Analysis of Migration in India

Anchit Gupta
Indraprastha Institute of Information Technology, Delhi
India

Ankit Agarwal
Indraprastha Institute of Information Technology, Delhi
India

ABSTRACT

Migration network is one of the networks which were less studied in the area of Graph Machine Learning. Specially in India as most diversed country in world shows large migration and recorded each decade. So, in this we represent the migration happening between different states and union territories of India as a network and then apply different techniques to identify some pattern in the migration. We apply weighted link prediction methodology to get the projected counts of migrants for different areas of the country. Next we obtain the node embeddings of the migration network using Graph Convolution Network (GCN) by incorporating node features like in-migrants count, out-migrants count, population, etc. of the state/UT which are represented as nodes. Then we showed that count of migrants to a place is related to the literacy rate, GDP of the place. This kind of analysis can be of great use for the Government agencies involved in planning the development of cities for the growing population.

KEYWORDS

datasets, migration network, weighted pagerank, weighted Link Prediction, graph neural network

ACM Reference Format:
Anchit Gupta and Ankit Agarwal. 2020. Analysis of Migration in India. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

We all know that it is quite common for people to move to areas which are better in terms of employment, business, education opportunities. Many of the recent research works have applied graph mining and machine learning to different kind of problems and obtained very good results. So, in our work we have collected and consolidated migration data of India according to the census of 2011. Now we have modelled the migration from one place to another as a directed network. In the last few years researchers have proposed many different graph embedding techniques where embeddings for each of the nodes are obtained while taking into account the relationship the different nodes, node attributes. So, over here we also try to obtain the best possible node embeddings and use them to solve the problems of node classification, weighted link prediction, community detection.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Number of Records in each file	5654
Total Files in Single file Final Dataset	72
Total Records Final Dataset	350 * 3
Total States/UT's as nodes in each network	35
Number of features from it can be mapped	29

Table 1: Main Dataset Summary

2 MOTIVATION

The graph mining techniques have been applied in various reasons from detection of disease in the network to check the user behaviour in the social media platform. It have been great help for the people working in various domains specially in Computer Science field. So, using the same motivation that applying state-of-the-art graph mining techniques in the Inter state migration network of India to get the useful insight which can be helpful in future decision making task.

3 DATASET

For our project we needed the migration counts between the states or UT's(Union Territories) along with their reason for migration. The only source of such data is the Indian census data of year 2001 and 2011. The statewise in-migrants data along with their source place is given as a spreadsheet where the total count of migration as well the counts segregated according to their reason for migration is mentioned. The challenge with this dataset is that these files contain 3 consecutive rows corresponding to each source place i.e. rural, urban, total. For the problem in hand we do not need the rural and urban counts separately. So, the required rows had to be manually selected and a separate spreadsheet created for each of the reasons of migration. Dataset has been created separately for the years 2001 and 2011.

Another data that is used is the state wise GDP, literacy rate, crime rate, employment rate. This information has been taken from Wikipedia articles.Also, the main census data have been taken which includes the population of each state, literacy rate, density, area, decadal growth etc are columns which are present in it.

4 LITERATURE REVIEW AND PREVIOUS WORK

There is no previous work which uses graph mining techniques specifically on migration network. Some of the distantly related works are discussed in this section. However none of these works is on the migration network in India. One of the works [3] which used

the japan cities data shows that the population vs migration follows the power law and indeed it's a scale free network. And in another paper[1] which uses the U.S.A migration data to find the patterns in the data in different scenarios such in major recession session in 2009-2010 and in 2007-2008 to find out the impact on migration during such difficult times. Their work inferred that people had a tendency to move to areas with lower rentals and lower cost of living.

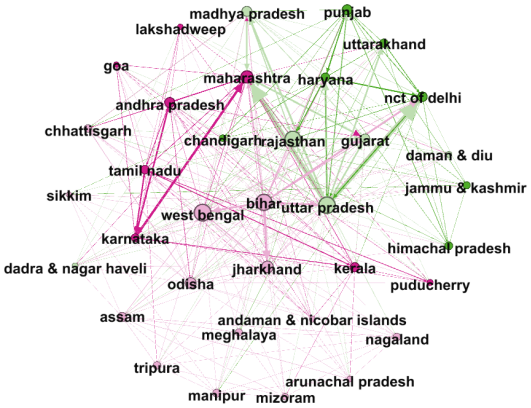

Figure 1: 2011 Total Migration Flow Network

We also developed the network on the basis of inter-state migration on basis of education and business. In these network we observe that the most of the people are moving to states which are having high in-degree (in 2001) in 2011 this kind of prove that their is possibility of the preferential attachment as in graph can be seen with state Maharashtra. Also, we saw that the people are not showing any general trend in 2011 census while in 2001 people have a high migration rate from the states which have relatively low literacy rate and the graph which is generated is near to clique as their is always some people moving from one state to another in any cause. We also happen to see the factors which are leading the migration such as the people are moving to states which are tend to have more HDI in 2001 then 2011.

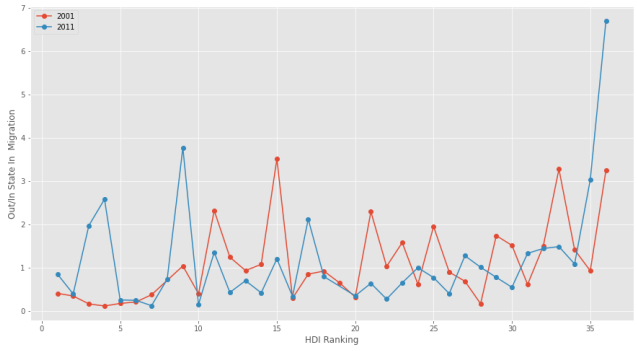

Figure 2: 2011 Total Migration Flow Network

5 METHODOLOGY

5.1 Feature Engineering

The dataset that is being created can't be used directly. As there are many factors come into play.

- **States:** The state names which are used in the different year data are very different as in course of decade the many states got renamed. Like Uttrakhand to Uttranchal. Also, many states comes into existence like Jharkhand which were not present in 2001 but in 2011. So, we have to rename each and every file according to our usage as every time renaming may cost us value able data.
- **Missing Data:** Some data is not present in the dataset. For this we have used the state reports on the migration for later years, meanwhile these values are not absolute these are given as percentage. We have taken those percentage and convert those into relative values. This problem is mainly consist on the data extracted from the census data of the year 2001.

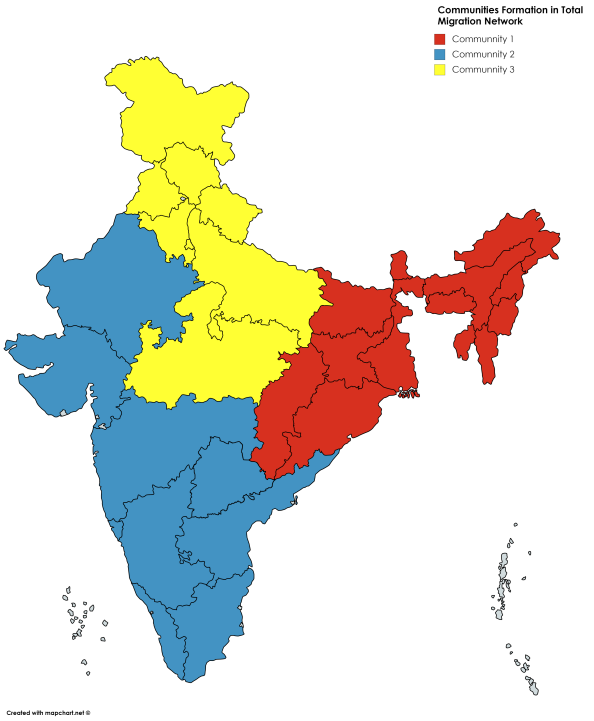

Figure 3: 2011 Total Migration Community Graph

5.2 Methods

This section describes the various methods that we have used on the migration network.

5.2.1 ***Network Analysis.*** We have done exploratory analysis of the data by representing the data using the different graphs and drawing conclusions using different algorithms like PageRank, HITS algorithm. We have created separate migration networks for

Feature Name	Description
Weighted Indegree	Sum of weights of incoming edges
Weighted Outdegree	Sum of weights of outgoing edges
PageRank	Simple PageRank Scores
Population Size	No. of people Residing in that state
Area	Area of the State
Density	Population Density

Table 2: Parameters Used in Embedding

the different reasons of migration like education, business, females migration. By considering the network with edge weights we have visualised the migration using a random walker and calculated the PageRank of the different nodes. We use HITS algorithm to find out the relation between the in and out migration of a state. Thereafter we use the modularity maximization algorithm to detect communities in the network. We have special attention to the network based on migration of females and migration due to education.

5.2.2 ***Node Classification using GCN.*** Now based on the literacy rate of the state we assign labels to each state denoting the class to which the state belongs i.e. high literacy rate or a low literacy rate. Similarly we assign labels based on the GDP of the state. We have used the deep learning method is used to for classification purpose. The node embeddings obtained using [2]GCN are used as the feature vectors of the nodes. First, we experimented by using all the node attributes mentioned in Table 2 while creating the node embeddings but the results obtained were not good. So, we used only 4 node attributes that are population size, area, density and decadal growth for getting the embeddings. The classification task was performed using a 2 layer shallow neural network.

5.2.3 ***Weighted Link Prediction.*** One of the major problems with the traditional link prediction methods like Common Neighbour, Adamic Adar and Resource Allocator is that they do not give any consideration to the edge weights i.e. the graph is treated as unweighted. So, we use method described by [3]Xiao in which he uses the weights of links for link prediction task by modifying the traditional Common Neighbour, Adamic Adar and Resource Allocator. It first creates a simple model which uses the existing methods and remove constant 1 and replace by weights from node pair to that common node.

$$s_{xy}^{WMI-CN} = \sum_{z \in O_{xy}} (W_{xz} + W_{zy}) I(L_{xy} : z) \tag{1}$$

$$s_{xy}^{WMI-AA} = \sum_{z \in O_{xy}} \frac{W_{xz} + W_{zy}}{1 + log(S_z)} I(L_{xy} : z) \tag{2}$$

$$s_{xy}^{WMI-RA} = \sum_{z \in O_{xy}} \frac{W_{xz} + W_{zy}}{S_z} I(L_{xy} : z) \tag{3}$$

$$I(L_{xy} : z) = \sum_{z \in O_{xy}} \left(-log \frac{M^T}{M} + \frac{N_\triangle}{N_\triangle + N_\wedge} \right) \tag{4}$$

where O_{xy} represents the common neighbor set of node pair (x, y). z is the common neighbour. W_{xz} is the weight of link (x, z). S_z denotes the strength of node z, i.e., the sum of weights of links directly connected with node z, which is defined as $S_z = \sum_{z \in (z)} W_{zz}$. And $I(L_{xy} : z)$ is the conditional self-information of the event that node pair (x, y) have one link given that their common neighbor z is available. It is used to calculate the link likelyhood.

Now to enhance the accuracy form it we introduce the a free parameter *a*. Using this parameter we try to enhance the prediction from the model. We know in a network in link prediction problem sometimes a weak ties can play a important role even more than that of strong ties. So, using *a* we try to control relative influence of weak ties. So, equation becomes:

$$s_{xy}^{WMI-CN_a} = \sum_{z \in O_{xy}} (W_{xz}^a + W_{zy}^a) I(L_{xy} : z) \tag{5}$$

$$s_{xy}^{WMI-AA_a} = \sum_{z \in O_{xy}} \frac{W_{xz}^a + W_{zy}^a}{1 + log(S_z)} I(L_{xy} : z) \tag{6}$$

$$s_{xy}^{WMI-RA_a} = \sum_{z \in O_{xy}} \frac{W_{xz}^a + W_{zy}^a}{S_z} I(L_{xy} : z) \tag{7}$$

6 RESULTS

When we perform community detection using modularity maximization, there are a total of 3 communities formed communities formed are continuous regions as shown in figure 3. The 3 communities that are formed are the north, south-west and east regions of India. The interesting point to note here is that the communities formed according to the geographical location even when there is no mention of the geographical location of the state in the dataset used. One other observation is about the migration between states belonging to different communities. Most of the higher values of migration have a state belonging to community 2 as the destination state. The reason for this that most of highly developed states lied in community 2. Also, the inter community migration involving community 2 as the destination is greater as compared to communities 1 and 3 as the destination.

Next analysis that we do is on the migration of women specifically. For this we networks of women migration for education and business purpose. We do this by using the PageRank algorithm to analyse the random behaviour of the women in the network. As shown in figure 5 of 2011 education migration data it can be clearly seen that Maharashtra is one of the popular destinations of migration. Also from figure 4 of 2001 education migration data

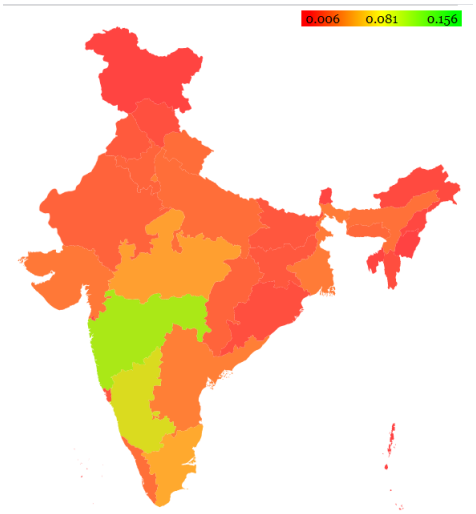

Figure 4: Education Female Flow 2001

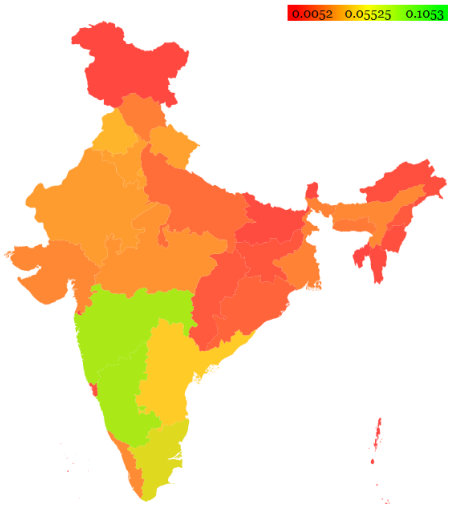

Figure 5: Education Female Flow 2011

Target Label	Network	Accuracy Scores
Literacy Rate	Total	0.52
	Education	0.61
	Business	0.54
GDP	Total	0.51
	Education	0.55
	Business	0.53

Table 3: GCN Results

Top 5 State 2011	Top 5 State 2001
Maharashtra	Maharashtra
Karnataka	Karnataka
Tamil Nadu	Tamil Nadu
Andhra Pradesh	Madhya Pradesh
Chandigarh	West Bengal

Table 4: Top 5 State in Page Rank Score in Education Migration Data

we have seen the there is rise of Rajasthan as popular destination of migrants travelling for the purpose of education. It is due to of the rise of the Kota city in Rajasthan as an education hub and some emerging women colleges.

Also in table 4 Maharashtra, Karnataka and Tamil Nadu are the states which maintain their position while West Bengal and Madhya Pradesh dropped down to positions 15th and 10th in 2011 from 2001. One of the reasons might be that in these states in between 2001-2011 the crime rate became 7-8 % more than the National Average. We also see that according PageRank scores, Gujarat in business migration network rose from 16th position in 2001 to 2nd in 2011. The reason being the high rank of Gujarat among all states according to the ease-of-doing business index in 2011. In the link prediction task we have used the weighted linked prediction method by Xia Y. We were able to obtain the ROC of the 0.69 in the business migration network . Most interesting part of using this algorithm is that how the AUC ROC value increases when the influence of the weak ties in the network is increased as seen in Figure 6 that the highest score is obtained when a=-1 where the weakest tie is most important tie in the calculation.

So, we can say that even though ties happening or flow of migration may be random in the network but the weak ties also hold a strong position while performing link prediction on the migration network.

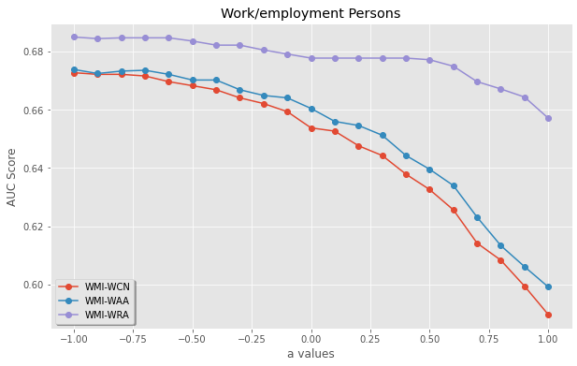

Figure 6: AUC ROC of Employment Link Prediction Task

We use the node embeddings obtained using GCN to perform the task of node classification based on the literacy rate and GDP values. The accuracy scores are shown in Table 3. From these results it can be inferred that the migration into a state does have an impact on the literacy rate, GDP, crime rate but just by using data from 2001 and 2011 it is a bit difficult to give a strong statement about its impact.

7 CONCLUSION AND FUTURE WORK

Here in our work we have drawn insights from the migration network using graph mining methods and made some interesting observations using community detection, link prediction, GCN for node embeddings. Using link prediction we have inferred about the role of weak ties in the migration network. This work can also be extended to the district level from state level and better results can be obtained once the data of Indian census 2021 becomes available. And also analysis can be showed on the inter country migration of people and to be connected with other country migration to extend the analysis part.

REFERENCES

[1] Travis Goldade, Batyr Charyyev, and Mehmet Gunes. 2018. Network Analysis of Migration Patterns in the United States. 770–783. https://doi.org/10.1007/978-3-319-72150-7_62

[2] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[3] S. Tomita and Y. Hayashi. 2006. Spatial Analysis of Centralization and Decentralization in the Population Migration Network. In *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation - Volume 60* (Tokyo, Japan) *(APVis '06)*. Australian Computer Society, Inc., AUS, 139–142.