Data and Methods Section

Yilun Dai

1. Data used

This research uses data from national and international databases as well as digitally-obtained big data. There are four sets of annual data: Chinese TV and film data, China's total fertility rate data, the mean age of first marriage data, and the number of students studying abroad, all of are from 1995 to 2016.

I used Chinese TV and film data, specifically, the number of films and TV series produced in Mainland China with elements of "early love" (referred to as the number of films and TV series below), as an indicator of public tolerance of "early love". I chose film and TV data because China has a strict censorship on contents of film and TV: according to State Administration of Press, Publication, Radio, Film and Television of the People's Republic of China, the works that are in theatres or on TV should have content consistent with mainstream social norms and values, and the Administration will delete any content it considers inappropriate or ban the film or the TV series with such content. Therefore, the number of films and TV series related to "early love" could reflect the society's attitude towards "early love". The data comes from Douban, a Chinese films and TV database that is similar to Rotten Tomatoes. Douban has over 20,000 films and over 15, 000 TV series on record (data from Zhihu), and is constantly updated with latest films, comments and ratings. Therefore, it possesses two important characteristics of Big Data described by Salganik: big enough and "always on" (i.e., is constantly updated).

An alternative to Douban's data is the first love age data from China Marriage and Love Relationship Survey, a nation-wide digital survey conducted by Peking University and baihe.com (a Chinese match-making website) in 2015. According to the survey, the mean age of first love is

decreasing by generation. Those who were born in 70s and 80s have a mean age of first love that is greater than 18 years old, while there is a sharp decrease in the number when it comes to those who were born in the 90s (15.18). There is a further decrease in mean age of first love for those who were born during or after 1995 (12.67). While this part of survey data reveals the decreasing trend of first love age, it is generation wise rather than annual data, and therefore not sufficient for the analysis of this survey.

I will use China's annual Total Fertility Rate (TFR) data from the World Bank in this research. China's Total Fertility Rate had been decreasing sharply from the mid-1980s (2.675 in 1986) to late 1990s (1.494 in 1999) due to the One Child Policy. As shown in Figure 2, Total Fertility Rate witnessed a slight increase in the 21st century, and has been around 1.6 since the late 2000s.

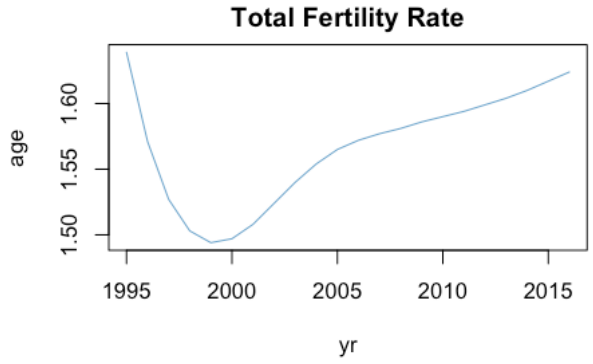

Figure 1: China's TFR from 1995 to 2016

This research uses multiple data sets from the National Bureau of Statistics of the PRC (referred to as the Bureau of Statistics in this paper). The mean age of first marriage from late 1980s to 2010 comes from the sixth China Population Census. While factual marriage (having held wedding and cohabiting without obtaining certificate of marriage) has no legal status since 1994 according to the Marriage Law of China, the sixth Census took factual marriage into account. The legal marriage age in China is 20 for female and 22 for male. However, the census has taken into

account those who get married before 20, since it is likely that at least the majority of this population get married before 1994. The mean age of first marriage data from 2010 to 2016 comes from the annual report of the Bureau of Statistics. Since only female first marriage age is available for this part of data, and male's mean first marriage age and female's mean first marriage age follow the same trend (Trading Economics), this research will use the mean age of first marriage of female. Figure 3 demonstrates that female mean age of first marriage is postponing from 22.85 in 1995 to 26.00 to 2016.

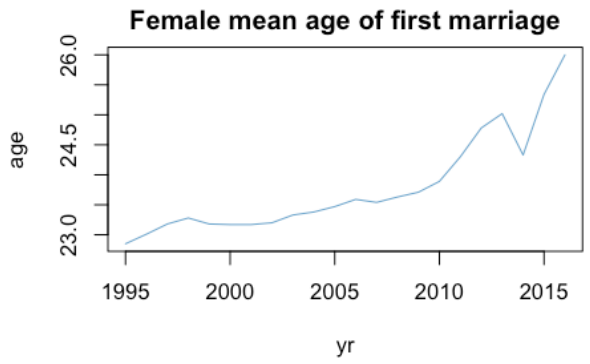

Figure 2: China's female mean age of first marriage from 1995 to 2016

The annual number of students studying abroad also comes from the Bureau of Statistics. Because College Entrance Exam in China is often seen as "sealing the deal" for the life track of teenagers, teenagers are told by their parents and teachers to spare no effort studying and not to get distracted by anything irrelevant to studying for exams. "Early love" is seen as a distraction that will result in poor performance in exams, especially failure in College Entrance Exam. What if College Entrance Exam is no longer the one and only way to high-quality higher education? Will "Early Love" still be regarded as a distraction that impedes teenagers' entrance into undergraduate institutions? Therefore, I will use the annual data of number of students studying abroad. We could see from Figure 4 that the number of students studying abroad fluctuate around

20,000 before 21st century. This number increases to 100,000 at the beginning of 21st century, but it does not begin to increase sharply until 2007.

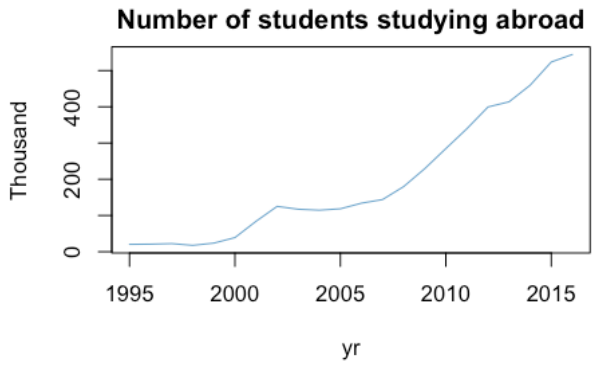

Figure 3: Number of Chinese students studying abroad from 1995 to 2016

2. Data Collection

The World Bank and the Bureau of Statistics both provide public access to datasets; I need to conduct web crawling to obtain films and TV data from Douban. Douban has categorized films and TV series by tags, including country of production, genre and theme. Films and TV series with all of the three tags "Mainland China", "Romance" and "School or Adolescence" are crawled. I used Python's web crawling tool, Beautiful Soup, to collect each film or TV series' name and year of production. I then used python to clean the data and obtain the number of films and TV series on teenage romance that are produced in Mainland China each year. Figure 4 is a histogram of number of films and TV series by year.

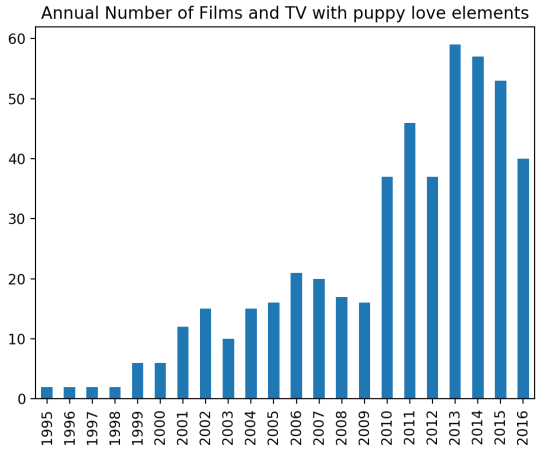

Figure 4: Number of Films and TV series on puppy love produced in China by year

3. Methods used and Model Selection

In this research, I will construct time series for the four data sets and analyze possible causal relationships between variables in R Studio. To test the likely relationships between the number of films and TV series and another variable, I will construct a 2-variable vector autoregression model (VAR) between the number of films and TV series and TFR, the number of films and TV series and female's mean age of marriage, and the number of films and TV series and the number of students studying abroad respectively. A two-variable VAR model of order p is a two-variable autoregression in which two equations are estimated; in each equation, the variable on the left-hand side is regressed on p-lags of itself and p-lags of the other variable. For each of the three VAR models I construct, in the first equation, I will regress the other variable on p lags of itself and p lags of the number of films and TV series. In the second equation, I will regress the number of films and TV series on p lags of itself and p lags of the other variable. According to Diebold, one key advantage of a VAR model is that it allows for cross-variable dynamics (pp. 499).

After finishing evaluating VAR models, I will conduct Granger causality test for each of the models. The theory of Granger causality was developed in the 1960s by Clive Granger, and it states that if a variable Y1 "Granger causes" a variable Y2, then the past values of Y1 should improve the prediction of Y2 compared to using the past values of Y2 alone. Granger causality test will reveal the direction of the causal relationship should there exists any.

I will use R's VARselect() function to select the optimal value of p for each model. Taken into account the disadvantage of a more complex model (according to the parsimony principle), I selected the value of lags based on SC. I constructed a VAR (2) model for the number of films and TV series and TFR (equation (1) and (2)), a VAR(3) model for the number of films and TV series and the female mean first marriage age (equation (3) and (4)), and a VAR(5) model for the number of films and TV series and the number of students studying abroad (equation (5) and (6)).

VAR (2) model:

$$TFR_t = 1.524 * TFR_{t-1} - 0.639 * TFR_{t-2} + 1.052 * 10^{-4} * NumFilmTV_{t-1} + 2.404 * 10^{-4} * NumFilmTV_{t-2} + \varepsilon_t \quad (1)$$

$$NumFilmTV_t = 194.014 * TFR_{t-1} - 97.684 * TFR_{t-2} + 0.606 * NumFilmTV_{t-1} + 0.08967 * NumFilmTV_{t-2} + \varepsilon_t \quad (2)$$

VAR (3) model:

$$age_t = 0.743 * age_{t-1} - 0.675 * age_{t-2} + 0.701 * age_{t-3} + 0.002 * NumFilmTV_{t-1} + 0.034 * NumFilmTV_{t-2} - 0.022 * NumFilmTV_{t-3} + \varepsilon_t \quad (3)$$

$$NumFilmTV_t = 10.377 * age_{t-1} + 21.713 * age_{t-2} - 18.055 * age_{t-3} + 0.328 * NumFilmTV_{t-1} + 0.129 * NumFilmTV_{t-2} - 0.030 * NumFilmTV_{t-3} + \varepsilon_t \quad (4)$$

VAR (5) model:

$$students_t = 1.584 * students_{t-1} - 0.647 * students_{t-2} + 0.477 * students_{t-3} - 0.437 * students_{t-4} + 0.702 * students_{t-5} - 0.226 * NumFilmTV_{t-1} -$$

$$1.487 * NumFilmTV_{t-2} - 1.962 * NumFilmTV_{t-3} - 1.370 * NumFilmTV_{t-4} - 2.037 * NumFilmTV_{t-5} + \varepsilon_t \quad (5)$$

$$NumFilmTV_t = 0.308 * students_{t-1} + 0.009 * students_{t-2} + 0.363 * students_{t-3} - 0.084 * students_{t-4} + 0.375 * students_{t-5} - 1.225 * NumFilmTV_{t-1} - 1.408 * NumFilmTV_{t-2} - 1.696 * NumFilmTV_{t-3} - 0.436 * NumFilmTV_{t-4} - 2.256 * NumFilmTV_{t-5} + \varepsilon_t$$

4. Results

After evaluating the three models, I performed the Granger causality test for each model. Table 1 shows the p-value and the direction of causality of each test. The first row shows the p-value for the first equation in each of the three models (equation (1), (3), and (5)), and the second row shows the second equation in each of the three models (equation (2), (4), and (6)). The smaller p-value of the two indicates the direction of the Granger causality, i.e., past values of the number of films and TV series helps predict the value of TFR at a significan level of less than 0.005; past values of the number of films and TV series helps predict the mean age of first marriage at a significant level of 0.005 to 0.01; past values of the number of students studying abroad helps predict the number of films and TV series relevant to "early love" at a significant level of 0.01 to 0.05.

Equation	*Exogenous Variables*		
	TFR	*Mean Age*	*Number of Students*
Variable ~ lags(NumFilmTV)	$3.623*10^{-5}$ ***	0.003387 **	0.4076
NumFilmTV ~ lags(variable)	0.4968	0.1927	0.02798 *

Table 1: Note: * indicates a causality at a significance level of 0.01 ~ 0.05; ** indicates a causality at a significance level of 0.005 ~ 0.01; *** indicates a causality at a significance level of less than 0.005.

Reference:

Diebold, F.X. (2015), Forecasting, Department of Economics, University of Pennsyl- vania,

http://www.ssc.upenn.edu/~fdiebold/Textbooks.html

Salganik, Matthew J. 2017. Bit by Bit: Social Research in the Digital Age. Princeton, NJ:

Princeton University Press. Open review edition.