Summary

Haoyu Sheng

2/22/2017

Abstract

The kane2017 package utilizes the data published in Williams Online Catalog to assess the number of students with Chinese/Cantonese/Japanese origins and compare their academic achievement with other students. This is done for 15 years from 2002 - 2016 by scanning the data directly from the text version of the webpages. A sequences of statistical comparisons will be utilized to examine the academic achievement of (partial) Asian International/American students.

Introduction

College admission, espeically recently, has been suspected of raising a higher standard towards Asian students. Though it is unclear that the admission standard has been raised at Williams, looking at the Asian students ratio of attaining Latin honors will be a useful tool in assessing the overall academic performance of Asian American students and can thus further reflect on the bar set during admission. The package kane2017 employs string manipulation(splitting, pattern detection) to split each string into its last name.

The data set is further manipulated by comparing each of the last names to a pre-existing list of possible last names of Asian origin. But processing whether if these last names are of Asian origin, we can then figure out the ratio of Latin honors among Asian students.

Data

The information used to construct the dataset is taken from the Registrar office. The text file is manually converted from the catalog through coying and pasting.Then the text file is converted into Data Frames with a column indicating the year after each student's name. In the end, I compile a data frame "DataSummary.RData" that includes all entries of names from 2002-2016 and the years associated with all the names.

In addition, the Chinese last name list is compiled through all possible pinyin combinations provided by yabla.com, the Cantonese last name list is provided by wikipedia and the Japanese last names are provided by tekeli.li. Though Asia is a diverse continent that includes many more ethnicities besides Chinese, Cantonese, and Japanese, for the sake of clarity and brevity, I will use "Asian"" to indicate individuals whose last names match with the reference list I found online.

Before I start analyzing the results, I feel the urge to briefly explain the process.

Though the input are already quite concise, since last names and honors are the only things we need, we need to cut down the strings.

For example, for input: *+Kathleen Malone Palmer, with highest honors in Neuroscience, readHonor will use

unlist(str_split(data[j],","))

gsub(".*", "", name)

```
to turn the original string into Palmer.
```

After obtaining a list of last names, we then compare the last names with the given list of last names. Now we have chineseLastName, japaneseLastName, cantoneseLastName, and totalVeri, which is the collection of the previous three.

The verifiers and student entries are previously stored through RData and passed into readHonor(data, verifier) readHonor will then return a data.frame as below

```
readHonor(students, totalVeri)[1:5,]
```

```
##   name honor year sigmaXi phiBetaKappa
## 1   Lo Summa 2002    TRUE         TRUE
## 2   Li Magna 2002    TRUE         TRUE
## 3 Song Magna 2002    TRUE         TRUE
## 4 Wang Magna 2002    TRUE         TRUE
## 5 Hong   Cum 2002   FALSE        FALSE
```

after readHonor returns a data frame, I translate the data frame into a simpler numerical form that is better suited for calculations.

```
honorStats(readHonor(students, totalVeri))[1:5,]
```

```
##   year total summa magna cum sigmaXi phiBetaKappa
## 1 2002    20     1     3   3       4            4
## 2 2003    20     1     3   3       4            4
## 3 2004    16     1     0   3       1            1
## 4 2005    18     1     3   2       3            2
## 5 2006    13     0     0   4       1            0
```

However, for the purpose of this project, which is intended to compare the academic achievements between Asian students and non-Asian students and obtain the ratio of the percentage of Latin honors attained by Asian students and percentage by non-Asian students, I created a helper method **totalStats**

```
totalStats(students)[1:5,]
```

```
##   year total summa magna cum sigmaXi phiBetaKappa
## 1 2002   534    11    70 106      98           58
## 2 2003   534    11    70 106      98           58
## 3 2004   509    11    67 100      96           58
## 4 2005   532    10    70 111      98           58
## 5 2006   506    10    67 104      96           58
```

With the aid of **totalStats** and **honorStats**, I was able to calculate the ratio between the percentage of honors attained by Asian students and the percentage of honors attained by non-Asian students through **ratioIndicator**

```
ratioIndicator(honorStats(readHonor(students,totalVeri)))[1:5,]
```

```
##   year   summa     magna       cum   sigmaXi phiBetaKappa
## 1 2002 2.57000 1.3350649 0.9994444 1.0936170    1.9037037
## 2 2003 2.57000 1.3350649 0.9994444 1.0936170    1.9037037
## 3 2004 3.08125 0.4001623 0.7083333 0.3243421    0.5405702
## 4 2005 3.17284 1.5029240 0.9261261 0.9017544    1.0198413
## 5 2006 0.00000 0.0000000 0.8570187 0.3991903    0.0000000
```

In the above table, a ratio that is greater than 1 indicates that there is a higher proportion of Asian students who are earning a particular honor than that of non-Asian students. In addition, to better understand the ratio, I created another helper method that enables calculations of ratio by using the sum of all honors throughout the years.

```
statsSum(honorStats(readHonor(students,totalVeri)))[1:5,]
```

```
##   honor    ratio
## 1 summa 2.572207
## 2 magna 1.312864
```

```
## 3   cum 1.229118
## 4 sigma 1.204438
## 5   phi 1.427551
```

statsummary

The function statsummary is used to generate different analytical graphical representation of the honor data and ratio collected and generated by the helper methods described above. The function takes in one parameter that indicates the type of analysis desired.

> *statsummary(type)* The parameter type has 5 important ratio measurement input values:"OR cum", "OR sigma", "OR phi", "OR all", and "Overall Honor Ratio"

Separate graphs of "OR summa" and "OR magna" are not implemented since

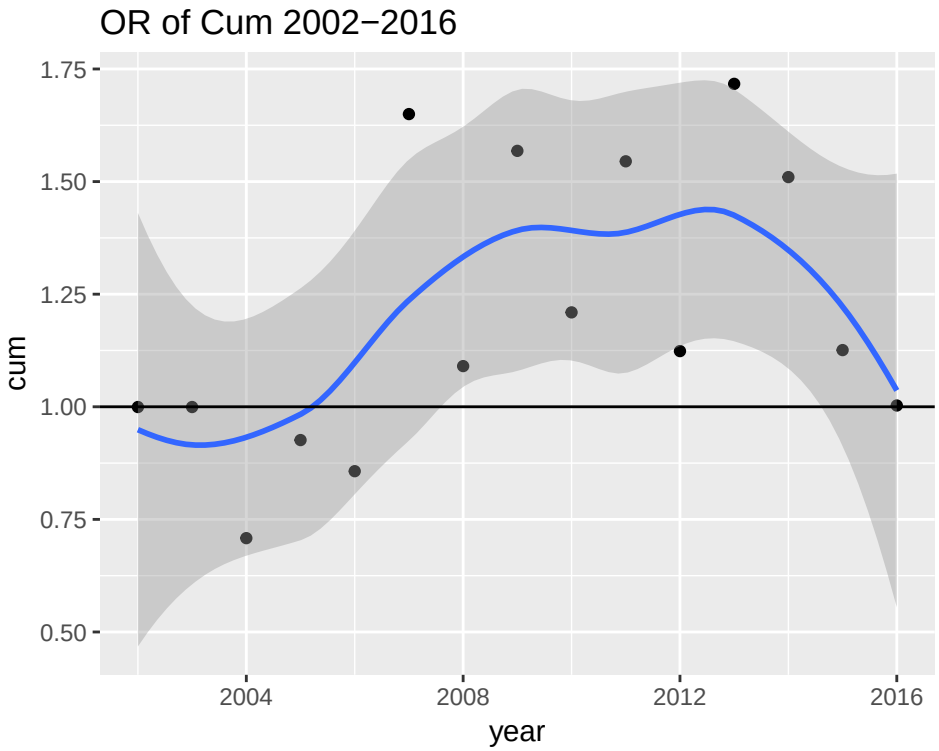
i) the size of measurement for percentage is too small

ii) they are shown in "OR all" and

iii) "OR cum" simply offers better measure for academic achievement since summa and magna, though having small sizes, are also counted within the category Cum Laude & Above.

```
statsummary("OR cum")
```

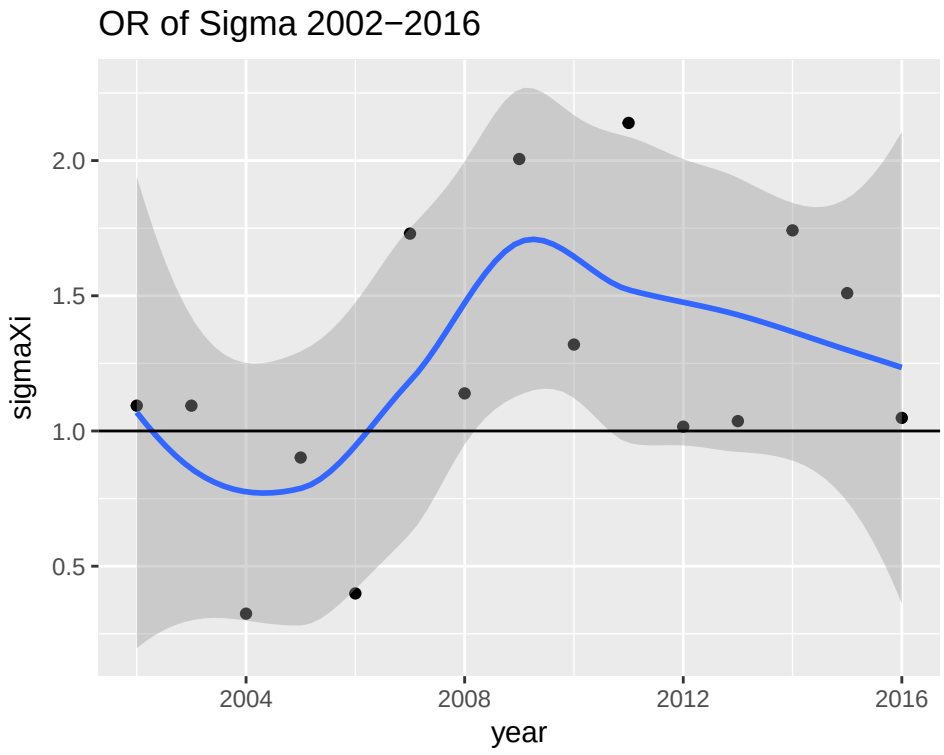
```
## `geom_smooth()` using method = 'loess'
```



If we take a close look at the graph of **statsummary("OR cum")**, the plots experiences a drastic increase from 2006 to 2007. If we look at the distribution of the dots, we can find out that before 2006, all the ratio values are less than or equal to 1, whereas after 2006, all the ratio values are greater than or equal to 1.

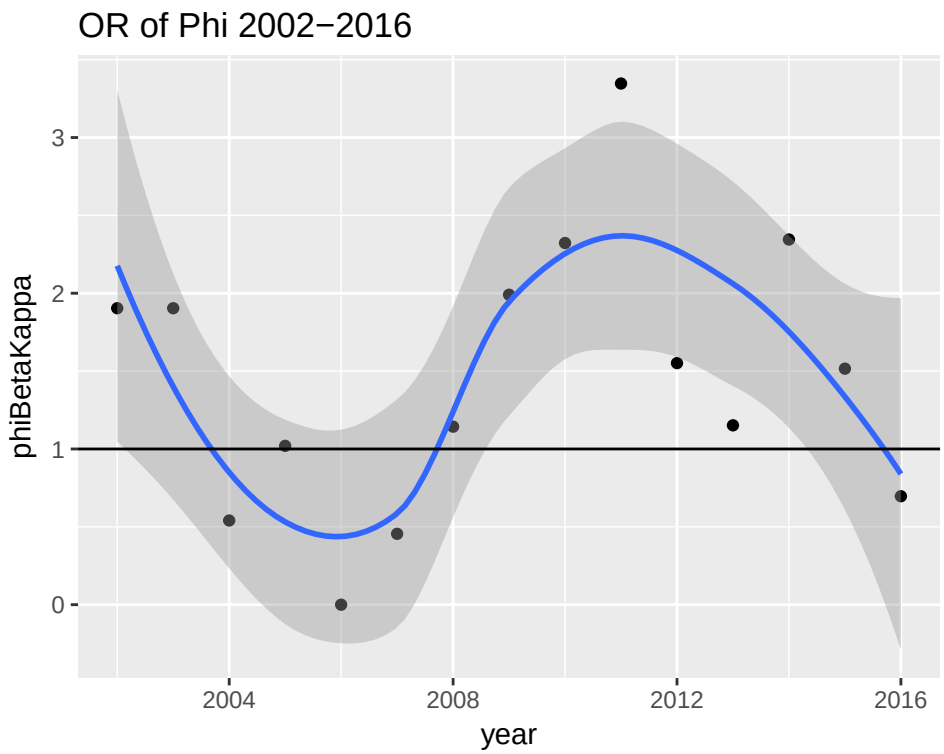
```
statsummary("OR sigma")
```

```
## `geom_smooth()` using method = 'loess'
```




```
statsummary("OR phi")
```

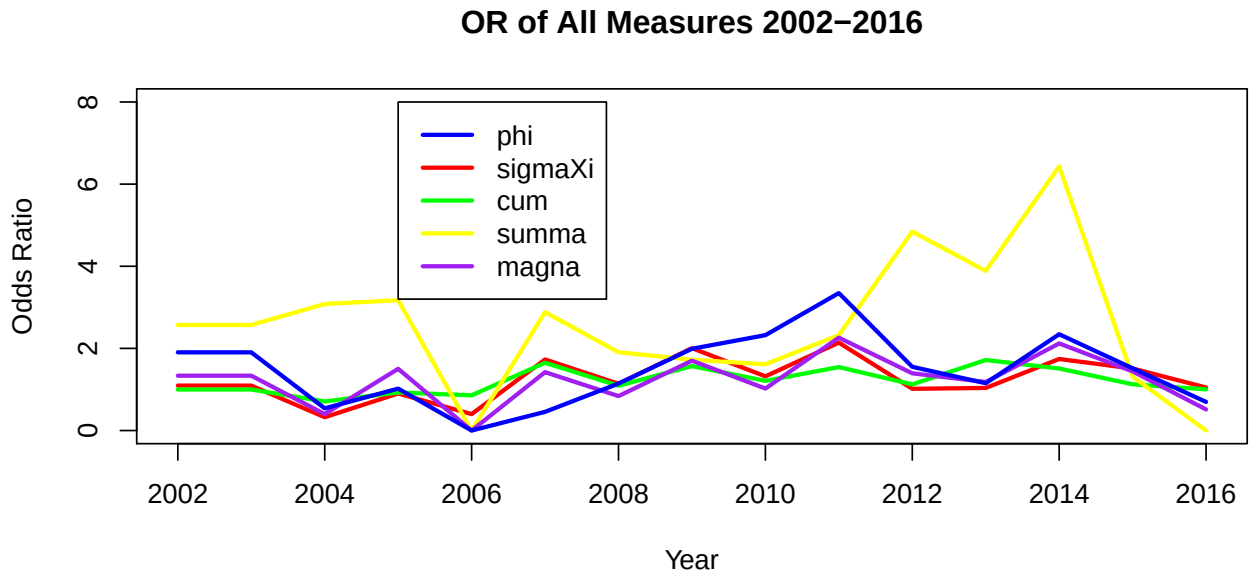
```
## `geom_smooth()` using method = 'loess'
```



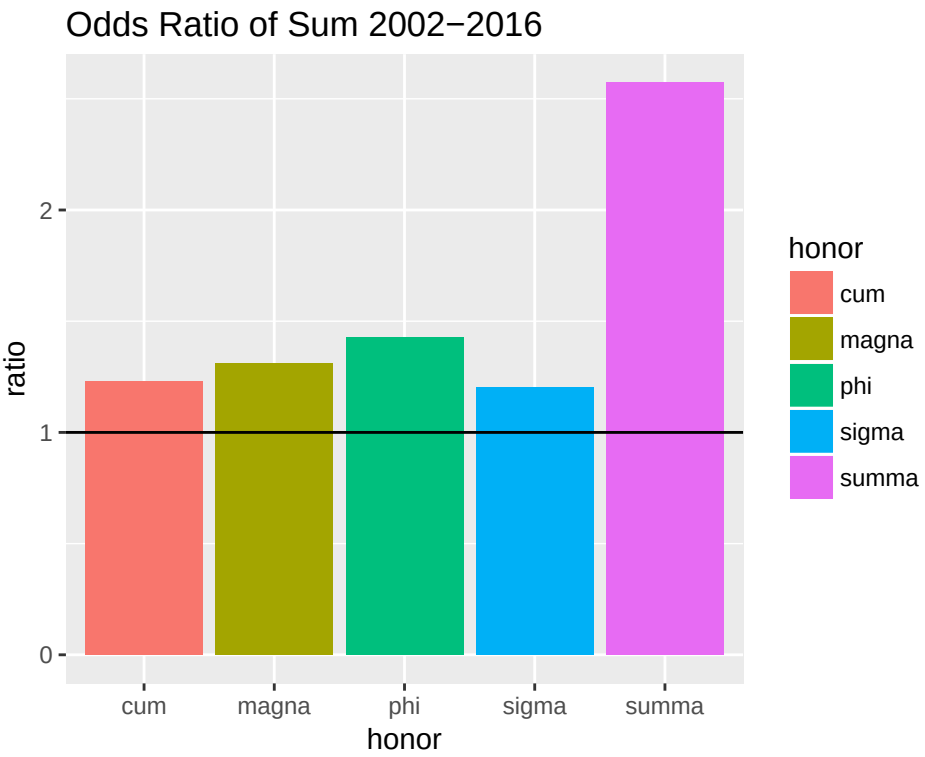
In both **statsummary("OR sigma")** and **statsummary("OR phi")**, the dots are mainly distributed above the horizontal line (y=1), which is indicative of a better performance by the Asian students. Interestingly enough, a somewhat similar trend can be observed here that both graphs have a steep slope at year 2007.

However, though they are similar to the first graph in this steepness of slope at around the same time period, **statsummary("OR sigma")** is more similar to graph 1 in a way that the Asian students are performing a lot better after year 2007 while the graph of **statsummary("OR phi")** resembles more of an undamped oscillation.

```
statsummary("OR all")
```




```
statsummary("Overall Honor Ratio")
```



Also, if we add up the honors in each category for all the Asian students and compare these numbers with the honors in each category for all the non-Asian students throughout the years, we can find that in all honor categories, Asians are doing better overall from 2002 to 2016 in all categories as all ratios are above 1.

Analysis

From the above analysis, it can be somewhat inferred that the Asian student population is doing significantly better after the years 2006 and 2007. Such academic performance, in my opinion, can be due to one of the following reasons:

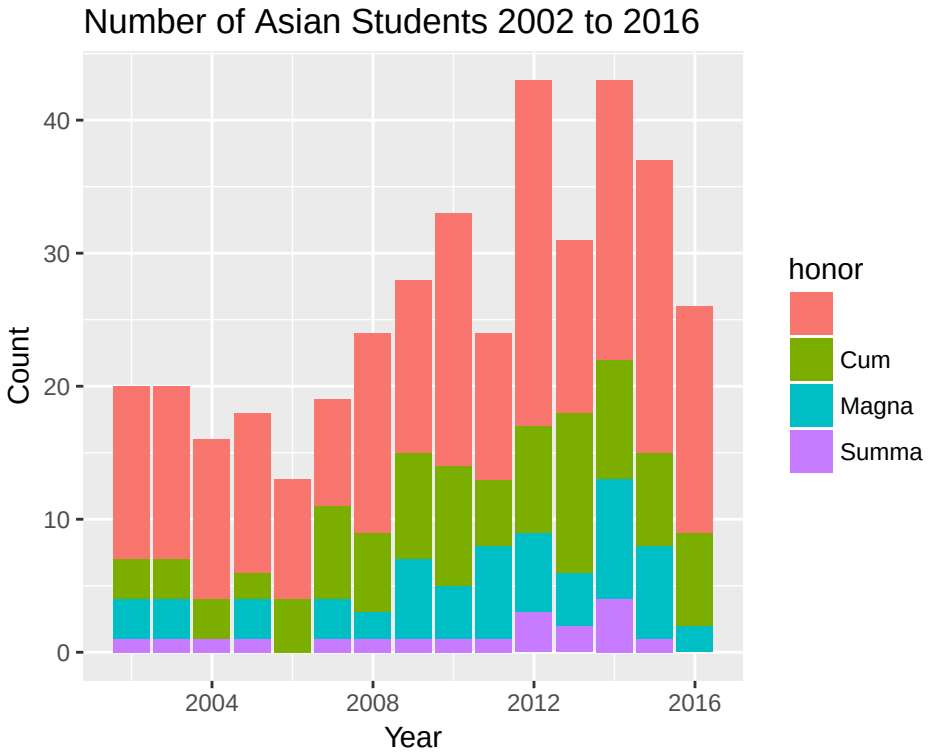
1) Asian students are innately more motivational
2) Asian students are facing higher admission standards due to affirmative actions
3) Asian students are facing higher admission standards since there are simply more applicants of Asian descent.

If 1) holds, then all academic performances by Asian students should be consistently better than that by non-Asian students. However, if we take a look at **OR of Cum2002-2016**, Asian students didn't start doing significantly better until the year 2007. Thus, it's clear that there is some implicit change in the competitiveness of Asian admission standards.

Since in 2), affirmative actions are too vague of a measure, we will avoid digging deep into this option, but will keep it as one of the possibilities.

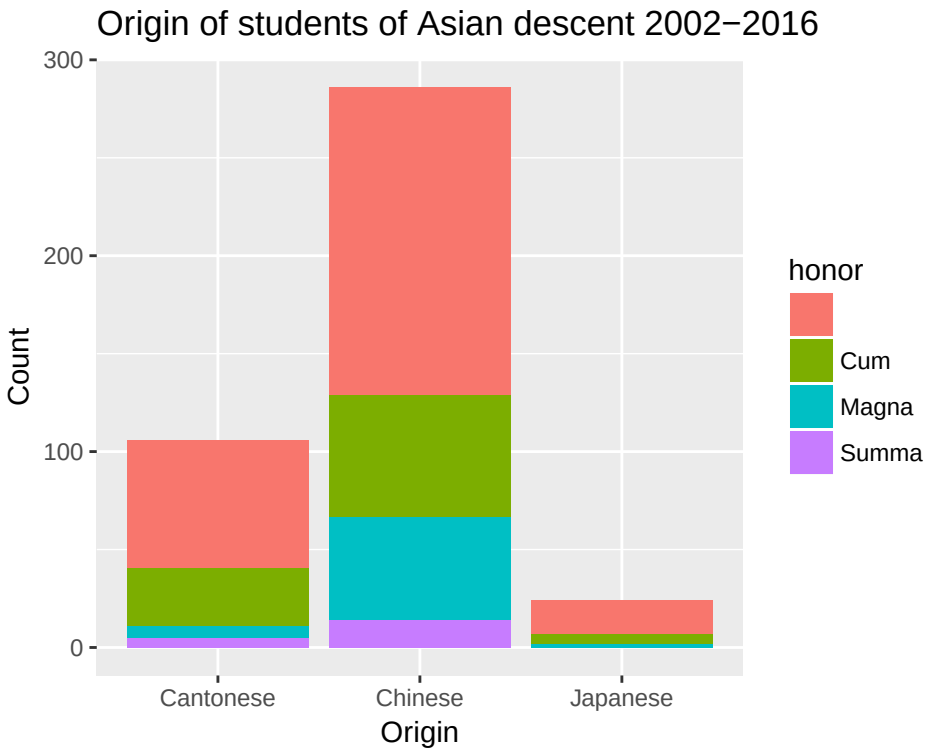
Before looking into 3), the below graph of count of students of Asian descent from 2002-2016 will show us some interesting findings.

```
statsummary("summary by count")
```



After the year 2007, the number of Asian students, though relatively stable, is overally increasing. Coincidentally, according to foreignpolicy.com, the number of Chinese students abroad has been drastically increasing since 2004-2005 academic year (which is around when the class of 2008 started college) and has increased nearly fivefold now. Therefore, though the Asian applicants from within the U.S should stay relatively stable, this influx of international applicants can increase the academic standards for admissions. A further examination can be done when countries of origins can be provided.

```
statsummary("summary by origin")
```



Conclusion

The summary statstics indicates that according to the Asian last name reference sheets, students of Asian origins tend to be significantly overachieving than non-Asian students after the year 2007, with an overall average Cum Laude odds ratio significantly better than that of non-Asian students. Also, the "Odds Ratio of Sum" graph indicates an overall total advantage of Asian students' academic performances. Thus, it can be inferred that Asian students tend to be more academically achieving. Though more data will be needed to evaluate what causes the rise in admission academic standards, it is clear through the analysis section that the higher academic standard is very likely to exist.

However, this package is still incomplete in many ways. First, distinguishing last names from its linguistic composition can be extremely ambiguous. It doesn't take into account some of the identical last names different cultures can share. Also, it doesn't take into account students of mixed racial identity. Therefore, the next step will be to incorporate more last name recognizing templates as well as trying to distinguish last names of ambiguous identities base on first names.