot be gauged. According to Fearon and Laitin (2003), the presence of rough terrain and the instability of the central government should favor insurgency and civil war. Their paper also pointed out that a newly independent state, which suddenly loses the coercive backing of the former imperial power, may not have their new military capabilities ready for a war. Moreover, the oil export countries tend to have weaker state apparatuses, since the rulers have less need for a elaborate bureaucratic system to raise revenue, which results in a weaker government force to fight the war. Collier and Hoeffler (2004) also pointed out that the primary commodities are associated with poor governance. Therefore, I use log of % mountainous, oil export, unstable government and new state as indicators to measure the perceived chance of winning the civil war for the minority group. Fearon and Laitin (2003) also suggested that, with a territorial base separated from the state's center by water or distance, the political and military technology of insurgency will be favored. Both my argument about $\frac{\beta}{\alpha}$ and their argument agree on that a country with a contiguous region face a higher risk of civil war.

In addition, there is no direct measure of the cultural difference $|x_A - x_B|$. Fearon and Laitin (2003) designed a dummy variable based on the descriptive annual report on Religious Freedom by U.S. State Department about policy discrimination, yet did not find its coefficient to be significantly different from 0. I use the same source, but code the variable in a slightly different way. Instead of focusing on the government's attitude of different religious group, I code 1 on the countries with the inter-group religious tension regardless to the government's position. Since we can hardly find quantitative estimate of the cultural difference, we can only use this dummy to approximate $|x_A - x_B|$.

α, $\frac{d}{1+d}$ and $\frac{y_m^A}{y^A}$ (or $\frac{y_m^B}{y^B}$) also showed up as control variables in the theory part, even though without a clear monotonic relationship with the probability of civil war onset. To capture the characteristics of the determinants of civil war onset, I will incorporate all these in our empirical test. α cannot be estimated numerically, and the polity score from the polity IV project will be included as an approximation. I will use ethnic fractionalization to estimate $\frac{d}{1+d}$. Since the median-mean income ratio is only available for OECD countries, we will use the GINI coefficient as an approximate for the income inequality within each country. Fearon and Laitin (2003) stated that lower civil liberty (polity score), high income inequality, and ethnic diversity lead to a higher risk of civil war.

For simplicity, assume that there is only one major minority group of each country which would raise secession issues. Let S denote their decision of declaring independence and initiating civil war, and it can take only two values $S = 0, 1$. $S = 1$ means that there is a civil war onset. At any point in time, country i belongs to one of these two states, denoted by S_i. Suppose the choice of beginning a war by country i can be described by the index model

$$S_i = \begin{cases} 1 & \text{if } F(W_i) + \eta_i \geq 0 \\ 0 & \text{if } F(W_i) + \eta_i < 0 \end{cases}$$

where W is a set of observed variables influencing the observed choice of starting a civil war, and $F(W_i)$ is the net change in expected utility after initiating the war. Members of W would involve variables, related to the key factors discussed in our theory model. Other unobserved country specific factors are summarized by the random variable η_i. Throughout, we assume that η and W are uncorrelated.

3.5.2 Multivariate Results

To study the key factors inducing civil wars, I update the data set used by Fearon and Laitin (2003) to 2009. The GINI coefficients and the dummy indicating the religious tension are also attached to each country-year. I coded a variable onset as 1 for all country-year in which a war started and 0 for all others according to the PITF case list. I include the pure ethnic war and the complex war with ethnic or religious factors involved. Thus, my coding rule does not match Fearon and Laitin (2003) or Collier and Hoeffler (2004) case by case, since they coded a civil war even if it is irrelevant to any ethnic or religious factors. Model 1 in Table 3 shows the results of a logit analysis using onset as the dependent variable and the independent variables are specified earlier in this section.[2] Prior war is a control variable indexing whether the country had a distinct civil war ongoing in the previous year. As the data are grouped duaration data, the possibility of temporal dependence between observations may need to be considered. For example, the growth rate in population or GDP per capita is highly persistent over the successive peaceful years. To allow for this kind of persistence, the speficication of Table 3 adds the lagged war dummy.

Per capita income (measured in 1985 US dollars in thousand and lagged one year) is strongly significant, so the probability of civil war onset is higher with lower per capita income. Holding other variables at their median values, a country in the tenth percentile on income has a 7.7% chance of ethnic civil war breakout over a decade, compared to 5.2% chance for a country with median income and a 0.6% changce for a country at ninetieth percentile ($573, $2205, $10756, respectively). One intersting observation is that western countries are the ones with higher income and lower chance of civl war break out. So we introduce a dummy variable as western, and find that the estimate of model 1 drops to -0.209. However, the coefficient is still significant different from 0, and the conclusion remains that countries are less likely to have civil

[2]The main result will not change much with the porbit test.

war with higher income. The coefficient of population is significant different from zero, with a positive sign. Holding other variable at medians, the risk of civil war onset over a decade is 3.6% at the tenth precentile population (1.45million) versus 7.8% at the ninetieth precentile population (58.7million). We do control the ethnic diversity, so it is statistically significant not due to the ethnic composition. The effect of noncontiguous state is statistically insignificant.

The estimates for new state is strongly significant. Holding other variables at their median values, in the first two years of independence a country has a 2.8% chance of civil out break, compared to 0.5% chance for a country established for more than 2 years. One way to interpret this result is that the government is the new state is not fully established, and the minority would think it is easier to fight in this situation. Our measure of internal tension between religious groups is associated with systematically higher risks of civil war onset. The median country had a 5.2% chance of civil war over a decade, whereas the same country with internal religious tension would have an estimated 9.0% chance. The estimate of religious fractionalization is also significant, which is not explained by my theory model.

The other variables do not come close to statistical significance. We summarize the expected sign and the estimate of the determinants of ethnic civil war onset on Table 3 below.[3]

3.5.3 Robustness Checks

Different regions of the world share a variety of historical, cultural and economic tradition, and some regions seem more prone to civil war. Therefore, we may wonder if any of the variables in the multivariate analysis just proxy for such factors. I add all (but one) regional dummies to our current logit analysis, the coefficients and significance levels are little affected. The regional dummies are joint significant

[3] $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 3.5.1: Logit Analysis of Determinants of Ethnic Civil War Onset, 1945-2009

		Expected Sign	Estimate
Prior War			-0.105
Per Capita Income	GDP Per Capita	$-$	-0.250^{***}
Population	log(population)	$+$	0.215^{*}
$\frac{\beta}{\alpha}$	Non-contiguous State	$+$	0.166
Perceived Chance of Winning	% Mountainous	$+$	0.039
	Oil Exporter	$+$	0.619^{***}
	Unstable Government	$+$	0.134
	New State	$+$	1.721^{***}
$\lvert x_A - x_B \rvert$	Religious Tension	$+$	0.554^{*}
α	Polity IV Score		0.016
$\frac{d}{1+d}$	Ethnic Fractionalization		0.769
	Religious Fractionalization		1.453^{*}
$\frac{y_m^B}{y^B}$	GINI Coefficient		-0.002

in a likelihood ratio test $p = 0.0011$. Including the region dummies individually reveals that a median sub-Saharan African countries has a rate of civil war onset 9.5% significantly higher than a median countries 5.2% not in this region over a decade. As discussed before, it could be that sub-Saharan African countries tend to have lower per capita income. None of the other regional dummies has a significantly higher risk of civil war, as the country characteristics already included in the model.

If added to my test on Table 3, dummy variables marking each decade (but one) are jointly significant in a likelihood test ($p = 0.0021$). There is a general upward trend in civil war risk after 1970s. Adding a dummy for the 1970s, and a variable marking the years indicates that from 1970, the odds of civil war rose 3% per year. Part of the explanation is that the income gap between under developed nations and the industrial nations kept growing, which made the under developed nations relatively poorer in comparison to 40 year ago. As discussed before, income effect plays a significant role in predicting the risk of civil war outbreak.

3.6 Conclusion

This paper's contributes to the study of the nature of secession by introducing the effect of civil war. Different from the classical literature which allowed for free secession, we assume that the federal government will intervene any insurgency with military means. Then, the corresponding economic cost of war distorts the political choice of the minority regions. Additionally, the minority region may not gain independence even after the civil war, which further depresses their political incentive of secession. To maximize their utility and minimize the minority region's willingness to declare independence, the majority region prefers to introduce a punishment phase to the minority region after winning.

Another major observation from this paper is on the fiscal policy of the federation or the newly independent regions. I find that the public good is under-provided at a suboptimal level after secession. However, with an asymmetric income distribution, the public good provision is altered towards the social optimal level. The social planner will always prefer the federation, since the marginal utility of public good is higher.

To illuminate the difference between how the majority and the minority are treated in the fiscal policy, I only include the residents from the majority region as the direct recipients of the public service. This answers the question why the oppressed minorities are still willing to remain under unification. As far as the marginal spillover per unit tax revenue is higher than the marginal benefit of providing their own public service, the public good creates economic incentive for the minority to stay together. This also explains why the low income minority rather leaves the high income federation to be independent. If they are treated equally as the recipients, the poor minority should be happy to enjoy the public good subsidized by their wealthy neighbors. Furthermore, the majority is always better off with the fiscal transfer from the minority. The model with increasing return to scale in production also provides an economic

incentive for the majority. However, it leads to a conclusion that larger countries are less civil war prone, which contradicts the result of most empirical works.

The empirical estimate supports the prediction of my theory model, in particular that the countries with a larger population, a lower per capita income, serious tensions among religious groups are at a higher risk of civil war. Other than these factors, the state weakness marked by a recent independence or being an oil exporter also favors insurgency. Among all regions, the Sub-Saharan African countries are more likely to have a civil war onset, which is consistent with their low per capita income and frequent inter-religion conflicts. If our analysis is correct, then the policy makers should pay more attention to their economic growth and try to ease tensions among religious groups. In specific terms, the international organizations should develop programs that improve the internal coordination among religious or ethnic groups and make aid to governments of developing countries to surge their growth.

This simple model cannot catch all relevant factors in the very complex question of ethnic groups and country formation. In particular, there is no multi-period bargaining between the minority and the majority. This means that the model cannot explain why some civil wars last longer than others and how this will affect the political choice of both groups. Furthermore, even within a single period game, our assumption of the homogeneous ideological preference with a region can be removed. In that case, the elected representative may not be the median income voter any more, and would try to alter the political result on secession for his own interest.

3.7 Appendix

Given $w(N) > \frac{y^A}{y^A+dy^B}$, region B will vote for secession, if

$$
\begin{aligned}
&p(\frac{1}{2}\alpha^2 + w(N)y^B)+ \\
&(1-p)(-(x_A-x_B)^2 + \alpha\beta\frac{w(N)(y^A+dy^B)}{y^A} + (w(N)y^B - \frac{(1-w(N))y^A}{d})(1-\frac{w(N)(y^A+dy^B)\alpha^2}{2y^{A2}})) \\
&\geq -(x_A-x_B)^2 + \alpha\beta\frac{y^A+dy^B}{y^A} + y^B(1-\frac{y^A+dy^B}{y^A}\frac{\alpha^2}{2y^A})
\end{aligned}
$$

Since region B is certainly better off to win the war or implement x_B, so the LHS will increase with p and $|x_A - x_B|$. To show the net change in expected utility will decrease with y^B, given $\frac{y^A}{y^B}$ fixed, I find

$$\frac{\partial(LHS-RHS)}{\partial y^B} = pw(N) + (1-p)(w(N) - \frac{(1-w(N))y^A}{dy^B}) - 1 < w(N) - 1 < 0$$

To show the LHS increases with $w(N)$, we have

$$
\begin{aligned}
\frac{\partial LHS}{\partial w(N)} &= py^B + (1-p)(\beta\alpha\frac{y^A+dy^B}{y^A} + \frac{y^A+dy^B}{d}(1+\frac{\alpha^2}{2y^A} - \alpha^2 w(N)\frac{y^A+dy^B}{y^{A2}})) \\
&= py^B + (1-p)(\frac{y^A+dy^B}{d}(\frac{d\beta\alpha}{y^A} + 1 + \frac{\alpha^2}{2y^A} - \alpha^2 w(N)\frac{y^A+dy^B}{y^{A2}})) \\
&> (1-p)(\frac{y^A+dy^B}{d}(\frac{d\beta\alpha}{y^A} + 1 + \frac{\alpha^2}{2y^A} - \alpha^2\frac{y^A+dy^B}{y^{A2}}))
\end{aligned}
$$

From the incentive constraint of region B paying for tax, we know $d\beta < \alpha\frac{y^A+dy^B}{2y^A}$, which is equivalent to $\frac{d\alpha\beta}{y^A} < \alpha^2\frac{y^A+dy^B}{2y^{A2}}$. Thus,

$$\frac{\partial LHS}{\partial w(N)} > (1-p)(\frac{y^A+dy^B}{d}(1+\frac{\alpha^2}{2y^A} - \alpha^2\frac{y^A+dy^B}{2y^{A2}}))$$

Since $\tau < 1$ under unification, we then have $\alpha^2\frac{y^A+dy^B}{2y^{A2}} < 1$. As a result,

$$\frac{\partial LHS}{\partial w(N)} > (1-p)(\frac{y^A + dy^B}{d}(\frac{\alpha^2}{2y^A})) > 0$$

Then, we can conclude that the higher $w(N)$ is, the more likely the civil war will happen.

Bibliography

Becker, B., and T. Milbourn, 2011, "How Did Increased Competition Affect Credit Ratings?," *Journal of Financial Economics*, 101(3), 493–514.

Benabou, R., and G. Laroque, 1992, "Using Priviliged Information to Manipulate Markets: Insiders, Gurus, and Credibility," *The Quaterly Journal of Economics*, 107(3), 921–958.

Berkowitz, D., 1997, "Regional income and secession: Center-periphery relations in emerging market economies," *Regional Science and Urban Economics*, 27, 17–45.

Bolton, P., X. Freixas, and J. Shapiro, 2012, "The Credit Rating Game," *Journal of Finance*, 67(1), 85–111.

Bolton, P., and G. Roland, 1997, "The Break-up of Nations: A political Economy Analysis," *Quaterly Journal of Economics*, 112, 1057–1090.

Buchanan, J., and R. Faith, 1987, "Secession and the Limits of Taxation: Toward a Theory of Internal Exit," *American Economic Review*, 77, 1023–1031.

Chen, Y., 2011, "Perturbed Communication Games with Honest Senders and Naive Receivers," *Journal of Economic Theory*, 146(2), 401–424.

Chen, Y., N. Kartik, and J. Sobel, 2008, "Selecting Cheap-Talk Equilibria," *Econometrica*, 76(1), 117–136.

Collier, P., and A. Hoeffler, 2004, "Greed and Grievance in Civil War," *Oxford Economic Papers*, 56, 563–595.

Crawford, V., and J. Sobel, 1982, "Strategic Information Transmission," *Econometrica*, 50(6), 1431–1451.

Ely, J. C., and J. Valimaki, 2003, "Bad Reputation," *The Quarterly Journal of Economic*, 118(3), 785–814.

Faure-Grimaud, A., E. Peyrache, and L. Quesada, 2009, "The Ownership of Ratings," *The RAND Jounral of Economics*, 40(2), 234–257.

Fearon, J. D., and D. D. Laitin, 2003, "Ethnicity, Insurgency and Civil War," *American Political Science Review*, 97(1), 75–90.

——— , 2008, "Religion, Terrorism, and Public Goods: Testing the Club Model," *Journal of Public Economics*, 92, 1942–1967.

——— , 2011, "Sons of the Soil, Migrants, and Civil War," *World Development*, 39, 199–211.

Fudenberg, D., and J. Tirole, 1991, *Game Theory*. MIT Press, Cambridge, MA.

Gradstein, M., 2004, "Political Bargaining in a Federation: Buchanan meets Coase," *European Economic Review*, 48, 983–999.

Kuhner, C., 2001, "Financial Rating Agencies: Are They Credible?-Insight into The Reporting Incentives of Rating Agencies in Times of Enhances Systemic Risk," *Schmalenbach Business Review*, 53, 2–26.

Lee, J., and Q. Liu, 2013, "Gambling Reputation: Repeated Bargaining with Outside Options," *Econometrica*, 81(4), 1601–1672.

Lizzeri, A., 1999, "Information Revelation and Certification Intermediaries," *The RAND Jounral of Economics*, 30(2), 214–231.

Mailath, G., and L. Samuelson, 2001, "Who Wants a Good Reputation," *Review of Economic Studies*, 68(2), 415–441.

Mathis, J., J. McAndrews, and J.-C. Rochet, 2009, "Rating The Raters: Are Reputational Concern Powerful Enough to Discipline Rating Agencies?," *Journal of Monetary Economics*, 56(5), 657–674.

Morris, S., 2001, "Political Correctness," *Journal of Political Economy*, 109(2), 231–265.

Oates, W., 1972, "Fiscal Federalism," *Harcourt Brace*.

Olofsgard, A., 2003, "Incentive for Secession in the Presence of A Mobile Group," *Journal of Public Economics*, 87, 2105–2128.

Ottaviani, M., and P. Sorensen, 2006a, "Professorional Advice," *Journal of Economics Theory*, 126(1), 120–142.

——— , 2006b, "Reputational Cheap Talk," *The RAND Journal of Economics*, 37(1), 155–175.

Ottaviani, M., and P. N. Sorensen, 2006c, "The Strategy of Professional Forecasting," *Journal of Financial Economics*, 81(2), 441–466.

Sharfstein, D. S., and J. C. Stein, 1990, "Herd Behavior and Investment," *The American Economics Review*, 80(3), 465–479.

Skreta, V., and L. Veldkamp, 2009, "Rating Shopping and Asset Complexity: A Theory of Rating Inflation," *Journal of Monetary Economics*, 56(5), 678–695.

Sobel, J., 1985, "A Theory of Credibility," *The Review of Economic Studies*, 52(4), 557–573.