NTDS - A Network Tour of Inter-county Migration in the US

Fatima Moujrid **Xiaoyan Zou** **Paul Mansat** **Anshul Toshniwal**

I. Introduction

Immigration and migration is a very hot topic nowadays. During the last 2016 US Presidential Election, immigration has been one of the main polarizing topic between Democrats and Republican, the two main political forces in the US. It appears that migration is a subject that deeply concerns people, and their political orientation is partly based on their belief about migration/immigration.

Other than the fact that migration is an important issue during election, the study of migration in a specific region also reflects the socio-economic background of the region – and to some extend its political orientation. Recent economic studies [3], have noted that income inequalities in the US are increasing, leading to the emergence of an upper-class that concentrate most of the American wealth. Although the specific definition of the upper-middle and middle class is controversial (some economists believe that a class is not only defined by its income but also by its cultural and educational background), there is a general consensus that some specific region in the US concentrate the best Public Services. These regions concentrate the highest incomes, leading to social inequalites between region in the US. The migration between each of these region is not homogeneous: high incomes citizen are most likely to move in region that offer the best Public Services, but where the housing prices is high. On the other hand, lowest incomes who can't afford houses in these expensive region move to less desirable region (in terms of Public Service quality). For instance, in 2014 it has been noted that 40% of the US upper-class lives near a school where the average score to the SATs is higher then the average. So migration is both a sensitive topic during Election and can also be an indicator of the socio-economic background of a region and we believe that by studying migration between counties in the US we can predict the result of Presidential Election. Through this project, we will try to predict the election result of a county in the US by considering various features characterizing the migration flow between counties in the US using two approaches Graph convolutional neural networks and graph signal processing.[1]

II. Data Acquisition

This project relies on two kind of information: (1) the migration between counties in the US and (2) the result of the 2016 Presidential Election at the county level. The first set of information is found on the IRS (the federal Internal Revenue Service agency) website and the second is brought by a data-set elaborated by the Guardian (a British newspaper). The IRS data-set comes up with many information fields for each county. The data-set is indexed according to counties. Each county is defined by a unique identifier called FIPS. It is a 5-digit number, the first two define the state in which the county is located, and the last ones define the county within the state.

For each county, the IRS data-set provides: the number of US citizens who migrated into a specific county, the number of non-US migrants as well as the total number of migrants who moved into the county, referred to later as migration flow.

Each migration flow is further characterised by some key features such as the agi (i.e average gross income) of the migrants, the number of migrants who pay taxes and the number of migrants who are tax exempt. The number of migrants who pay taxes is refereed to as the *return* feature and the number of people who do not pay taxes will be called the *exempt* feature.

To use the information described, data extraction, cleaning and discarding of irrelevant features such as ethnicity, age and state, is applied to the IRS data-set but is not described here. Further information on that subject can be found on the documentation of the cleaning code in the project notebook.

The Guardian data-set is defined by three features: the FIPS code of the county, the percentage of voters in the county who voted Democrat, and the number of voters who voted for the Republicans. With the help of the FIPS code we were able to merge the Presidential result election data-set and the migration data-set.

III. Data Exploration

To get an idea of the structure of the graphs created, their degree distribution is plotted. Two graphs are built by setting a threshold to keep counties where 70% of migrants are tax exempt is the *exempt* graph and 35% are paying taxes in the *return* graph. Consequently, the *return* graph has more percentage of nodes having higher degrees relatively to *exempt* graph. The degree distribution also implies that most of the counties have degree less than 50 connections. The migration graph, as expected, has higher degree nodes as it has edges between all the counties which have a migration between them. The degree distribution of both graphs follow a power law and hence are scale free. Due to its the high threshold, the *exempt* graph has fewer edges compared to the *return*

[1]Project done for course Network Tour of Data Science in EPFL. Codes can be found at: https://github.com/zxyzz/ntds$_{project}$

graph. Consequently, the *exempt* graph is much sparser. As the returns graph contains higher degree nodes,it is expected to have a higher average clustering coefficient and a larger giant component size. Calculating the quantities explicitly confirm the predictions.
The returns graph and the exemption graph are simulated using *Barabási–Albert* (BA) network and *Erdős–Rényi* (ER) network,as shown in the figures 1 and 2:

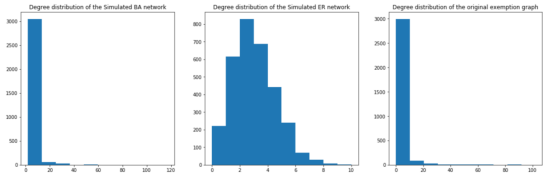

Figure 1: From left to right: the BA , ER simulation of the degree distribution of *exempt* graph and the actual degree distribution of the graph

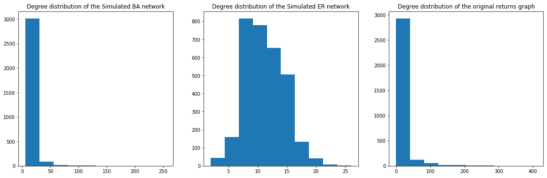

Figure 2: From left to right: the BA , ER simulation of the degree distribution of *return* graph and the actual degree distribution

In both the cases it was observed that the BA network provided a better fit for the degree distribution. As the ER network is a random network. The BA network also provided a good fit for the clustering in the case of exemption graph but not so in the returns graph.It could be because the returns graph is more dense and has a more structure in it as compared to the simulated BA network.

IV. Data exploitation

A. Presentation of the graph

In the construction of the graphs, the following consideration is made: a migration flow has a high proportion of people paying taxes if more than 40% of migrants are paying taxes. On the other hand, the flow is highly tax exempt if 80% of the citizen in the migration flow are exempted from paying taxes. The boundaries were chosen so that it corresponds to roughly the fourth quantile for each distribution.
This dichotomy allows us to study the correlation between a voting pattern and a type of migration in the county; and yields to the creation of two graphs:

- $graph\_nonRGB\_returns$: the nodes of this graph are the county and are associated to a signal that is majority vote per county in 2016 election : Republicans take value $+1$ and Democrats -1. There is a edge between node i and j if (1) there is an actual migration between county i to county j and (2) if the latter migration flow has a proportion of people paying taxes greater than a specified threshold.
- $graph\_nonRGB\_exempt$: this graph has the same architecture as the aforementioned graph, only that we are now filtering according to the proportion of people that are exempted from paying taxes.

B. First attempt at predicting election result

At this stage, a possible correlation between socio-economic inequalities between counties in the US and the voting pattern is investigated: the party that would ensure that people that have access to the best Public Services retain this privilege is most likely to perform well in the wealthiest region. These regions would be conservative and would be characterized by migration inflow where people have above average incomes and thus a high proportion of people paying taxes. On the other hand, we also emit the hypothesis that counties with inflow that have a high proportion of people tax exempt vote for the party that support bills that would promote better Public Services for all. These counties would vote for the party that promotes a welfare state.

To evaluate the correctness of this hypothesis we created two graphs: the first one considers only the migration flows between counties were more then 38% of the migrants are paying taxes, known as *return* graph. The second graph considers only the migration flow between counties where more then 70% of the migrants are paying taxes, known as *exempt* graph. If the first stated hypothesis is correct then on the first graph we should be able to see that most of the connections of the graph are concentrated toward counties that are conservative (i.e Republicans). On the other hand, if the second hypothesis is correct, most migrations will have as destination a Democrat county.
In the *exempt* graph visualization with *Gephi*, as shown in figure 3, no particular structure that would corroborate the second hypothesis arises. Edges are from Republican to Democrat and from Democrat to Republican in no particular order. However, in the observation of the *return* graph, in figure 4 is more conclusive: most of the flows with a high proportion of people paying taxes have as destination a Democrat county. This observation rejects the first emitted hypothesis, however it is still an insightful observation: it appears that from the return graph if one studies the degree of the node, one could be able to tell if its a Democrat or Republican county.

C. Prediction based on degree of county

The aforementioned observation tells us that by studying the degree of a node, we might be able to predict the label (i.e Republican or Democrat) of that node. The driving force behind our first prediction algorithm is quite simple : we believe that we can split the nodes into two categories. The first category being nodes with high degree and the second category being nodes with low degree. These two categories will then mapped to reciprocally Democrat or Republican. However, the problem remains on finding the correct threshold to construct our graph (remember that our graphs are constructed using a threshold on the proportion of migrants paying taxes or not) and what should be the degree that defines the limit between

the two aforementioned category. This limit is from now on referred as the "cut". This problem of finding the best possible tuple of hyper-parameters is a cross-validation problem [2]. Hence, we implemented a cross validation algorithm that finds the best possible cut and threshold for this problem and computes the accuracy of predicting the county election result in such a way. The result of this implementation points out that a cut at degree equal 6, would rightfully predict half of the Democrat's counties and 92% of the Republicans giving an overall accuracy of 85%. This is not great accuracy score, as one could simply say that all counties are Republican and get an overall 81% accuracy because of the heavy tailed nature of the labels signal.

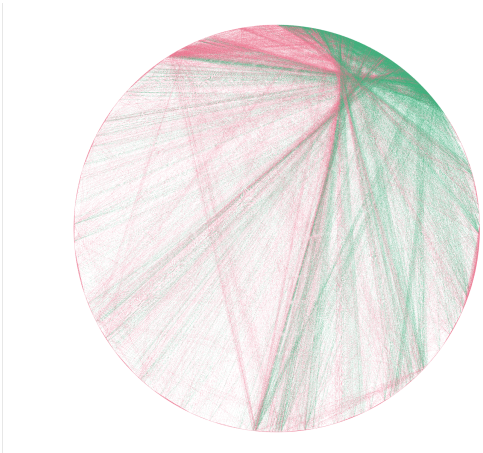

Figure 3: Gephi with Circular Layout of the *exempt* Graph

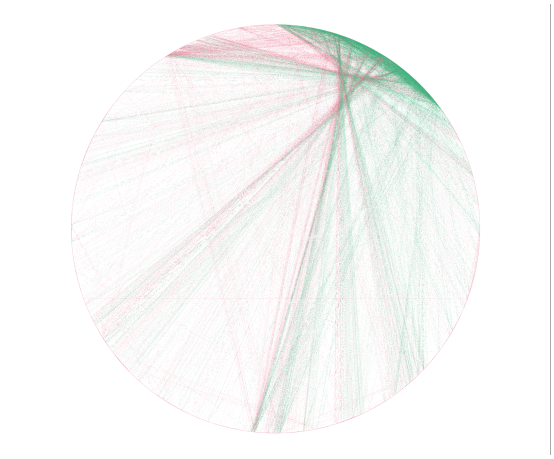

Figure 4: Gephi with Circular Layout of the *return* Graph

D. Prediction based on degree neighboring nodes

The previous technique based on predicting the label of a node based on its absolute degree proved to perform poorly because half of the Democrat nodes are wrongly predicted. Hence, a second approach is implemented: predict the label of a node based on the average degree of its neighbors. In previous prediction algorithm, to distinguish between democrat/republic nodes,a high threshold on the proportion of people taxes is set. Consequently,a high proportion of nodes in both categories is edge free. Alternatively, the neighboring nodes are studied to decrease the threshold. Even-though more Republican nodes will have connection, we believe that still by making the average of all of there connections we will get a lower average degree than for Democrat nodes. The approach is done for both graphs (to which we applied the same cross-validation discussed before). Consequently,a poor accuracy score of 78% is reached. To remedy to the heavy tailored nature of the labels (much more republic nodes than democrat),a log penalizing loss function such as the log is introduced. we are computing the absolute difference between the number of nodes above the cut for Democrats minus the number of Republican nodes above the cut, we are penalizing the Republican nodes with the log function. By doing so, we are forcing the "loss function" (based on the degree of the nodes) to make sure that the number of nodes above the cut is kept small for Republican . An overall accuracy of 87% is reached.

E. Features Graphs

At this stage, after investigating the possibility of predicting the majority vote of a county based on the degree of its neighbours. A spectral analysis of the two graph used previously is performed in the sake of improving the accuracy. Considering that the unemployment rate and the Average gross income (AGI) are two major economical factors that would drive an immigration flow from a county to another.
Therefore, based on the two graphs, namely, the *return* and *exempt* graphs, the unemployment rate and the Agi of each county are added. Based on this feature graph, an unweighted adjacency is built, where a edge is set between a pair of counties if more than 38% of the immigrants are paying taxes in the *return*. Similarly, a second adjacency matrix is built based on the *exempt* graph , where two nodes are connected if at least 56% of the immigrants are exempted from taxes.

F. Spectral analysis and signal processing

In order to predict the vote pattern of a particular county,a signal based on the labels of each node (contains only -1 and 1) is partially omitted for 20% of the counties selected randomly. Consequently, the entries of these counties are substituted by 0
The laplacian of the *return* adjacency is built and its eigen values and vectors are computed. A low pass filter based on the eigen decomposition of the laplacian is implemented. It is applied on the masked signal. The main idea behind this choice is the high smooth performance of the low pass filter on a signal. As it maintains a very low difference between a node and its neighbors. Therefore, The filtered signal values are no more limited into two values relatively to the ground

[2]We refer to cross-validation, but we are not splitting the data to create a validation set and a proper training set, hence talking about cross-validation here might be an over-statement. However, the term still encapsulate the idea that we are trying to find the best possible tuple (cut, threshold) for this prediction.

truth labels signal. A threshold is set so that if the label of node in the filtered signal is positive it takes 1. Otherwise it takes -1.
Additionally, a heat kernel which is supposed to take in a vector of eigenvalues of the laplacian and output a vector of evaluations of the heat kernel at those eigenvalues, is used.
The same procedure is applied on the *exempt* feature graph.In order to evaluate the performance of both filters on each graph. A cross-validation analysis is conducted. In fact, considering the heavy-tailed nature of data, F1 score is used to evaluate the accuracy. This analysis is performed with 20 iterations. Only the mean and variance over these iterations are considered and displayed in the table below:

Graph	Low pass		Heat kernel	
	Mean	variance	Mean	variance
Return feature graph	89%	0.008	88%	0.01
Exempt feature graph	91%	0.007	95%	0.007

Table I: Results of the cross-validation of the accuracy analysis

Relatively to the prediction using the degree distribution of a county, both the heat kernel and the low pass filter has yielded into identical results, as far as the *return graph* is concerned. The vote prediction accuracy significantly increased using the *exempt graph*.

G. Graph convolutional network

The accuracy of vote pattern with low pass filtering the masked labels signal and through applying a heat kernel did not perform better than the neighbours degree method for the *return graph*. Therefore, another prediction model known as Graph convolutional neural network (GCNN) is employed to predict the majority vote (Republic or Democrat) of a county based on the unemployment rate, the Agi. The model is applied on both graphs, namely the *return* and *exempt* graphs. Using the GCNN implementation developed in [2].The classifier is feeded with a training set randomly selected from the shuffled overall dataset respectively with a matrix $X^{tr}_{migration1} \in R^{N_{return} \times 2}$ and $X^{tr}_{migration2} \in R^{N_{exempt} \times 2}$, containing the unemployment rate and the AGI information. N_{return} being the number of counties where 38% of immigrant are paying taxes. N_{exempt} the number of counties where more the 56% of the immigrants are exempted from taxes. Similarly, the labels signal is randomly split into a training/test set on which the GCNN is performed. Once trained the model is then used to predict the majority of the remaining counties in the testing sets. The output is evaluated through F1 score which is computed over the testing labels and the testing fraction of the data.
The training is used with 500 epoch, a decay rate of $5 \times 10^{-}6$, a learning rate of 0.2,a polynomial order of 3. The dropout rate is set at 0.8 to reduce over-fitting during the training and improve generalization error.
The results of the GCNN classifier are displayed in the table below:

The graph convolutional neural network shows more or less the same accuracy score for both graphs. It has brought some enhancement for the *return graph*. However, the outcome of GCCN might be improved by optimizing the parameters of the model such as the learning rate and the degree of the polynomial order. Note that that the adjacency matrix of both graph are highly sparse. Therefore feeding the model with a richer data might improve the result.

	Return feature graph	Exempt feature graph
Accuracy-mean	92%	91%
Accuracy-var	0.008	0.007

Table II: Results of the Graph convolutional neural network on both graphs

H. Similarity Graphs

The prediction performance of the three investigated methods over graphs based on raw feature are still not satisfying. Therefore, similarity graph based on the Gaussian kernel are used.The studied data-set provides the origin of the immigrating flow from a county to another. To take advantage of this affinity, three similarity graphs are constructed: The *total* graph which consider the whole flow between counties, the *US* graph that consider the migration of US citizen and the third being the *foreigner* graph which includes the inter-county flow constituted of non-US citizens.
The Gaussian kernel coupled with euclidean distance as shown in the equation below, is used to obtain the wights w_{ij} of the adjacency matrix for each graph. The features used in this case are slightly different from the previous ones.The considered feature for the three graphs are: the proportion of immigrating people exempted from taxes relatively to those who are paying in particular county. The second feature is the normalized AGI of the county:

$$w_{ij} = exp(\frac{-||x_i - x_j||}{\sigma}) \tag{1}$$

σ is the kernel width. The sparsity and connectivity of the adjacency matrix is ensured through the threshold ϵ set at 0.2 in the *total* immigration matrix, and to 0.5 in the *US* and *for* matrices.
Similarly the previous section, the same labels signal is used. The laplacian eigen decomposition along with a low pass filtering and heat kernel are implemented on each adjacency matrix. Moreover the (GCNN) is performed on the three similarity graphs. The accuracy results based on F1 score for the three methods are displayed in the table below:

Graph	low pass filter		Heat kernel		GCNN	
	mean	var	mean	var	mean	var
total	0.93	0.006	0.93	0.009	0.92	0.008
US	0.92	0.007	0.93	0.009	0.93	0.008
foreigner	0.85	0.03	0.83	0.020	0.83	0.025

Table III: Results of the cross-validation of the accuracy analysis using similarity graphs

Using similarity between counties based on the AGI and the *exempt/return* ratio didn't bring a significant enhancement

to the accuracy score relatively to the results obtained with the return and exempt graphs without involving any similarity kernels. For the total migration between counties and the migration of US citizens. The accuracy drops using the similarity graph of foreigners between counties. This is probably due the fact that the *foreigner* is very small. It contains 455 nodes whereas the *total* and *US* contains 2963 nodes.

I. Similarity graphs structure

In this part, The possibility of any division or isolation into particular communities is according to the voting pattern is investigated in the three graphs. Additionally, plots of the labels signal on each graph using the NetworkX force-directed layout is implemented similarly to [1].The spring layout is used for the force directed layout . This means that each node tries to get as far away from the others as it can, while being held back by the edges which are assimilated to springs, having a spring constant related to their corresponding weight. The plots of the majority vote signal on the total immigration , US citizen and foreigners immigration graphs between counties is displayed in the figures below:

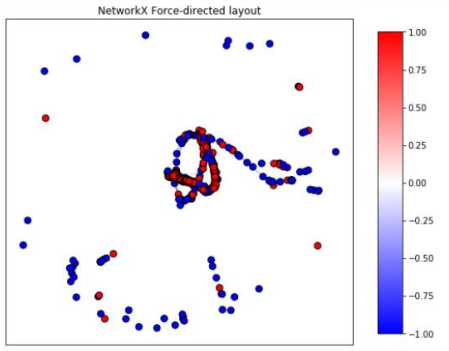

Figure 5: NetworkX force-directed layout of the total migration graph

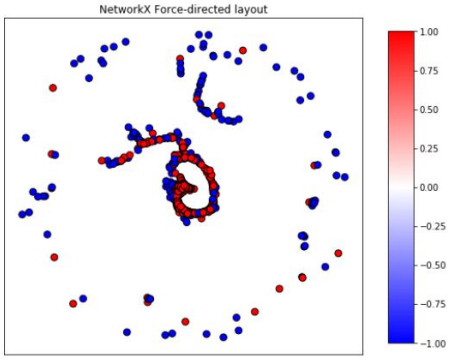

Figure 6: NetworkX force-directed layout of the US citizens migration graph

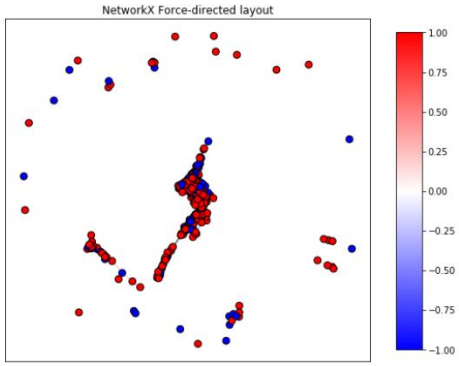

Figure 7: NetworkX force-directed layout of the foreigner individuals migration graph

V. Conclusion

This work is aimed at predicting the majority vote of a county (Republic vs Democrat) based on the economical conditions of the migrants in their destination county. The studied data provides information onthe unemployment rate,the average gross income,the number of migrants paying taxes or exempt from paying taxes as well as the origin of the immigrants(US citizen vs foreigner). The initial consideration and visualization of a highly taxable migrants and highly tax exempt migrants graphs labeled according to 2016 election results points out that immigrants that are paying taxes are moving mainly between democrat counties. Consequently, the expected two poles (clusters) based on the two-party system in the US could not be identified. However, a possible democrat/republic classification through the node (ie.county) neighbours degree is highlighted.

Graph Signal processing techniques like low pass filtering and heat kernel on the presidential results signal coupled with graph Fourier method are used. Omission of some values of the signal and subsequent filtering were performed. The comparison of the heat kernel and low pass filtering to the ground truth signal showed a better prediction accuracy for the *exempt* relative to the *return* accuracy. The GCNN , as a learning method is also implemented for the same goal. GCNN has rather improved the prediction accuracy of the *return* graph.

The three methods are also used for three other graphs constructed according to the origin of the immigrant:The three methods performed poorly on the *foreigner* graph. The three methods perform better with larger and richer data-sets. There is possible rooms for result improvement by optimizing the parameters of the GCNN. The spring layout for adjacency matrix of the three graphs, where each nodes is labeled based on the presidential results is done. There is again no possibility to identify distinct clusters .

References

[1] M.Defferrard. *A network tour of data science (EPFL) 2018: Course material: milestone 4: graph signal processing*. URL: https://github.com/mdeff/ntds_2018/blob/

master / milestones / 4_graph_signal_processing_student_solution.ipynb. (accessed: 17.12.2019).
[2] M.Defferrard. *A network tour of data science (EPFL) 2019: Course material: assignment 2: learning with graphs*. URL: https://github.com/mdeff/ntds_2019/blob/master / assignments / 2_learning_with_graphs_solution . ipynb. (accessed: 04.12.2019).
[3] Richard V. Reeves. "Dream Hoarders". In: (2017).

VI. Annexe

The plot of the laplacian embedding for the three similarity graphs. The second and fourth eigen vectors of the laplacian is used to do the embedding, are displayed below:

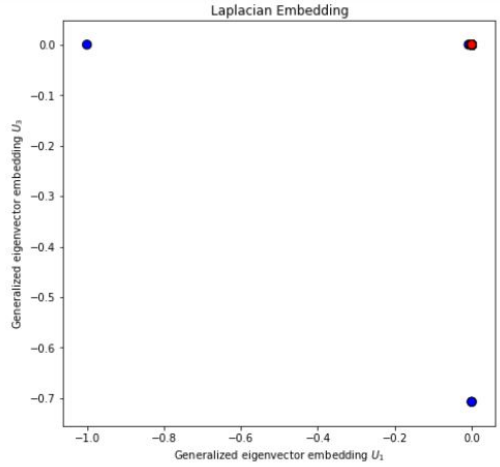

Figure 8: 2D Laplacian embedding of the total immigration graph

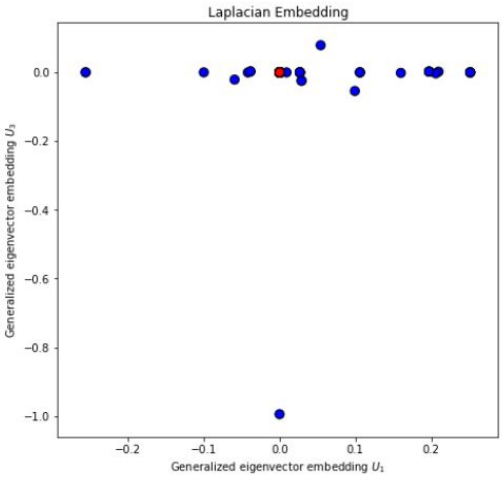

Figure 9: 2D Laplacian embedding of the US citizens immigration graph

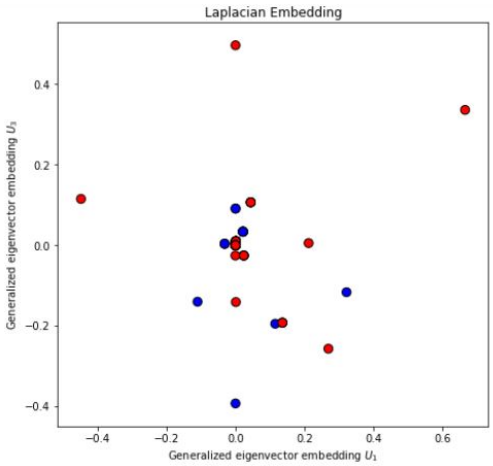

Figure 10: 2D Laplacian embedding of the foreigner individuals immigration graph