Homework#2

Delores Tang

10/14/2018

1. Imputing age and gender

(a) In order to impute age and gender from SurvIncome to BestIncome, I would like to fit a generalized linear model to variables in SurvIncome.

```
# Linear Regression in SurvIncome
Agelm <- lm(Age ~ Weight + T_inc, data = SurvInc)
summary(Agelm)
```

```
##
## Call:
## lm(formula = Age ~ Weight + T_inc, data = SurvInc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9129  -3.7610   0.0717   4.0397  21.9223
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.421e+01  1.490e+00  29.666   <2e-16 ***
## Weight      -6.722e-03  9.803e-03  -0.686    0.493
## T_inc        2.520e-05  2.263e-05   1.114    0.266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.941 on 997 degrees of freedom
## Multiple R-squared:  0.001267,   Adjusted R-squared:  -0.0007361
## F-statistic: 0.6326 on 2 and 997 DF,  p-value: 0.5314
```

```
Genderlm <- lm(Gender ~ Weight + T_inc, data = SurvInc)
summary(Genderlm)
```

```
##
## Call:
## lm(formula = Gender ~ Weight + T_inc, data = SurvInc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70371 -0.13714 -0.00253  0.13815  0.59659
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.761e+00  5.110e-02  73.600  < 2e-16 ***
## Weight      -1.953e-02  3.362e-04 -58.098  < 2e-16 ***
## T_inc       -5.250e-06  7.760e-07  -6.765 2.28e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2037 on 997 degrees of freedom
```

```
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8341
## F-statistic:  2513 on 2 and 997 DF,  p-value: < 2.2e-16
```

Therefore, the equations that could be used to impute Age and Gender to BestIncome will be:

$$\text{Age} = 44.21 - 6.722 \times 10^{-3} \times \text{Weight} + 2.52 \times 10^{-5} \times \text{Total Income}$$

and

$$\text{Gender} = 3.761 - 1.953 \times 10^{-2} \times \text{Weight} - 5.25 \times 10^{-6} \times \text{Total Income}.$$

(b) Using the equations obtained from question (a), I imputed Age and Gender to BestIncome based on the SurveyIncome dataset. Since gender has to be binay, I set all imputed gende that is greater than 0.5 as 1, and others as 0. Also, I assumed that the sum of labor and capital income in BestIncome data is equivalent to the variable Total Income in SurveyIncome data.

```
BestInc$Age <- 44.21 - 0.006722 * BestInc$Weight + 2.52e-05 * (BestInc$L_inc +
    BestInc$C_inc)
BestInc$Gender <- 3.761 - 0.01953 * BestInc$Weight - 5.25e-06 * (BestInc$L_inc +
    BestInc$C_inc)

# Since Gender has to be binary
for (i in 1:length(BestInc$Gender)) {
    if (BestInc$Gender[i] < 0.5) {
        BestInc$Gender[i] <- 0
    } else {
        BestInc$Gender[i] <- 1
    }
}
```

(c) For age,
mean: 44.89, sd = 0.219, min = 43.98, max = 45.70, no. of observation = 10000;
For Gender,
mean: 0.4614 sd = 0.499, min = 0, max = 1, no, of observation = 10000.

```
summary(BestInc$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.98   44.74   44.89   44.89   45.04   45.70
```

```
summary(BestInc$Gender)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4614  1.0000  1.0000
```

```
sd(BestInc$Age)
```

```
## [1] 0.2191322
```

```
sd(BestInc$Gender)
```

```
## [1] 0.4985327
```

```
length(BestInc$Age)
```

```
## [1] 10000
```

```
length(BestInc$Gender)
```

```
## [1] 10000
```

(d) Correlation matrix:

```r
res <- cor(BestInc)
round(res, 2)
```

```
##        L_inc C_inc Height Weight   Age Gender
## L_inc   1.00  0.01   0.00   0.00  0.92  -0.17
## C_inc   0.01  1.00   0.02   0.01  0.23  -0.05
## Height  0.00  0.02   1.00   0.17 -0.05  -0.13
## Weight  0.00  0.01   0.17   1.00 -0.30  -0.78
## Age     0.92  0.23  -0.05  -0.30  1.00   0.07
## Gender -0.17 -0.05  -0.13  -0.78  0.07   1.00
```

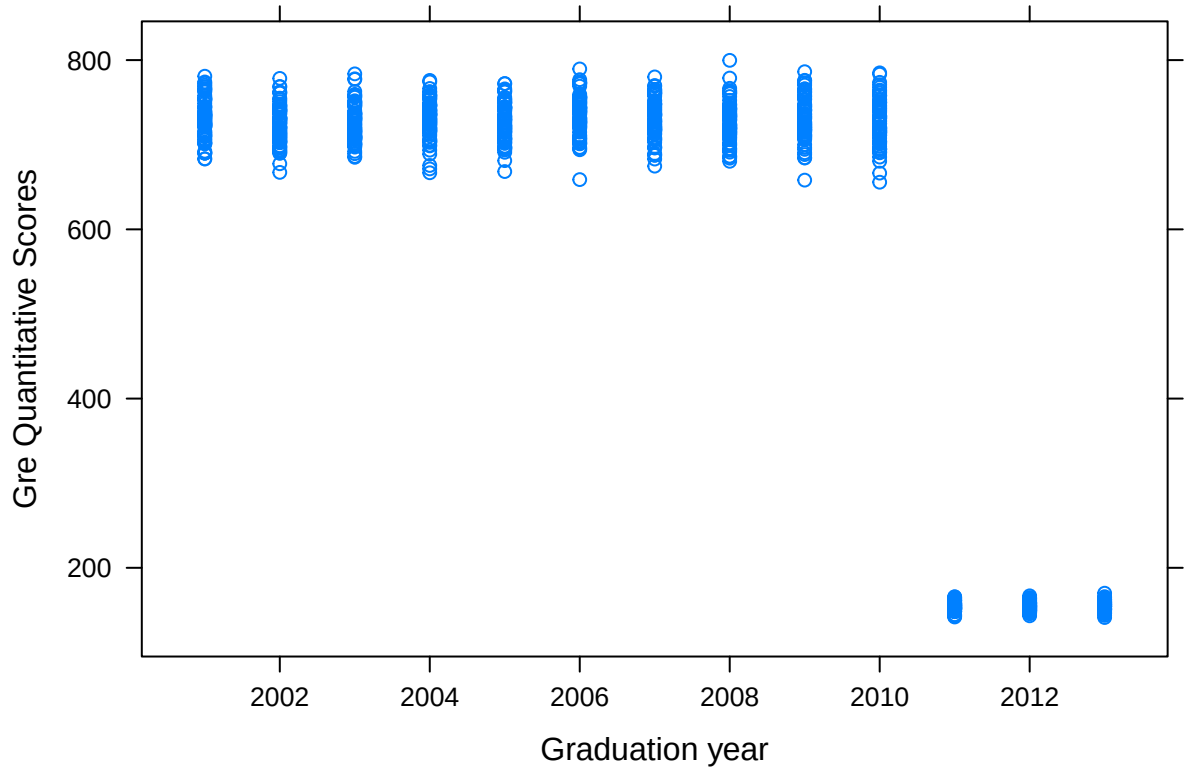
2.(a) Coefficients for intercept: 89541.293, Std.Error: 878.764
Coefficients for gre score: -25.763, Std.Error: 1.36

```r
salmod1 <- lm(salary ~ gre, data = IncIntel)
summary(salmod1)
```

```
##
## Call:
## lm(formula = salary ~ gre, data = IncIntel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28761  -7049   -293   6549  37666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89541.293    878.764  101.89   <2e-16 ***
## gre           -25.763      1.365  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10460 on 998 degrees of freedom
## Multiple R-squared:  0.2631, Adjusted R-squared:  0.2623
## F-statistic: 356.3 on 1 and 998 DF,  p-value: < 2.2e-16
```

(b) Due to the change in GRE's scoring system, a system drift on the data occurred at the year 2010-2011. Therefore, as indicated by the scatterplot, people's GRE scores dropped significantly due to this change.

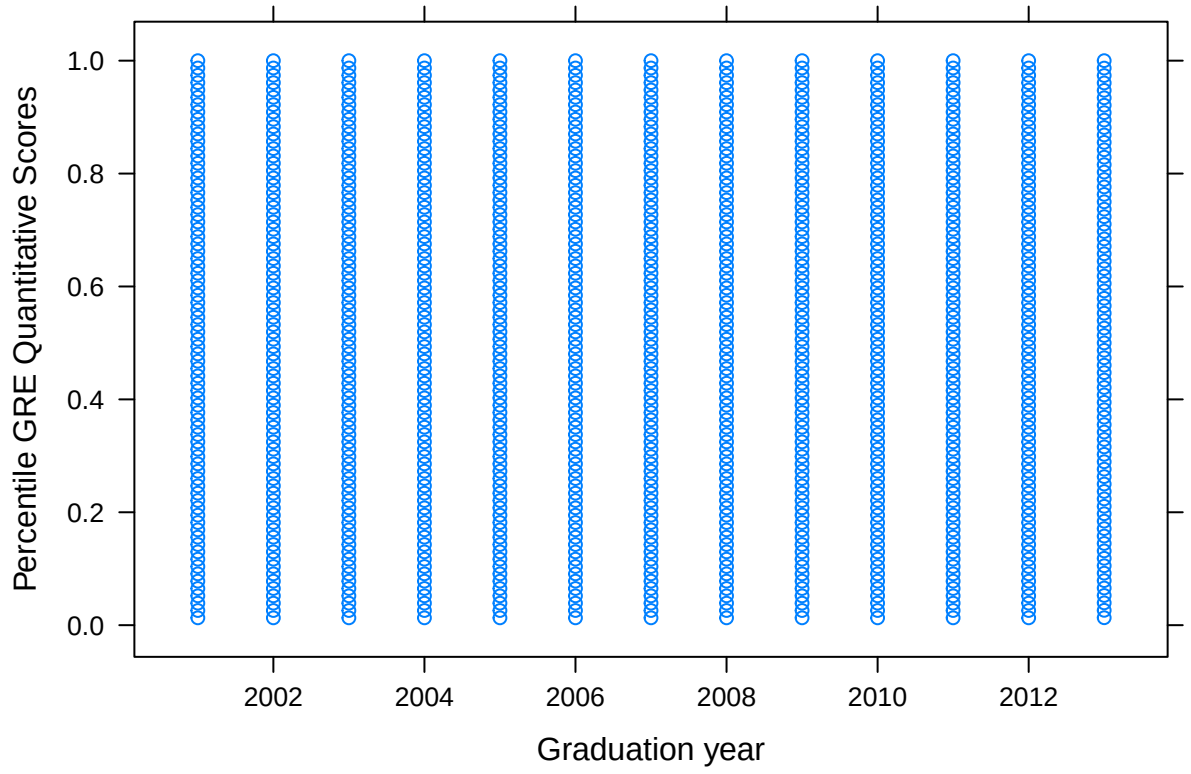
```r
# Scatter plot of Gre scores vs. Graduation year
xyplot(gre ~ year, xlab = "Graduation year", ylab = "Gre Quantitative Scores",
    data = IncIntel)
```



To accurately test our hypothesis, we would have to rescale GRE scores to eliminate the effect of the data drift by using the percentile a person gets in the year he takes GRE test, instead of the general GRE quantitative score, to indicate his or her academic performance on a GRE test.

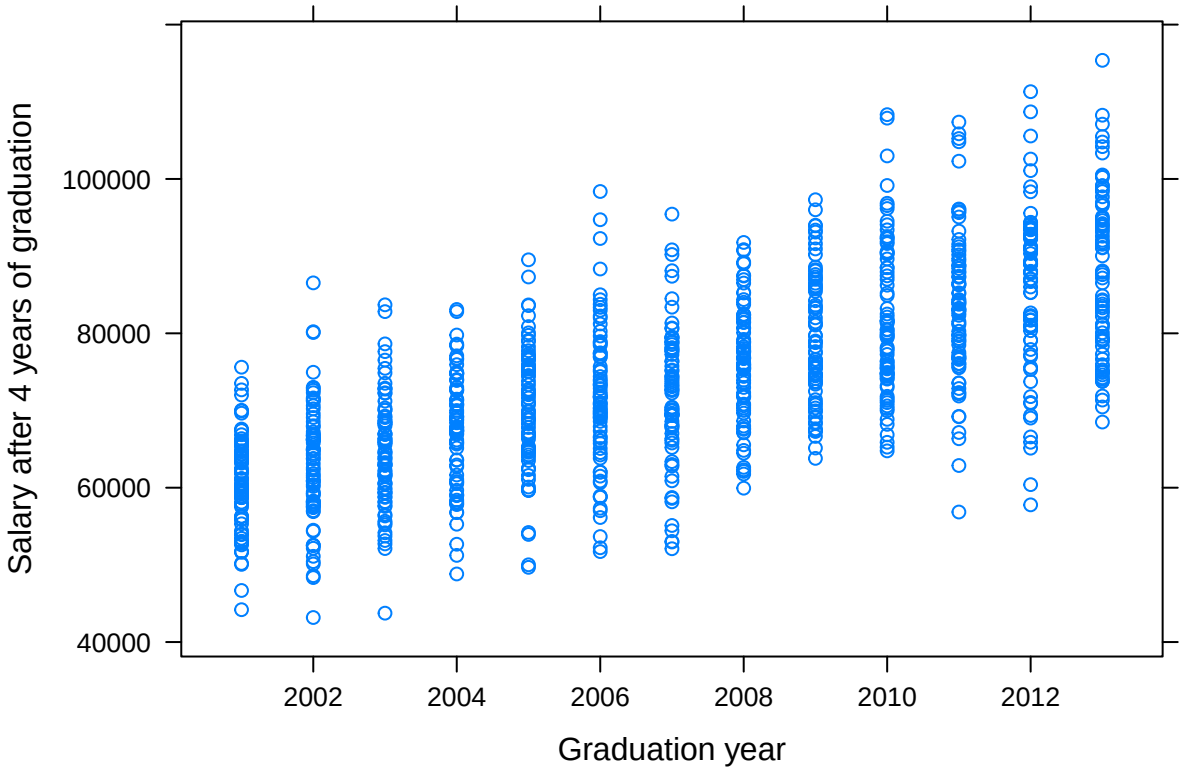
```
# Define a function that extract a list for each year's gre scores
grefunc <- function(yr) {
    grelist <- list()
    for (i in (1:nrow(IncIntel))) {
        if (IncIntel$year[i] == yr) {
            grelist <- append(grelist, IncIntel$gre[i])
        }
    }
    return(grelist)
}

# Make a dictionary that sort each year's participants' gre scores
gredic <- list()
for (yr in (2001:2013)) {
    grelist <- grefunc(yr)
    gredic <- append(gredic, list(grelist))

}

# Create a new variable gre_perc that indicates the participant's percentile
# GRE in that year
for (yr in (2001:2013)) {
```

```
    for (i in (1:nrow(IncIntel))) {
        if (IncIntel$year[i] == yr) {
            dicyr <- IncIntel$year[i] - 2000
            gredic_yr <- unlist(gredic[dicyr])
            gredic_yr_rank <- ecdf(gredic_yr)
            IncIntel$gre_perc[i] <- gredic_yr_rank(IncIntel$gre[i])
        }
    }
}

xyplot(IncIntel$gre_perc ~ IncIntel$year, xlab = "Graduation year", ylab = "Percentile GRE Quantitative
```



(c)

```
# Scatterplot of the original salary vs. year
xyplot(salary ~ year, xlab = "Graduation year", ylab = "Salary after 4 years of graduation",
    data = IncIntel)
```



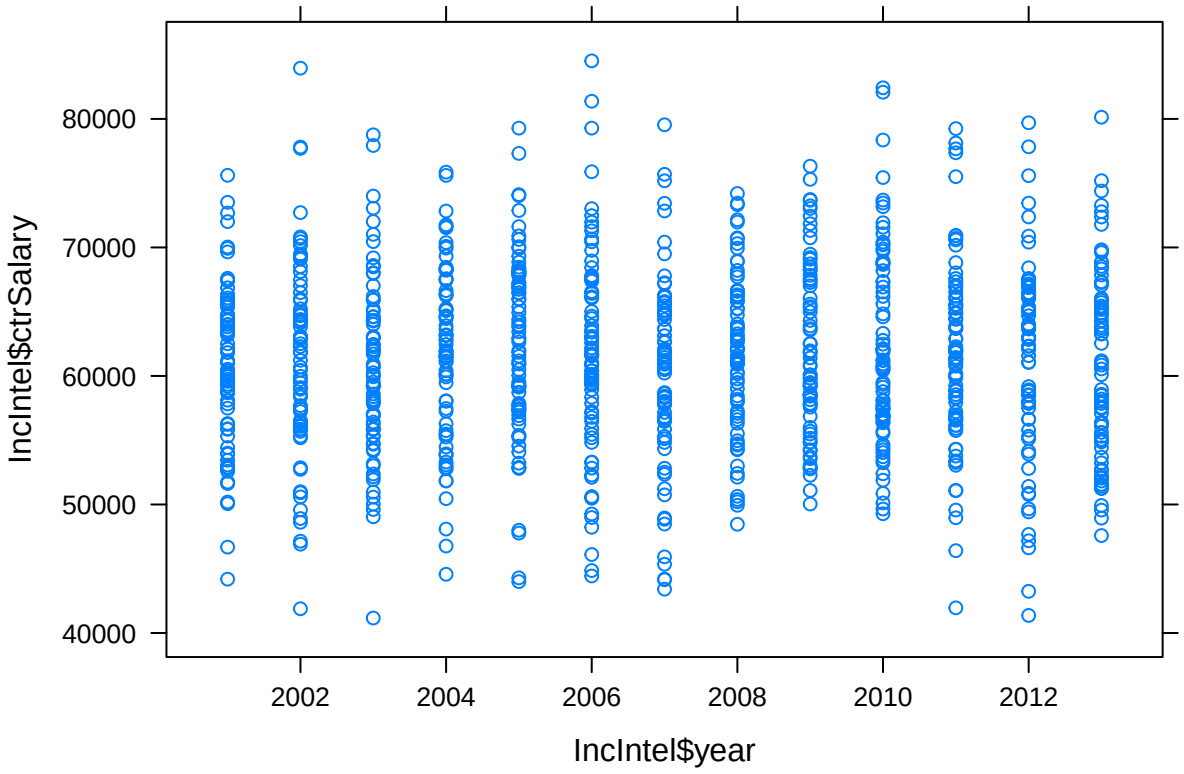
As indicated by the scatter plot, salary inflates gradually over the years. In that case, years is a stationarity that could confound our hypothesis testing. Therefore, we could calculate the average annual growth rate of income, and balance out this annual rate of salary growth by dividing a person's salary by the growth rate to the power of time (in years).

```
# Control for stationarity Average income per year
avg_inc <- aggregate(select(IncIntel, salary), list(IncIntel$year), mean)
avg_inc
```

```
##    Group.1   salary
## 1     2001 60710.71
## 2     2002 63034.40
## 3     2003 64518.74
## 4     2004 67773.49
## 5     2005 70492.59
## 6     2006 71678.24
## 7     2007 72133.65
## 8     2008 76432.58
## 9     2009 79030.63
## 10    2010 81741.30
## 11    2011 83563.85
## 12    2012 86012.59
## 13    2013 87300.52
```

```
## Average growth rate
avg_growth <- mean((avg_inc$salary[13:2] - avg_inc$salary[12:1])/avg_inc$salary[12:1])
IncIntel$ctrSalary <- IncIntel$salary/((1 + avg_growth)^(IncIntel$year - 2001))
```

```
xyplot(IncIntel$ctrSalary ~ IncIntel$year)
```



(d) Estimated coefficients:
Intercept: 61643.2, Std.Error: 455.9;
GRE percentiles: -411, Std.Error: 782.0.

Now the coefficient for GRE percentile, unlike that in (a), turned to be insignificant ($p > 0.05$).

Due to the change we made in (b), the estimated values of both coefficients varied signifiantly as we adjusted GRE variable to a new GRE percentile variable that ranges from 0 to 1.

Similarly, after controlling for system drift in GRE scores and the stationarity of time, the results of linear regression model suggests that GRE quantitative score is not a significant predicting factor of one's salary after 4 years of graduation. It suggests that the alternative hypothesis is likely to be rejected.

```
adjustLm <- lm(ctrSalary ~ gre_perc, data = IncIntel)
summary(adjustLm)
```

```
##
## Call:
## lm(formula = ctrSalary ~ gre_perc, data = IncIntel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20174.9  -4772.1    105.6   4791.6  23171.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  61643.2      455.9 135.216   <2e-16 ***
## gre_perc      -441.0      782.0  -0.564    0.573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7138 on 998 degrees of freedom
## Multiple R-squared:  0.0003185,  Adjusted R-squared:  -0.0006831
## F-statistic: 0.318 on 1 and 998 DF,  p-value: 0.5729
```

3. Assessment of Kossinets and Watts (2009)

The article by Kossinets and Watts (2009) is primarily interested in the social phenomenon of homophily, the tendency people generally have to associate with those who are similar to themselves. Based on the review of the literature, the authors introduced two possible explaining theories for such phenomenon: choice homophily and induced homophily based on individualistic and structuralistic thinkings. Acknowledging the fact that individual preference and external structural factors interplay to affect individuals' relationships with others, they intended to analyze the interaction between the two forces in a dynamic model. Therefore, in this current paper, Kossinets and Watts aimed to focus on university community to answer their research question: how do structural proximity and individual preference interplay to account for the observed homogeneity in one's formed social bondings and structural positions.

The university community was a population of interest by Kossinets and Watts (2009). Their data was obtained from three sources: the total number of 7,156,162 e-mail exchanges of 30,396 university students, faculty, and staff in a period of one year (270 days); participants' individual attributes; and their records of course registration. They reported descriptive statistics, mathematical methods that they used, and their definitions of variables in the method section. They defined some basic variables, such as the intensity of two individuals' interaction by computing the instantaneous strength using the sliding window filter technique, and proposed rules for tie formation to follow the mechanisms of cyclic closure and focal closure.

One potential problem that I have noticed in the data cleaning process of this research is that the authors have utilized the same methodologies of computing variables on the whole population of professors, students, and staff. For example, they have noticed that the CDF of

shared bulk messages over all pairs of individuals were S-shaped and, in turn, they were able to approximate the threshold value g_* to be 140. In another word, if two individuals share at least 140 bulk email messages per semester, then they were said to be sharing an implicit focus. However, analyzing implicit focus by using shared bulk email messages might work well on university students but probably not professors and staff. As the authors pointed out themselves, their observation of having more than 99th percentile of shared bulk messages ($g_* > 140$) was observed to closely match the effect of sharing a class for student pairs. But, while acknowledging the fact that sharing classes is an essential way for students to form social ties, it could be impertinent to generalize this definition of implicit focus to pairs of professors and staff members. This potential weakness in data cleaning could give inaccurate reports to the evolution of social network for student-professor/staff pairs or professor-professor (staff-staff) pairs.

Moreover, it is always worth to note that, even in university settings where emails are extensively used in social connections, many interpersonal interactions exist out of classes and email exchanges. This current model by Kossinets and Watts overlooked other ways of social interaction by only paying attention to in-school email exchanges and course registration records. One way that they could address this weakness is to also incorporate survey methods to find out other means of communication between college students outside of email and class settings. For example, if a majority of the population uses Facebook or Instagram as a way to form social interactions, analyzing data from these sources might give researchers a better understanding of the phenomenon of homophily in that university.