UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EMPIRICAL ANALYSIS IN LABOR MARKET AND INTERNATIONAL FINANCE

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECONOMICS

by

Zijing Zhu

September 2020

The Dissertation of Zijing Zhu
is approved:

_____
Professor Grace Gu, Chair

_____
Professor Kenneth Kletzer

_____
Professor Brenda Samaniego de la Parra

_____
Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Empirical Analysis in Labor Market and International Finance

by

Zijing Zhu

This dissertation presents three empirical studies on topics in regional growth, international finance, and global commodity prices. The first chapter analyzes the labor market reallocation following the reduction of transportation costs. The second chapter discovers developing countries' external debt holdings abilities with different exchange rate regimes, and the third chapter reveals how commodity prices are affected by information from the news in high-frequency trading.

The first chapter discovers how labor is reallocated across cities as they become more connected to the High-Speed Railway (HSR) network in China. Specifically, by adapting graph theory from computer science, this chapter constructs a continuous connectivity index that captures the changes of 285 cities' indirect connections to the HSR network over ten years. Additionally, using China City Statistic Yearbook that covers these cities' labor market outcomes from 2003 to 2017, this chapter finds that increasing the indirect connectivity to the HSR network has insignificant effects on total employment. However, there are heterogeneous effects on employment in different industries. This chapter suggests that cities with higher indirect connectivity to the HSR network have more employment in skilled and non-service industries but less employment in service industries. Moreover, with or without direct connections to HSR generates different effects on labor reallocation. Cities have HSR contributes to the increase in

employment in skilled industries, while cities near HSR has more increase in non-service employment.

The second chapter examines how exchange rate regimes affect external debt holdings for developing countries, both at normal times and during default episodes. In order to uncover the connection of more floating exchange regimes and external debt holdings, this chapter collects 57 countries' external debt level and exchange rate regimes from 1970 to 2007, including 232 default episodes. Using GMM regressions, the results reveal that compared to countries with more fixed exchange rate regimes, countries with more floating exchange rate regimes hold less external debts, especially during default episodes. Moreover, switching from fixed exchange rate regimes to floating regimes further reduces external debt holding, this can be driven by the abandonment of fixed exchange rate regimes commitment.

The third chapter is a group project with Yifei Sheng and Yunxiao Zhang. To uncover the news impact on the price of WTI crude oil futures, the third chapter applies supervised and unsupervised machine learning algorithms to conduct news sentiment and topic analysis. With the assumption that the crude oil futures market is efficient enough to respond quickly to new information, this chapter obtains high-frequency price and news from the Bloomberg terminal. Using results from logistic regression and K-means clustering, this chapter defines the positive score and topic for each news article as inputs for the final logistic regression. The regression results show that the "World Crude Oil" news is more positively correlated with price increase than other topics. Moreover, the "WTI Crude Oil" news has the highest correlation with the price increase as the positive score increases.

To myself,

per aspera ad astra

through hardships to the stars

Acknowledgments

Pursuing a Ph.D. in economics is never an easy journey for me. Whenever I start doubting myself, I always remind myself why I am here. I would read my personal statement I used applying graduate schools, where I summarized, ”All these precious experiences in the past made me who I am now. I was never afraid of pursuing my dream, and I am certainly not now. I hope I will create more unforgettable memories in graduate school.” Finishing graduate school, I can claim that I have had great memories filled with happiness and satisfaction, despite the obstacles I encounter ALL THE TIME. Just as the saying goes, “Tough times never last, but tough people do.” Whatever I have learned here is much beyond knowledge from books. Five years are around one-fifth of my age, and I have never regretted spending them exploring my passions. I have learned so much, yet I still have so much to learn. I think I am ready for my next journey.

I want to thank my advisor, Grace, for her excellent guidance and unbelievable patience throughout the years. She has made this journey so much more comfortable for me by helping me throughout passing my qualification exam and finishing my dissertation. I want to thank my committee members, Ken and Brenda, for their efforts to read my dissertation and give me valuable suggestions. I want to thank all other professors in the department to provide me with the top knowledge in economics and give great feedback to my research. I want to thank our program coordinator Sandra, who has always been there for any small errands. I want to thank my cohorts, who are pursuing the same goals together with me. I will miss the struggles we face together, as well as the happiness we share. It is never easy to say goodbye.

However, we always have to sail for a new journey. We all meet because our passions overlap. I want to share my favorite quote with everyone who shares the same passion with me: "Nobody grows old merely by living a number of years. We grow old by deserting our ideals. Years may wrinkle the skin, but to give up enthusiasm wrinkles the soul." May we meet again one day.

Chapter 1

The Growth of High-Speed Railway Network and its Effects on Labor Reallocation in China

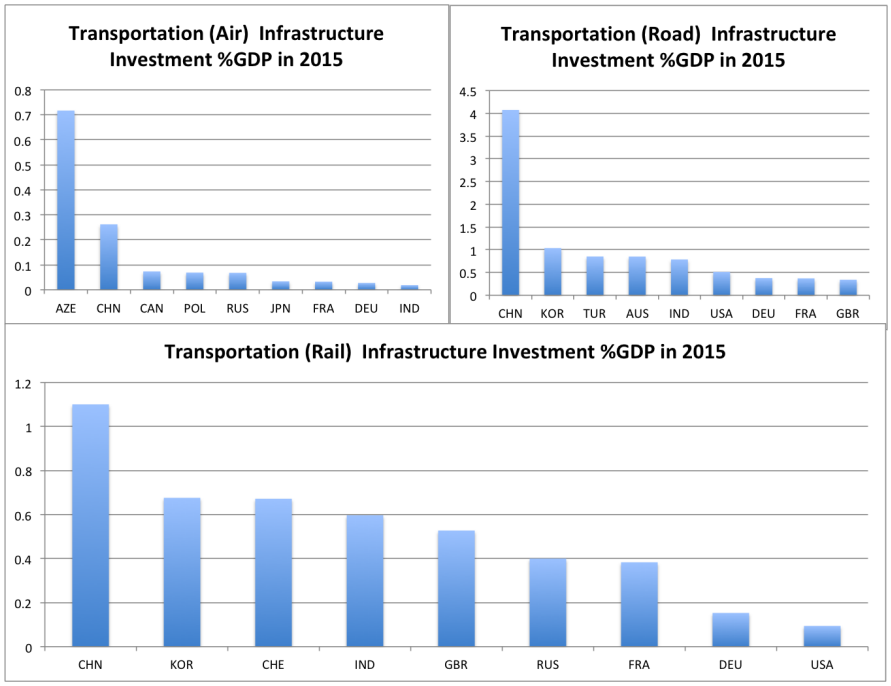
1.1 Introduction

Developing countries spend huge sums on transportation infrastructure projects that shape their cities every year. Baum-Snow et al. (2015) point out that 20% of the World Bank lending goes to transportation infrastructure for developing countries. Figure 1.1 shows ranked countries' transportation infrastructure investment as a percentage to GDP in 2015. China ranked first for spending on both highways and railways, and second in air-related transportation investments expenditures. According to the definition of high-speed rail, the first high-speed rail in China is the Beijing-Tianjin Intercity Railway, which has operated since 2008[1]. Over the decades, China's High-Speed Railway (HSR) has formed a *Four Verticals and Four*

[1] It is debatable to claim the Qinhuangdao-Shenyang High-speed Railway, operated in 2003, to be the first high-speed rail since its track could not operate over 250 km/h(the desired speed) at the beginning of the operation.

Horizontals network that connects south to north, east to west. Between 2008 and 2018, HSR has transported seven billion passengers, with a daily ridership of around 8.3 million. The introduction and development of HSR deeply shaped the Chinese economy in different aspects. This paper will focus on the HSR network's effects on the Chinese labor market.

Figure 1.1: Rank of Transportation Infrastructure Investment to GDP in 2015



Note: This graph shows the Top 9 countries who have the most investment to GDP ratio on transportation infrastructure in 2015. The types of transportation infrastructure include airline transportation, road transportation and railway transportation.
Source: The World Bank

By reducing the inter-city travel time, HSR has changed the accessibility and increased total ridership in China (Lin(2017), Cheng et al(2015), Cao et al(2013)). The increased convenience of shorter commute time and advantages of rail travel increase labor mobility among cities. Does the reduction in labor market frictions lead to labor reallocation among

cities connected to HSR, and if so, how is labor reallocated? This paper answers this question by examining how a city's connections to China's High-Speed Railway(HSR) relate to its employment levels in different industries. HSR passenger surveys from Jianbin(2011), Wu et al.(2013), Olivier et al(2014) show that passengers who take high-speed rail have a higher monthly salary, but its still unknown how HSR impacts employment across various industries. By conducting a model assuming firms sourcing tasks across cities, combining with HSR passenger surveys, this paper hypothesizes that HSR primarily impacts the employment in skilled and non-service industries positively for cities exposed to HSR.

This paper contributes to existing literature in analyzing the effects of improving transportation infrastructures worldwide, including transporting freights (Eaton and Kortum(2002), Donaldson and Hornbeck(2016), Grossman and Rossi-Hansberg(2008)) and passengers (Lin(2017), Lin(2019), Zheng and Kahn(2013), Zheng and Kahn(2018), Xu(2018), Cheng et al(2014), Rus and Nombela (2007), Heuermann et al(2018), Bernard et al(2016)). The biggest contribution of this paper is treating connectivity as a continuous variable. Rather than treating the connection as a dummy variable that only reveals the differences between having and not having the transportation infrastructure in the city, this paper further examines the effects of an increase in a city's connectivity to the HSR network. As the network keeps growing, new cities get connected, and old cities increase connectivity. An increase in the connectivity includes direct connections with more cities, and indirect connections with more cities. While direct connections reveal one city's accessibility to certain nearby cities, the indirect connectivity reveals one city's accessibility to nearby cities' nearby cities. The accessibility to nearby cities may not change significantly by the introduction of HSR since HSR has no absolute advantage compared

to buses and cars for short distance. The accessibility to cities further away is more important when measuring one city's connectivity. Besides, direct connectivity and indirect connectivity are usually highly correlated since an HSR line usually connects a large group of cities. Due to these reasons, this paper measures each city's indirect connectivity over the sample year, which is calculated as a spillover effect from this city's five closest cities. The detailed measurement of direct and indirect connectivity is described in Section 1.5.1.

Empirically, two types of endogeneity challenge the identification of causal effects. The timing of operating HSR can be correlated with unobserved labor market potentials, and the selection bias of cities connected to HSR. The introduction of HSR, at least in the beginning stage, was not induced by economic conditions for a certain area but rather based on the existing tracks, which were built a long time ago. As the requirement for HSR train speed increased, new tracks were built in different regions. The construction date may be affected by the economic potentials of certain regions, the open date for a certain HSR line, however, was largely affected by the engineer difficulties of construction (Lin (2017)). After controlling the location-time fixed effect, the timing of operation should be identified. Moreover, this paper measures the city's indirect connections to the HSR network through its nearby cities. For each city i, after eliminating its all direct connections to its nearby cities, the city i's nearby cities direct connections were solely driven by the nearby cities economic conditions, which is unlikely to be correlated with city i.

Another concern is the selection bias of cities with direct HSR connections. According to the Ministry of Railway of China, the goal of building the HSR network is to connect all capital cities for each province in China. By the end of 2018, over 100 cities connect to

the HSR network. Many of them are byproducts of connecting the capitals and make capitals more accessible. One can interpret this ultimate goal by viewing cities with or without HSR in terms of their distance to the nearest capital, rather than their economic conditions. Section 1.5.2 presents more evidence supporting these statements.

After a city directly or indirectly connected to HSR, this city may have a labor inflow or outflow or have inflow and outflow at the same time. Thus, the overall effect on employment can be ambiguous. To fully uncover the labor reallocation among cities, Section 1.4 constructed a model with firms sourcing tasks among cities to produce final goods. The model suggests that reducing transportation cost among cities can increase wages for cities with direct transportation improvement, as well as its nearby cities. As labor income increase, the living costs increases if the housing supply is inelastic. Thus, labor in different industries will react differently to the decrease in transportation cost. Labor who have higher income, like labor in skilled industries, or labor in industries with higher demand, such as non-service industries which usually receive sourcing tasks, will move to have labor inflow, while labor in other industries may move out due to higher living cost.

By conducting regressions analyzing the effects of increasing indirect connectivity on employment level and industry compositions, this paper finds that there is no significant change in overall employment level. However, increasing indirect connectivity affects industry compositions differently with different exposures to HSR. Increasing one unit of indirect connectivity in HSR increases skilled employment by 17%, and non-service employment by 13%, and it decreases service employment by 8%. After distinguishing cities with HSR or get exposed to HSR from nearby cities, this paper discovers that the increase in skilled employment and the

decrease in service industries mostly come from cities that have HSR, while the increase in non-service employment comes from cities near HSR. Therefore, labor in skilled industries are more likely to move in cities with HSR due to higher wage, and labor in non-service industries are more likely to move toward cities near HSR due to higher demand.

Moreover, this paper also discovers the fact that HSR increases the specialization patterns in terms of industry compositions. The paper reveals that with an increase in indirect connectivity, a city's employment in an industry increases more if this city and its nearby cities have very different compositions in this industry. For example, the skilled intensive city will attract more skilled workers if its nearby city is not skilled intensive. The results suggest that a city connects to HSR has an increase in job opportunities, not only for labor in this city, as well as its nearby cities, due to a reduction in labor market frictions.

The paper is formatted as follows: Section 1.2 discusses the development the stylized facts regarding HSR in China; Section 1.3 reviews related literatures; Section 1.4 refers to the theoretical backgrounds; Section 1.5 describes the methodologies in measuring connectivity and solving endogenous issues in this paper; Section 1.6 shows some evidence that backs up the hypothesis and regression results; Section 1.6 presents the empirical results and Section 1.7 concludes.

1.2 Facts about HSR in China

1.2.1 The Development of HSR in China

The HSR network in China is the largest network in high-speed rail, covering almost two-thirds of the world's commercial high-speed rail tracks[2]. The 2014 Railway Safety Management Regulation classifies the rails in China into three types depending on their speed and usages. According to the Regulation, only trains running above 250 km/h(155 MPH) in passenger dedicated lines are called the high-speed rail. Trains with speed between 160 km/h to 250 km/h, which also share lines with freight transportation, are classified as fast rail, and the rest are the normal rails. Compared with fast rail and normal rail, high-speed rails run much faster by building a seamless track and ballastless track. At the early stage of the development of HSR, HSR trains were running on the existing tracks for normal trains. However, with the increasing demand for faster trains, it became essential to start building new tracks to support higher speed. Technically, the ballastless tracks can ensure trains to run over 300 km/h. Based on the different definitions of rails, this paper excludes the fast rail, which is different from some of the literature. Most of the fast rail lines transport passengers and freight. Even though the speed for fast rail is still significantly higher than normal rail, this paper only examines the economic outcomes in the labor market, which is induced by passenger travel.

By the end of 2018, HSR has extended to 31 of China's 33 provincial-level administrative divisions and reached 29,000 kilometers (18,000 miles) in total length. According to the

[2]Source: https://en.wikipedia.org/wiki/List_of_high-speed_railway_lines_in_China

Mid-to-Long Railway Plan[3], the Four Verticals (connect north with south) and Four Horizontals (connect east with west) were mostly completed by 2014. Since then, the goal has been to increase the total length of HSR to approximately 30,000 km by 2020 and start building new rails that extend the network to be "Eight Verticals and Eight Horizontals." The figures in Appendix A.0.1 show the development of the HSR network in China over the years from 2008 to 2018. In the figures, each blue circle indicates a city that was connected to the HSR network at that year. The black link between cities is HSR tracks. The size of the circle stands for the employment size in this city at that year[4]. Based on the figures, it is clear that HSR lines were first built in the eastern and southern parts of China. Geographically speaking, eastern and southern regions are mostly plains with more comfortable weather conditions, which makes building new tracks much easier. Due to the same reasons, the eastern and southern parts of China have higher population intensities, which makes building HSR more profitable. Starting from 2014, HSR was extended to the western and northern parts of China and connected more cities in different regions. At 2018, all cities connected to HSR are in the same network[5]

1.2.2 The Usage and Passengers

Compared to normal rails, HSR reduces at least half of the intercity travel time. HSR also provides better conditions during the trip, which including the improvement of the equipment, like the seats, restrooms, and the hygiene in the train. Compared to airplanes, HSR is

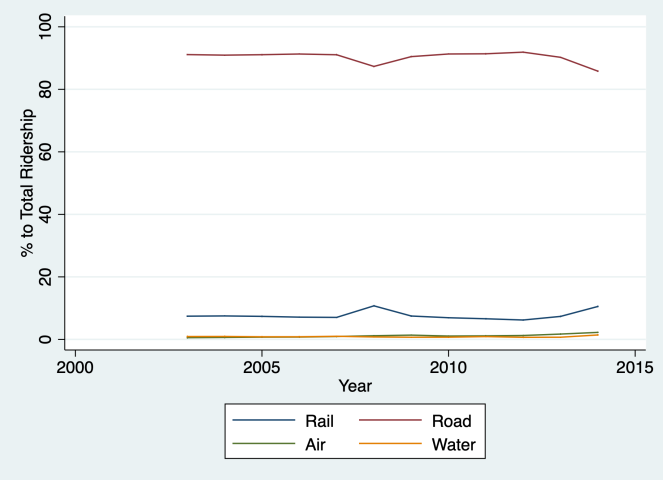
[3]Published by the ministry of railway of China

[4]Except in 2018, due to data availability, the size of employment is the city's employment level in 2017.

[5]Meaning that a city connected to HSR can reach any other city connect to HSR with HSR, no matter in which region.

much cheaper, and the price is much more stable. Besides reducing in intercity travel time, HSR can also reduce the intracity travel time. Unlike airports, HSR stations are usually built near the center of the cities. They are highly connected with intracity transportation, such as subways, taxis, and buses, which deliver passengers to destinations throughout the city.

Figure 1.2: Ridership of Different Transportation Methods



Note: This graph shows the changes in the usage of different transportation methods over the years, including the railways, airlines, highways and road, and waterways.
Source: China National Statistic Yearbook

Figure 1.2 shows the average usage of different transportation methods over the years. As shown in Figure 1.2, more than 80% of the trips are made by cars, buses, and coaches. Moreover, railway accounts for around 10% of the ridership. The figure also indicates the fact that more short trips are made since cars, buses, and coaches are dominant transportation methods for shorter distance. According to Lin(2017), HSR ridership steadily increased from 300 million in 2010 to 830 million in 2014. Nowadays, HSR is a strong competitor for both long and short distance trips. (Lin (2017)).

The fact that high-speed rails are comparatively more expensive backs up the hypothesis that high-income groups are more likely to choose high-speed rail as their transportation method. According to passenger surveys(Jianbin(2011), Wu et al.(2013), Olivier et al(2014)), the average monthly income for passengers who take high-speed rail ranges from 4300 to 6700 yuan(600 to 1000 dollars), which is classified as the high income group in China[6]. Besides, the surveys also reveal that a large portion of the passengers are traveling for business purposes. The surveys support the hypothesis that the development of HSR should have heterogeneous effects on employment in different industries. Passengers who take HSR are more likely to be employed in relatively high-income industries.

1.3 Literature Review

This paper is closely related to two strands of literature. The first strand measures the economic outcomes from the development of transportation infrastructures, including highway roads, railways, and airplanes. This strand includes studies in different transportation infrastructures, in different countries, transporting goods or passengers, and affecting the economy in different aspects. The second strand of literature examines the reduction in labor market frictions, and their effects on labor mobility and labor reallocation across cities. Abundant research address labor's choices of where to live and where to work given certain frictions in the labor market, which this paper is closely related to theoretically.

[6]Classified by the Chinese Statistics Bureau

1.3.1 Transportation Infrastructure

There is a large amount of research to examine the economic outcomes from a reduction in transportation cost in transporting freights and passengers. To analyze the trade patterns among different countries, Eaton and Kortum (2002) build a model that explains how geographic distance and transportation cost hinders tradings among countries. Donaldson and Hornbeck (2016) examines the increase in market access from the introduction of railways in the US. For the economic outcomes, they examine when agricultural goods are more accessible nationwide through railways, how land prices, the leading factor prices, were affected. Donaldson (2018) also exam the railroads in India and its effects on reducing trade costs and interregional price gaps. They discovers that the railroads in India has increased interregional and international trade, as well as its real income levels, thus regions with railroads experience welfare gain. Michaels (2006) use the United States Interstate Highway System to identify the labor market consequences of a higher trade openness. The paper finds that the relative demand for skilled manufacturing workers increases in counties with a high endowment of human capital, which is consistent with the predictions of the Heckscher-Ohlin model.

My paper contributes to the literature that examines the economic outcomes from an improvement in transporting passengers. There are researches study the economic outcomes from highways. Baum-Snow et al. (2017) study how configurations of urban railroads and highways influenced urban form in Chinese cities since 1990. Their findings differ by the shapes of railroads and highways, which indicate that radial railroads decentralize industrial activity and

ring roads decentralize both industrial and service sector activity. Faber (2014) studies China's National Trunk Highway System, which is a large scale natural experiment that connects small peripheral counties to large metropolitan agglomerations. He finds that the system affects the diffusion of industrial and total economic activity to peripheral regions, or reinforce the concentration of production in space. The estimation results suggest that network connections have led to a reduction in GDP growth among non-targeted peripheral counties. Baum-Snow et al, and Faber inspire this paper in analyzing the transportation network development and heterogeneous effects from different network and industries.

As for the introduction of HSR in China and other countries, numerous studies focus on the economic impacts of HSR developments, including a study by Bernard et al. (2016) on the impact of Shinkansen line on Japanese firms. Several recent studies also examine the macro impact of China's HSR system. Zheng and Kahn (2013) study its effect on housing prices in secondary cities, which is consistent with McDonald and Osuji (1995). They find empirical evidence that residential land values increase due to a new elevated transit line in Chicago. Ke et al. (2017) evaluate its effects on per capita GDP growth of targeted cities. In micro level, some studies examine the micro effects on firm behaviors. Xu (2018) states that HSR has increased firm productivity by an increasing firm to firm matching efficiency, and by lowering the labor migration cost, HSR has increased the overall welfare in China. Lin et al. (2019) studies the interregional flows of private investments and finds there is an increase in cross-city investment for city pairs connected with HSR.

Besides analyzing the overall effects of HSR, some literature have focused on the heterogeneous effects in different cities. Qin (2017) examines the distributional effect in the core

and peripheral areas. The paper suggests that the affected counties on the upgraded railway lines experienced reductions in GDP and GDP per capita following the upgrade, which was largely driven by the concurrent drop in fixed asset investments. He et al. (2017) examine the different economic outcomes among initially rich and poor counties reacting to the new expressways in China. While both of them have an increase in GDP, initial poor counties have worse pollution after the introduction of expressways. They point out that the different outcomes among counties reflect the balance between economic activities and environment quality. Cheng et al. (2014) explores the effect of HSR on employment rates and specialization in both major cities and hinterlands in Europe (North-west European network) and China (Pearl River Delta area). They find that the effects differ widely depending on the different stages of economic development. In more developed regions, introducing HSR has a convergent impact. While some regions experience rapid development, the impacts are divergent.

This paper focuses on the heterogeneous effects of introducing HSR on the Chinese labor market. Few studies have revealed the effects of transportation infrastructures on the local labor market. This paper is closely related to Lin (2017), who examines how HSR has shaped urban employment and specialization patterns. She finds that HSR in China increases city-wide passenger flow and employment. She also finds industries with a higher reliance on nonroutine cognitive skills benefit more from HSR-induced market access to other industries. Lin analyzes the effect of a dummy variable, that is, the effect of introducing new HSR lines. In contrast, this paper explores the increase in connectivity to HSR, a continuous variable, on employment in different industries.

Other researches have analyzed the effects of new infrastructure for air travel and how

it has impacted labor reallocation and collaboration in various industries. Giroud and Mueller (2015) argue that direct airline routes that reduce travel time between headquarters and factories can induce labor and capital reallocation and thus increase productivity. Catalini et al. (2016) study how routes introduced by Southwest Airlines during 1991 to 2013 led to a 50% increase in scientific collaborations as researchers were better able to meet face to face to work on projects and publications. They also found that lower fares and the flow of people and ideas benefited younger scientists and that increased opportunities for low cost travel helped them be productive than their local peers. Similarly, Dong et al. (2018) show that after a city is connected to the HSR network, there is significant productivity increase in terms of quantity and quality of journal publications. This paper suggests the lower transportation cost and increased connectivity of HSR leads to reallocation of labor and other heterogeneous effects among various industries, especially high-skilled and tourism.

This paper also discusses economic agglomeration by analyzing the specialization patterns after HSR exposure. There are many literature analyzing economic agglomeration among regions and cities, which include Ottaviano et al. (2002), Rosenthal and Strange (2004), Moretti (2014), Allen and Arkolakis (2014), and literature studying the economic agglomeration of firms from spillover effect and positive externalities, including Duranton and Puga (2004), Moretti (2004), Combes et al. (2012), Greenstone et al. (2010). By analyzing how city clusters with very similar and very different industry compositions react differently regarding a decrease in transportation cost among them. this paper also contributes to the literature supporting economic agglomeration.

1.3.2 Labor Market Frictions and Commuter's Choices

Many literature have focused on analyzing economic activities within cities. Fujita and Krugman (1995), Lucas and Rossi-Hansberg (2002), Lemoy et al. (2012), Ahlfeldt et al. (2015) have built theoretical models study the size and internal structures of cities. In their models, transportation cost plays a very important role in shaping commuter's decision on where to live and where to work. Gibbons and Machin (2005) evaluates the benefits of rail access to households in London. They find that households valuate rail access more than other local amenities. Parry and Small (2009) and Small et al. (2006) uncover the optimal price of public transportation considering overall welfare gains.

Empirically, many studies on urban transportation infrastructure focus on measuring the reductions in transportation costs from different kinds of transportation infrastructure projects. For example, Cao et al. (2013), Behrens and Pels (2012), Heuermann et al. (2018) show how HSR improved intercity commuting from reducing travel time and increasing each cities' accessibility. How does decrease in transportation cost change commuters' choices in where to live and work? Many empirical studies have analyzed the urban residences' behavior changes responding to transportation infrastructure improvements. Baum-Snow et al. (2005) provide empirical evidence from sixteen cities in analyzing their residents change in commute modes responding to expanded rail networks. Their results suggest that new rail lines have been more successful at drawing new riders in denser, more centralized cities, and primary social benefit associated with the new rail lines comes from the significant reduction in commute time.

Baum-Snow (2007) also examines highways in the US and points out that the construction of new limited access highways has contributed to central city population decline. Similarily, Duranton and Turner (2007) finds that population and employment increase regarding the growth of roads and public transit in major cities in the US. With public transportation improvement, residence tend to live further away from the center cities or CBD of a city to avoid higher living cost.

Does the changes in commuters' behavior beneficial for the overall economy? Who is benefited from the improvement of public transportation infrastructure? Tsivanidis (2018) answers these questions by assessing the BRT system in Colombia. His paper finds that the BRT system increases welfare and output, and the high-skilled workers benefit slightly more. The heterogeneous benefits between the high-skilled and low-skilled workers are the results considering the usage of this transportation method, how easily individuals substitute between commutes, whether the system connects workers with employment opportunities, and equilibrium adjustment of housing and labor markets. For similar reasons, this paper evaluates the employment changes regarding the connection to HSR in different industries. Section 1.4 provides a theoretical support.

1.4 Model

To understand how is labor reallocated with respect to a increase in connectivity to HSR network, a simple model is presented in this section. This model intents to reveal that first,

increase in connectivity to HSR network is a reduction in transportation cost between cities; second, labor across cities are reallocated pursuing higher wage; third, there will be heterogeneous effects on the employment in different industries.

1.4.1 Model Environment

There are N cities in the country producing final goods **Y**. Following Grossman and Rossi-Hansberg (2008), each final good y produced by city j is producing from a continuum of tasks $i \in [0,1]$, with CES technology and constant returns of scale:

$$y_j = [\int_1^0 x(i)^{\frac{\sigma-1}{\sigma}} di]^{\frac{\sigma}{\sigma-1}} \tag{1.1}$$

In each city j, there are limited supply of land $\bar{H}_j$, l_j number of workers with wage equals to $w_j(i)$. The wage is different across cities and depends on the tasks. The whole economy N is endowed with an inelastic supply of workers who are perfectly mobile across cities.

1.4.2 Consumption

A representative household in city j 's utility is defined by the final good consumption C_j, housing H_j, and city amenity δ_j:

$$U_j = \delta_j C_j^{\alpha} H_j^{1-\alpha} \tag{1.2}$$

Assume that household's revenue equals to R_j with final goods price normalized to one and housing price equals to q_j, the indirect utility function for each household in city j is:

$$V_j = b\delta_j q_j^{\alpha-1} R_j \tag{1.3}$$

where $b = \alpha^{\alpha}(1-\alpha)^{1-\alpha}$.

From Cobb-Douglas utility function, total final goods consumption equal to α portion of total income, while total housing expenditure takes $(1-\alpha)$ portion of total income:

$$C_j^* = \alpha * R_j \tag{1.4}$$

$$H_j^* * q_j = (1-\alpha) * R_j \tag{1.5}$$

Suppose housing expenses are redistributed as a lump-sum to consumers, thus, for each household:

$$R_j = w_j{}^{7} + (1-\alpha) * R_j \tag{1.6}$$

The total income for city j, summing all households income R_j is:

$$v_j * l_j = R_j * l_j = w_j * l_j + (1-\alpha) * R_j * l_j \tag{1.7}$$

Combining housing market clearing condition and Equation 1.7:

$$q_j = \frac{(1-\alpha)}{\alpha} * \frac{w_j * l_j}{\bar{H}_j} \tag{1.8}$$

[7] $w_j = \int_0^1 w_j(i)di$

Equation 1.8 indicates that if one city's housing supply is inelastic at $\bar{H}_j$, the increase in labor income, either from an increase in w_j or from an increase number of workers l_j, the housing price will increase. The rising housing price makes working in city j less attractive. Unless there is an expansion in housing supply, eventually some workers in city j will mover out as housing price keep growing.

1.4.3 Production

For each city j, firms are in perfect competitive market. For each final good y_j produced in a firm at city j, this firm can produce some tasks in city j, and sourcing some tasks outside city j, like in city k, if city k produces these tasks relatively cheaper. This way, final goods can be produced taking advantage of both city j's high productivity and city k's cheaper labor cost. Consider the iceberg transportation cost $\tau_{jk}(\tau_{jk} > 1)$, the cost of sourcing task i in city k for firms in city j is:

$$g_{jk}(i) = \frac{\tau_{jk} * w_k(i)}{z_j} \tag{1.9}$$

z_j is the productivity of city j. Following Giroud and Mueller (2015), city j's efficiency in producing good j is the realization of a random variable Z_j from its city-specific probability distribution $F_j(z) = Pr[Z_j \leq z]$, and city j's efficiency distribution is Frechet:

$$F_j(z) = \exp(-T_j z^{-\theta}) \tag{1.10}$$

where $T_j > 0$ and $\theta > 1$. The distributions are independent across cities. In this distribution, T_j denotes the average absolute advantage of city j in producing the final goods, while θ represents the comparative advantage across cities. θ is the same across all cities, and it reflects the amount of variation within the distribution. A bigger θ implies less variability, meaning there are less differences in efficiency for all cities to produce the same product.

Firms in city j will look for city k to source task i at the lowest cost $g_{jk}(i)$. Based on the distribution of productivity in city j, the cost of city j source task i in city k follows the distribution:

$$G_{jk}(g) = Pr[g_{jk}(i) \leq g] = Pr[\frac{w_k(i)}{z_{jk}(i)}\tau_{jk} \leq g] = Pr[z_j \geq \frac{w_k(i)}{g}\tau_{jk}] \tag{1.11}$$

$$G_{jk}(g) = 1 - \exp(-T_i(w_k(i)\tau_{jk})^{-\theta} g^{\theta}) \tag{1.12}$$

The distribution of city j'cost in producing task i is:

$$G_j(g) = Pr[g_{jk}(i) \leq g, k = 1,2,...N] = 1 - \prod_{k=1}^{N}[1 - G_{jk}(g)] \tag{1.13}$$

$$G_j(g) = 1 - \exp(-\Phi_j g^{\theta}) \tag{1.14}$$

where $\Phi_j = \int_0^1 (T_j \sum_k^N (w_k(i)\tau_{jk})^{-\theta}) di$

Looking at the unit price for the final goods with tasks producing at home city and sourcing to other cities, assuming the production of final goods follow the CES production:

$$g_j = [\int_1^0 g_j(i)^{\sigma-1} di]^{\frac{1}{\sigma-1}} \tag{1.15}$$

If $\sigma < 1+\theta$, and firms in perfect competitive market that do not make any profits, the unit cost of producing a final good in city j is:

$$g_j = c_j = \gamma \Phi_j^{-\frac{1}{\theta}} \tag{1.16}$$

where γ is Gamma Function[8].

1.4.4 Transportation Cost

Transportation cost is defined as τ_{jk} in this model. It plays a very important role in deciding what types of tasks and how many tasks will city j sourcing with city k. Before the development of transportation infrastructures, transportation cost is purely driven by geographic distances, where cities sourcing tasks with nearby cities, ignoring further cities with cheaper cost. With the development of transportation infrastructures, it is very likely that more cities are trading with cities further away while still reducing cost.

In perfect competitve market, firms producing final goods with the same unit cost, thus:

$$c_j = \gamma \Phi_j^{-\frac{1}{\theta}} = \gamma [\int_0^1 (T_j \sum_k^N (w_k(i)\tau_{jk})^{-\theta}) di]^{-\frac{1}{\theta}} = \bar{c} \tag{1.17}$$

[8] $\gamma = [\Gamma(\frac{\theta+1-\sigma}{\theta})]$

Equation 1.17 indicates that $\Phi_j = \bar{\Phi}$ for $j = 1, 2, ...N$. In this case, when T_j increases, meaning that city j's productivity increases, $w_k(i)$ increases not only for city j, but also for cities that city j sourcing tasks with. The cities that city j sourcing with are more likely to be nearby cities since if τ_{jk} is too high, it is too costly for firms in city j to source any task.

Equation 1.17 also suggests that if τ_{jk} decreases, for example, by introducing HSR, transportation cost is lower, $w_k(i)$ will increase for city j and city k compares to other cities. With a higher wage, city j and city k will attract more workers working in task i. In summary, transportation cost affects wage in connected cites compare to other non-treated cities pairs, and it affects wage differently across industries, depending on the sourcing tasks between firms in the cities. Thus, labor will be reallocated among cities with or without HSR. Pursuing higher wage, labor will move to cities with higher connectivity to HSR. Two types of industries will be affected the most, first, industries with sourced tasks, which are more likely to be industries with lower labor cost, like non-service industries; second, industries with high income, like the skilled industries. Labor in this industries are the main passengers to take HSR due to their higher income.

To look at a city's total labor income, following Giroud and Mueller 2015, define X_{jk} to be the cost of city j spending on city k for sourcing tasks, and X_j to be the total cost on producing the final goods, X_{jk} can be written as:

$$X_{jk} = \frac{T_j \tau_{jk}^{-\theta} \int_0^1 w_k(i)^{-\theta} di}{\Phi_j} * X_j = \frac{T_j \tau_{jk}^{-\theta} \int_0^1 w_k(i)^{-\theta} di}{\bar{\Phi}} * X_j \tag{1.18}$$

City k can produce task i for city j, as well as many other cities, the total labor income in city k is:

$$Y_k = w_k * l_k = \sum_j^N X_{jk} = \bar{\Phi} \int_0^1 w_k(i)^{-\theta} di \sum_j^N T_j \tau_{jk}^{-\theta} X_j \tag{1.19}$$

This paper defines $\sum_j^N T_j \tau_{jk}^{-\theta} X_j$ in Equation 1.19 as the spillover effects city k receives from nearby cities that affects their labor income. City k gets higher spillover effects from nearby cities when nearby cities have higher productivity, or when the transportation costs between them are lower. Empirically, this paper reveals productivity of a city by its real GDP before sample period. Plus, for transportation cost, a continuous measurement of indirect connectivity to HSR is measured for each city in China across years. The indirect connectivity for a city shows the change in transportation cost between this city and all other cities through this city's nearby cities. The detailed measurement is shown in Section 1.5.

In summary, this model predicts that after connected to HSR, there will be a decrease in transportation cost between cities in the network. As spillover effects from nearby cities are measured by transportation cost to high productivity cities, an increase in a city's spillover effects increase its total labor income. Thus, in order to get higher wage, workers are reallocated across cities with and without HSR. However, as a city's total labor income increases, the living cost increases as well if its housing supply is inelastic. Some labor will move out of the city due to this reason. Therefore, connecting to HSR affects labor reallocation differently across industries. For industries that are more likely to have sourced tasks, there will be workers move in, while other industries will have workers move out due to the increasing living costs.

1.5 Methodology

1.5.1 Measure Connectivity

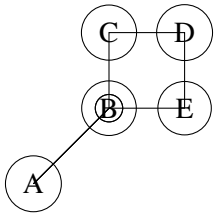
The key contribution of this paper is to measure how cities' connectivity to HSR changes over time. This paper adopts the graph theory to calculate connectivity for 285 cities over ten years. When a city connects with other cities by HSR, they form a network. Based on graph theory, each city is treated as a node, and HSR is the link between nodes. A network is a representation of connections among a set of items. The network of HSR is a set of connections among cities. As the network of HSR evolves over the years, each city's connectivity to the network changes. Rather than treating connectivity to HSR as a dummy variable, this paper analyzes connectivity as a continuous variable. After a city gets connected to the HSR network, its connectivity to HSR is not zero anymore; however, as time goes by, its connectivity to HSR can be increasing continuously by connecting to more cities in the network. Measuring connectivity can offer insight into a particular city's role in larger trends in transportation, employment, and economic impact. As for graph theory, connectivity relates to the "importance" of a node.

A node acquires importance based on various aspects. First, important nodes should have many connections. Second, important nodes should be served as a bridge that connects other nodes. Third, important nodes should reach other nodes in the same network with shorter distance[9].

[9]The distance here does not referred as geographic distance, but the number of links.

Figure 1.3 shows an example of a typical network. In Figure 1.3, there are five

Figure 1.3: Example



Nodes: A, B, C, D, E, which are considered five cities in this paper. Each city is connected with 1 to 3 other cities with Links. Depending on how many connections each city has, each city will have a measurement for Degree as direct connections. For example, here, city B is connected with A, C, and E directly. In this case, the Degree of B is 3. Degree measures the first aspect of a node's importance. For each city, the more Degree it has, the more connected it is.

In network literature, distance is defined as the shortest path between two nodes. For example, in Figure 1.3, to go from A to E, there are several paths. One path can be from A to B to E, or A to B to C to D to E. The distance between A to E is 2, which is calculated from the shortest path, A to B to E. In the same example, B is acting as a bridge of connecting node A and E. In addition, B is the essential point for node A to reach node C and D. Node B is also important for node C to reach node E, however, it is not the only bridge between node C and E because there is another path, from C to D to E for node C to reach E. Betweenness for a node, for example node v, measures the number of shortest paths for any node pair in the same network that has to pass a node v, normalized by the total number of shortest paths for this node pair. Betweenness measures the second aspect of an important node. For each node, the higher the betweenness, the more connected it is.

In a network, each node will have a measure of distance to each other node in the same network. Eccentricity is defined as the largest distance between any node v and all other nodes in the same network. For example, the largest distance of node A is from A to D, which is 3. In this example, eccentricity for each node is as follows: $\{A:3, B:2, C:2, D:3, E:2\}$. After knowing eccentricity for all nodes in a network, it is clear to find the Radius and Diameter of a network. The Radius for a network is the minimum of Eccentricity, while the Diameter is the maximum of a network. For nodes with Eccentricity equals to Radius is in the Center of a network, while nodes with Eccentricity equal to Diameter is in the Periphery of a network. Center and Periphery, as dummy variables, measure the third aspect of node importance. If a node is in the Center, it is more connected than other nodes. In contrast, if a node is in the Periphery, it is less connected than other nodes. When a network is large, and Eccentricity for different nodes varies by a fair amount, a node may be neither in Center nor in Periphery.

Table 1.1 summarize the variables this paper uses to measure connectivity. Before

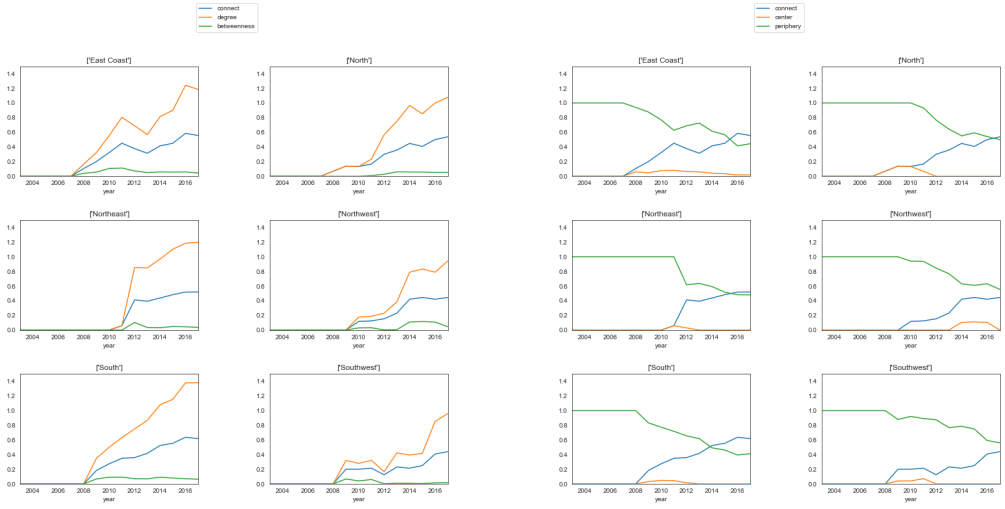
Table 1.1: Different Measurements of Connectivity

Variable	Definition	Calculation	Relation
Dummy:			
Center	Nodes with eccentricity equal to radius	1 if center,0 if not	+
Periphery	Nodes with eccentricity equal to dismeter	1 if periphery, 0 if not	-
Continuous:			
Degree	# of neighbors	d(v)	+
Betweenness	# of shortest paths that pass node v	$\sum_{s,t \in N} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$	+

Eccentricity (node v): the largest distance between n and all other nodes in the same network
Radius: the minimum Eccentricity of a network
Diameter: Maximum distance between any pair of nodes in the same network
Neighbors(node v): nodes that have direct connections with node v

cities are connected to HSR, their connectivity is counted as 0. For each variable, based on their correlation with connectivity, this paper assigns zero in Degree, Betweenness, Center, and 1 in Periphery for cities that are not in the HSR network. Figure 1.4 shows the connectivity change over time in different regions in China using different measurements. The blue curve in the graph is the dummy variable, which equals one if a city is connected to HSR. From Figure 1.4, the most connected regions are east and alone the coast as well as the south. This corresponds with the development of HSR in China.

Figure 1.4: Heterogeneity in Connectivity Across Regions



Note: This graph shows the changes in connectivity from 2008 to 2017 for cities in six regions in China.The panel in the left shows the connectivity measurement of continuous variables, and the panel in the right shows the connectivity measurement of dummy variables.
Source: All measurement of connectivity is calculated by author. Detailed explanation is specified in section 1.5

1.5.2 Dealing with Endogenous Issues

There are two main endogenous issues in this paper that could bias the estimations. One is the unobserved contemporaneous shocks that could affect the labor market outcomes and the timing of operating a new HSR at the same time. Another concern is the selection bias of cities to be connected to HSR. In other words, there can be unobserved variables that differ from cities with and without HSR, and these variables can affect the labor market outcomes, thus bias the estimations.This section will discuss in detail how this paper is dealing with these concerns.

First, the timing of operating an HSR line depends on the construction date and the open date. At the early stage of HSR network's development, it has been common to use the existing track for normal rails to operate high-speed rails. As stated in Lin (2017), 12 out of 45 HSR lines in 2014 were based on existing lines in 2005. In Dong et al. (2018)'s paper, they use the tracks for military usage that existed in 2005 as the instrumental variable for HSR tracks due to the strong correlations between the two. As the requirements of speed increases, the government needs to build new tracks for operating HSR. It is possible that the government recognizes economic potentials of a region and connects its cities with HSR; however, while the tracks might be constructed across the region at the same time, the date of initial operation can vary significantly.The open date depends on construction progress, which is largely affected by engineering difficulties. Besides, following Donaldson and Hornbeck (2016), Giroud and Mueller (2015), Lin (2017), this paper includes location*time[10] fixed effect to control for heterogeneous

[10]Specifically, this paper uses region*year fixed effect and uses province*year fixed effect as a robustness check, the results are consistent.

development of regions over time.

Another endogenous concern for this paper is the selection bias for the city with HSR. According to Ministry of Railway of China, the ultimate goal of establishing the HSR network is to connect all capital cities for each province with high-speed trains. Thus, it is safe to claim that cities that are not capitals of their provinces are byproducts of connecting province capitals. In this case, the cities with and without HSR connections may not differ substantially in economic conditions but are instead selected or not based on their distances to the nearest province capital. In Xu(2018), he builds an instrument variable for HSR network based on the connections between capital cities and the least cost routes depend only on engineer difficulties and geographic conditions.

In terms of how to deal with the remaining concerns over endogenous issues, this paper uses the unbiased measurement for connectivity from the city's economic conditions. The selection of cities could bias the estimation in the regressions, especially when connectivity is affected by the economic conditions of a city. For example, since Beijing is the capital of China, we can assume Beijing should be the Center of the HSR network, and it should have a larger set of direct connections compare to other cities. However, based on this paper's measurement, one city's connectivity is largely affected by its location in the network. Its direct connections to other cities define its location, and it is highly correlated with its nearby cities' direct connections to other cities. In other words, one city's connectivity is highly correlated with its surrounding cities' connectivity, which is unbiased from its own economic condition.

In this case, instead of using one city's direct connectivity, I calculate each city's indirect connectivity induced by nearby cities' direct connectivity, weighted by their distance

and pre-sample real GDP conditions. A similar approach, which is called the Market Access approach, is first introduced by Donaldson and Hornbeck (2016) in their paper estimating the railroads in the US, and followed by Lin (2017) in estimating HSR in China. A county's market access increases when it becomes cheaper to trade with another county, particularly when that other county is richer. Donaldson and Hornbeck (2016) measure changes in market access to all cities in the country with and without the transportation method. This approach is estimating the aggregate impacts of a transportation project; in such case, market access is influenced by changes beyond the local area in the railroad network. This paper is not using this market access approach because the measurement of connectivity is adapted from graph theory. Rather than measuring the change of travel time and cost, this paper is measuring different aspects of how a city is becoming more and more important in a network. Instead of calculating market access change over all cities in the country, this paper only focuses on how connectivity has changed from the five nearest cities. However, this paper is adapting the idea from the market access approach that real GDP of nearby cities matter in terms of their spillover effects.

Specifically, instead of measuring the direct connectivity of a city in a network, the indirect connectivity from five nearby cities is measured and calculated as weighted averages, which is called the spillover connectivity in this paper. Before doing that, following Giroud and Mueller (2015), this paper divides all cities into three groups concerning their connections to HSR. As shown in Table 1.2, the three groups are: the treated group, the other group, and the control group. For cities in the treated group, they have direct connections with other cities in the HSR network. In the other group, cities do not have direct connections themselves, but at least one of their nearby cities has direct connections with other cities. The control group

contains cities without HSR themselves, and neither are their nearby cities. As shown in Figure 1.5, city A, B, C, D, E are in the treated group, F is in other group and G is in control.

Table 1.3 shows the summary statistics across years for cities in the three groups

Table 1.2: The Three Groups

Treated Cities		Control Cities
Treated	Other	
Connected to HSR in year t-1, nearby cities can be connected or not Ex: A,B,C,D,E	Nearby cities connected to HSR in year t-1, not connected itself Ex: F	Not connected to HSR, no nearby cities get connected in year t-1 Ex: G

Note: In most of the cases(719 episodes), has HSR means near HSR, only in 15 episodes, the city has HSR but not near HSR

Figure 1.5: Example

defined above. In order to ensure the sample cities for each group are the same over years, I classify cities into eventually control group, eventually other group and eventually treated group based on their HSR connection in 2017, which is the end of the sample period. From Table 1.3, it is evident that for major variables that represent major economic conditions of a city, there is not difference among the three groups over years on average. Figure 1.6 shows the trends in different economic variables for the three groups, while the red vertical line represents the first introduction of HSR in 2008. Summarizing all these variables for all three groups, it is noticeable that before 2008, all three groups shared the similar trends for eight variables, after 2008, however, there has been trend change for the control group while the treated group and the other group continuously shared the similar trends for all variables. Table 1.3 and Figure 1.6 reduce the concerns for selection biases among cities with and without HSR.

In order to make sure one city's indirect connectivity is not induced by its economic

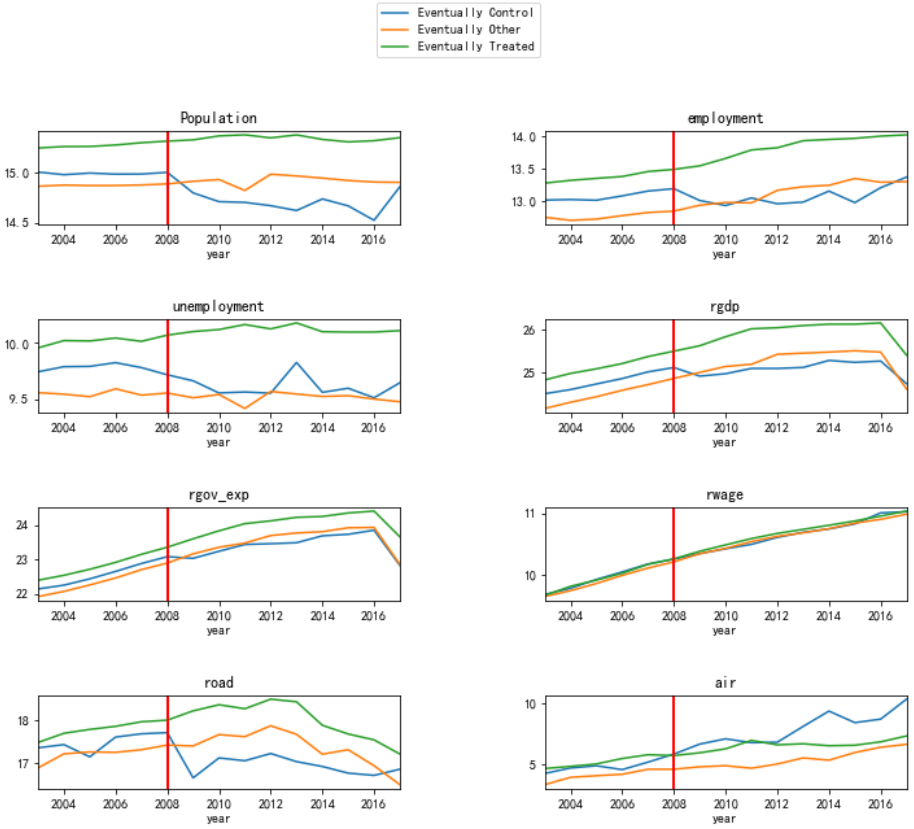
Table 1.3: Summary Statistics for Cities in Different Groups

Variables	Eventually control	Eventually Other	Eventually Treated
Count	435	1269	1752
Population	14.89	14.91	15.32
	0.81	0.62	0.58
Employment	13.06	13.02	13.67
	0.69	0.58	0.86
Unemployment	9.72	9.53	10.08
	0.74	0.66	0.79
Rgdp	24.89	24.99	25.64
	0.91	0.82	1.02
Rgov_exp	22.84	23.13	23.58
	0.93	0.81	1.04
Rwage	10.19	10.4	10.44
	0.46	0.46	0.51
Airplane_usage	5.84	5	6.1
	6.32	6.08	7.12
Aoad_usage	17.3	17.31	17.93
	1.66	1.62	1.42

Note: All variable are in log value

conditions in any way, for the treated group, when calculating their indirect connections from the nearby cities, their nearby cities' direct connection to them will be eliminated. In such a case, their indirect connections are only induced by their nearby cities' direct connections to cities excluding themselves. Figure 1.7 shows two examples of such eliminations. The left two panels are showing the case when city E is connected with city B and D. When calculating city B and D's direct connections, I eliminate B and D's connection to E, in such case, the network is changing from the first panel to the second panel on the left side. In the right two panels, city E only has one connection with city D, and so does city D, when calculating the indirect connectivity of city E and D, this paper eliminates this direct connection between them. After the elimination, however, the network has become completely disconnected. Even though cities

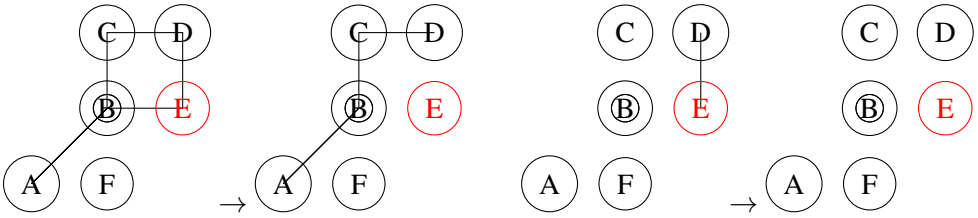
Figure 1.6: Group Trend Comparison



Note: This graph plots different economic backgrounds for cities in different groups from 2003 to 2017. The different measurement of economic backgrounds are total population, totol employment and unemployment level, real GDP, real government expenditures, real wage, the road ridership and airline usage.
Cities are divided into three groups, including the treated group, the other group, and the control group. Here the cities are grouped by its final group in 2017.

A, B, C, D, E are still in the treated group, their measurement of connectivity will be the same as cities in the control group, which is zero. By eliminating the city's direct connections in the network, the indirect connections, defined as nearby cities' direct connections, are purely induced by nearby cities' connections with all other cities in the network, thus further easing the endogenous concerns.

Figure 1.7: Taking off Direct Connection for Treated and Other Group



$$X\_Spill_{i,t} = \sum_{c}^{5} Weight_{c,i} * X_{c,t} \tag{1.20}$$

In addition, Equation 1.20 shows how one city's indirect connectivity, $X\_Spill_{i,t}$ is calculated. For each city i, this paper groups city i's five nearest cities by geographic distance as nearby cities of city i. For each nearby city c in this group, this paper calculates its own direct connectivity $X_{c,t}$ in the network. For city i in the treated group, $X_{c,t}$ will be the direct connections after eliminating the direct connection to city i.

Equation 1.21 shows how $Weight_{c,i}$ is calculated for each nearby city. There are two components in calculating the weights for city c in a nearby city group for city i, representing two forces that affect the spillover effects of connectivity from a nearby city. The first component is the real GDP ratio before the sample started in 2002. The real GDP ratio is calculated

by using the real GDP of city c in 2002, divided by the sum of real GDP for the city i's nearby city group. The intuition is that among all nearby cities, the city i will be benefited more from a connectivity increase from a richer city, for example, a capital city of a province.

The second component is city c's relative geographic distance compare to the closest city, city c1. In this case, for every city i, the weight for its closest city, city c1, will only be the real GDP ratio, since the second component will be one. If all city c in the nearby city group have a similar distance to go to city i, then the second component should be close to one for all of them. However, if there is large heterogeneity in the distance among the nearby city group for city i, then the city that is relatively far away from city i, will have little effects on the city i's connectivity. The intuition behind this is that geographic distance matters for spillover effects. If we assume that city i's closest city, city c1, regardless of its real GDP, has the most impact on the city i's connectivity. The impact of other nearby cities should decrease regarding their relative distance away from city i to city c1. The two forces work in opposite ways in deciding the weight of a nearby city's spillover effects. The richer and closer city always has a higher spillover effect in terms of connectivity.

$$Weight_{c,i} = \frac{log\_rgdp_{c,2002}}{\sum_{c \in N} log\_rgdp_{c,2002}} * \frac{log\_GeoDist_{i,c1}}{log\_GeoDist_{i,c}} \tag{1.21}$$

Table 1.4 represents summary statistics for geographic distances for nearby cities to the target city. In each column, the average distances of the closest city to the fifth closest city are reported. From Table 1.4, we can see there is not much heterogeneity in terms of geographic distance from the closest to the fifth closest. Figure 1.8 shows the distributions in bar charts,

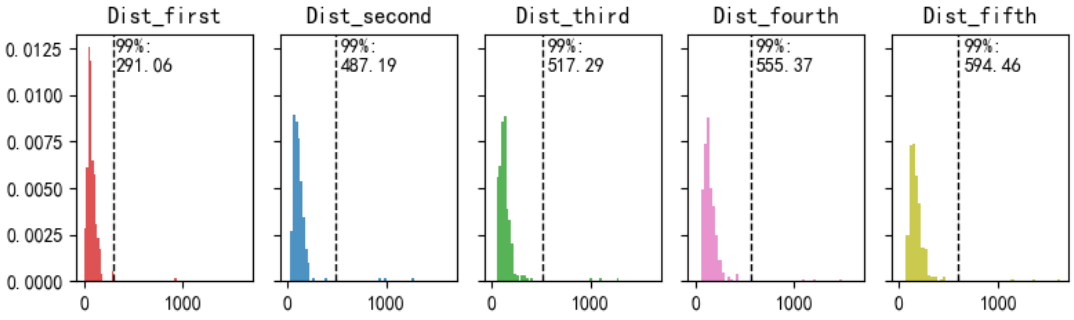
which indicates that even though there are outliers, in general, there is not much heterogeneity among nearby cities. This statement is also well demonstrated in Figure 1.9. Figure 1.9 shows the distributions of all city c from the closest to the fifth cloest and the sum of the weights. The weights are all distributed around 0.2 with a summation around 1. The distributions of distance and weights ensure the credibility of $Weight_{c,i}$ in calculating the spillover effect. There are no heterogeneous distance distributions among cities that would bias the results. Besides, there are no heterogeneous distance distributions among nearby cities for each city i that would have made some cities' weights in the nearby city group invalid.

Table 1.4: Distance(km) Summary Statistics

	Dist_first	Dist_second	Dist_third	Dist_fourth	Dist_fifth	total	aver_distance
count	282	282	282	282	282	282	282
mean	82.16	115.52	139.39	160.28	179.67	677.02	135.40
std	66.82	109.83	116.74	129.15	140.78	542.49	108.50

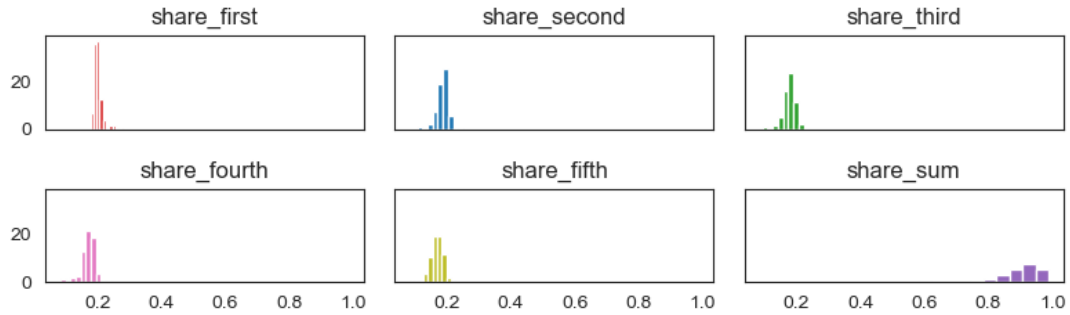
Note: 1 km = 0.62 miles

Figure 1.8: Frequency Histogram of Distance(km) in Order



Note: This graph shows the distributions of distance for each city to its nearby cities, from the closest to the fifth closest cities.

Figure 1.9: Distributions of Weights



Note: This graph shows the distributions of the weights, considering both distance and real gdp ratio in 2003, for each city to its nearby cities, from the closest to the fifth closest cities.

1.5.3 Comparing Direct and Indirect Connectivity

Table 1.5 presents summary statistics comparing direct and indirect connectivity at all observations. The second row shows the means for all connectivity measurements. The indirect connectivity is similar and smaller than direct connectivity for all connectivity measurements. Moreover, the standard deviation, which shows in the third row, is much smaller for indirect connectivity. The fact can be explained by the elimination of direct connection among nearby cities and the target city in the treated group. Since the endogenous connections are removed in spillover effects, the estimations from regressions will not be biased.

Figure 1.10 shows the change of indirect connectivity among different groups over

Table 1.5: Summary Statistics for Measuring Connectivity

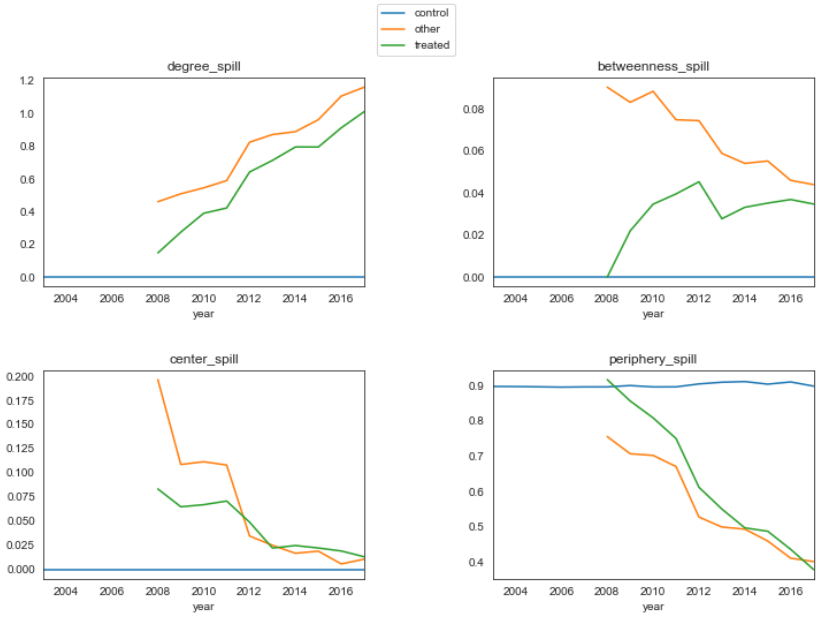
	degree_spill	degree	betweenness_spill	betweenness	center_spill	center	periphery_spill	periphery
count	3031	3031	3031	3031	3031	3031	3031	3031
mean	0.3	0.37	0.02	0.03	0.02	0.02	0.76	0.84
std	0.48	0.8	0.05	0.1	0.06	0.13	0.23	0.37
min	0	0	0	0	0	0	0	0
max	2.41	5	0.35	0.67	0.49	1	0.99	1

years. It is apparent that the treated group has less indirect connectivity compared to the other group. This is also because of the eliminations in direct connectivity for some cities in the

treated group.

Figure 1.11 shows how direct and indirect connectivity change over time for four big

Figure 1.10: Change of Connectivity among Groups



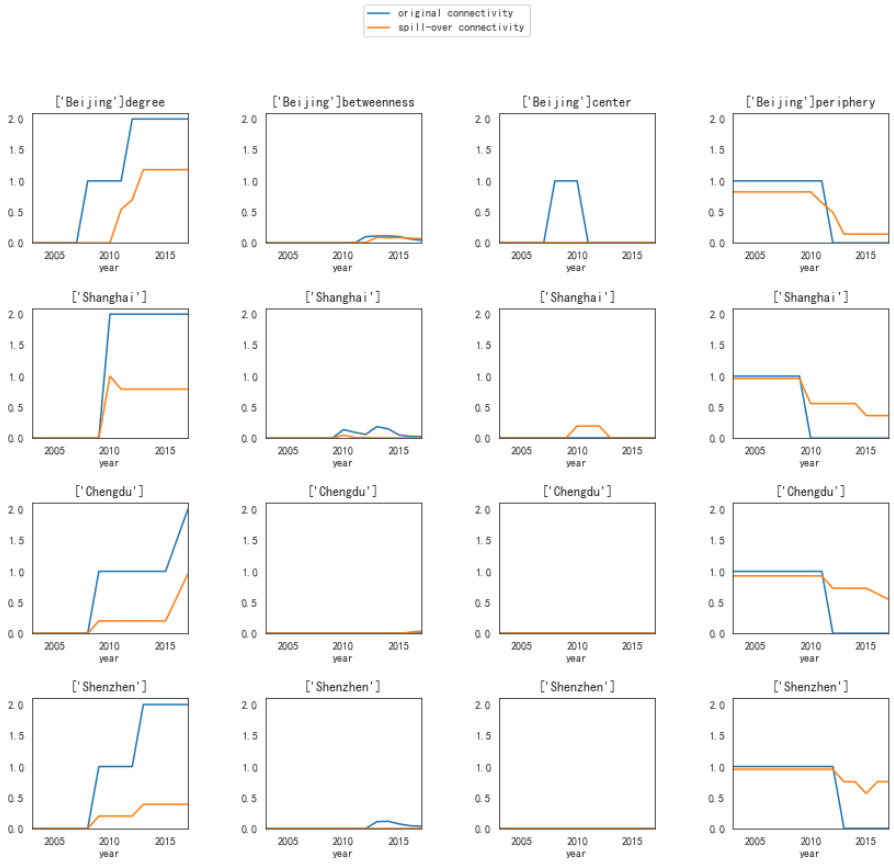
Note: This graph plots how differently the three groups' indirect connectivity change from 2003 to 2017. The indirect connectivity is always higher for the treated group.

cities in different regions of China. As showing in Figure 1.11, direct connectivity and indirect connectivity share similar trend for all cities, the magnitudes, however, are different from each other. The indirect connectivity is smaller than direct connectivity for all cities.

1.5.4 Comparing Industry Compositions

Besides the measurement of connectivity described in Section 1.5.1, there is another layer in terms of measuring connectivity, which is the difference in industry compositions in connected cities. If one city is connected with cities that have similar industry compositions,

Figure 1.11: Examples from the Big Cities



Note: This graph plots the comparisons of direct and indirect connectivity change for four big cities in China.

the indirect connectivity should affect labor reallocation differently than when it is connected to cities with very different industry compositions. For example, if one city is skilled intense, another is tourism intense, tourism workers in skilled intense city may have a larger application pool; if one city is skilled intense, another is also skill intense, skill workers in both cities will have a larger application pool. In other words, for each city i, there are measurements for the differences and similarities of industry compositions between city i and its nearby cities.

$\frac{1}{5}\sum_{i\neq c}^{5} D_{IND_{k,c,t_0},IND_{k,i,t_0}}$ and $\frac{1}{5}\sum_{i\neq c}^{5} D_{IND_{-k,c,t_0},IND_{-k,i,t_0}}$ measure the connectivity change by being connected with cities with similar or different industry compositions.

$$D_{IND_{k,c,t_0},IND_{k,i,t_0}} = |\frac{IND_{k,c,2003}}{IND_{c,2003}} - \frac{IND_{k,i,2003}}{IND_{i,2003}}| \tag{1.22}$$

$$D_{IND_{-k,c,t_0},IND_{-k,i,t_0}} = |\frac{IND_{-k,c,2003}}{IND_{c,2003}} - \frac{IND_{-k,i,2003}}{IND_{i,2003}}| \tag{1.23}$$

Equation 1.22 indicates the differences in city i's industry composition with five of its nearby cities' industry compositions. The differences in the industry decomposition is calculated in year 2003 due to endogenous concerns. After adding the absolute values of all the differences in industries' employment level, the difference is normalized by the number of nearby cites, which is 5. Similar method is applied calculating Equation 1.23, except here the measurement indicates whether city c has similar industry composition in other types of industries with nearby cities.

1.6 Data, Regressions, and Empirical Results

1.6.1 Data

The major data source for this paper is the China City Statistical Yearbook for city-level macroeconomic outcomes. Table 1.6 shows employment in different industries for each city, both in level and industry compositions, as the two main dependent variables. The 19 industries are divided into four classifications outlined by Lin (2017), including skilled employ-

ment, tourism-related employment, other-service employment, and other non-service employment. The specific classification is shown in Table A.1 in the Appendix. The main independent variable is connectivity measured in Section 1.5.1 and some other variables.

Table 1.6: Data Definitions and Sources

Variables	Definition	Data Source
Dependent Variables:		
Em_i,t,k	Log value of employment for industry k for city i in year t	City Statistical Yearbook
Em_composition_i,t,k	Percentage of employment in industry k divided by total employment in for city i in year t	Author's Calculation
Independent Variables:		
Connectivity_i,t	Explained in	Author's Calculation
Other Variables:		
rgdp_i,t	Log value of real GDP for city i year t	City Statistical Yearbook
population_i,t	Log value of population for city i year t	City Statistical Yearbook
unemployment_i,t	Log value of unemployment for city i in year t	City Statistical Yearbook
rwage_i,t	Log value of average wage for city i year t	City Statistical Yearbook
rgov_exp_i,t	Log value of government expenditure for city i year t	City Statistical Yearbook
road usage_i,t	Log number of passengers who used road to transport in city i year t	City Statistical Yearbook
airplane usage_i,t	Log number of passengers who used airplane to transport in city i year t	City Statistical Yearbook

City i: excluding cities with missing values, in total there are 282 cities in prefecture-level in the sample
Year t: range from 2003 to 2017

1.6.2 Regressions and Results

1.6.2.1 Benchmark Regressions

Benchmark regressions analyzing how spillover effect in measuring connectivity in different ways are shown in Equation 1.24. $Y_{i,t}$ is the dependent variable, which is the logged employment level in city i at time t. $X\_Spill_{i,t-1}$ is the spillover connectivity, including degree, betweenness, center and periphery, in city i at time t-1. By adding city and region times year fixed effect, city-specific and region to year fixed characters and trend will be controlled. Table **??** shows the results for this regression.

$$Y_{i,t} = \beta * X\_Spill_{i,t-1} + \alpha_i + \phi_l * \theta_t + \varepsilon_{i,t} \tag{1.24}$$

Table 1.7 shows how overall employment level reacts to the increasing in indirect connectivity. According to Table 1.7, none of the row shows a significant increase or decrease in employment level with an increase in indirect connectivity. The results suggest that, rather than a sole direct labor flow, there are a combination of labor inflow and outflow with respect to an increase in connectivity. In other word, some labor are seeking jobs and working outside city i when the city i is exposed to a higher connections to other cities with HSR, while some labor from other cities start to look for jobs and work in city i. The inflows and outflows canceled out with each other, thus the coefficients here in Table 1.7 are insignificant.

In order to further analyze the labor reallocation, this paper examines the heterogeneity of labor's decisions to move inside or outside a city with higher indirect connectivity. There are three hypothesis that this paper is testing: first, whether being in treated group or other group compared to control groups matters in determining the outflow and inflow; second, whether labor in different industries will have different behaviors reacting to city getting higher connections; third, whether a city's difference with nearby cities' industry compositions matter for city's labor market being affected by indirect connections. To test the hypothesis, this paper run more regressions in the following section.

Equation 1.25 intent to test whether being in different group, has HSR itself, near HSR, or has no connection to HSR, affects employment level differently. $Y_{i,t}$ is the logged employment level in city i at time t. $X\_Spill_{i,t-1}$ is the spillover effect city i receive at time t-1. Compare to Equation 1.24, there are two extra dummy variables intersecting with the spillover effects. $Treated_{i,t-1}$ is a dummy variable equals to one if city i has HSR itself in time t-1. Similarly, $Other_{i,t-1}$ equals to one if at least one of city i's nearby cities has HSR in time t-1. Based

Table 1.7: Spillover Effects on Employment Level

	Dependent variable:			
	employment			
	(1)	(2)	(3)	(4)
degree_spill	0.002			
	(0.025)			
betweenness_spill		−0.061		
		(0.108)		
center_spill			−0.106	
			(0.095)	
periphery_spill				−0.040
				(0.048)
Observations	3,031	3,031	3,031	3,031
R^2	0.952	0.952	0.952	0.952
Adjusted R^2	0.948	0.948	0.948	0.948

Note: standard errors are robust and cluster to province *p<0.1; **p<0.05; ***p<0.01

on the sample that I have, in most[11] of the cases, when one city is directly connected to HSR, it also has nearby cities directly connected to HSR. However, a city can only being in one group, if it is in the Treated group, it cannot be in the Other group even though its nearby cities also have direct connections to HSR. Thus, in Equation 1.25, β_1 is measuring the difference between treated group and control group; β_2 is measuring the difference between other group and control group; and $\beta_1 - \beta_2$ is measuring the difference between treated and other group.Table 1.8 shows the regression results.

[11] In 719 episodes, has HSR means near HSR, only in 15 episodes, the city has HSR but not near HSR

$$Y_{i,t} = \beta_1 * X\_Spill_{i,t-1} * Treated_{i,t-1} + \beta_2 * X\_Spill_{i,t-1} * Other_{i,t-1} + \alpha_i + \phi_l * \theta_t + \varepsilon_{i,t} \quad (1.25)$$

Table 1.8: Spillover Effects on Employment Level in Different Groups

	Dependent variable:			
	employment			
	(1)	(2)	(3)	(4)
degree_spill:treated	0.046			
	(0.033)			
degree_spill:other	−0.017			
	(0.023)			
betweenness_spill:treated		0.332		
		(0.258)		
betweenness_spill:other		−0.202**		
		(0.092)		
center_spill:treated			0.067	
			(0.181)	
center_spill:other			−0.187**	
			(0.093)	
periphery_spill:treated				0.095***
				(0.027)
periphery_spill:other				0.003
				(0.032)
Observations	3,031	3,031	3,031	3,031
R^2	0.953	0.953	0.953	0.953
Adjusted R^2	0.948	0.948	0.948	0.948

Note: standard errors are robust and cluster to province *p<0.1; **p<0.05; ***p<0.01

In Table 1.8, there are four regressions with respects to different measurement of connectivity. Column one shows that there is no significant difference for degree to affect employment level among groups. For betweenness, one unit increase in betweenness, decreases employment 20% more in other group compart to control group. For treated group, there is an positive but insignificant effect in employment level compare to control group. Column 3 shows that 1 unit increase in the spillover effect of being in the center in the network, there will be a 19% decrease in cities in other group, compared to control group. The results indicate that labors in cities do not have HSR, but near HSR, will more likely to move out of the cities, especially when their nearby cities are a bridge of connecting other cities, or in the center of the network. Cities in treated group, however, have mixed effects in terms of labor outflow and inflow. Labor could move outside the treated city or move in from cities in other group, or control group. Column 4 reports the coefficient relates periphery spillover effect compare to employment. Note that periphery is a measurement that has negative relationship with connectivity. Compare to other cities, cities in periphery receive less connectivity from the network since it is far away for these cities to reach out other cities in the network. However, the cities are still in HSR network, which is much better than cities in control group that has no exposure to HSR. When there is one unit decrease in periphery spillover effect, cities in treated group decrease employment levely by 9% compare to other group, and compared to control group.The result suggests that when cities' nearby cities move more towards network center, more workers will move out from treated cities since nearby cities are becoming more connected in the network.

It is noticeable that the magnitudes for the changes in employment is very high responded to one unit increase in connectivity for all measurements. The reason is that one unit

change in the connectivity measurement is very high in my sample. According to the summary statistics in Table 1.5, the mean for degree spillover effect is 0.3 and for periphery spillover effect, the mean is 0.7, while other two measurements' means both at 0.02. Thus, it is very uncommon that there will be one unit increase in indirect connectivity.

1.6.2.2 Industry Level Regressions

To further analyze the heterogeneity in labors' reallocate decisions, and test the hypothesis that workers in different industries will make decisions differently with respect to the introduction of HSR, this paper runs regressions in industry level. According to the passenger surveys cited in the previous section, passengers who are taking HSR are in high income group of China due to the higher expenses. Moreover, Lin (2017) examines how HSR has changed employment differently for cities with famous tourist attractions. Based on these information, this paper further tests how an increase in connectivity can affect employment differently in different industries.

$$Y_{i,t,k} = \sum_{k}^{4} \beta_k * X\_Spill_{i,t-1} * ind_k + \alpha_i + \gamma_k * \theta_t + \phi_l * \theta_t + \varepsilon_{i,t} \tag{1.26}$$

Equation 1.26 shows the regression. Compare to the benchmark regressions, Equation 1.24 and Equation 1.25, there are extra dummy variables intersecting the independent variable $X\_Spill_{i,t-1}$. Table A.1 in Appendix shows the detailed explanations of 19 industries in China, as well as the corresponded US industries. Following Lin (2017), this paper also divides the

19 industries into 4 groups: skilled, tourism-related, other service and other non-service. Each dummy variable represents a industry group. $Y_{i,t,k}$ is the dependent variable, and there are two measurement here. One is the logged employment level in industry k, in city i at time t, another one is the industry composition for industry k, in city i at time t. Table 1.9 reports the results for logged empolyment level in different industries, and Table 1.10 report the effects in industry compositions.

Table 1.9 has four columns, indicating different measurement of connectivity respec-

Table 1.9: Spillover Effects on Different Industries' Employment Level

	Degree	Betweenness	Center	Periphery
Other_ns	0.13^{***}	0.71	0.12	-0.37^{***}
	(0.04)	(0.41)	(0.20)	(0.07)
Other_s	-0.08^{*}	-0.36	-0.17	0.22^{***}
	(0.03)	(0.21)	(0.16)	(0.07)
Skill	0.17^{***}	0.55	-0.07	-0.38^{***}
	(0.04)	(0.35)	(0.13)	(0.08)
Tourism	0.08	-0.12	-0.46	-0.28^{**}
	(0.05)	(0.57)	(0.31)	(0.10)
Num. obs.	12124	12124	12124	12124
R^2	0.94	0.94	0.94	0.94
Adj. R^2	0.94	0.94	0.94	0.94

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.
standard errors are robust and cluster to province

tively. While each column name indicating the connectivity measurement, each row represents the employment level for each industry. For example, the first column reports that, when there is a one unit increase in degree spillover effect, there will be a 13% increase in other non-service industries' employment, and a 17% increase in skilled industries' employment, and the other service industries' employment will have a 8% decrease. The coefficient for tourism-related industries is positive but insignificant. Column two and three show consistent sign in the co-

efficients for all industries, but the results are all insignificant. For the last column, it shows the results of how switching to periphery in the network can affect employment in different industries differently. As being in periphery means a decrease in connectivity, the signs in the coefficients are consistent with column one as well. Compared to cities in periphery of a network, or cities without HSR[12], cities relatively closer to the center in the network has 37% more employment in other non-service industries, and 38% more employment in skilled industries, plus, 28% more in tourism-related industries. However, there will be 22% less employment in other service industries.

Table 1.10 reports the results for the same regressions with different dependent variable. The $Y_{i,t,k}$ here is different industry compositions, calculated by employment in industry k, divided by total employment, both at city i time t, times a hundred. The coefficients are consistent with Table 1.9. The first column shows that when degree spillover effect increase one unit,

Table 1.10: Spillover Effects on Industry Compositions

	Degree	Betweenness	Center	Periphery
Other_ns	3.66^{*}	21.17	6.69	-10.48^{***}
	(1.49)	(11.25)	(6.56)	(2.51)
Other_s	-5.05^{***}	-24.67^{**}	-5.88	13.53^{***}
	(1.22)	(8.85)	(6.18)	(1.97)
Skill	1.32^{**}	3.16	-0.34	-2.77^{**}
	(0.41)	(3.95)	(1.99)	(0.95)
Tourism	0.08	0.34	-0.48	-0.27
	(0.07)	(0.64)	(0.34)	(0.16)
Num. obs.	12124	12124	12124	12124
R^2	0.83	0.83	0.83	0.84
Adj. R^2	0.83	0.83	0.82	0.83

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.
standard errors are robust and cluster to province

[12]This paper assigns periphery to be equal to one before a city get connected to HSR

other non-service industry composition increase by 3.66% and skilled industry composition increase by 1.3%. Column one also shows that these industries' increases in employment is due to the decrease in other service department, and the decrease equals to roughly 5%. The change in tourism-related industries are not significant. Column 4 shows that being in periphery or has no HSR, decrease other non-service composition by 10.5% and decrease skilled employment by 2.8%, but increase other service composition by 13.5%. The other two columns show mostly insignificant coefficients, but the signs are consistent with other columns in Table 1.10.

The results show that there are heterogeneity in employment in different industries responding to an increase in indirect connectivity. Specifically, other non service industries and skilled industries receive employment inflow, but other service industries have employment outflow, and employment in tourism-related industries have both inflows and outflows, and the effects cancel each other out.

To further analyze whether the heterogeneity in different industries will also vary in different groups, this paper intersect the group dummy variables like in Equation 1.25. Equation 1.27 shows the detail regression. Besides dummy variables indicating different industries, there are two extra dummy variables represents cities in treated or other group, like in the previous Equation 1.25. Thus, this equation examines how is labor reallocated differently across industries and across groups. Table 1.11 shows the regression results.

$$\begin{aligned} Y_{i,t,k} = & \sum_{k}^{4} \beta_{k,treated} * X_{i,t-1} * ind_k * Treated_{i,t-1} \\ & + \sum_{k}^{4} \beta_{k,other} * X_{i,t-1} * ind_k * Other_{i,t-1} \\ & + \alpha_i + \gamma_k * \theta_t + \phi_l * \theta_t + \varepsilon_{i,t} \end{aligned} \tag{1.27}$$

In Table 1.11, each column shows different employment results for different measurement of connectivity, and each row represents different group across industries. The top four rows show regression results for industries employment level in the Treated group, while the bottom four show for the Other group. For the treated group, responds to one unit increase in degree, skilled employment in the Treated group increase 24% more compares to the Control group, and 10% more compares to the Other group; other service employment decrease 18% more than the Control group, and 16% more than the Other group; tourism-related employment increase 18% more than the Control group, and 16% more than the Other group; for other non-service industries, there is no significant effect in the Treated group, but there is a 14% increase in the Other group compare to the Control group.

Other columns in Table 1.11 shows the same regression with different connectivity measurement. The signs are consistent with column 1 for all significant coefficients. Table 1.11 and the following tables will not report the measurement of periphery due to its complexity in interpreting its relationship with connectivity[13].

Table 1.12 reports results from the same regressions as Table 1.11, except the dependent variable is industry compositions. The signs are consistent with Table 1.11.

The results in this section show that there are different labor reallocation in different industries across groups. In the Treated group, there are more skilled and tourism-related worker inflows and service workers outflows compared to the Other group. In the Other group, there are more non-service worker inflows. Referring Table 1.9 and Table 1.10, the total increase in non-service industries' employment is assigned more in the Other group, while the

[13]Compare to cities with HSR, periphery equals to one means less connectivity. However, compare to cities without HSR, periphery equals to one means this city is in the HSR network, thus higher connectivity.

Table 1.11: Spillover Effects on Different Industries' Employment in Different Groups

	Degree	Betweenness	Center
Other_ns_t	0.12	0.08	0.03
	(0.06)	(0.59)	(0.29)
Other_s_t	-0.18^{***}	-1.10^{**}	-0.66^{***}
	(0.05)	(0.36)	(0.17)
Skill_t	0.24^{***}	1.05	0.08
	(0.05)	(0.57)	(0.21)
Tourism_t	0.18^{*}	1.13	0.63
	(0.07)	(0.63)	(0.48)
Other_ns_o	0.14^{***}	0.94^{*}	0.18
	(0.04)	(0.45)	(0.22)
Other_s_o	-0.02	-0.10	0.09
	(0.03)	(0.26)	(0.20)
Skill_o	0.14^{***}	0.38	-0.14
	(0.04)	(0.30)	(0.14)
Tourism_o	0.02	-0.57	-1.02^{**}
	(0.06)	(0.65)	(0.36)
Num. obs.	12124	12124	12124
R^2	0.94	0.94	0.94
Adj. R^2	0.94	0.94	0.94

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.
standard errors are robust and cluster to province

increase in skilled industries' employment is allocated in the Treated group. Moreover, the decrease in service industries comes more from the Treated group.

The results suggest that has HSR and near HSR induce different patterns in labor reallocation. Cities near HSR attracts more non-service worker than cities with HSR, and cities with HSR attracts more skilled and tourism-related workers than cities near HSR. The results are consistent with the passenger surveys indicating that high-income businessmen and travelers are the main passengers in HSR.

Table 1.12: Spillover Effects on Industry Compositions in Different Groups

	Degree	Betweenness	Center
Other_ns_t	4.57^{*}	15.94	11.73
	(1.95)	(15.80)	(7.90)
Other_s_t	-7.52^{***}	-34.48^{**}	-17.43^{*}
	(1.61)	(12.36)	(6.97)
Skill_t	2.72^{***}	16.67	4.53
	(0.74)	(8.83)	(3.60)
Tourism_t	0.23^{*}	1.87^{*}	1.18^{*}
	(0.11)	(0.95)	(0.58)
Other_ns_o	3.13^{*}	23.03	4.12
	(1.52)	(12.66)	(7.86)
Other_s_o	-3.64^{**}	-21.19^{*}	0.02
	(1.25)	(10.49)	(7.45)
Skill_o	0.52	-1.63	-2.82
	(0.38)	(3.28)	(1.75)
Tourism_o	-0.01	-0.20	-1.33^{***}
	(0.09)	(0.80)	(0.35)
Num. obs.	12124	12124	12124
R^2	0.83	0.83	0.83
Adj. R^2	0.83	0.83	0.82

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.
standard errors are robust and cluster to province

1.6.2.3 Industry Level Regressions with Differences Industry Compositions

In this section, this paper discusses the difference in industry compositions among each city cluster. A city cluster is defined as city i and its five nearest cities by geographic distances. For each city i, its indirect connectivity, i.e.,the spillover effect, is calculated by its nearby cities direct connectivity, as explained in Section 1.5. The question this section discusses is whether city i's nearby cities industry compositions matter in terms of the magnitude of their spillover effect towards city i. If city i is closer to cities with similar industry compositions, the labor reallocation will be different if city i is closer to cities with very different industry compositions. In Equation 1.28, the differences in industry composition is measured by the av-

erage absolute differece of city i's employment in industry k and all nearby cities employment in the same industry, all in year 2003. Choosing the initial sample year as the time to calculate industry composition differences ensures there is no endogenous issues. By intersecting average difference in industry composition in a city cluster,$\bar{D}_{k,i,2003}$, Equation 1.28 examines how labor reallocates differently with different industry compositions in a city cluster. Table 1.13 and Table 1.14 show the results of the regression with different dependent variable, one with logged employment level in different industries, and another one with industry compositions.

$$\begin{aligned} Y_{i,t,k} = &\sum_{k}^{4} \beta_k * X_{i,t-1} * ind_k * \bar{D}_{k,i,2003} \\ &+ \alpha_i + \gamma_k * \theta_t + \phi_l * \theta_t + \varepsilon_{i,t} \end{aligned} \tag{1.28}$$

where:

$$\bar{D}_{k,i,2003} = \frac{1}{5} \sum_{c \neq i}^{5} \left| \frac{IND_{k,c,2003}}{IND_{c,2003}} - \frac{IND_{k,i,2003}}{IND_{i,2003}} \right|$$

Table 1.13 shows spillover effect can be passed through differently to employment level in different industries with different industry compositions in a city cluster. Each column shows different measurement of connectivity, and for each row, the first industry is the industry whose employment gets affected by HSR, while the second industry is the industry that composition differences are calculated. The first columns shows that as degree increases one unit, skilled workers increase 7% more if the city i's skill industry composition is one unit different from its nearby cities. Similar results are found for other industries. The results suggests that with the same exposure in HSR, a city's industry k's employment increase more, if this city's nearby cities have very different industry composition in industry k. This results indicating a specialization pattern with the introduction in HSR. As cities get more connected by HSR, a

Table 1.13: Spillover Effects on Employment Level intersect Industry Composition Differences

	Degree	Betweenness	Center
Other_ns_Skill	-0.04	-0.13	0.03
	(0.03)	(0.29)	(0.18)
Other_s_Skill	0.00	0.09	-0.06
	(0.02)	(0.21)	(0.16)
Skill_Skill	0.07^{**}	0.41^{*}	0.18
	(0.02)	(0.21)	(0.14)
Tourism_Skill	-0.01	-0.27	0.02
	(0.03)	(0.21)	(0.24)
Other_ns_Tourism	-0.01	-0.58	-0.42
	(0.11)	(1.11)	(0.55)
Other_s_Tourism	-0.20^{***}	-1.39^{*}	-0.24
	(0.06)	(0.67)	(0.45)
Skill_Tourism	0.03	-0.35	0.02
	(0.06)	(0.59)	(0.48)
Tourism_Toursim	0.40^{***}	3.29^{***}	1.33^{*}
	(0.08)	(0.75)	(0.66)
Other_ns_Others	-0.01	-0.01	-0.08
	(0.02)	(0.19)	(0.09)
Other_s_Others	0.06^{***}	0.53^{***}	0.35^{**}
	(0.01)	(0.13)	(0.11)
Skill_Others	-0.01	-0.10	-0.03
	(0.02)	(0.15)	(0.07)
Tourism_Others	-0.02	-0.01	-0.14
	(0.02)	(0.28)	(0.12)
Other_ns_Otherns	0.03	0.10	0.09
	(0.02)	(0.18)	(0.08)
Other_s_Otherns	-0.05^{***}	-0.46^{***}	-0.32^{**}
	(0.01)	(0.12)	(0.10)
Skill_Otherns	0.00	0.07	-0.01
	(0.01)	(0.13)	(0.07)
Tourism_Otherns	0.01	-0.12	0.03
	(0.02)	(0.23)	(0.11)
Num. obs.	12124	12124	12124
R^2	0.95	0.94	0.94
Adj. R^2	0.94	0.94	0.94

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.
standard errors are robust and cluster to province

city specialized in skilled industry, will attract more workers from other cities to work here as a skilled worker.

Other rows in the first column of Table 1.13 also show interesting results. For example, with the same amount of exposure in spillover effect from HSR, a city's service workers are more likely to flow out if its nearby cities are specialized in tourism-related industries or non-service industries. Other columns in Table 1.13 show the results for the same regression with different connectivity measurements. The signs and significance level are consistent with the first column.

Table 1.14 shows the same regression with a different dependent variable. All the signs and significant levels of the coefficients are consistent with Table 1.13. One extra finding is that a city's non-service workers flow out if its nearby city is specialized in service industries.

Table 1.14: Spillover Effects on Industry Composition intersect Industry Composition Differences

	Degree	Betweenness	Center
Other_ns_Skill	-1.30	-6.49	0.69
	(0.80)	(7.69)	(5.16)
Other_s_Skill	0.34	2.29	-2.85
	(0.65)	(6.10)	(4.63)
Skill_Skill	0.99^{**}	5.13	2.40
	(0.32)	(2.87)	(2.17)
Tourism_Skill	-0.03	-0.93	-0.24
	(0.06)	(0.50)	(0.41)
Other_ns_Tourism	2.53	9.32	-6.44
	(2.32)	(30.32)	(15.27)
Other_s_Tourism	-5.02^{**}	-27.38	-0.45
	(1.59)	(24.48)	(12.06)
Skill_Tourism	1.75	12.03	4.41
	(0.93)	(12.45)	(6.20)
Tourism_Toursim	0.74^{***}	6.02^{***}	2.48
	(0.14)	(1.63)	(1.42)
Other_ns_Others	-1.25^{***}	-8.74	-8.03^{**}
	(0.34)	(5.28)	(2.53)
Other_s_Others	1.56^{***}	11.67^{*}	9.88^{***}
	(0.28)	(4.93)	(2.57)
Skill_Others	-0.27	-3.22^{*}	-1.54
	(0.16)	(1.52)	(1.25)
Tourism_Others	-0.05	0.29	-0.30^{*}
	(0.03)	(0.58)	(0.14)
Other_ns_Otherns	1.47^{***}	9.91^{*}	8.04^{***}
	(0.29)	(4.51)	(2.12)
Other_s_Otherns	-1.53^{***}	-11.26^{**}	-8.97^{***}
	(0.23)	(4.13)	(2.23)
Skill_Otherns	0.04	1.69	0.76
	(0.15)	(1.73)	(1.12)
Tourism_Otherns	0.02	-0.34	0.17
	(0.03)	(0.45)	(0.12)
Num. obs.	12124	12124	12124
R^2	0.84	0.83	0.83
Adj. R^2	0.83	0.83	0.83

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.
standard errors are robust and cluster to province

1.7 Conclusion

This paper analyzes the labor reallocation among cities with their increasing connectivity to the HSR network in China. By adopting the graph theory, this paper contributes to the existing literature by calculating the changes in *indirect* connectivity to HSR for 285 Chinese cities from 2008 to 2017. The convenience of HSR has lowered the transportation costs among cities and reduced labor market frictions allowing for the freer movement of both skilled and unskilled workers. Before connecting to HSR, each city is different in terms of its specialized industries (manufacturing, tourism, etc.). After connecting to HSR, each city has different exposures to HSR and therefore, the impacts of connectivity can be different for both the labor markets and industries of the respective cities. This further proves the heterogeneous impacts of HSR on labor reallocation.

There are four major findings in this paper. First, this paper finds that the overall employment level is not affected significantly by the increase in *indirect* connectivity to HSR. The overall number of employed workers does not necessarily change in the cities; instead, there is labor reallocation between the cities.

Second, despite the insignificant overall effects of HSR connectivity on employment levels, there are heterogeneous effects on employment levels in different industries. Specifically, when *indirect* connectivity to HSR increases, employment in skilled industries and non-service industries increases as well, while service employment decreases, and employment in tourism-related employment remains ambiguous.

Third, besides heterogeneous effects in industry composition, there are also hetero-

geneous effects regarding different levels of exposure to HSR. This paper discovers that when *indirect* connectivity increases, cities directly connected to the HSR network experience an increase in skilled industries, while cities without direct connections to HSR but *near* HSR have an increase in non-service industries.

Fourth, this paper reveals the fact that HSR increases the specialization patterns in terms of industry compositions. After the indirect connectivity to HSR increases, a city specialized in the skilled industry will experience an increase in the skilled industry, especially if its nearby cities are not specialized in skilled industries. This is consistent with economic agglomeration theories.

Chapter 2

Exchange Rate Regimes and External Debt Holdings for Developing Countries

2.1 Introduction

External debt is the debt owed to nonresidents' repayable in currency, goods, or services. Total external debt is the sum of publicly guaranteed, and private nonguaranteed long-term debt, use of IMF credit, and short-term debt. Compared to other debts borrowed from domestic investors with local currency, external debt is more directly related to the exchange rate. The real payment of the external debt can fluctuate enormously due to the volatility of the exchange rate, especially under the floating exchange rate regimes. This paper evaluates the differences in external debt holdings among developing countries with different exchange rate regimes, both in normal times and default times.

Studying external debt holding ability is of great importance. As stated discussed in

related papers, e.g., Dias (2011), Olivier and Anastasia (2006), borrowing from other countries is essential for economic development, especially being in an open economy. External borrowing spurs investment and economic growth (Greenidge 2010). Besides, external debt is a major source through which developing countries finance their budget deficit (Qayyum et al. (2014)). A large number of empirical and theoretical works have focused on analyzing the relevance between external debt and economic development (Checherita-Westphal, Rother(2012), Reinhart and Rogoff(2009,2010,2011)). However, few studies connect external debt with exchange rate regimes. Cespedes, Chang, and Velasco (2004) discuss the relationship between exchange rates, balance sheets, and macroeconomic outcomes in a small open economy model. Using a panel dataset with 57 developing countries' external debt holdings and exchange rate regimes from 1970 to 2007, this paper analyzes the relation empirically.

This paper also examines the relations between external debts and exchange rate regimes both in normal times and during default. The portion of a country's debt borrowed from foreign lenders includes borrowing from commercial banks, governments, or international financial institutions. These loans, including interest, must usually be paid in the currency in which the loan was made. In order to earn the needed currency, the borrowing country export goods to the lender's country. A debt crisis will occur if creditors lose confidence in believing a country's ability to repay the external debt, especially after an economic downturn. Thus, the relations between exchange rate regimes and external debt holdings can be different during a debt crisis.

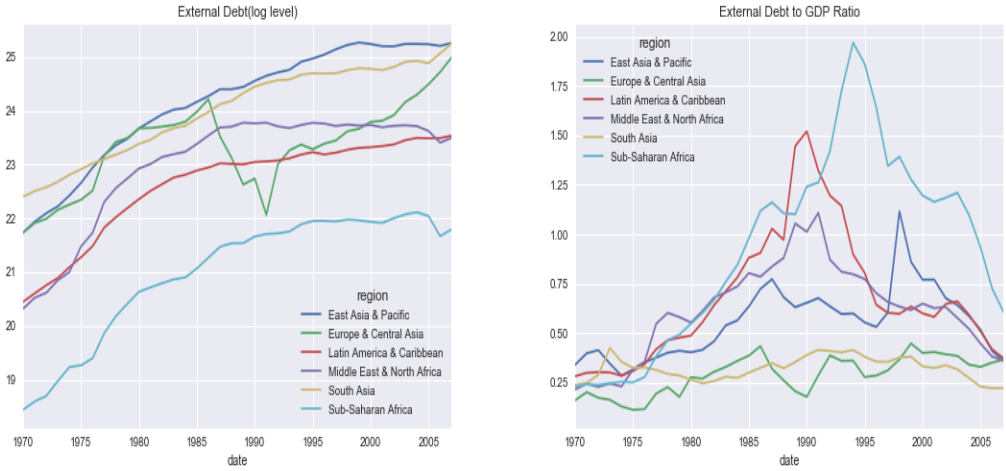
In order to control other factors that can also affect one country's ability to hold external debts, this paper classifies the factors that affect external debt into three groups: the demand

side, the supply side, and the cost borrowing(Greenidge(2010)). From the demand side, a country needs high external debt when its production is high. At such a time, they need to borrow more to invest or finance economic activities. Alternatively, if its production is low, they also need to borrow more to recover. Plus, when the government deficits from extra spendings, they tend to borrow extra. On the supply side, when a country does not have enough income or other capital sources, like FDI, to finance their spending or needs for developing, they will need to borrow. Moreover, from the perspective of the cost of borrowing, one country tends to borrow aboard if it is cheaper than borrowing domestically.

This paper is closely related to papers discussing exchange rate regimes, external debts, and sovereign default. In regards to connecting external debt with exchange rate regimes, Na, Grohe, Uribe, and Yue's (2014) 's paper theoretically proves the relationship between default and exchange rate policy. They state that during default episodes, countries currently in floating exchange rate regimes can significantly support more external debt than countries in fixed exchange regimes. Collectively, this paper empirically examines how a country's exchange rate regimes affect its external debt both in normal and default times. This paper studies the external debt level and external debt to ratio both in normal times and during default. Substantial studies have focused on measuring the cost of default in different fields, including direct output loss (Sturzeneger (2002), De Paoli, Hoggarth, and Saporta (2006), Yeyati, Panizza(2009)), contraction in trade sectors (Borensztein and Panizza(2008), Rose(2005), Martinez and Sandleris(2008)), exclusion from the world capital market(Borensztein et al. (2007), Arellano and Kartashova (2007), Fuentes, Saravia(2009)). This paper also exams whether a country's default history decreases its ability to borrow.

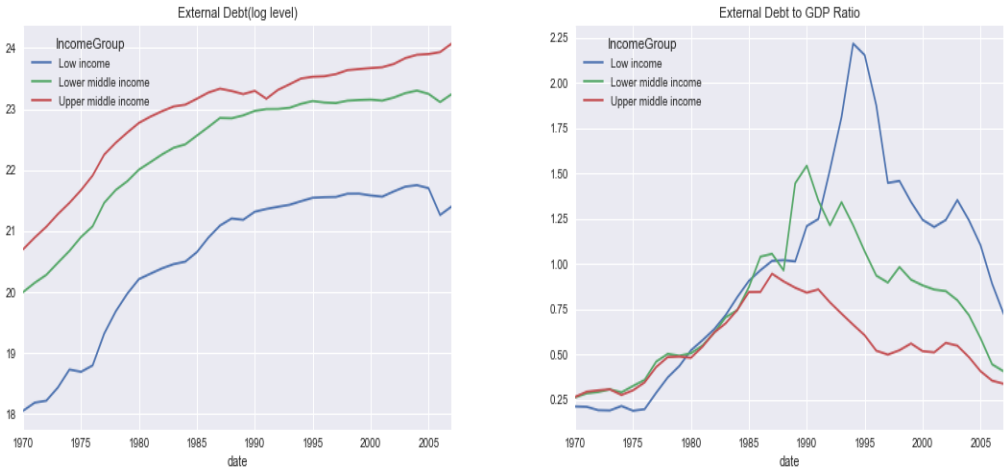
To analyze the relationship between external debt and exchange rate regimes, this paper uses a panel dataset which contains 57 countries. The list of the country is in Appendix Table B.3. Their external debts are captured in two ways, one dependent variable is external debt level in the logarithm term, and another is external debt to GDP ratio. Figure 2.1 and Figure 2.2 show the growth of external debt in developing countries by regions and by different income groups. The country list of these two classifications of the countries is in the Appendix. The graphs show that the external debt level keeps increasing for all regions and income groups. However, the external debt to GDP ratio kept increasing until the 90s. In the 90s, countries except for Europe and Asia started to have the ratio decrease, decreasing for countries in all income groups. Since the external debt level keeps increasing during the 90s, the explanations for the sudden decrease of external debt to GDP ratio can be due to an even higher increase of GDP level.

Figure 2.1: External Debt in Different Regions



Note: This graph shows the growth of external debt level (in 2005 USD) across the continents
Source: WDI

Figure 2.2: External Debt in Different Regions



Note: This graph shows the growth of external debt to GDP ratio (in 2005 USD) across the continents
Source: WDI

There are endogeneity issues in the regressions. First, the exchange rate regimes and default may have reserve causalities with external debt. The level of external debt can also affect the decision of default and the choices of exchange rate regimes. Second, the lagged term of external debt gives rise to autocorrelation. Third, the fixed effects of the country's characteristics, like geography and demographics, may be correlated with the explanatory variables. The fixed effects are contained in the error term in the regressions. Lastly, this paper has 57 countries(large N) and 37 years(small T).To ease the endogeneity concerns, besides clustering standard errors with regions for each country, this paper also uses the Generalized Method of Moments(GMM) estimator. The results reveal that compared to countries with more fixed exchange rate regimes, countries with more floating exchange rate regimes hold less external debts, especially during default episodes.

Specifically, the regressions results from the panel dataset show there are no signifi-

cant differences of external debt holdings for countries in different exchange rate regimes during normal times. However, for one unit increases of exchange rate regime(more floating) during default, it decreases external debt level by 0.9% more than in normal times. Moreover, this paper distinguishes the difference between changing exchange rate regimes and switching exchange rate regimes. Switching exchange rate regimes from fix to float increases external debt to GDP ratio by ten percentage points. However, in default times, the switch increases external debt to GDP ratio by 30 percentage points. For the debt level, a switch of exchange regime from float to fixed reduces debt level by 2%. The switched sign between different dependent variables depends on whether exchange rate regimes are changing or switching. Switching is a bigger change in exchange rate regimes. GDP will be depressed due to the abandonment of fixed exchange rate regimes, at least in the short run.

For robustness check, first, this paper runs the same benchmark regressions but with the lagged regime, float and default. Also, this paper adds more control variables for other robustness checks. The robustness regression results are consistent with benchmark results. In extension, this also checks how regime change or switching affects the growth rate of external debt level or external debt to GDP ratio. Correspondingly, this paper uses the growth rate of other control variables. Since the lag term of external debt is no longer an explanatory variable, this paper uses the GLS estimator with countries' fixed effects and clustering standard errors by the regions of the countries. The results show that increasing exchange rate regimes to be more float, or switching from fixed to floating regimes decreases the growth rate of external debt level and the growth rate external debt to GDP ratio in normal times. However, it increases the growth rate of external debt to GDP ratio during default. The results are consistent with the

benchmark and robustness regressions.

The remainder of the paper is structured as follows. Section 2.2 presents the relative literatures. Section 2.3 shows the empirical methodology and data. Section 2.4 discusses the results from benchmark regressions and robustness checks. Section 2.5 shows some extensions. Section 2.6 concludes.

2.2 Literature Review

My paper is related with two streams of research: (i) studies on sovereign default with exchange rate regimes; (ii) studies on a country's ability in supporting external debt;

In regards to sovereign default, this paper follows the strand of strategic sovereign default literature, which assumes that the government can strategically default on its debt as mentioned in Eaton and Gersovitz (1981) and Arellano (2008). When the benefits of default exceed the benefits of repaying debt, countries choose to default.

During default episodes, being in different exchange rate regimes affects a country's economic performance with its recovery (Na, Grohe, Uribe and Yue(2014)). In their paper, it is argued that, under the optimal policy, developing countries have limited commitments of debt contracts and downward wage rigidity in the labor market, not only that, but large devaluations accompany the default. The real exchange rate depreciates as tradable goods become more expensive than non-tradable ones, which induces agents to switch their expenditure away from tradable and toward non-tradable. This redirection of aggregate spending stimulates labor de-

mand since the non-tradable sector is labor-intensive and prevents the emergence of involuntary unemployment. Under the currency peg, the real exchange rate depreciates insufficiently due to the real wages, which are considered the labor costs of firms, remaining too high. The fixed high labor costs are caused by the combination of downward nominal wage rigidity and the currency peg.

Besides, they state that fixed exchange rate economies can support less external debt compared to economies with optimal floating rates. Default has two benefits under currency peg: one is to spur the recovery in the consumption of tradable by freeing up resources through the repudiation of external debt. These resources would otherwise be devoted to servicing the external debt. The second benefit is that default lessens the unemployment consequences of the external crisis. Furthermore, they state that country premium is twice as high under peg as is under the optimal devaluation policy. The difference is explained by first, in the peg economy, the typical default occurs in more severe contractions in the traded sector than the less severe contractions under the optimal devaluation policy. Second, in the peg economy, the steady and significant increase in unemployment makes default more attractive.

Furthermore, Grohe and Uribe(2016) analyzes the inefficiencies that arise from the combination of fixed exchange rates, nominal rigidity, and free capital mobility(as seen in the European currency union). The model predicts that the combination of currency peg and free capital mobility creates a negative external anomaly that causes borrowing during booms and high unemployment during contractions. The optimal exchange rate policy eliminates unemployment and calls for large devaluations during the crisis. The existence of the external anomaly creates a need for government intervention. The first best allocation can be brought

about via exchange rate policy or via labor or production subsidies. These subsidies occur at the level of the firm financed by income taxes levied at the household level. Capital control tax restricts capital inflows in good times and subsidizes external borrowing in bad times. The benevolent government has an incentive to levy taxes on external debt during expansions as a way to limit nominal wage growth.

Some benefits brought about by devaluation consist of the lowering of the unemployment rate and the improvement of the trade balance. Current literature has proved another benefit of floating exchange rate regimes during default episodes, which is improving trade balance. Kirwoluzky, Muller, and Wolf(2014) study an environment in which default takes the form of a re-denomination of debt from foreign to domestic currency. In their model, debt re-denomination and devaluation lowers the real burden of debt, making fiscal policy sustainable. Empirically speaking, Paoli, Hoggarth, Saporta(2006) discovers that output losses from twin crises appear to be bigger when a debt crisis is accompanied by a banking crisis rather than a currency one. The reason for the low output losses is that a currency crisis involves a sharp depreciation of the domestic currency, leading to the silver lining of stimulating exports.

Looking aside from the benefits of devaluation after default, there are risks when applying devaluation during the debt crisis. If countries' debts are mostly in foreign currency, devaluation can be a bad choice after default. With that in mind, there are benefits to consider when staying in fixed exchange rate regimes. Literature refers to "twin crises" as the situation when the debt crisis happens at the same time with banking or currency crisis. "Twin crises" that consist of simultaneous currency and debt crisis is always caused by the inappropriate composition of debt for past-default countries. Jeanne and Guscina(2006) shows that a much higher

proportion of sovereign debt-issued both domestically and abroad, is more commonly denominated in foreign currency in past-defaulters than in non-defaulters. The paper also points out the safe debt structure for emerging economies to follow. They state that instead of having short-term, foreign currency debt, countries with safe debt structure should be holding long-term and domestic currency debt.

For the composition of countries' debt, Eichengreen, Hausmann and Panizza (2005) report that between 93% and 100% of all developing country debt is issued in foreign currencies, depending on the measure used. Moreover, outside the main financial centers and Europe, developed countries have between 70% and 90% of their obligations in foreign currencies. The debts tend to be concentrated in a handful of currencies. Dias, Richmond, and Wright (2011) constructed a sample of long term debts owed by 100 developing countries from 1979 to 2006. At any given time, countries had borrowed in about 75 different currencies. However, almost 70% of all debt in 2000 was denominated in U.S. Dollars, and the five most important currencies (Dollar, Yen, Euro, Special Drawing Right, and Deutschmark) accounted for more than 90% of the total.

It is harder for countries to repay their debt when their currencies are devalued since the real repayment will increase. Moreover, this is referred to as the "original sin" of developing countries by Eichengreen and Hausmann(1999). They point out that less developed countries are more vulnerable to international financial crises than developed countries because of the currency composition of their debt. Namely, for less developed countries", the domestic currency cannot be used to borrow abroad or to borrow long term, even domestically. Instead, many emerging market countries borrow in foreign currency.

Cespedes, Chang and Velasco(2004) find that when facing an adverse shock if an economy has a large debt denominated in foreign currency, a weaker local currency can also exacerbate debt- service difficulties and wreck the balance sheets of domestic banks and firms. This channel may cause devaluations to be contractionary, not expansionary. As documented by Ricardo Hausmann et al. (2001) and Guillermo Calvo and Carmen Reinhart (2002), balance sheet effects have emerged as a prime reason why many central banks are reluctant to allow their currencies to devalue in response to external shocks.

Only a limited amount of papers study the countries' ability to support external debt. Greenidge,Dra- kes, Graigwell(2010) finds the major contributing factors of the build-up of the Caribbean Community's foreign debt stock are the output gap, real effective exchange rate, exports, real interest rate, and current deviation of government expenditure from its trend value. Of all the contributing factors, the output gap, the real cost of foreign borrowing, the real effective exchange rate, and exports are inversely related to the level of external indebtedness. At the same time, the current deviation of government expenditure from its trend value is positively associated.

2.3 Data and Empirical Methodology

2.3.1 Data

Table 2.1 shows the definition and source of all the variables. The sign behind the variables indicates the expected relationship between this variable and external debt. (?) shows the relationship is ambigulous.

For countries default history, I obtain the data from Reinharts website[2]. She defines

Table 2.1: Variables Definitions and Data Sources

Variables	Variable Definition	Data source
	IMPORTANT VARIABLES:	
Regime	Exchange rate regimes, from 1 to 6, the higher means more float	IMF
Default	Dummy, equals to one when countries are claimed to be in default or restructures episodes	Reinhart's website [1]
Float	Dummy, equals to one when exchange rate regime >=4	Author's calculation
	COST OF OF BORROWING:	
Real IR(-)	U.S. prime lending rate adjusted by U.S. inflation	WDI and author's calculation
Rdep (?)	real exchange rate depreciation, where $RealER_i = NER_{i,US} * \frac{CPI_{US}}{CPI_i}$	WDI and author's calculation
Ndep (?)	nominal exchange rate depreciation (Negative if Depreciation against US Dollars)	WDI and author's calculation
	DEMAND SIDE:	
GDP(log)(?)	The log value of Nomial GDP	WDI
Y_gap_t1(-)	Last period deviation of real GDP(current US$)	WDI and author's calculation
G_gap_t1 (+)	Last period deviation of general government final consumption expenditure (current US$)	WDI and author's calculation
Export (?)	Log value of exports of goods and services (current US$)	WDI
Export/GDP (?)	Export/GDP	WDI
Δtot(+)	the fluctuation of terms of trade	World Bank(DataMarket)
Balance(log)_t1(-)	Log value of last period government primary balance, calculated by tax revenues -spending	WDI and author's calculation
Balance/GDP_t1(-)	Last period government primary balance again GDP ratio	WDI and author's calculation
	SUPPLY SIDE:	
ED(log)_t1(+)	Log value of total external debt stocks (current US$)	WDI
ED/GDP_t1(+)	External Debt/GDP	WDI
FDI(log)(-)	Log value of foreign direct investment, net inflows (current US$)	WDI
FDI/GDP(-)	FDI/GDP	WDI
Reserves/Imports(-)	Total reserves expressed in terms of the number of months of imports(Reserves/(Imports/12))	WDI

default as the failure of a borrower to meet principal or interest payment on the due date. And she defines a default event as occurring when a country defaults on, or restructures, its total external debt. This dataset contains 57 countries 232 default episodes from 1970 to 2007. I then conduct a panel dataset, which contains default as a dummy variable. When the default dummy equals to 1, it means country i at time t is going through a default episode.

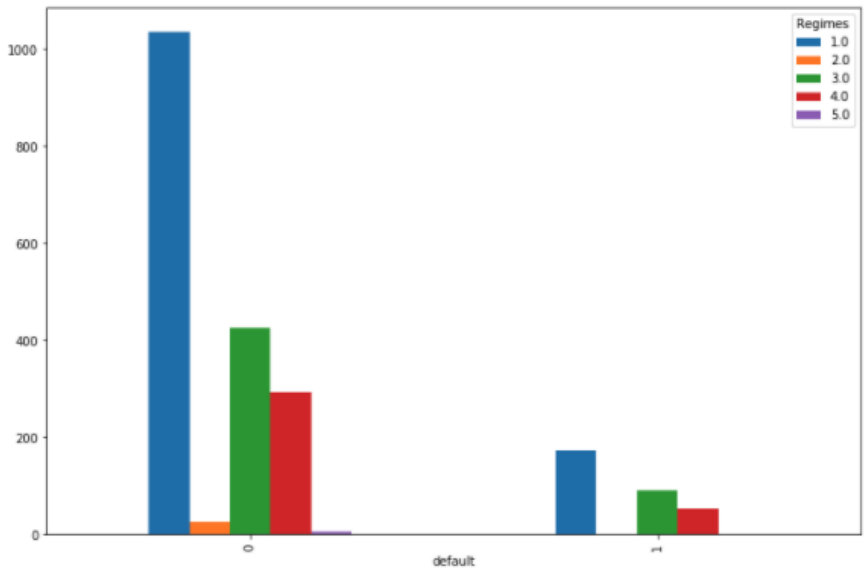
To see how being in different exchange rate regimes can affect the countrys abilities in supporting external debt, I use IMFs classifications of exchange rate regimes. In the coarse classification, the regimes range form 1 to 6 and as the number goes up, the regime becomes more floating. The detailed definition of exchange rate regimes are presented in Appendix Table

[2] http://www.carmenreinhart.com/data/

B.6.

Figure 2.3 shows the distribution of different exchange rate regimes during default

Figure 2.3: Exchange Rate Regimes' Distributions between Normal and Default Times



Note: 0 means normal times and 1 means during default.

Figure 2.4: External Debt Distribution in Different ER Regimes

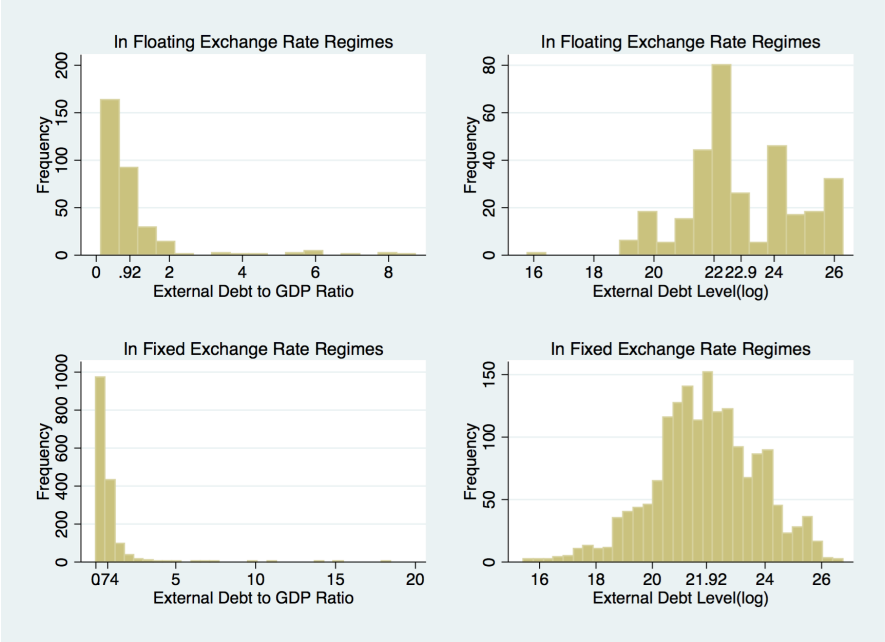
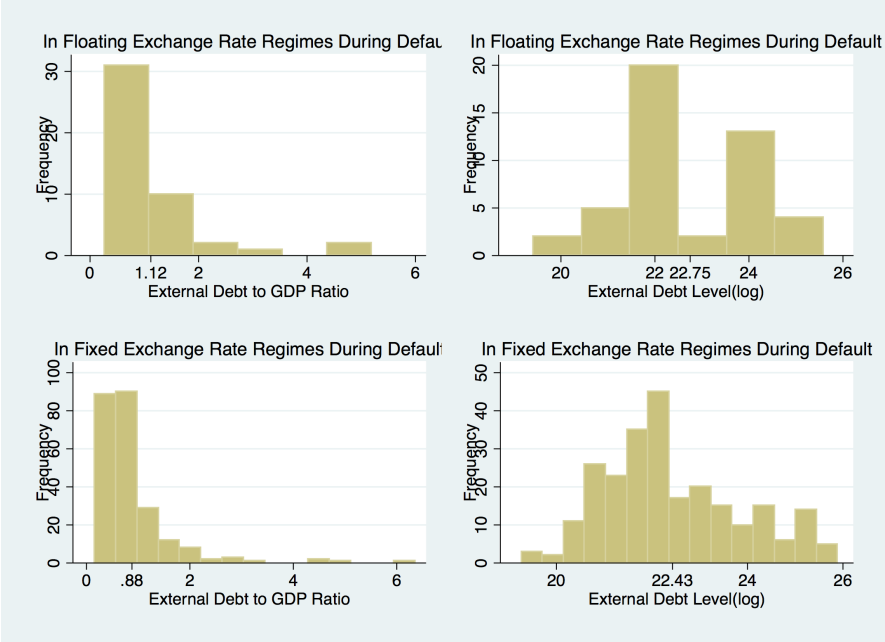


Figure 2.5: External Debt Distribution in Different ER Regimes(During Default)



and normal times. The distributions are similar between different times, with fixed exchange rate regime countries more than floating exchange rate regime countries. However, the external debt holdings are different. Figure 2.4 shows the external debt distribution in the dataset by different exchange rate regimes for both external debt to GDP ratio and the log value of the external debt level. The graph shows that compared to floating exchange rate regimes, first, the external debt distribution is more spread in fixed exchange rate regimes, and one reason is that there are more fixed regimes episodes than floating regimes. Second, the mean external debt log value is lower in fixed regimes for both measurements of external debt. Third, fixed regimes have higher values of external debt to GDP ratio From Figure 2.4. We can have a brief idea of how external debt distributions are different between different exchange rate regimes across the whole dataset. Figure 2.5 shows the same distributions, but only during default episodes. The

results do change when excluding normal times. For external levels, the mean of both exchange rate regimes are closer than in Figure 2.4.

For output gap and expenditure gap, I calculate the current percentage deviations in real output and government expenditure from their respective HP trend values. In order for that to be calculated, I must take the log difference between GDP and HP trend. As stated in Hercowitz(1986), the growth of external debt should be negatively correlated with variation in income and positively correlated with shocks to government expenditure.

To obtain annual real exchange rate for each country in my dataset, I calculate it by using nominal exchange rate times the price ratios. The price ratios are calculated by dividing US CPI with home country CPI ratio. The fluctuation of terms of trade is calculated by the percentage change of current terms of trade against last period's terms of trade. If the terms of trade fluctuation increases, this will decrease export revenue, thus increase external debt. For the cost of borrowing, I use the U.S. prime lending rate adjusted by U.S. inflation. If the cost of borrowing goes up, the external debt borrowing should go down. For exports, I only include export to GDP ratio because the log of export is highly correlated with the log of GDP. I dont include imports because import value is highly correlated with outputs. For reserves in terms of import goods and service, I got the data directly from WDI and it is calculated by Reserve/(Import/12). This variable measures how long the current reserves can finance imports without additional borrowing.For government primary balance, I took the government tax revenue and government spending data from WDI and calculate government primary balance by taking the difference.

Table 2.2 presents the summary statistics of all the variables.Table B.1 and Table B.2

in Appendix show the correlations between these dependent variables for both kinds of regressions. The two high ones are between log exports and log GDP, as well as nominal depreciation and real depreciation. Thats why in the regression of log values, I eliminate log export. Real depreciation and nominal depreciation is correlated but they are never included in the same regression.

Table 2.2: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
IMPORTNAT VARIABLES:					
Default	2,166	0.15	0.35	0	1
Regime	2,105	2.01	1.22	1	5
Float	2,166	.19	.39	0	1
COST OF BORROWING:					
Real IR	2,166	3.99	2.48	-1.27	8.7
Rdep(%)	1,615	0.43	13.63	-83.17	103.27
Ndep(%)	2,077	-8.57	26.47	-99.96	809.87
DEMAND SIDE:					
Y_gap_t1	1,796	-0.38	2.01	-20.36	4.67
G_gap_t1	1,924	-0.007	0.26	-0.99	4.87
Δtot (%)	2,104	2.21	37.75	-95.86	1301.95
Exports(log)	1,946	21.34	2.03	13.15	26.69
Export/GDP_t1	1,927	0.27	0.18	0.001	2.62
Balance(log)_t1	1,900	-9.55	17.87	-25.85	24.27
Balance/GDP	1,854	-27.38	352.89	-9632.53	387.45
SUPPLY SIDE:					
FDI(log)	1,668	18.37	2.48	9.21	24.75
FDI/GDP	1,801	0.02	0.06	-0.83	0.89
ED(log)	1,978	22.07	1.85	15.44	26.75
ED/GDP	1,894	0.77	1.15	0.03	18.47
Reserve/Import	1,404	3.75	3.37	0.03	56.73

2.3.2 Empirical Methodology

To estimate the effect of being in different exchange rate regimes have on external debt, the benchmark regression is of the general form:

$$ED_{i,t} = \beta_0 + \beta_1 Regime_{i,t} + \beta_2 D_{i,t} + \beta_3 D_{i,t} * Regime_{i,t} + \beta_4 ED_{i,t-1} + \beta_5 Y_{gap} + \beta_6 G_{gap} + \varepsilon_{i,t} \quad (2.1)$$

Where $ED_{i,t}$ includes the log value of external debt and external debt to GDP ratio. $D_{i,t}$ stands for default episodes, which equals to 1 when country i at year t are in default periods. $ED_{i,t}$ is the depended variable that includes log value of external debts and external debt to GDP ratio. For the interaction term, I use $D_{i,t} * Regime_{i,t}$ to try to evaluate whether during default times, more floating exchange rate regime countries can support more external debts compared to normal times.

For control variables, I include last period's Y_{gap} and G_{gap}, which are output gap and government expenditure deviation separately. $ED_{i,t-1}$ is the lag term of the log value of external debt or external debt to GDP ratio. For regression with the log value of external debt level, I also control for the log value of GDP. $\varepsilon_{i,t}$ is the error term. Table 2.4 shows the results of the regression.

In addition, for the second regression, I generate another dummy variable called float, which equals to 1 when country i at year t are in floating exchange rate regimes(greater than 4).Then I change the interaction term into $D_{i,t} * f_{i,t}$ to evaluate whether during default times, countries being in floating exchange regimes can support more external debts compare to normal times. Other variables are the same as Equation 2.1. Table 2.4 also shows the results of the regression.

$$ED_{i,t} = \beta_0 + \beta_1 Float_{i,t} + \beta_2 D_{i,t} + \beta_3 D_{i,t} * Float_{i,t} + \beta_4 ED_{i,t-1} + \beta_5 Y_{gap} + \beta_6 G_{gap} + \varepsilon_{i,t} \quad (2.2)$$

Equation 2.2 is either two dummy variables. $Float_{i,t}$ equals to 1 when country is in float exchange rate regime and $Default_{i,t}$ equals to 1 when country is during default episodes.

For the coefficients in Equation 2.2, Table 2.3 shows the meanings of each coefficient.

Table 2.3: Coefficients in Equation 2.2

	Float	Fix	Diff
Default	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$	$\beta 1 + \beta 3$
Normal	$\beta_0 + \beta_1$	β_0	β_1
Diff	$\beta_2 + \beta_3$	$\beta 2$	$\beta 3$

$\beta 2$ shows during default episodes, how being in fixed regimes affects external debt differently from being in default or normal times. $\beta_1 + \beta_3$ evaluates how exchange rate regime switch from fix to float affects external debt in default times. And β_1 evaluates how exchange rate regime switch from fix to float affects external debt in normal times. That's why the interaction term's coefficient, β_3, evaluates the difference of the effect that exchange rate regime switching form fix to float between default and normal times.

There are some endogeneities in my regressions. First, since including a dynamic component can better reflect the dynamic nature of external debt, I include the last period of external debt as an independent variable in the right hand side in both regressions. And this will generate endogenous problems because last period external debt highly correlated with the dependent variable. Second, default and exchange rate regimes may have reverse causalities with the dependent variable. Since the level of the external debt this period also affects the probabilities of default or the choice of exchange rate regimes. In this case, panel data regression

estimates from both fixed-effects and random-effects estimators will be biased and inconsistent. That's why I use Arellano-Bond Dynamic Panel GMM Estimations to solve the endogeneity problems, which are an instrumental variable estimator. In the estimation process, lags of the endogenous and exogenous variables are suitable instruments for the model. Both estimation results of benchmark regressions are shown in Section 2.4.

For robustness check, I add more control variables.

$$ED_{i,t} = \beta_0 + \beta_1 Regime_{i,t} + \beta_2 D_{i,t} + \beta_3 D_{i,t} * Regime_{i,t} + \beta_4 X_{i,t} + \varepsilon_{i,t} \tag{2.3}$$

$X_{i,t}$ are control variables. The source and definition of these variables are specified in Table 2.2. In addition to the benchmark regressions independent variables, Default, last period external debt, Y_{gap}, G_{gap}, I further include real and nominal depreciation, fluctuations in term of trade, the log value of external debt and the ratio of external debt to GDP, FDI to GDP ratio, exports and government primary balance to GDP ratio in the last period, exports to GDP ratio, and the log value of FDI, government primary balance in the last period, reserves in terms of imports and real cost of borrowing. For depreciation, I use both real depreciation and nominal depreciation to see whether the nominal exchange rate announced by the government is empirically implemented and whether they will affect the results differently. And as in last section, I use another intersection term $D_{i,t} * f_{i,t}$. Equation 2.4 shows the regression:

$$ED_{i,t} = \beta_0 + \beta_1 Float_{i,t} + \beta_2 D_{i,t} + \beta_3 D_{i,t} * Float_{i,t} + \beta_4 X_{i,t} + \varepsilon_{i,t} \tag{2.4}$$

The control variables can be divided into three groups: (i) the cost of borrowing; (ii) the demand for borrowing; (iii) the supply of borrowing:

The cost of borrowing first involves with the direct real cost and it is expected to

have negative relationship with external debt (Greenidge, Drakes and Craigwell(2010)). Second, nominal depreciation will increase the real debt repayment for foreign currency debt and is considered as another cost. This will depress external borrowing. However, since the external debt is in measured by US dollars, nominal depreciation will also increase external debt (Ngeco(2000)).

For the demand side of borrowing, in addition, nominal depreciation as well as real depreciation can also be considered as variables that affect the demand of borrowing. In terms of export, depreciation leads to higher export revenue and will allow for a decline in external indebtedness. Thus, adding the effect of increasing borrowing costs, the effect depreciation should be ambiguous. If there is a depreciation associated with an improved export performance, the improvement in export can offset the increase in the foreign-currency value of external debt.

Also, export is another fact that influence external debt in demand. There should be inverse effects between exports and external debt. The more export revenue a country earns allows it to service a greater portion of its external debt, causing the outstanding amount to decline. Alternatively, a fall off in the value of exports can force governments to borrow externally to finance their operations.

I consider the fluctuation of term of trade as another demand for borrowing. Fluctuations in terms of trade may cause a reduction in export revenues which, in turn, may encourage external borrowing (Ngeco(2000)).

In addition, the output gap and government expenditures affect the demand of borrowing Greenidge, Drakes and Craigwell(2010). A deviation from GDP trend increases the need of borrowing. However, if it is a big positive deviation, there will be a lower borrowing

because of lower demand. And if government expenditure is high last period, there will be more desires for countries to borrow at the current period.

Government primary balance in the last period also influences the demand of borrowing. If government primary balance is negative in the last period, they have more incentives to borrow and vice versa. Thus government primary balance in the last period is negatively correlated with external debt.

In terms of the supply side of borrowing, I include last period external debt, FDI and reserves in terms of imports values. Last period debt is positively related to current period's debt. Foreign direct investment is negatively related with external debt because when FDI increase, there are more supply of capital and decrease the need of borrowing (Helkie and Howatd (1990)).

The reserve in terms of imports value measures how long an economy can finance its imports by using its stock of reserves without seeking refuge in higher levels of external borrowing(Hajivassiliou(1987)). This ratio indicates both creditworthiness and low demand for new loans because the existing stocks of reserves can be used to do such financing. Thus, the ratio should be negative related with external debt.

Next section describes the definition and source of the variables described above.

2.4 Empirical Results

2.4.1 Benchmark Regressions

Equation 2.1 and 2.2 are the benchmark regression for different measurment of the dependent variables. Using Arellano-Bond dynamic panel GMM estimation model with robust standard error by different regions, Table 2.4 shows the results.

Table 2.4 shows the results from the benchmark regression with different estimations, which include the last period external debt, regime, and default or intersection term with the float or regime and default, regimes, default and last period GDP deviation and government expenditure deviation as explanatory variables. For external debt level, I use two measurements to run regression on Equation 2.3 and Equation 2.4. One is the log value of external debt, and another one is the external debt to GDP ratio.

Table 2.4 shows the results for GMM estimation. Columns 2 and 3 show the results for using the log value of external debt. For the interaction terms, column 2 and column 3 all show positive, but no significant results. However, column 1 shows that when during default and rescheduling episodes, countries switch exchange rate regimes one unit higher, i.e., to be more floating, it increases external debt level by 0.9% more than during normal times. However, when during default and rescheduling episodes, countries switch from fixed regimes to floating regimes, decrease external debt to GDP ratio, and 40 percentage points more than during normal times. Alternatively, it decreases the external debt level by 2%. The difference between the signs is that since the variable ”Regime” only captures small changes, it is like a one-unit switch. However, a regime switch from float to fix or vice versa is a big change, unless it is

Table 2.4: Benchmark Results Using GMM

VARIABLES	Reg2.1 ED(log)	Reg2.2 ED(log)	Reg2.1 ED/GDP	Reg2.2 ED/GDP
Regime*D	0.0096**		0.0493	
	(0.0025)		(0.0297)	
Regime	-0.0078		-0.0103	
	(0.0047)		(0.0060)	
Float*D		0.0219		-0.4158***
		(0.0121)		(0.0677)
Float		-0.0289*		0.1001***
		(0.0137)		(0.0216)
GDP(log)	0.0320***	0.0290***		
	(0.0027)	(0.0039)		
Default	-0.0148	0.0029	0.0851	0.2968***
	(0.0085)	(0.0100)	(0.0620)	(0.0445)
Y_gap_t1	-0.0019	-0.0038	0.0009	0.0024
	(0.0017)	(0.0025)	(0.0032)	(0.0066)
G_gap_t1	0.0258	0.0211	0.1282	0.1501
	(0.0391)	(0.0336)	(0.0946)	(0.1594)
ED(log) _t1	0.9431***	0.9399***		
	(0.0037)	(0.0033)		
ED/GDP _t1			0.9013***	0.9180***
			(0.0150)	(0.0115)
Constant	0.6312***	0.7574***	0.0641***	0.0097
	(0.0813)	(0.1257)	(0.0096)	(0.0186)
Observations	1,574	1,575	1,580	1,581
Number of country1	53	53	53	53
Sargan test-Chi-sq	928.15	1008.61	253.32	155.27
Sargan test-p-value	0.72	0.76	0.7	0.569
AR(1) test-p-value	0.02	0.02	0.02	0.14
AR(2) test-p-value	0.49	0.47	0.46	0.38

from 3 to 4. Since the variable float capture more of a big change, the sign can be different. For effect in normal times, coefficients in row 2 and 4 show the results. For regimes switching in normal times, the effect is negative but insignificant. However, for regimes switching from fixed to floating, the signs are different for different dependent variables. Specifically, a switch from fixed to floating regimes decreases the log level of external debt by 2% in normal times but increases external debt to GDP ratio by ten percentage points. This is because regime-switching, especially big switching, has a big effect on GDP levels, generating the difference. This result is corresponding with (Na, Grohe, Uribe and Yue(2014)), which provides a model that generates the results that countries in floating exchange rate regimes support more external debt in default times. Switching to be more floating can decrease the external debt level directly because countries have other means to support economic activities, like depreciating to increase export revenue, decreasing real wages to decrease firms' costs. This means helping the economy to survive without too much borrowing. The ratio is decreasing because the big switching of regimes may generate a depress in GDP quickly. When GDP decreases more than external debt level, the ratio increases.

For other explanatory variables, for all the columns, the last period external debt presents a high correlation with the current period value. To be specific, a 100% increase of last period external debt generates around 94% increase in current period external debt or 90 percentage points increase in the current debt ratio. For default, only the last column shows that default has a positive impact on external debt. Default increases external debt to GDP ratio by 30 percentage points. This is because, during default episodes, countries are more stressed with money and need to borrow more to recover their economy. Also, GDP decreases a lot during

default times.

GDP level is only included in the log level external debt regressions. A 100% increase in GDP level increases external debt by 3%. When GDP increases, it means the economy is getting better, and the economy needs more finance to support the economy. For the last period deviation of GDP and government expenditure, there are no significant.

At the end of the table, I report the sargan tests chi-square value and p-value. Sargan test has a null hypothesis of "the instruments as a group are exogenous". Therefore, the higher the p-value of the Sargan statistic, the better. Besides, this paper reports the autocorrelation test, the Arellano-Bond test for autocorrelation has a null hypothesis of no autocorrelation and is applied to the differenced residuals. The test for AR(1) process in first differences usually rejects the null hypothesis because: $\Delta e = e_{i,t} - e_{i,t-1}$ and $\Delta e_{i,t-1} = e_{i,t-1} - e_{i,t-2}$. Since they both have $e_{i,t-1}$, they should be correlated. However, the test for AR(2) in first differences is more important since it will detect autocorrelation in levels.

The benchmark regression results have some differences in other people's results. For output gap, Greenidge,Dra- kes, Graigwell(2010) using DOLS estimation, found that increased output gap leads to a reduction in the stock of external debt. For the government expenditure gap, they also found a positive correlation. In the next section, this includes more control variables to see whether the results will change.

2.4.2 Robustness check

2.4.2.1 Corresponding with Benchmark Regression

For the first robustness, considering the endogeneity problem between the regime, default with external debt, this paper uses the lag terms of exchange rate regime, float dummy, the lag term of default, and run the same regressions as the benchmark. Table 2.5 shows the results. The sign and significance are consistent with benchmark regression. The only difference is that default affects the external debt level negatively and significantly. This is because there is no external debt to GDP ratio, so default does not depress GDP to make the ratio more positive. During default times, countries are less likely to be able to borrow money from outside.

Using the same methodology but adding all three classifications of control variables. Table 2.6 shows the robustness check in the log term of external debt. Compared with benchmark GMM regression, the results are consistent. For interaction terms, they all lose significance. In normal times, switching regimes 1 unit up increases 1% of external debt level. If switching regimes from fixed to floating decreases external debt to GDP ratio by three percentage points. That shows that in normal times, more floating regimes countries do not need to borrow too much. Unlikely, the significance disappears for column 2 and 4.

The difference between these two columns is seen by including the nominal exchange rate, which is the announcement exchange rate by the government, which is quite different from the real exchange rate adjusted by the price index. Nominal depreciation lowering external debt means that nominal depreciation increases the real debt repayment through the balance sheet effect. However, real depreciation increase in borrowing is since countries with real depreciation

Table 2.5: Robustness Check Using GMM (Benchmark lag)

VARIABLES	Reg2.1 ED(log)	Reg2.2 ED(log)	Reg2.1 ED/GDP	Reg2.2 ED/GDP
Regime_t1*D_t1	0.0084		0.0788**	
	(0.0046)		(0.0261)	
Regime_t1	-0.0119*		-0.0070	
	(0.0055)		(0.0145)	
Float_t1*D_t1		0.0272		2.1834
		(0.0272)		(1.7453)
Float_t1		-0.0361***		0.1089
		(0.0074)		(0.0620)
Default_t1	-0.0139	-0.0033	0.0165	0.0310
	(0.0128)	(0.0146)	(0.0294)	(0.0366)
GDP(log)	0.0316***	0.0286***		
	(0.0022)	(0.0039)		
Y_gap_t1	-0.0016	-0.0027	0.0014	0.0014
	(0.0022)	(0.0019)	(0.0033)	(0.0062)
G_gap_t1	0.0195	0.0134	0.1203	0.1342
	(0.0255)	(0.0284)	(0.1006)	(0.1887)
ED(log)_t1	0.9451***	0.9419***		
	(0.0042)	(0.0033)		
ED/GDP_t1			0.9039***	0.9284***
			(0.0153)	(0.0038)
Constant	0.6034***	0.7221***	0.0554	-0.0064
	(0.0942)	(0.1341)	(0.0300)	(0.0148)
Observations	1,573	1,575	1,579	1,581
Number of country1	53	53	53	53
Sargan test-Chi-sq	1027.35	1008.61	100.68	95.27
Sargan test-p-value	0.72	0.643	0.99	0.99
AR(1) test-p-value	0.02	0.02	0.02	0.13
AR(2) test-p-value	0.49	0.47	0.46	0.47

benefit from boosted export sectors and can borrow more. Only the real exchange rate depreciation has a significant effect on the level of external debt. The real exchange rate depreciation takes the significant in the regression, showing that real depreciation increases external debt by 0.06%.

GDP still has positive and significant on external debt level. 100% increase in GDP increases the external debt level by 2%. The government spending gap from the last period increases the external debt level by 7%. This means that if the government spent too much in the last period, this would increase external debt in the current period significantly. These results are corresponding with Greenidge, Drakes, Graigwell(2010)s paper.

Reserve to Import ratio coefficients shows that when a country has more other supplies of capital, like their own reserves, they borrow less. Reserve to Import ratio increase 1 unit, generate around 1% decrease of external debt. When the real interest rate increases, the borrowing cost increases, and this will reduce external debt significantly by around 1%. The last period of external debt level is also of great importance to the current period of external debt. The fluctuation of terms of trade, primary government balance in the last period, output gap, default do not significantly affect the external debt level. These results are corresponding with the benchmark regression results.

Table 2.7 shows the same regressions with lagged regimes and float. The results are consistent with Table 2.6.Table 2.8 shows the same robustness check with external debt to GDP ratio as the dependent variable. The results are corresponding with the benchmark regressions and Table 2.6. A 1 unit switching of regimes to more floating decreases external debt to GDP

Table 2.6: Robustness Check Using GMM (Debt level)

VARIABLES	Reg2.3 ED(log)	Reg2.4 ED(log)	Reg2.3 ED(log)	Reg2.4 ED(log)
Regime*D	0.0082	0.0044		
	(0.0108)	(0.0116)		
Regimes	-0.0114*	-0.0060		
	(0.0049)	(0.0036)		
Float*			0.0044	-0.0013
			(0.0227)	(0.0224)
Float			-0.0319*	-0.0264
			(0.0125)	(0.0137)
Default	-0.0132	0.0017	0.0031	0.0096
	(0.0168)	(0.0164)	(0.0064)	(0.0081)
ED(log)_t1	0.9692***	0.9707***	0.9677***	0.9700***
	(0.0081)	(0.0099)	(0.0072)	(0.0094)
GDP(log)	0.0231***	0.0202**	0.0230***	0.0199**
	(0.0048)	(0.0063)	(0.0047)	(0.0063)
Y_gap_t1	-0.0002	-0.0015	-0.0004	-0.0021
	(0.0022)	(0.0022)	(0.0021)	(0.0020)
G_gap_t1	0.0756*	0.0694*	0.0711*	0.0665
	(0.0340)	(0.0319)	(0.0350)	(0.0351)
Reserve/Import	-0.0085	-0.0112*	-0.0087	-0.0114*
	(0.0060)	(0.0047)	(0.0060)	(0.0046)
Real IR	-0.0124***	-0.0100**	-0.0125***	-0.0099**
	(0.0026)	(0.0030)	(0.0024)	(0.0029)
Δtot	-0.0000	-0.0002	-0.0000	-0.0001
	(0.0000)	(0.0003)	(0.0000)	(0.0003)
Gov_Balance(log)_t1	0.0002	0.0004	0.0001	0.0004
	(0.0005)	(0.0004)	(0.0005)	(0.0004)
FDI(log)	-0.0004	-0.0003	-0.0004	-0.0002
	(0.0005)	(0.0004)	(0.0005)	(0.0004)
Ndep	-0.0003		-0.0002	
	(0.0003)		(0.0003)	
Rdep		0.0007**		0.0008**
		(0.0003)		(0.0003)
Constant	0.3507***	0.3650**	0.3692***	0.3777**
	(0.0853)	(0.1069)	(0.0720)	(0.1000)
Observations	1,004	871	1,005	872
Number of country1	43	41	43	41
Sargan test-Chi-sq	840.52	759.42	840.07	756.79
Sargan test-p value	0.99	0.91	0.99	0.92
AR(1) test-p value	0.10	0.12	0.10	0.12
AR(2) test-p value	0.48	0.44	0.48	0.43

Table 2.7: Robustness Check Using GMM (Debt Level,lag)

VARIABLES	Reg2.3 ED(log)	Reg2.4 ED(log)	Reg2.3 ED(log)	Reg2.4 ED(log)
Regime_t1*D	0.0135**	0.0087		
	(0.0049)	(0.0074)		
Regimes_t1	-0.0148**	-0.0103*		
	(0.0047)	(0.0040)		
Float_t1*D			0.0133	0.0059
			(0.0137)	(0.0153)
Float_t1			-0.0364**	-0.0329**
			(0.0114)	(0.0112)
loged_t1	0.9684***	0.9712***	0.9663***	0.9698***
	(0.0075)	(0.0095)	(0.0061)	(0.0087)
GDP(log)	0.0233***	0.0202**	0.0236***	0.0200**
	(0.0048)	(0.0063)	(0.0045)	(0.0063)
Default	-0.0256***	-0.0103	0.0038	0.0073
	(0.0020)	(0.0058)	(0.0082)	(0.0100)
Y_gap_t1	-0.0003	-0.0017	-0.0005	-0.0023
	(0.0023)	(0.0023)	(0.0021)	(0.0022)
G_gap_t1	0.0701*	0.0685**	0.0647	0.0610
	(0.0285)	(0.0262)	(0.0334)	(0.0326)
Reserve/Import	-0.0086	-0.0111*	-0.0091	-0.0116**
	(0.0059)	(0.0047)	(0.0058)	(0.0044)
Real IR	-0.0118***	-0.0095**	-0.0124***	-0.0098**
	(0.0024)	(0.0030)	(0.0023)	(0.0028)
Δtot	-0.0000	-0.0001	-0.0000	-0.0001
	(0.0000)	(0.0003)	(0.0000)	(0.0003)
Balance(log)_t1	0.0002	0.0004	0.0001	0.0004
	(0.0005)	(0.0004)	(0.0005)	(0.0004)
Ndep	-0.0002		-0.0002	
	(0.0003)		(0.0003)	
Rdep		0.0008**		0.0008**
		(0.0003)		(0.0003)
Constant	0.3617***	0.3575**	0.3811***	0.3813***
	(0.0824)	(0.1075)	(0.0597)	(0.0912)
Observations	1,012	871	1,013	872
Number of country1	43	41	43	41
Sargan test-Chi-sq	848.41	760.72	834.96	95.27
Sargan test-p value	0.97	0.92	0.99	0.93
AR(1) test-p value	0.10	0.12	0.10	0.12
AR(2) test-p value	0.49	0.44	0.48	0.47

ratio by 2 percentage points.

For robustness variables, I include export to GDP ratio and FDI to GDP ratio. Only FDI to GDP ratio reduces external debt to GDP ratio because it is another supply of capital, like reserves. Specifically, a 100 percentage points increase in FDI to GDP ratio, around 40 percentage points decrease generates in external debt to GDP ratio. As for exports, the sign is ambiguous and insignificant.

In addition, the sign for nominal and real exchange rate depreciation is flipped and again this is due the effect of depreciation on GDP that flipped the sign of the ratio. Also the last period external debt to GDP ratio drop 1 unit will reduce current value 70 to 80 percent. And the effect of real interest rate becomes insignificant but still negative. For the fluctuation of term of trade also becomes significant. Specifically, a 100 percentage points increase in the fluctuation of term of trade, around 0.04 percentage points decrease generates in external debt to GDP ratio. For government primary balance to GDP ratio, the coefficients are small, but negative and very significant. The flipped sign from expectation is because the depressed primary balance also decreases GDP. Other control variables have the similar significances and signs.

Compared to Greenidge,Drakes,Graigwell(2010)s paper, for external debt in the Caribbean Community, they find that 1% increase of output gap decrease external debt by 0.97 percent points and 1% increase of expenditure gap increase external debt by 0.27 percent point. They also find real depreciation reduces external debt.

Their results are similar to my results even though I contain more country samples in different areas. But for exports, they find exports have negative impact on external debt, which is different from my results. But they do mention in their paper that the relationship between

exports and external debt are ambiguous because there are two effects in different directions. On the one hand, exports revenue increase countries abilities in serving more external debt. On the other hand, sometimes exports increase because of deprecation, which has negative effect on external debt. As above, I also replace regime and float with their lagged term and run the same regression. Table 2.9 shows the results. And the results are consistent with Table 2.8.

Table 2.8: Robustness Check Using GMM (Debt to GDP Ratio)

VARIABLES	Reg2.3 ED/GDP	Reg2.4 ED/GDP	Reg2.3 ED/GDP	Reg2.4 ED/GDP
Regime*D	0.0077	0.0375		
	(0.0139)	(0.0210)		
Regimes	-0.0204***	-0.0150		
	(0.0045)	(0.0079)		
Float*D			-0.0484	0.0982
			(0.0993)	(0.0526)
Float			-0.0113	-0.0257
			(0.0150)	(0.0164)
ED/gdp_t1	0.7402***	0.8581***	0.8592***	0.8635***
	(0.0796)	(0.0088)	(0.0474)	(0.0117)
Default	-0.0077	-0.0580	0.0021	0.0204
	(0.0348)	(0.0485)	(0.0130)	(0.0166)
Y_gap_t1	-0.0015	-0.0004	-0.0020	-0.0003
	(0.0026)	(0.0004)	(0.0012)	(0.0005)
G_gap_t1	0.0220	0.0175	0.1022**	0.0182
	(0.0363)	(0.0111)	(0.0268)	(0.0129)
Export/GDP	0.2183	-0.0120	0.0412	-0.0203
	(0.1495)	(0.0346)	(0.0551)	(0.0377)
FDI/GDP	-0.2482	-0.4862**	-0.4397*	-0.4497*
	(0.1246)	(0.1631)	(0.1997)	(0.1771)
Reserve/Import	-0.0075*	-0.0051**	-0.0049	-0.0055**
	(0.0031)	(0.0014)	(0.0029)	(0.0017)
Real IR	-0.0013	-0.0016	-0.0044**	-0.0023
	(0.0019)	(0.0013)	(0.0015)	(0.0016)
Δtot	-0.0004***	-0.0014**	-0.0003**	-0.0012***
	(0.0001)	(0.0003)	(0.0001)	(0.0003)
Balance/GDP_t1	-0.0000	-0.0000***	-0.0000*	-0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Ndep	-0.0056***		-0.0056***	
	(0.0010)		(0.0010)	
Rdep		-0.0056***		-0.0058***
		(0.0012)		(0.0012)
Constant	0.1440***	0.1552**	0.0787**	0.1270***
	(0.0249)	(0.0406)	(0.0264)	(0.0313)
Observations	1,000	867	1,001	868
Number of country1	43	41	43	41
Sargan test-Chi-sq	365.15	1008.61	100.68	293.48
Sargan test-p value	0.91	0.71	0.99	0.74
AR(1) test-p value	0.09	0.01	0.014	0.01
AR(2) test-p value	0.49	0.47	0.46	0.466

Table 2.9: Robustness Check Using GMM (Debt to GDP ratio,lag)

	Reg2.3	Reg2.4	Reg2.3	Reg2.4
VARIABLES	ED/GDP	ED/GDP	ED/GDP	ED/GDP
Regime_t1*D	0.0219	0.0806		
	(0.0307)	(0.0588)		
Regimes_t1	-0.0169**	-0.0082*		
	(0.0053)	(0.0037)		
Float_t1*D			0.0685	0.0539
			(0.1383)	(0.1385)
Float_t1			-0.0125	-0.0179
			(0.0117)	(0.0111)
ed/gdp_t1	0.7414***	0.8627***	0.7454***	0.8804***
	(0.0794)	(0.0097)	(0.0820)	(0.0134)
Default	-0.0394	-0.1247	0.0291	0.0326
	(0.0789)	(0.0902)	(0.0442)	(0.0221)
Y_gap_t1	-0.0016	-0.0004	-0.0007	-0.0005
	(0.0023)	(0.0004)	(0.0033)	(0.0003)
G_gap_t1	0.0213	0.0158	0.0364	0.0117
	(0.0351)	(0.0115)	(0.0400)	(0.0126)
Export/GDP	0.2243	-0.0009	0.2367	-0.0338
	(0.1364)	(0.0339)	(0.1378)	(0.0366)
FDI/GDP	-0.2257*	-0.4391**	-0.2982**	-0.4229**
	(0.0889)	(0.1679)	(0.1036)	(0.1616)
Reserve/Import	-0.0073*	-0.0047**	-0.0078*	-0.0053**
	(0.0032)	(0.0015)	(0.0031)	(0.0017)
Real IR	-0.0015	-0.0023	-0.0024	-0.0025
	(0.0024)	(0.0022)	(0.0020)	(0.0016)
Δtot	-0.0001	-0.0014**	-0.0004**	-0.0010**
	(0.0002)	(0.0004)	(0.0001)	(0.0003)
Balance/GDP_t1	-0.0000	-0.0000***	-0.0000	-0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Ndep	-0.0058***		-0.0052**	
	(0.0012)		(0.0014)	
Rdep		-0.0058***		-0.0058***
		(0.0011)		(0.0010)
Constant	0.1302***	0.1292***	0.0963***	0.1173***
	(0.0200)	(0.0250)	(0.0220)	(0.0253)
Observations	1,000	867	1,001	868
Number of country1	43	41	43	41
Sargan test-Chi-sq	368.46	280.80	378.93	321.61
Sargan test-p value	0.92	0.91	0.90	0.81
AR(1) test-p value	0.09	0.01	0.01	0.1
AR(2) test-p value	0.46	0.47	0.43	0.48

2.5 Extension

In Extension, I eliminate the lag term of external debt by using the growth rate of external debt. It changes my origianl question by asking how exchange rate regimes switching affects the growth rate of external debt rather than the levels. Since this change, for all the control varibles, I also switching them to the growth rate.Here are the regressions:

$$\Delta ED_{i,t} = \beta_0 + \beta_1 Regime_{i,t-1} + \beta_2 D_{i,t-1} + \beta_3 D_{i,t-1} * Regime_{i,t-1} + \beta_4 \Delta X_{i,t} + \alpha_i + \varepsilon_{i,t} \tag{2.5}$$

$$\Delta ED_{i,t} = \beta_0 + \beta_1 Float_{i,t-1} + \beta_2 D_{i,t-1} + \beta_3 D_{i,t-1} * Float_{i,t-1} + \beta_4 \Delta X_{i,t} + \alpha_i + \varepsilon_{i,t} \tag{2.6}$$

Since there is no lag external debt, I use the the GLS estimator with fixed effect and standard error clustering in regions. And Table 2.10 and Table 2.11 show the regressions in Equation 2.5 and Equation 2.6. I use the lagged term of regime, float and default because of the existence of reverse causalities between external debt and them.

From Table 2.10 and Table 2.11, I find out that increasing exchange rate regime or switching from fixed to floating regimes decreases the growth rate of external debt level, as well as the growth rate external debt to GDP ratio in normal times. However, it increases the growth rate of external debt to GDP ratio during default. The results are consistent with the benchmark and robustness regressions.

Table 2.10: Externsion Using GMM (Debt Level,lag)

	Reg2.5	Reg2.6	Reg2.5	Reg2.6
VARIABLES	ΔED	ΔED	ΔED	ΔED
Regime_t1*D_t1	0.0056	0.0060		
	(0.0069)	(0.0067)		
Regime	-0.0406**	-0.0368***		
	(0.0102)	(0.0087)		
Float_t1*D_t1			0.0217	0.0091
			(0.0179)	(0.0162)
Float			-0.0831***	-0.0681***
			(0.0100)	(0.0063)
Default_t1	-0.0209**	-0.0191	-0.0224*	-0.0198
	(0.0080)	(0.0101)	(0.0110)	(0.0143)
$\Delta GDP(log)$	-0.0285**	-0.0324**	-0.0257**	-0.0298**
	(0.0096)	(0.0110)	(0.0077)	(0.0095)
ΔY_gap_t1	-0.0038	-0.0043	-0.0069	-0.0017
	(0.0115)	(0.0147)	(0.0136)	(0.0169)
ΔG_gap_t1	-0.0320	-0.0244	-0.0377	-0.0323
	(0.0336)	(0.0369)	(0.0329)	(0.0354)
$\Delta Export(log)$	-0.0012	-0.0142	0.0004	-0.0131
	(0.0096)	(0.0096)	(0.0089)	(0.0092)
$\Delta FDI(log)$	-0.0004	-0.0004	-0.0004	-0.0004
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
$\Delta Reserve/Import$	-0.0165**	-0.0170*	-0.0165**	-0.0168*
	(0.0052)	(0.0070)	(0.0054)	(0.0075)
$\Delta RealIR$	-0.0093	-0.0126	-0.0085	-0.0116
	(0.0070)	(0.0093)	(0.0067)	(0.0088)
Δtot	0.0000	0.0002***	0.0000	0.0002***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$\Delta Gov\_Balance(log)\_t1$	-0.0001	-0.0001	-0.0001	-0.0001
	(0.0001)	(0.0002)	(0.0001)	(0.0001)
$\Delta Ndep$	0.0006		0.0004	
	(0.0003)		(0.0003)	
$\Delta Rdep$		0.0007*		0.0006*
		(0.0003)		(0.0003)
Constant	0.1635***	0.1538***	0.0905***	0.0848***
	(0.0226)	(0.0197)	(0.0020)	(0.0005)
Observations	939	816	940	817
R-squared	0.0752	0.0765	0.0517	0.0553
Number of country1	43	38	43	38

Table 2.11: Externsion Using GMM (Debt to GDP Ratio, lag)

VARIABLES	Reg2.5 $\Delta ED/GDP$	Reg2.6 $\Delta ED/GDP$	Reg2.5 $\Delta ED/GDP$	Reg2.6 $\Delta ED/GDP$
Regime_t1*D_t1	0.0165**	0.0177		
	(0.0061)	(0.0096)		
Regime_t1	-0.0207**	-0.0179**		
	(0.0070)	(0.0062)		
Float_t1*D_t1			0.0455*	0.0610*
			(0.0195)	(0.0258)
Float_t1			-0.0445**	-0.0428**
			(0.0147)	(0.0112)
Default_t1	0.0099	-0.0088	0.0149*	-0.0041
	(0.0057)	(0.0117)	(0.0071)	(0.0145)
ΔY_gap_t1	0.0308	-0.0756	0.0294	-0.0745
	(0.0289)	(0.0491)	(0.0274)	(0.0477)
ΔG_gap_t1	-0.1808**	-0.0961*	-0.1881**	-0.1061**
	(0.0555)	(0.0399)	(0.0538)	(0.0405)
$\Delta Export/GDP$	0.7869**	0.7453***	0.7947**	0.7547***
	(0.2619)	(0.1732)	(0.2574)	(0.1729)
$\Delta FDI/GDP$	0.6890***	0.7484	0.6549***	0.7014
	(0.1399)	(0.4892)	(0.1445)	(0.4984)
$\Delta Reserve/Import$	-0.0135*	-0.0077	-0.0135**	-0.0074
	(0.0053)	(0.0066)	(0.0051)	(0.0061)
$\Delta RealIR$	-0.0055**	-0.0039	-0.0056**	-0.0040
	(0.0019)	(0.0023)	(0.0017)	(0.0022)
Δtot	-0.0001	-0.0008***	-0.0001	-0.0008***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\Delta Gov\_lBalance/GDP\_t1$	-0.0000***	-0.0000**	-0.0000***	-0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$\Delta Ndep$	-0.0026***		-0.0027***	
	(0.0003)		(0.0003)	
$\Delta Rdep$		-0.0020**		-0.0021**
		(0.0006)		(0.0006)
Constant	0.0302	0.0272	-0.0032	-0.0006
	(0.0162)	(0.0157)	(0.0036)	(0.0044)
Observations	939	816	940	817
R-squared	0.1219	0.1847	0.1156	0.1749
Number of country1	43	38	43	38

2.6 Conclusion

Using a dataset containing 57 countries' 232 default episodes from 1970 to 2007, this paper empirically analyzes the relationship between exchange rate regimes and external debts during default episodes and normal times. The results suggest that the effect of exchange rate regimes has on external debt is different between default times and normal times. The results differ from different measurements of external debt as well. One unit switch of exchange rate regimes decreases the external debt level in normal times but increases it during default. Switching from fixed regimes to floating regimes decreases the external debt level but increases external debt to GDP ratio. This difference in the sign is because switching exchange rate regime depresses GDP as well.

Chapter 3

News Sentiment and Topic Analysis on Crude Oil Future Prices

3.1 Introduction

Crude oil is a commodity, and as such, it tends to have large fluctuations at price than more stable investments such as stocks and bonds. Crude oil prices are influenced by a variety of factors. As Baumeister and Kilian (2014) conclude that the explanatory power of these factors vary over time and that different factors are important at different time horizons. Though the factors to explain the change of oil prices are inconclusive, Brandt and Gao (2019) mention that news information can provide a way to quantify macroeconomic and other events that could affect crude oil prices. In this paper, we analyze the contents of news articles to study how information about crude oil related news affects crude oil futures price.

News information could affect crude oil prices in different ways. Shiller (2015) ar-

gues that the news media plays an important role in setting the stage for market moves and provoking them. On one hand, news could convey information on the current significant variables that affect the price, reflecting the current market confidence. On the other hand, news could serve as an update of the changes in these significant variables, reflecting the expectations about future oil supply and demand conditions, which could affect the current crude oil prices.

In this paper, we consider broad web news from various sources. We analyze these web crude oil related news to uncover how this information affects the crude oil futures price using the intra-day high-frequency data. To do so, we use both supervised and unsupervised machine learning algorithms to study the impact of news sentiment and news topics on crude oil futures price increase. The important assumption in our analysis is that the crude oil futures market is nearly perfectly efficient, which means that the price can adjust quickly to any new information released to the public.

A key contribution of our paper is to demonstrate finer and more objective classifications of news effects on crude oil price change. First, we use unsupervised machine learning algorithms to group crude oil related news articles into different topics without providing prior knowledge on how each topic links to a particular set of words. Second, we conduct both news sentiment analysis and news topics analysis of each news article using intra-day high- frequency data, these results providing us a new index data indicating crude oil price increase or decrease for future studies based on textual analysis.

In this paper, first, we conduct news sentiment analysis with logistic regression to see how each word in an article can affect the increase or decrease of the crude oil price from September 2019 to December 2019. We find that among all the news words, 152 words have

coefficients smaller than -0.5, which are collected as the most negative words. Also, 159 words have coefficient over 0.5, we defined these words as the most positive words. Based on the news sentiment analysis results, we also construct a novel index indicating the sentiment score of a news article given the most positive (negative) words. Second, we categorize news articles discussing crude oil over the entire year of 2019 into topics using unsupervised machine learning algorithms K-means. The K-means algorithm generates 4 clusters of news topics. We rename and interpret each of them to be World Crude Oil topic, WTI Crude Oil topic, Financial Analysis topic, and Editorial Opinion topic. Each of the news articles would be assigned to be one of the topics. Finally, we estimate how the news topic and sentiment score would affect the crude oil future price using logistic regression in 5 minutes window. The results suggest that on average, "World Crude Oil" news has the highest correlation with a crude oil price increase. Moreover, the more positive news is under the topic "WTI Crude Oil," the higher probability that WTI crude oil futures price will increase within 5 minutes.

The remainder of this paper is organized as follows: Section 3.2 presents the related literature. Section 3.3 provides detailed data information. Section 3.4 shows the news sentiment analysis methodology and how the positive score is calculated for every news article. Section 3.5 details the construction of topic analysis. Section 3.6 presents our main empirical analysis. Section 3.7 concludes.

3.2 Literature Review

This paper mainly relates to three streams of research: (1) studies on general crude oil price; (2) studies on news effects of crude oil price; (3) studies on news textual analysis on general equity market. Broad literature has studied the explaining factor of crude oil price. Hamilton (2009) suggests that the real price of oil follows a random walk without drift. Rapaport (2013) distinguish between demand and supply driven component of crude oil returns by examining its correlation with the equity market. Baumeister and Kilian (2014) discuss an exhaustive set of oil pricing factors from the literature, conclude that the explanatory power of these factors varies over time and that different factors are important at different time horizons. These findings crucially depend on the underlying model structure and assumptions. This paper corroborates these effects from the assumption that the crude oil futures market is nearly perfectly efficient, updated news information is a direct way to reflect the market sentiment.

Our paper is part of a growing body of research using textual analysis to examine how news affects economic and financial variables. Most work in this literature deals with the general equity market and aggregate news about equities and the economy. For example, Garcia (2013) studies the effect of sentiment on asset prices using the New York Times between 1905 and 2005. They find that after controlling for other well-known time-series patterns, the predictability of stock returns using news content is concentrated in recessions. Soo (2015) develops a measure of sentiment across local housing markets by quantifying the positive and negative tone of housing news in local newspaper articles. Bi and Traum (2020) examines how newspaper reporting affects government bond prices during the U.S. state default of the 1840s.

Our paper is different from the aforementioned studies in that we consider a much broader set of news resources and news categories as inputs, we can capture both micro-level supply and demand factors and macroeconomic related factors, even geopolitical developments.

This paper also contributes to the literature about news effects on the crude oil prices. Most of the literature work with the effects of regularly scheduled macroeconomic releases on crude oil price. For example, Kilian and Vega (2011) propose a formal test of the identifying assumption that energy prices are predetermined for U.S. macroeconomic aggregates using daily energy prices on daily news from the U.S. macroeconomic data releases. In this aspect, this paper is different from the existing literature focusing on responses of crude oil prices to scheduled macro news announcement from U.S. Another difference from papers on the schedule announcement is that our news indices are at high frequency than the scheduled releases with a fixed frequency. A recent paper by Brandt and Gao (2019) uses sentiment scores for global news from RavenPack global macro package to see how news about macroeconomic fundamentals and geopolitical events affect crude oil markets. Our paper is different from this paper since we use machine learning algorithms to group news into different topics, which avoid prior knowledge and generate more objective classifications of news effects on crude oil price change.

3.3 Data

3.3.1 Data Source and Description

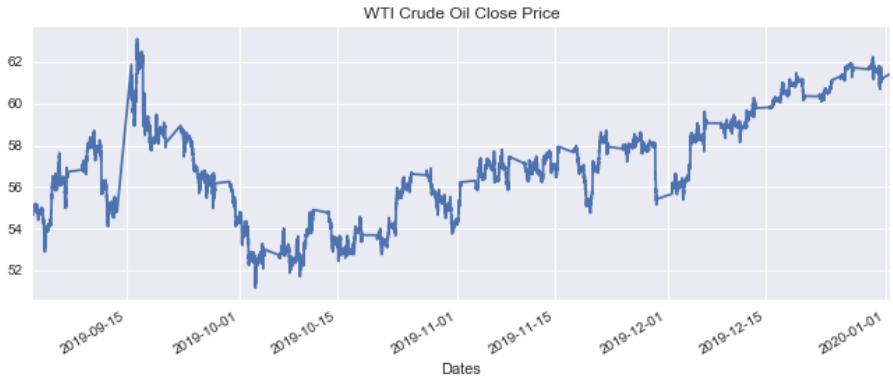
The data used in this paper are obtained through Bloomberg terminal, including the west texas intermediate (WTI) crude oil futures prices, and crude oil-related news articles. This paper focuses on WTI crude oil. WTI refers to oil extracted from wells in the US and sent via pipeline to Cushing, Oklahoma. This paper chooses WTI crude oil over Brent crude oil, which counts two-thirds of all crude contracts around the world, and Dubai crude oil, which is the major supply for the Asian market. The reason is that WTI crude oil has been the main benchmark for crude oil consumed in the United States. Thus its price is more related to the supply and demand conditions in the US market than in other markets worldwide.

The intra-day trading data is provided by Bloomberg, with price updating at high frequencies. However, the trading will be suspended every day at 14:00 to 15:00, as well as every Saturday. In order to match the frequency of news, this paper uses high-frequency intra-day trading futures prices for WTI crude oil. The price is updating every five minutes, which matches the first quantile of news frequency distributions[1]. This paper takes the sample period to be the last quarter of 2019. Figure 3.1 shows the WTI crude oil close price throughout the period. From Figure 3.1, it is clear that the close price has high volatility but no obvious trend during the last quarter of 2019, which makes it a perfect sample for this paper's analysis.

News articles are the major sources for text analysis. This paper analyzes contents of

[1] Shown in Figure 3.4

Figure 3.1: Close Price Fluctuation in Sample Period



each news article to find the news effect on WTI crude oil futures price. The news articles are collected through the Bloomberg terminal's news section. By specifying the news topics to be 'Crude Oil', including the keyword, 'WTI', Bloomberg terminal gives editorial recommended Bloomberg and web news from various sources. Figure 3.2 is a word cloud showing the most frequent words in the news articles collected by this paper. In 2019, news articles under the topic of WTI crude oil discussed the 'oil price', 'trade war', 'saudi arabia', 'texax intermediate', etc. More detail analysis in news contents will be presented in Section 3.4.

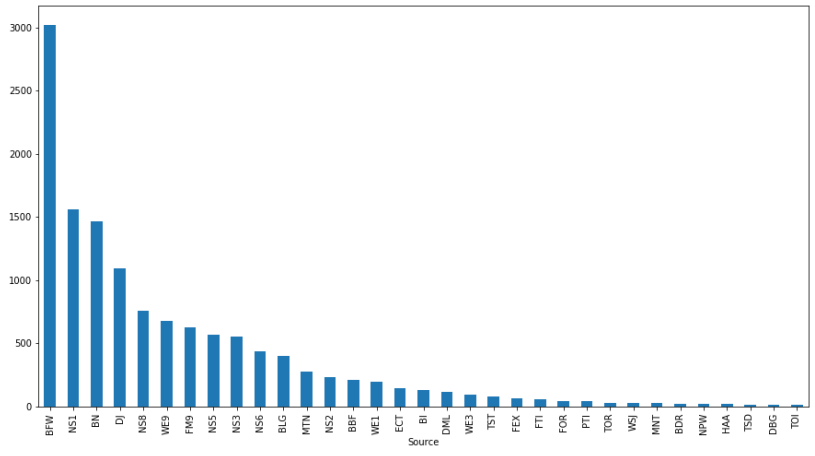
Figure 3.3 shows the major news sources for this paper's dataset. In total, there are

Figure 3.2: 2019 WTI Crude Oil Related News Frequent Words

75 news sources. Figure 3.3 plots the news sources with total news article released counts over ten. All news are global news written in English, reporting crude oil-related topics worldwide. Around 6000 news articles are web news, Bloomberg provides the news subject and the link of the websites. Web news content is collected through web scraping. Two-thirds of the web news is successfully obtained through web-scraping, with the rest replaced by its subject. Throughout 2019, there are 13183 news articles, and 4615 of them are released after September. Since this paper only has price data from September 2019 to December 2019, only news articles released after September will be included in news sentiment analysis and the final regressions, but all news articles in 2019 will be included in news topic analysis.

As for the frequencies of news articles, Figure 3.4 presents that, on average, around

Figure 3.3: Major News Source

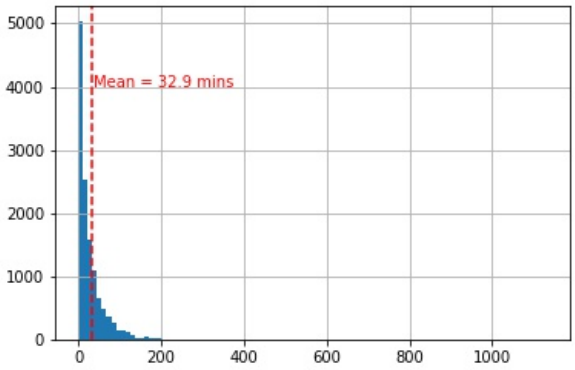


Note: News sources are extracted from each news body, provided by Bloomberg

every thirty minutes, there is one news article about WTI crude released. The news frequency distribution gives guidance on what should be price frequency to analyze the news effect. At-

tempt to uncover the news effect, this paper assumes the market is efficient, which means the market reacts fast enough to new information. This paper thus defines any news article's effect on WTI crude oil future price is reflected by how the price was changed within five minutes after this news has been released.

Figure 3.4: News Frequency Distribution



Note: This graph shows the distribution of how frequently a news has been released in the dataset. The x-axis is in the unit of minute, and the y-axis is the count. On average, around every 33 minutes, there release one news article about WTI crude oil in 2019.

To analyze how the price was changed, this paper establishes a dummy variable based on WTI crude oil future price, indicating a price increase or decrease episode. If price increases within five minutes after a news article has been released, the price dummy will be one. Otherwise, if price decreases or doesn't change, the price dummy will be zero. There is nearly no case for the price to remain the same within the five-minute slot throughout the dataset. Thus, when price dummy equals zero, it means the price has decreased within five minutes after this news release. For each news article, by looking for the price change within five minutes after

release, this paper matches news and the price dummy. Figure 3.5 examines whether the data is balanced by comparing the number of price increase episodes with price decrease episodes. Roughly speaking, the data is balanced, with increasing price episodes slightly larger.

Figure 3.5: Price Dummy Distribution

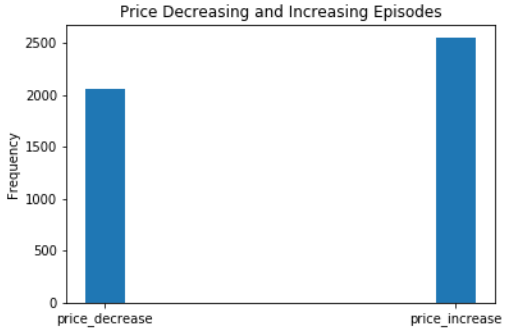
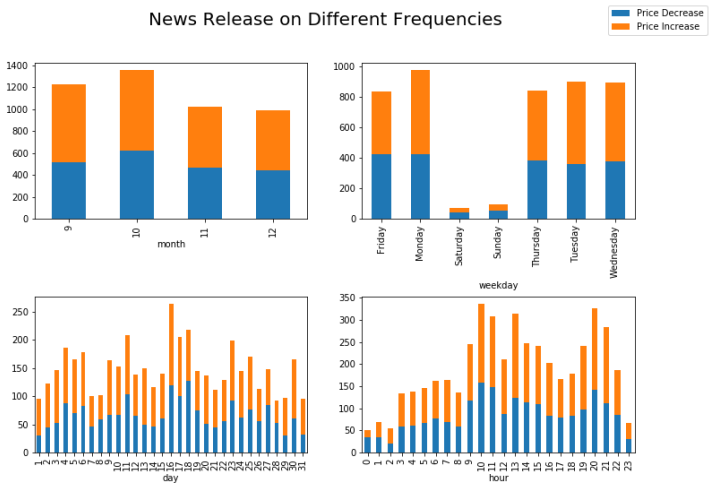


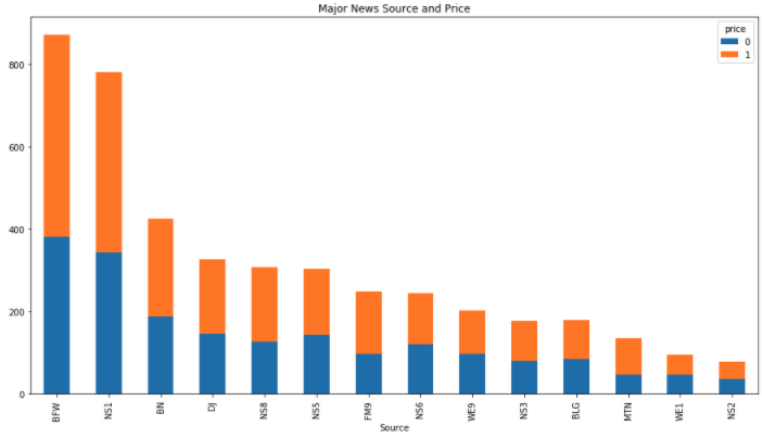
Figure 3.6: Matching News and Price



Except examining the balance of episodes across all datasets, this paper also presents the price sensitivity in different release times of the news articles. In other words, whether news released at certain times, like in the morning or at the end of the month are more likely to have a

biased impact on price. Figure 3.6 shows the price decrease and increase episodes distributions across different months, days of the week, date of the month, and hours of the day. The figure indicates that over one thousand news on the topic of crude oil news were released every month. Moreover, most of the news are released on weekdays rather than on the weekends, and most of them are released during the daytime rather than night. However, there are no obvious patterns for the date of the news release. Comparing the number of news following price increase and price decrease episodes, Figure 3.6 presents that there are no significant differences in these two episodes at any time frequencies. Similarly, Figure 3.7 attempts to uncover the relationship between price and news sources. By examining the major news sources, it is clear that all these new sources have almost even numbers of the price increase and decrease episodes across the dataset. Thus, it is essential to analyze the news contents, rather than the time of the news release, or the news sources, to understand the positive or negative news effect on crude oil futures prices.

Figure 3.7: Matching News and Price



3.3.2 Analyze Text Data

In order to analyze the news contents, this paper preprocesses news before further analysis with regressions. Text data are unique data types that need transformation before fitting into a regression model. This paper follows the standard text mining procedures to extract the useful features from the news contents, including tokenization, removing stopwords and lemmatization.

The First step of preprocessing text data is to break every sentence into individual words, which is called tokenization. Taking individual words rather than sentences breaks down the connections between words. However, it is a common method to use to analyze large sets of text data. It is efficient and convenient for computers to analyze the text data by examines what words appear in an article and how many times these words appear, and this analysis is sufficient enough to give insightful results.

After tokenization, each news article will transform into a list of words, symbols, digits, and punctuation. The next step is to remove useless information. For this analysis, symbols, digits, and punctuation are not very useful, so that this paper removes them. Furthermore, this paper removes stopwords. Stopwords are words that frequently appear in many articles, but without significant meanings. Examples of stopwords are 'I', 'the', 'a', 'of'. These are the words will not intervene in the understanding of articles if removed. Besides using the standard English stopwords provided by the NLTK library[2], this paper also includes other lists of stopwords, provided by Loughran and McDonald(2016). These lists of stopwords are widely used in economic analysis, including dates and time, more general words that are not economic

[2] NLTK is a python package for text analysis. It contains a list of English stopwords.

meaningful[3].

Removing stopwords, along with symbols, digits, and punctuation, each news article will transform into a list of meaningful words. However, in order to count the appearance of each word, it is essential to remove grammar tense and transform each word into its original form. For example, if we want to calculate how many times the word 'open' appears in a news article, we need to count the appearances of 'open', 'opens', 'opened'. Thus, lemmatization is an essential step for text transformation. Lemmatization is taking each word into its original lemma. Another way of converting words is called stemming, which is taking the linguistic root of a word. The reason why this paper chooses lemmatization over stemming is that after stemming, some words become hard to read. For interpretation purposes, the lemma is better than the linguistic root. After lemmatization, each news article will transform into a list of words that are all in their original forms.

Figure 3.8 and 3.9 shows an example of news article before and after text prepro-

Figure 3.8: Original News Article

```
'     Total Refinery, Alpiq, Trade Woes: European Energy Pre-Market  2019-08-26 06:12:07.16 GMT      By John Viljoen
(Bloomberg) -- The following may affect European energy  shares today:  * Note, U.K. markets closed due to holiday
News  * Watch These European Stocks as U.S.-China Trade Woes Escalate  * ALPH SW: Alpiq First Half Adjusted Ebitda CH
F55 Mln  * FP FP: Total Reduces Some Gonfreville Refinery Unit Rates Amid  Strike  * LSNG RM: Lenenergo Second Quarte
r Net Income 3.91 Bln Rubles,  +36% Y/y  * ORSTED DC: Ørsted, Eversource Submit Massachusetts Wind Farm  Proposal
Commodities  * WTI Crude: -1% to $53.64/bbl  * Brent Crude: -0.9% to $58.83/bbl  * Natgas: +1.1% to $2.175/Mmbtu    A
genda  * N.A. Energy Weekly Agenda  * Oil daybook Europe  * Earnings:  ** Other  *** Alpiq Holding AG (ALPH SW)  ***
Maha Energy AB (MAHAA SS)    For more energy wraps in Europe, click here.  For more energy sector wraps in the U.S.,
click here.    To contact the reporter on this story:  John Viljoen in Cape Town at jviljoen@bloomberg.net  To contac
t the editor responsible for this story:  Blaise Robinson at brobinson58@bloomberg.net    '
```

Figure 3.9: News Article after Text Preprocessing

```
'total refinery alpiq trade woe european energy pre market john viljoen bloomberg follow affect european energy share
today note market close due holiday news watch european stocks hina trade woes escalate alph alpiq half adjust ebitda
mln total reduces gonfreville refinery unit rates amid strike lsng lenenergo net bln rubles orste rste eversource sub
mit massachusetts wind farm proposal commodity crude brent crude natgas mmbtu agenda energy agenda oil daybook europe
earning alpiq holding maha energy energy wrap europe click energy sector wrap click contact reporter story john viljo
en cape town contact responsible story blaise robinson'
```

cessing. After tokenization, removing unnecessary words and lemmatization, the original news

[3]Loughran and McDonald's list can be found at https://sraf.nd.edu/textual-analysis/resources/StopWords.

articles only contains informative words that are ready for further transformation, which will be discussed in Section 3.4.

3.4 News Sentiment Analysis

This section will discuss the news sentiment analysis methodology and how the positive score is calculated for every news article. First, this paper analyze the sentiment of each unique word using a logistic regression. The estimated logistic regression coefficient for each unique word represents its sentiment. This paper defines the effect's direction of each unique word by the sign of its coefficient, and the size of the effect by the absolute value of its coefficient. Moreover, by selecting the words with the highest absolute value in coefficients, this paper defines the most positive and negative words indicating a price increase or decrease episode. Lastly, this paper calculates the positive score for each news article based on how many positive and negative words this news article contains.

3.4.1 News Sentiment Analysis on Unique Words

News sentiment analysis is the analysis that uncovers the predicting power of each unique word in indicating a price increase or decrease episode. In order to do so, this paper uses a supervised machine learning algorithm, which is called logistic regression. Logistic regression is a classification algorithm that deals with binary classification problems. Binary classification

has exactly two classes to choose between. In this paper, there are positive and negative classes indicating price increase or price decrease. Logistic regression is a linear classifier, it is a transformation from a linear function:

$$f(x) = b_0 + b_1 * x_1 + ... + b_n * x_n \tag{3.1}$$

where b_0, $b_1 ... b_n$ are the estimators of the regression coefficients for a set of independent variable $x = (x_1, x_2 ... x_n)$. The logistic regression function $p(x)$ is the sigmoid function of $f(x)$:

$$p(x) = \sigma(f(x)) = \frac{1}{1 + exp^{-f(x)}} \tag{3.2}$$

After transformation, $p(x)$ will be in the range of $[0, 1]$, which can be interpreted as probability. Generally, $p(x)$ is interpreted as the predicted probability that $f(x)$ given x is equal to one, and $1 - p(x)$ is the probability that $f(x)$ is zero. In this paper, $p(x)$ is defined as the probability that WTI crude oil futures price increases within five minutes after news article x_i's release.

Applying logistic regression to conduct news sentiment analysis, this paper treats each news article as a observation, and the contents in news article as the features, and estimates β_{w0},

β_{w1}, ... β_{wj} from the following equation:

$$\begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \dots \\ Y_i \end{pmatrix} = \begin{pmatrix} X_{0,w0} & X_{0,w1} & \dots & X_{0,wj} \\ X_{1,w0} & X_{1,w1} & \dots & X_{1,wj} \\ X_{2,w0} & X_{2,w1} & \dots & X_{2,wj} \\ \dots & & & \\ X_{i,w0} & X_{i,w1} & \dots & X_{i,wj} \end{pmatrix} * \begin{pmatrix} \beta_{w0} \\ \beta_{w1} \\ \dots \\ \beta_{w_j} \end{pmatrix} \tag{3.3}$$

where i stands for each news article as a new observation, and wj is the jth unique word in all news articles. On the left hand side, Y_i is the price change dummy described in the previous section. Specifically, the value of Y is decided by the following conditions:

$$Y_i = \begin{cases} 1, & \text{if } price_{t+5} - price_t > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

On the right hand side, the first term is a sparse matrix, with each row stands for each news article and each column stands for each unique word. There are over 20,606 unique words that has ever shown in 4616 news articles, which indicates the shape of the sparse matrix. Each value $X_{i,wj}$ of the sparse matrix is denoted as the tfidf value for each unique word wj in each news article i. Tfidf is short for term frequency, inverse document frequency. It is a common feature engineering method for text analysis and is widely used in literature. For example Bi and Traum(2020), Fraiberger(2019), Shapiro(2018) have used tfidf to extract text features for

different analysis. Specifically, tfidf is calculated by:

$$X_{i,wj} = \frac{1 + log(t_{i,wj})}{1 + log(\sum_i^N t_{i,wj})} * log(\frac{N}{\sum_{wj} t_{i,wj}}) \tag{3.5}$$

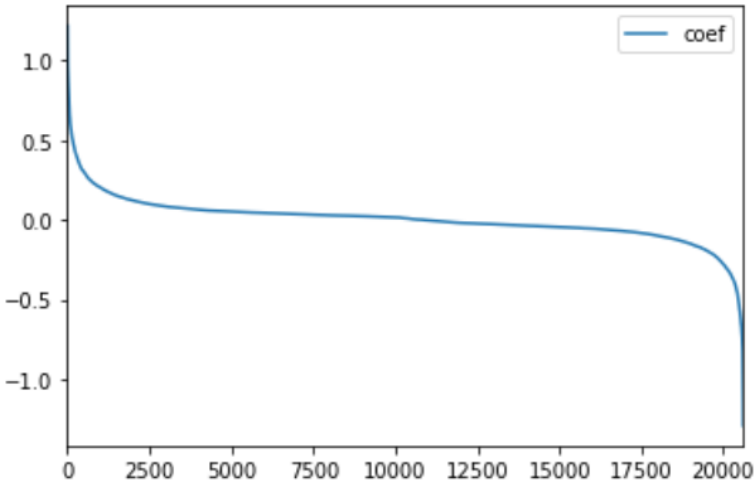
where $t_{i,wj}$ is the frequency of word wj appears in news article i. By examining the equation, it is clear that the first term is the calculating the term frequency and the second term is calculating the inverse document frequency. The first term is evaluating how many time the word wj appear in news article i, normalized by the length of news article i. The higher term frequency indicating a higher tfidf value, presenting the fact that the word wj plays a very important role in news article i by appearing significant times. However, the effect of wj will be weaken if wj also appears in many other news articles besides i, which means it is a common word for this topic. This process is captured by the second term, which is the inverse of how many news articles wj appears divided by the total number of news articles. In this case, N equals to 4616. Combining two effects, a word wj with high tfidf values in news article i means that wj appears many times in news article i, and only appears in few other news articles.

After transforming all preprocessed news articles into the sparse matrix, all data are ready for regression. Figure 3.10 shows the estimation results from Equation 3.3 fitting a logistic regression model. Each point in the x-axis stands for a unique word collecting from all news articles, and there are 20,606 of them. The y-axis stands for the sign and the size of the coefficient for each word. Figure 3.10 indicates that most of the unique word by itself has very limited effect in affecting price, with coefficient very close to zero. However, there are some words that have coefficients with absolute value over 0.5, which are defined as the most positive

and negative words by this paper.

For the most positive and most negative words, based on the logistic regression re-

Figure 3.10: The Effect of Words



Note: Each point in the x axis stands for a unique word, collecting from all news articles

sults, 152 words have coefficients smaller than -0.5, and they are collected as the most negative words. On the other side, 159 words have coefficient over 0.5, and are defined as the most positive words. In total, positive and negative words account for around 1.5% of all unique words. Figure 3.11 shows the word clouds of the most positive and most negative words. The font size of each word in the word clouds indicates the size of its coefficient in absolute value. Figure 3.11 shows that words like 'analyst','investing' are showing the greatest effect in indicating a crude oil futures price increase episode, and words like 'partner','hike' contribute the most to crude oil futures price decrease in a news article.

Figure 3.11: The Word Clouds

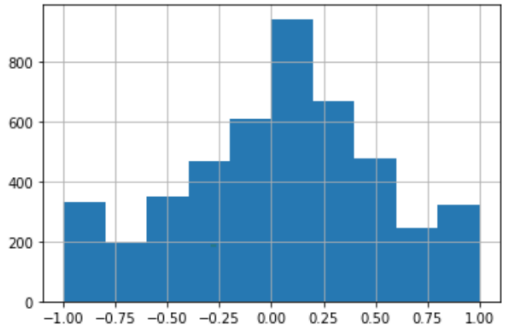
3.4.2 Positive Score for Each News Article

After defining the most positive and negative words in terms of contributing to price increase or decrease episodes, this section will discuss how to calculate the sentiment score for each news article. This paper defines the sentiment score of a news article based on how many positive and negative word it contains. Specifically, this paper calculates the positive score for each news article by comparing how many times positive words appear with the appearance of the negative words. Thus, the positive score of a news article i is defines as follow:

$$Score_i = \frac{N_{pos} * F_{pos} - N_{neg} * F_{neg}}{N_{pos} * F_{pos} + N_{neg} * F_{neg}} \tag{3.6}$$

where N_{pos} is the total counts of positive words in news article i, and F_{pos} is the total frequency of each positive/negative word in news article i. For example, if news article i has two positive words, and one appears 10 times and another one appears 15 times, N_{pos} will be two and F_{pos}

Figure 3.12: The Positive Score Distribution



will be 25. To calculate the positive score, this paper finds the total occurance of positive words and negative words in each news article, takes the difference, and normalizes it by their summation. Figure 3.12 shows the positive score distribution across all news articles.

The positive score of a news article takes a value between $[-1, 1]$. A positive value means this news article is positive, while a negative value means this article is negative. As the absolute value of positive score increases, it becomes more positive or more negative depending on the sign. Figure 3.12 shows that most of the news article has a positive score close to zero, indicating the distribution is roughly a normal distribution with mean at zero, if ignore the two ends. If the positive score for a news article is one, it only has positive words and vice versa. Among all 4616 news articles, only around 600 articles are in such cases. Thus, this paper still assume the monotonic positive correlation between the positiveness of a news article and its positive score.

In summary, this section uncovers the effect of each unique word collection from all news articles in the last quarter of 2019, attempts to estimate the coefficients for each unique word on how it affects the crude oil futures prices. In order to do so, this paper uses a supervised machine learning algorithm called logistic regression to solve this binary classification problem.

Logistic regression estimates the coefficients for all unique words and selects the most positive and negative words based on the coefficients' sign and size. The most positive and negative words help identify the sentiment score for each news article. By calculating the total occurrence of positive and negative words in a news article, this paper defines the positive score for each news article, which is useful for the regression in Section 3.6.

3.5 Topic Analysis

Instead of reading each article and manually separating our news sample into different topics, we use an unsupervised machine learning algorithm, K-means, to detect common patterns in news articles and group them into clusters, i.e., topics.

3.5.1 K-means Algorithm

K-means is one of the most commonly used clustering algorithms in machine learning. Unlike supervised learning, which first defines a list of keywords in each topic and classify each article that includes those keywords into a specific topic, researchers first need to determine the number of clusters (topics), then the K-means algorithm will assign each sample to the cluster where its distance from the centroid of the cluster is minimized. The algorithm of K-means is described as followed:

1. Specify the number of K clusters. [4]

[4] In theory, the Elbow method helps to determine the optimal K values, by plotting K on the horizontal axis and

2. Initialize the centroid point $\mu_k (k \in K)$ of each cluster with a random value.

3. Calculate the squared Euclidean distance of each sample x_i to the centroid point of each cluster.

$$||x_i - \mu_k||^2$$

4. Assign the sample x_i to the closest cluster k where the squared Euclidean distance is minimized.

$$\min_k ||x_i - \mu_k||^2$$

5. Update μ_k by taking the mean of sample points assigned to cluster k.

6. Repeat 3-5 until the sum of the squared distances overall K clusters is minimized.

$$\min \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2$$

where c_k denotes sample points in cluster k.

Following Bi and Traum(2019), we use the tfidf vector of each article (i.e., row vector of the tfidf matrix discussed in the previous section) as a sample point for training the algorithm, so that each article will be assigned uniquely to a cluster. The training set covers the whole sample period (20190101 - 20191231), which includes a total of 13,183 news.

When determining the value of K, the Elbow method doesn't provide us an optimal

sum of minimized distances of each cluster on the vertical axis. The optimal K is found when the slope of Elbow curve flattens, i.e., when the y value converges. However, in practice, the Elbow curve does not converge so that users need to determine the value of K based on their purpose. Our result of Elbow method can be provided upon request.

choice of K (the Elbow curve doesn't converge). Therefore, we run the algorithm several times with different K values and thus choose $K = 4$ based on the results that provide us with the most meaningful and best interpretability of news topics.

3.5.2 Clustering Results

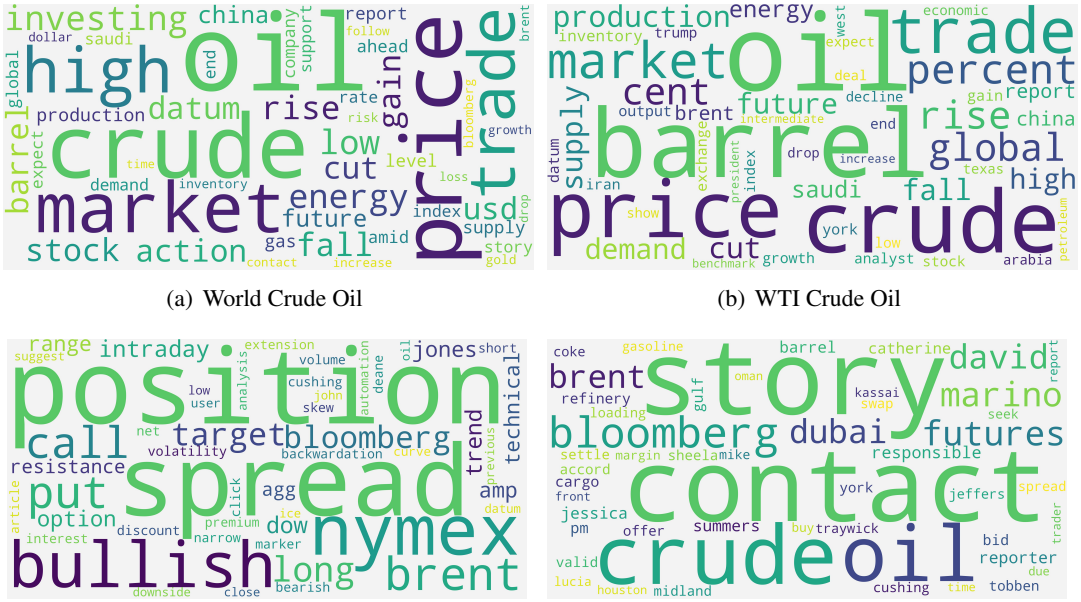
The K-means algorithm generates 4 clusters of news topics. Figure 3.13 plots the word clouds of top 50 important words[5] for each topic, with the more important ones shown in a bigger font. Given our specification of four news topics, we are able to interpret each topic to be: (a) - "World Crude Oil", (b) - "WTI Crude Oil", (c) - "Financial Analysis" and (d) - "Editorial Opinion". The "World Crude Oil" cluster has keywords related to the global oil market, such as market, trade, energy, China, USD, etc. "WTI Crude Oil" cluster includes keywords like price, supply, demand, production, which reflects more about the information of WTI crude oil in the North American region[6]. "Financial Analysis" contains keywords position, spread, call, put, option, relating to the price analysis, and use of derivatives for crude oil. The last cluster is "Editorial Opinion," with keywords story, contact, reporter, etc.

Table 3.1 shows the number of news in each topic. More than 80% of news throughout year 2019 are assigned to the "World Crude Oil" and "WTI Crude Oil" topics. 13.86% news are related to "Editorial Opinion", which is more likely to be a summary or report of already released new information. There are only 563 news classified into "Financial Analysis" topic,

[5]For each cluster, we sum each word's tfidf value over all the articles within this cluster, rank them from the highest to the lowest, and then pick the top 50 words for word cloud plot.

[6]There are some words shown in both of these two topics, which interprets these two topics to be unclear and need to be improved in future work.

Figure 3.13: Word clouds of Clustering Results



(a) World Crude Oil

(b) WTI Crude Oil

(c) Financial Analysis

(d) Editorial Opinion

which includes both analysis of price change and analysts' forecast of future price movement. As we can see in the next section, the last two topics are less important indicators of price changes.

Table 3.1: Number of News in Each Topic

Topic Index	Topic Name	news_cnt	%
a	World Crude Oil	6,386	48.44%
b	WTI Crude Oil	4,407	33.43%
c	Financial Analysis	563	4.27%
d	Editorial Opinion	1,827	13.86%
total		13,183	100%

3.6 Test News Topic and Sentiment Score on Futures Price Change

In this section, we estimate the relation of news topic and sentiment score with futures price change using logistic regression. By assuming that the WTI crude oil futures market is nearly perfectly efficient, investors responding quickly to any new information released to the market, futures price should adjust quickly in line with its intrinsic value. For example, a good news article about the discovery of new oil wells should push down the futures price immediately, as investors are competing with each other to find this price decrease opportunity. Since the market is nearly efficient, the futures price can reflect this new information very quickly. Given this assumption, we can test the relation between crude oil news released and its futures price change.

We first define our dependent variable to be one if futures price after 5 minutes is higher than the price when news released; and zero otherwise[7], as stated in Section 3.3. In future work, we will use different time spam, calculating price change for robustness checks. To test how crude oil news relates to price changes, we use the logistic regression, which estimates the influence of x on Y's probability for each news article. The regression equation we estimate follows:

$$Prob(Y_i = 1|x) = \sigma(\beta_0 + \beta_1 Score_i + \sum_k^3 \beta_k Topic_k + \sum_k^3 \beta_{k,i} Topic_k * Score_i) \tag{3.7}$$

[7]The price data we use is high-frequency data, with 5 minutes interval. In our final regression sample (2019Q4), there doesn't exist price unchanged within each 5 minutes interval.

where

$$Y_i = \begin{cases} 1, & if \quad price_{t+5} > price_t \\ 0, & otherwise \end{cases} \tag{3.8}$$

and *Topic*, a categorical variable, is {0: "World Crude Oil"; 1: "WTI Crude Oil"; 2: "Financial Analysis"; 4: "Editorial Opinion"}. In the regression, three dummy variables are generated indicating which topic this news article is in, and topic 0('World Crude Oil') is chosen as the benchmark topic. As described in section 3.4, *score* evaluates the degree of news sentiment. The higher the *Score* is, the more positive the news is.

The LHS of equation (3.7) denotes the probability of price increase ($Y = 1$) given all x. As discussed in Section 3.4.1, on the RHS, $\sigma(\cdot)$ is a Sigmoid function. In the parenthesis of $\sigma(\cdot)$ is the linear combination of independent variables that we are interested.

Our final logistic regression sample covers only the last quarter of 2019, with 4,616 observations due to the availability of high-frequency futures price data. Table 3.2 reports the regression results. The coefficient on constant is interpreted as the impact of news in topic 0 ("World Crude Oil") on the probability of price increase. On average, news in the "World Crude Oil" topic is more likely to be correlated with price increases. If the news is related to "WTI Crude Oil," the coefficient becomes 0.0159 (0.3267 minuses 0.3108), meaning that "WTI Crude Oil" news on average doesn't have a significant impact on a price increase. For topic 2 & 3 ("Financial Analysis" and "Editorial Opinion"), news has relatively less positive impacts on price increases, as their coefficients (-0.2821 and -0.2581) are both statistically significant and smaller than the coefficient of topic 0 (0.3267).

After considering the impact of sentiment score for each news under different top-

ics, the coefficient on *Score* can be interpreted as a correlation between the degree of topic 0 ("World Crude Oil") news sentiment and the probability of price increase. The result is statistically significantly positive (1.2642), meaning that the more positive "World Crude Oil" news is, the higher probability that price will increase within 5 minutes. The coefficient on the interaction term $Topic1 * Score$ (0.6034), is also positive and statistically significant, which implies that the sentiment score of news under topic 1 ("WTI Crude Oil") is more correlated with the probability of price increase than news under topic 0 ("World Crude Oil"). Coefficients of the interaction terms $Topic2 * Score$ and $Topic3 * Score$ are negative but not statistically significant. Thus we reject the hypothesis that news sentiment under topic 2 and 3 is statistically different from those under topic 0.

To sum up, on average, "World Crude Oil" news has the highest correlation with a price increase. Nonetheless, the more positive news is under the topic "WTI Crude Oil," the higher probability that WTI crude oil futures price will increase within 5 minutes.

3.7 Conclusion

With the development of machine learning algorithm and its use in economics literature, it is worthwhile to apply machine learning algorithm to understand news impact on financial asset price movement. In this project, we use both supervised and unsupervised machine learning algorithms to learn impact of news sentiment and news topics on crude oil futures price increase. By assuming that crude oil futures market are nearly perfectly efficient (price

Table 3.2: Test for News Effect on Price Change
$Prob(Y = 1|X) = \sigma(\beta_0 + \beta_1 Score + \beta_2 Topic + \beta_3 Topic * Score)$

Dependent Variable	**Coef.**	
Score	1.2642	***
	(0.082)	
Topic1 "WTI Crude Oil"	-0.3108	***
	(0.073)	
Topic2 "Financial Analysis"	-0.2821	*
	(0.161)	
Topic3 "Editorial Opinion"	-0.2581	**
	(0.105)	
Topic1 * Score	0.6034	***
	(0.201)	
Topic2 * Score	-0.1910	
	(0.429)	
Topic3 * Score	-0.0868	
	(0.205)	
Const.	0.3267	***
	(0.044)	
No. Observations	4,616	
Pseudo R-squ.	0.071	
Robust Std. Err.	YES	

*Notes: The number in the parenthesis reports the t-statistics. *** indicates* $P < 0.01$. *** indicates* $P < 0.05$. ** indicates* $P < 0.1$.

adjusts quickly to any new information released to the public), we use high frequency data for estimating the news impact. The results show that "World Crude Oil" news on average is positively correlated with price increase, and the more positive "WTI Crude Oil" news is, the higher probability that crude oil futures price will increase within five minutes. The implication of our project is that using the coefficients in our last regression results, we are able to construct a news index, which can be used further in estimating the magnitude of price change, together with other macro and micro economic control variables.

There are a lot of work can be done in the future to improve our results. For example,

we will modify our filter of stop word in order to improve our sentiment results to be better consistent with human cognition. We can also try Neural Network algorithm to learn news topic, which as a supervised machine learning algorithm might provide more precise learning than the K-means algorithm that we used in the project.

Bibliography

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Allen, T. and Arkolakis, C. (2014). Trade and the Topography of the Spatial Economy. *The Quarterly Journal of Economics*, 129(3):1085–1140.

Arellano (2008). Default risk and income fluctuations in emerging economies. *American Economic Review*, pages 690–712.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies*, 21(2):785–818.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The Quarterly Journal of Economics*, 122(2):775–805.

Baum-Snow, N., Kahn, M. E., and Voith, R. (2005). Effects of urban rail transit expansions: Evidence from sixteen cities, 1970-2000 [with comment]. *Brookings-Wharton Papers on Urban Affairs*, pages 147–206.

Baum-Snow, N., Loren, B., Vernon, H. J., A, T. M., and QinghuanZhang (2017). Roads, railroads, and decentralization of chinese cities. *The Review of Economics and Statistics*, 99(3):435–448.

Baumeister, C. and Kilian, L. (2014). A general approach to recovering market expectations from futures prices with an application to crude oil.

Behrens, C. and Pels, E. (2012). Intermodal competition in the london-paris passenger market: High-speed rail and air transport. *Journal of Urban Economics*, 71(3):278–288.

Bernard, A. B., Moxnes, A., and Saito, Y. U. (2016). Production Networks, Geography and Firm Performance. CEP Discussion Papers dp1435, Centre for Economic Performance, LSE.

Bi, H. and Traum, N. (2019). Sovereign risk and fiscal information: A look at the us state default of the 1840s. *Federal Reserve Bank of Kansas City Working Paper*, (19-04).

Borensztein, E. and Panizza, U. (2008). The costs of sovereign default. *IMF Working Paper*.

Brandt, M. W. and Gao, L. (2019). Macro fundamentals or geopolitical events? a textual analysis of news events for crude oil. *Journal of Empirical Finance*, 51:64–94.

Byrne, J. P., Fazio, G., and Fiess, N. (2013). Primary commodity prices: Co-movements, common factors and fundamentals. *Journal of Development Economics*, 101:16–26.

Calvo, G. and Reinhart, C. (2000). Fear of floating. *NBER Working Paper Series*.

Cao, J., Liu, X., Wang, Y., and Li, Q. (2013). Accessibility impacts of chinas high-speed rail network. *Journal of Transport Geography*, 28:12–21.

Cashin, P., Liang, H., and McDermott, C. J. (2000). How persistent are shocks to world commodity prices? *IMF Staff Papers*, 47(2):177–217.

Cashin, P., McDermott, C. J., and Scott, A. (2002). Booms and slumps in world commodity prices. *Journal of development Economics*, 69(1):277–296.

Catalini, C., Fons-Rosen, C., and Gaul, P. (2016). Did cheaper flights change the direction of science? Technical Report 1520.

Cespedes, L. F., Chang, R., and Velasco, A. (2000). Balance sheets and exchange rate policy. Working Paper 7840, National Bureau of Economic Research.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260.

Checherita, C. and Rother (2012). The impact of high government debt on economic growth and its channels : An empirical investigation for the euro area. *European Economic Review*.

Cheng, Becky, L., and Roger, V. (2014). High speed rail networks economic integration and regional specialisation in china and europe. 2.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6):2543–2594.

David, E. and Jon, K. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA.

Deaton, A. and Laroque, G. (1992). On the behaviour of commodity prices. *The review of economic studies*, 59(1):1–23.

Demers, E., Vega, C., et al. (2008). Soft information in earnings announcements: News or noise?

Dias, D. A., Richmond, C. J., and Wright, M. L. J. (2011). The stock of external sovereign debt. *NBER Working Paper*, 17551.

Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Donaldson, D. and Hornbeck, R. (2016). Railroads and American Economic Growth: A Market Access Approach. *The Quarterly Journal of Economics*, 131(2):799–858.

Dong, X., Zheng, S., and E, K. M. (2018). The role of transportation speed in facilitating high skilled teamwork. Working Paper 24539, National Bureau of Economic Research.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In Henderson, J. V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4 of *Handbook of Regional and Urban Economics*, chapter 48, pages 2063–2117. Elsevier.

Duranton, G. and Turner, M. A. (2007). Urban growth and transportation. Working Papers tecipa-305, University of Toronto, Department of Economics.

Eaton, J. and Gersovitz, M. (1981). Debt with potential repudiation : Theoretical and empirical analysis. *Review of Economic Studies*, 48:289–309.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Eichengreen, B. and Hausmann, R. (2007). *Currency Mismatches , Debt Intolerance , and Original Sin Why They Are Not the Same and Why It Matters*. NBER Working Paper.

Elder, J., Miao, H., and Ramchander, S. (2013). Jumps in oil prices: the role of economic news. *The Energy Journal*, 34(3).

Faber, B. (2014). Trade integration. market size and industrialization: Evidence from china national trunk highway system. Technical Report dp1244.

Fuentes, M. and Saravia, D. (2010). Sovereign defaulters : Do international capital markets punish them ? *Journal ofDevelopment Economics*.

Fujita, M. and Krugman, P. (1995). When is the economy monocentric?: von thunen and chamberlin unified. *Regional Science and Urban Economics*, 25(4):505–528.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.

Gibbons, S. and Machin, S. (2005). Valuing rail access using transport innovations. *Journal of Urban Economics*, 57(1):148–169.

Giroud, X. and Mueller, H. M. (2015). Capital and labor reallocation within firms. *The Journal of Finance*, 70(4):1767–1804.

Greenidge, K., Drakes, L., and Craigwell, R. (2010). The external public debt in the caribbean community. *Journal of Policy Modeling*.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Grohe, S. and Uribe (2016). Downward nominal wage rigidity , currency pegs , and involuntary unemployment. *Journal of political economy*.

Grossman, G. M. and Rossi-Hansberg, E. (2008). Trading Tasks: A Simple Theory of Offshoring. *American Economic Review*, 98(5):1978–1997.

Hajivassiliou, V. (1897). External debt repayments problems of ldc's. *Journal of Econometrics*, pages 205–230.

Hamilton, J. D. (2009). Understanding crude oil prices. *The Energy Journal*, 30(2).

He, G., Xie, Y., and Zhang, B. (2017). Balancing Development and the Environment in a Changing World: Expressways, GDP, and Pollution in China. HKUST IEMS Working Paper Series 2017-43, HKUST Institute for Emerging Market Studies.

Helkie, W. and Howar, D. (1991). Board of governors of the federal reserve system. *International Finance Discussion Papers*.

Hercowitz, Z. (1986). On the determination of the external debt:the case of israel. *Journal of Monetary Economics*, 18:121–145.

Heuermann, F, D., Schmieder, and F, J. (2018). The effect of infrastructure on worker mobility: Evidence from high-speed rail expansion in germany. Working Paper 24507, National Bureau of Economic Research.

Jeanne, O. and Guscina, A. (2006). Government debt in emerging market countries: A new data set. *IMF Working Paper*.

JHausmann, R. and Rodrik, D. (2002). Economic development as self-discovery.

Ke, X., Chen, H., Hong, Y., and Cheng, H. (2017). Do china's high-speed-rail projects promote local economy?new evidence from a panel data approach. *China Economic Review*, 44(C):203–226.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3):1053–69.

Kilian, L. and Vega, C. (2011). Do energy prices respond to us macroeconomic news? a test of the hypothesis of predetermined energy prices. *Review of Economics and Statistics*, 93(2):660–671.

Kriwoluzky, A., Gernot, J. M., and Wolf, M. (2014). Exit expectations in currency unions.

Lemoy, R., Raux, C., and Jensen, P. (2012). Exploring the polycentric city with an agent-based model. working paper or preprint.

Levy, E. and Panizza, U. (2010). The elusive costs of sovereign defaults. *Journal of Development Economics*.

Lin, Y. (2017). Travel costs and urban specialization patterns: Evidence from chinas high speed railway system. *Journal of Urban Economics*, 98(C):98–123.

Lin, Y., Qin, Y., Sulaeman, J., Yan, J., and Zhang, J. (2019). Facilitating investment flows: Evidence from chinas high-speed passenger rail network.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Lucas, R. and Rossi-Hansberg, E. (2002). On the internal structure of cities. *Econometrica*, 70(4):1445–1476.

McDonald, J. and Osuji, C. I. (1995). The effect of anticipated transportation improvement on residential land values. *Regional Science and Urban Economics*, 25(3):261–278.

Michaels, G. (2006). The Effect of Trade on the Demand for Skill - Evidence from the Interstate Highway System. CEP Discussion Papers dp0772, Centre for Economic Performance, LSE.

Moretti, E. (2004). Workers' education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review*, 94(3):656–690.

Moretti, E. (2014). Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority. *The Quarterly Journal of Economics*, 129(1):275–331.

Na, S., Schmitt-grohe, S., and Uribe, M. (2015). *Center for Quantitative Economic Research WORKING PAPER SERIES A Model of the Twin Ds : Optimal Default and Devaluation*.

Ottaviano, G., Tabuchi, T., and Thisse, J.-F. (2002). Agglomeration and trade revisited. *International Economic Review*, 43(2):409–435.

Ouyang, A. Y. and Rajan, R. S. (2014). What determines external debt tipping points? *Journal of Macroeconomics*, 39(PA):215–225.

Paoli, B. D., Hoggarth, G., Saporta, V., Paoli, B. D., Hoggarth, G., and Saporta, V. (2009). Output costs of sovereign crises : some empirical estimates working paper no . 362 output costs of sovereign crises : some empirical estimates. *NBER Working Paper*, 362.

Parry, I. W. H. and Small, K. A. (2009). Should urban transit subsidies be reduced? *American Economic Review*, 99(3):700–724.

Qayyum, U., Din, M.-u., and Haider, A. (2014). Foreign aid , external debt and governance. *Economic Modelling*, 37:41–52.

Qin, Y. (2017). No county left behind? The distributional impact of high-speed rail upgrades in China. *Journal of Economic Geography*, 17(3):489–520.

Rapaport, A. (2010). Supply and demand shocks in the oil market and their predictive power. *Available at SSRN 2472379*.

Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a time of debt. *NBER Working Paper*.

Rose, A. K. (2005). One reason countries pay their debts : renegotiation and international trade. *Journal of Development Economics*, 77:189–206.

Rosenthal, S. and Strange, W. (2004). Evidence on the nature and sources of agglomeration economies. In Henderson, J. V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 49, pages 2119–2171. Elsevier, 1 edition.

Rus, G. and Nombela, G. (2007). Is investment in high speed rail socially profitable? *Journal of Transport Economics and Policy*, 41(1):3–23.

Shapiro, A. H., Sudhof, M., and Wilson, D. (2020). Measuring news sentiment. Federal Reserve Bank of San Francisco.

Shiller, R. J. (2015). *Irrational exuberance: Revised and expanded third edition*. Princeton university press.

Small, K. A., Winston, C., Yan, J., Baum-Snow, N., and Gmez-Ibez, J. A. (2006). Differentiated road pricing, express lanes, and carpools: Exploiting heterogeneous preferences in policy design [with comments]. *Brookings-Wharton Papers on Urban Affairs*, pages 53–96.

Soo, C. (2015). Quantifying animal spirits: news media and sentiment in the housing market. *Ross School of Business Paper*, (1200).

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Tsivanidis, J. N. (2018). The aggregate and distributional effects of urban transit infrastructure: Evidence from bogots transmilenio. *University of Chicago*.

Xu, M. (2018). Riding on the new silk road: Quantifying the welfare gains from high-speed railways.

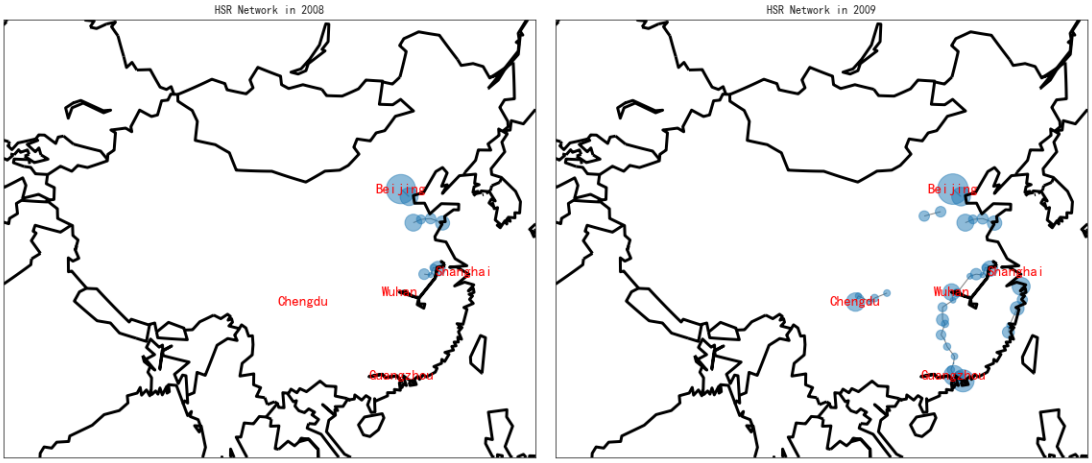
Zheng, S. and Kahn, M. E. (2013). Understanding china's urban pollution dynamics. *Journal of Economic Literature*, 51(3):731–72.

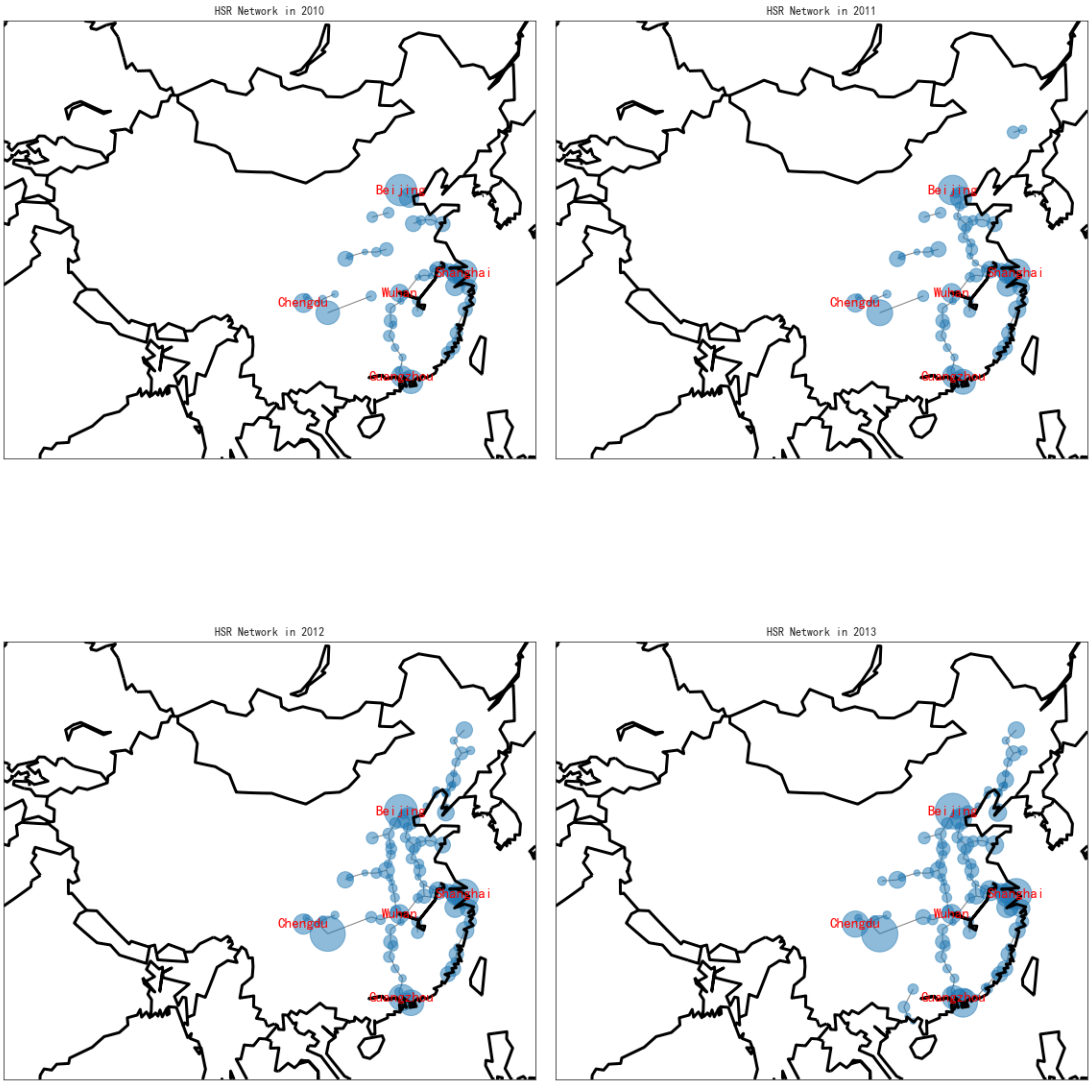
Appendix A

Chapter One

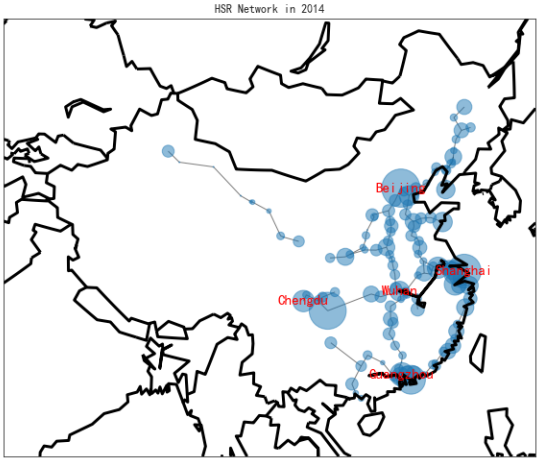
Appendix

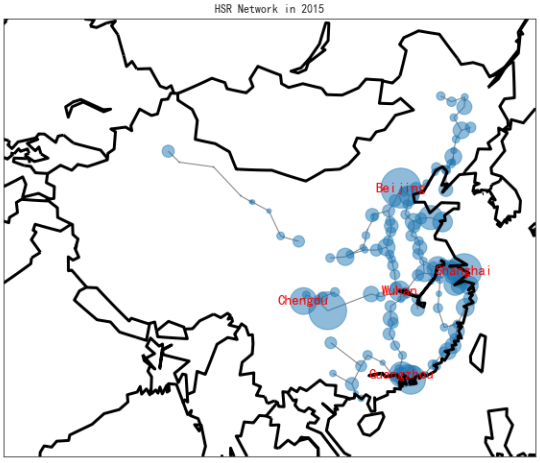
A.0.1 The Network of HSR in China

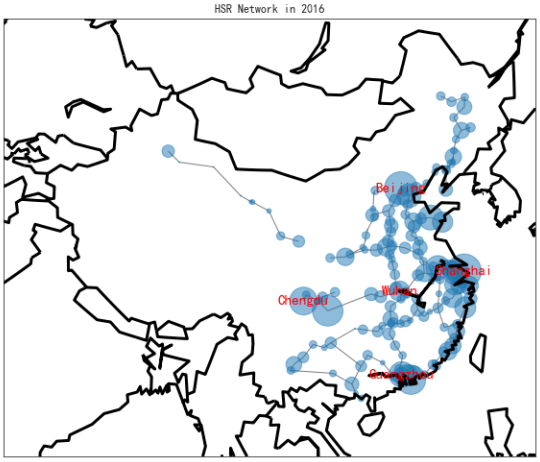


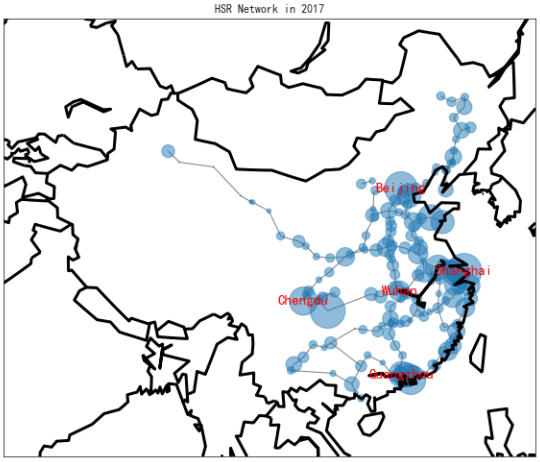


A.0.2 Industries Classification


HSR Network in 2014
Beijing
Chengdu
Wuhan


HSR Network in 2015
Beijing
Chengdu
Wuhan


HSR Network in 2016
Beijing
Chengdu
Wuhan


HSR Network in 2017
Beijing
Chengdu
Wuhan

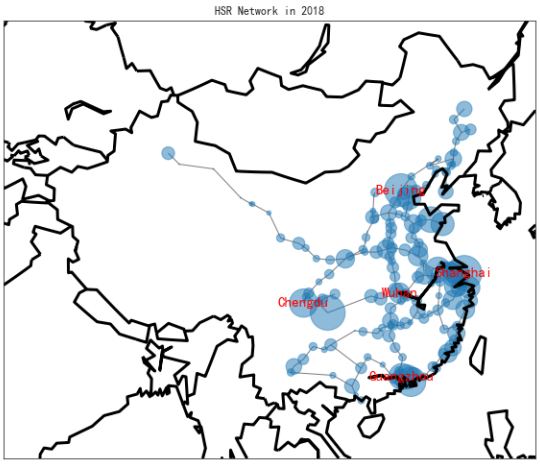


Table A.1: 19 Industries in China

Table ID	Chinese Industries	US industries	NAICS
Skilled Employment:			
ind7	Information transmission, computer service and software	Information	51
ind10	Finance and insurance	Finance and insurance	52
ind11	Real estate	Real estate and rental and leasing	53
ind12	Leasing and Business service	Professional,scientific and technical services	54
ind13	Scientific research,technical service	Professional,scientific and technical services	54
ind18	Culture,sports and entertainment	Arts,entertainment and recreation	71
Tourism-related employment:			
ind9	Hotels and catering service	Accommodation and food service	72
Other service employment:			
ind6	Transportation,warehousing and post	Transportation,warehousing	48
ind8	Wholesale and retail trade	wholesale and retail trade	42,43
ind14	Management of water conservancy,environment and public facilities		
ind15	Household services, repair and other service		
ind16	Education	Educational Service	61
ind17	Health, social work	Health care and social assistance	62
ind19	Public Management and Social 0rganization		
Other non-service employment:			
ind1	Agriculture, forestry,animal production and hunting,fishing	Forestry,fishing,hunting and agriculture support	11
ind2	Mining and quarrying	Mining	21
ind3	Manufacturing	Manufacturing	31
ind4	Production and Distribution of Electricity, Gas and Water	Utilities	22
ind5	Construction	Construction	23

Appendix B

Chapter Two

Appendix

B.0.1 Correlation Table

Table B.1: Correlations of Independent Variables(Log Values)

	Regimes	Default	Y_gap_t1	G_gap_t1	Δtot	Real IR	Ndep	Exports(Log)	FDI(Log)	Rdep	Reserves	GDP(Log)	Balance(log)
Regimes	1												
Default	-0.0054	1											
Y_gap_t1	-0.1638	0.0164	1										
G_gap_t1	-0.0849	0.0006	0.094	1									
Δtot	0.0333	0.038	-0.0261	-0.0537	1								
Real IR	0.1078	0.1188	0.0353	0.1936	-0.0362	1							
Ndep	-0.1828	-0.2416	0.0189	-0.0898	-0.0335	-0.2532	1						
Exports(Log)	0.3012	-0.0455	-0.2082	0.015	-0.0008	0.0379	-0.0218	1					
FDI(Log)	0.1902	-0.0943	-0.0730	0.0287	0.0164	-0.0012	-0.0186	0.2458	1.0000				
Rdep	-0.0521	-0.0951	-0.0854	-0.1818	-0.0534	-0.1631	0.8113	0.0578	0.0975	1			
Reserve/Import	0.1708	-0.1431	-0.1737	-0.0679	0.0266	-0.1184	0.0339	0.2774	0.1783	0.044	1		
GDP(Log)	0.3215	0.0011	-0.171	-0.0077	-0.0239	0.0405	-0.119	0.8113	0.2168	0.0007	0.2953	1	
Balance(log)	0.2079	-0.0276	-0.0147	-0.0787	-0.0547	0.0029	-0.0899	0.1020	0.1996	-0.0124	0.1087	0.2655	1

Table B.2: Correlations of Independent Variables(To GDP Ratios)

	Regimes	Default	Y_gap_t1	G_gap_t1	Δtot	Real IR	Ndep	Exports(ratio)	FDI(Ratio)	Rdep	Reserves	Balance/GDP
Regimes	1											
Default	0.0027	1										
Y_gap_t1	-0.1647	0.009	1									
G_gap_t1	-0.0861	-0.005	0.1043	1								
Δtot	0.0171	0.0374	-0.0196	-0.0484	1							
Real IR	0.126	0.1433	0.0335	0.1916	-0.0398	1						
Ndep	-0.1868	-0.2499	0.0191	-0.1	-0.04	-0.2502	1					
Exports/GDP	-0.1601	0.0002	0.115	-0.071	0.0918	-0.0067	0.1061	1				
FDI/GDP	-0.0025	-0.1199	0.0003	0.0003	0.0387	-0.0407	0.1304	0.3523	1			
Rdep	-0.0366	-0.112	-0.0841	-0.1904	-0.0632	-0.1617	0.55	0.0284	0.0866	1		
Reserve/Import	0.1871	-0.1557	-0.1756	-0.0673	0.03	-0.1232	0.0269	-0.1704	0.0639	0.047	1	
Balance/GDP	-0.0574	0.0292	0.1954	-0.0676	-0.0023	-0.0219	-0.0745	0.0314	-0.0477	-0.0596	-0.0665	1

B.0.2 Countries in Dataset

Table B.3: 57 Countries Defaulted from 1970 to 2007

Albania	Dominican Republic	Jamaica	Poland
Algeria	Ecuador	Jordan	Romania
Argentina	Egypt	Liberia	Russia
Benin	El Salvador	Madagascar	Senegal
Bolivia	Equat Guinea	Mali	South Africa
Brazil	Gabon	Mauritania	Tanzania
Burkina Faso	Gambia,	Mexico	Togo
Cameroon	Ghana	Morocco	Turkey
CAR	Guatemala	Nicaragua	Uganda
Chad	Guinea	Niger	Uruguay
Chile	Guinea-Bissau	Nigeria	Venezuela
Colombia	Guyana	Pakistan	Zambia
Congo, Rep.	Honduras	Panama	
Costa Rica	India	Peru	
Cote D'Ivoire	Indonesia	Philippines	

B.0.3 ER Regimes Classification

Table B.4: Countries Classified by Regions

Europe & Central Asia(5)	Albania	Poland	Romania
	Russia	Turkey	
Middle East & North Africa(4)	Morocco	Jordan	Egypt
		Algeria	
Latin America & Caribbean(19)	Argentina	Bolivia	Brazil
	Chile	Colombia	Costa Rica
	Dominican Republic	Ecuador	El Salvador
	Guatemala	Guyana	Honduras
	Jamaica	Mexico	Nicaragua
	Panama	Peru	Uruguay
	Venezuela		
Sub-Saharan Africa(25)	Benin	Burkina Faso	CAR
	Cameroon	Chad	Congo, Rep.
	Cote D'Ivoire	Equat Guinea	Gabon
	Gambia, The	Ghana	Guinea
	Guinea-Bissau	Liberia	Madagascar
	Mali	Mauritania	Niger
	Nigeria	Senegal	South Africa
	Tanzania	Togo	Uganda
	Zambia		
South Asia(2)	India	Pakistan	
East Asia & Pacific(2)	Indonesia	Philippines	

Table B.5: Countries Classified by Income Groups

Low Income(16):	Benin	Burkina Faso	Cameroon	
GNI per capita less	Chad	Equat Guinea	Gambia, The	
than $1,025 or less	Guinea	Guinea-Bissau	Liberia	
in 2015	Madagascar	Mali	Niger	
	Senegal	Tanzania	Togo	
	Uganda			
Lower middle income(19):	Bolivia	CAR	Congo, Rep.	Cote D'Ivoire
GNI per capita	Egypt	El Salvador	Ghana	
between $1,026 and $4,035	Guatemala	Honduras	India	
	Indonesia	Jordan	Mauritania	
	Morocco	Nicaragua	Nigeria	
	Pakistan	Philippines	Zambia	
Upper middle income(19):	Albania	Algeria	Argentina	Brazil
GNI per capita	Colombia	Costa Rica	Dominican Republic	
between $4,036 and $12,475	Ecuador	Gabon	Guyana	
	Jamaica	Mexico	Panama	
	Peru	Romania	Russia	
	South Africa	Turkey	Venezuela	
High Income(2):	Chile	Poland	Uruguay	
GNI per capita above $12,476				

Table B.6: The IMF's Classifications of ER Regimes

Fine	Coarse	Decription:
1	1	No separate legal tender
2	1	Pre announced peg or currency board arrangement
3	1	Pre announced horizontal band that is narrower than or equal to +/-2%
4	1	De facto peg
5	2	Pre announced crawling peg
6	2	Pre announced crawling band that is narrower than or equal to +/-2%
7	2	De factor crawling peg
8	2	De facto crawling band that is narrower than or equal to +/-2%
9	3	Pre announced crawling band that is wider than or equal to +/-2%
10	3	De facto crawling band that is narrower than or equal to +/-5%
11	3	Moving band that is narrower than or equal to +/-2% (i.e., allows for both appreciation and depreciation over time)
12	3	Managed floating
13	4	Freely floating
14	5	Freely falling
15	6	Dual market in which parallel market data is missing.

Source:IMF Website
Consider float regimes when greater or equal to 12 in Fine, greater or equal to 4 in Coarse.
I only use Coarse classification in this study.