NTDS- A network tour of inter-county immigration in the US

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

I. Problem formulation

Immigration has been a true reflection of the political and economic changes in the united states. In this project,the focus is thrown on the inter-counties immigration.The aim is to study how the immigration flow is impacted by the several social and economical properties of a specific county, such as the fiscal mechanisms. This includes the taxes and the exemptions the immigrants are submitted to. Moreover, one question is raised, can the immigration flow, in particular, predict the election votes(Republic vs democratic) is a county.

II. Data Acquisition

The data is extracted from the Harvard database, it contains 1863 counties (ie.nodes). Data points include the state to which each county belongs, the inflow immigration of each county, for the years 2009, 2012 and 2016. The proportion of these immigrants that are either submitted or exempted from taxes in their destination county is also provided. Moreover, the data set contains the Adjusted gross income(AGI) of these individuals. Additionally, the data sets contain some social and economic factors proper to each county such as the median age and the evolution of the number of people of a white Hispanic origin. The number of votes for the republican and Democratic parties are provided for each county, for the period of 2012-2016. In fact, based on our problem formulation, only the data specific to immigration flow is kept. Namely, the number of immigrants exempted from taxes, paying taxes, as well as the average AGI for each county.The columns that do not provide a piece of valuable information in our case is discarded.

III. Data Exploration

As far as the data exploration is concerned, in order to handle the multiple features of each data point (ie.county), an adjacency matrix is created for each of the following features: The proportion of individuals that migrate and are paying taxes (referred to as $returns$), the proportion of individuals migrating and that are not paying taxes (referred to as $exempt$),the AGI (referred to as avg). In the aim to detect what motivate the immigration between counties. Two similarity graphs are built such as the nodes are the counties and the edges are respectively,$returns$, $exempt$. At this stage no kernel is used as the $returns$, $exempt$ proportion out of the total immigration flow fit perfectly a Gaussian law. Additionally, the nodes are labeled based in the elections results of 2016, either Democratic or republican. The adjacency matrices are threshold-ed in order to keep only counties where more than 75% of immigrants are exempt from paying taxes for the first graph and more than 35% of immigrates are paying taxes are kept.

IV. Data exploitation

Based on the labelling of the nodes (Democratic/republic), the possibility of clustering the data accordingly is investigated. Therefore, a PCA, ISOmap .The three methods are performed over the data in 2D.

V. Discussion

At this stage of the project, The visualization with GEPHI tools of the return and exempts graphs show that immmigration flow with a high proportion of returns are mainly between Democratic counties. There is nearly no returns immmigration that is coming from Republican county. Whereas It appears that immigration flow characterized by a high proportion of exemption has no particular structure : they can be between Republican to Democrat county or vice versa, with no particular preference. The embedding of data didn't provide any valuable information as it can be seen that the two classes are very overlapped no matter what dimension reduction method is used.

VI. Further Improvements

Since the current data set does not show a clear link between the inflow immigration and the majority vote in a county.A new data set is used, from Internal Revenue Service (IRS). The three feature studied before are still provided by this dataset.Moreover, it provides information about People who migrated are either US citizen or foreigner. The RBF weighted adjacency matrix for the $exempt$ and $returns$ features are built for foreigners and US citizens in a flow as well as for the total flow. Additionally, the result of votes of each county in the 2016 elections are collected from The Guardian newspaper and merged with the (IRS) data set. A signal of 1(Republican) and -1 (Democratic) out of this election votes in each county is constructed. A proportion of this signal is masked (substituted by zeros) and then applied to the previously constructed graphs. A low pass filter will be used. Accordingly if any of masked nodes got a positive value after the filtering, this node is then republic if not it is democratic. An accuracy analysis will be performed relatively to the original signal.

VII. Intermediate results

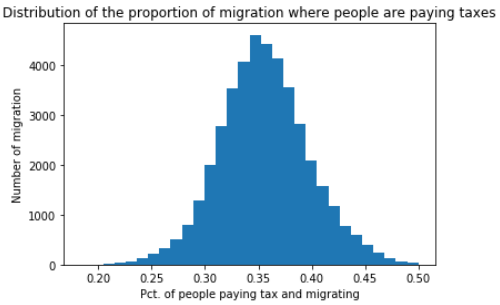

Figure 1: Distribution of the proportion of immigrants that are paying taxes relatively to the total immigration inflow in county

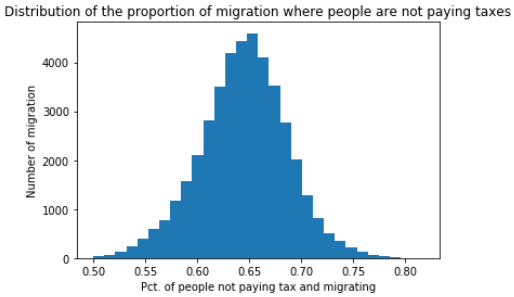

Figure 2: Distribution of the proportion of immigrants that are paying taxes relatively to the total immigration inflow in county

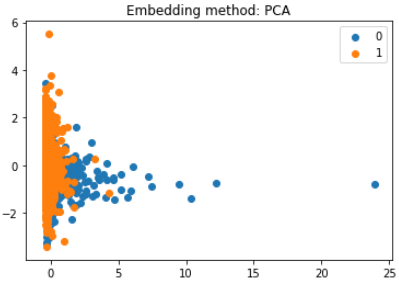

Figure 3: 2D embedding using PCA method

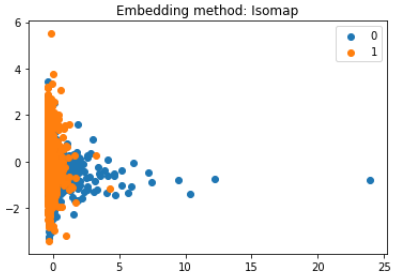

Figure 4: 2D embedding using ISOMAP method

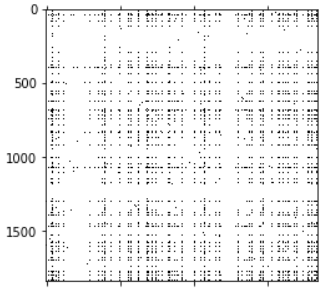

Figure 5: The adjacency matrix of the IRS Data set for the return feature

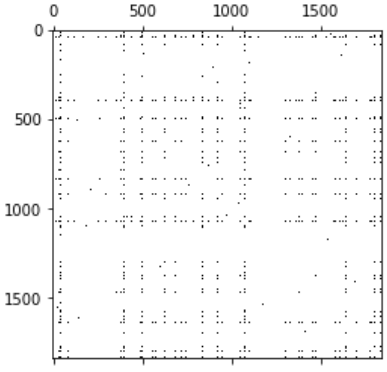

Figure 6: The adjacency matrix of the IRS Data set for the exempt feature