Internal Migration between US States - A Social Network Analysis

EXTENDED ABSTRACT[1]

Gunther Maier, Michael Vyborny

Vienna University of Economics and Business Administration (WU-Wien)
Vienna, Austria

Abstract

In this paper we use the novel (at least in regional science) technique of social network analysis and apply it to one of the most analyzed topics in the discipline, US internal migration. We want to see whether social network analysis can yield any new insights into this well known process. We want to compare the technique to more conventional methods of analysis in migration. The paper will give an overview of recent literature about internal migration between US states and summarize the main findings. It will then present an overview of social network analysis, define key concepts and describe the main components of the technique. This discussion will also involve a discussion of currently available software for social network analysis. Then, we will apply the technique to the official data about internal migration between US states as published by the US bureau of the census, to see whether the technique can reproduce the main results of the traditional techniques and whether it can yield any new insights.

1. Introduction

This extended abstract is supposed to give an overview of the paper for the European congress in regional science 2005 in Amsterdam we are currently working on. Due to competing requests for our time and the unusually early deadline for the conference papers, we were unable to submit the final version of the paper in time for the CD-ROM. Nevertheless we are convinced that social network analysis (SNA), the method used in our paper, is a potentially very useful technique in regional science and should be considered more widely in the discipline. Therefore, we decided to invest the time into writing this extended abstract, in order to include at least a rough sketch of our work into the conference CD-ROM. The final version of the conference paper will be available at the time of the conference at http://wuw-sre.wu-wien.ac.at/dpprice.html .

In this paper we want to bring the method of social network analysis to one of the most fundamental topics of regional science, the analysis of migration. Our main concern is not so much to generate new insights into migration, but to explore, to what extent SNA is suitable for the application to a topic like migration. In order to test the approach, we use migration between the 50 states and the District of Columbia in the US. The basic information we use in this paper is the migration table 1995-2000 as published by the US-census (http://www.census.gov/population/cen2000/phc-t22/tab03.xls).

In the following section we will discuss the relationship between migration and networks. We will see that different migration tables could be used for the analysis. In section 3 we will describe the basic concepts of SNA and give an overview of the key methods. In section 4 we will apply some of these methods to internal migration between US states.

2. Migration and networks – basic terms and concepts

Our analysis will be confined to internal migration between the 50 states of the US and the District of Columbia between 1995 and 2000. No flows of population from or to foreign countries will be taken into account. Internal migration only considers flows that originate and terminate in different US states. This information is generated such that in 2000 the respondents of the census are asked where they lived five years ago. When the respondent

[1] The final version of this conference paper is available online at http://wuw-sre.wu-wien.ac.at/dpprice.html

lives in state *B* at the time of the interview and reports to have lived in state *A* five years ago, this person is a migrant from state *A* to state *B*, irrespective of whether the person has resided in one or more other states during this five year period. Internal migration deals with the redistribution of a constant population. The sum of people counted at their state of residence in 1995 and in 2000 gives the same number. People who were born, who died, or who migrated to the US from abroad during this five year interval do not show up in this information.

This information can be arranged in form of a matrix. The migration matrix arranges rows and columns named by states and in each field of the matrix shows the number of people who migrated – according to the above definition – between the state that labels its row and the one that labels its column. Thus, the migration matrix shows some form of interaction between the states. Since SNA focuses on the relationship between actors, it also deals with interaction and might therefore be suited for analyzing migration. The interactions analyzed in SNA are typically characterized in matrix form as well. In SNA this interaction is usually called a "network". This term is meant in an abstract sense as a set of actors and the relationships between them. In this abstract form the term has nothing to do with social networks like networks of friends and relatives, which are important influencing factors for migration behavior. We apply the term "network" in this abstract sense, despite the fact that SNA can of course be applied to the analysis of such social networks and has actually been named after this type of application.

The migration matrix showing the number of people who have migrated between states is the raw data set of our analysis. One important technical aspect in migration analysis as well as in SNA is the treatment of the main diagonal of the matrix. The migration matrix of the US census shows in its main diagonal the number of people who lived in the same state at both time periods. Therefore, the row and column sums show the distribution of the population in the matrix at the beginning and at the end of the time period. Since SNA typically ignores diagonal information, we set these cells to zero in all our analyses. Therefore, the row and column sums give the number of people migrating from and migrating to the states. If we denote the migration matrix with the main diagonal set to zero by *M*, we get the following relationships, where *OM* is the vector of out-migrants and *IM* is the vector of in-migrants. *I* is the unit vector, i.e., a column vector with all ones.

$$OM = M * I, \qquad\qquad IM = I' * M$$

The migration matrix *M* represents the first case of a network that we can analyze by SNA. In terms of graph theory, *M* characterizes a values digraph, since the values in matrix elements M_{ij} and M_{ji} are typically different. The graph is valued since the values in the matrix elements represent more than just a relation exists or does not exist.

One of the problems with the migration matrix *M* is that the states are very different in population size. The largest state in terms of population, California, has with 33.8 Mio almost seventy times the population of Wyoming (493,782), the state with the lowest population. This makes a comparison between the states problematic. Therefore, the migration matrix has been standardized in a number of ways. We can divide each element of the matrix by its respective row sum, such that each row in the standardized matrix adds up to the value one. This removes the disturbing influence of the different size origins. The elements in each row of this matrix show, where the migrants leaving a specific state go. They can be interpreted as the probability with which a migrant leaving a certain state will end up in the other states. We denote this matrix by *mo*, the migration matrix standardized to out-migration. Similarly, we can standardize the matrix by dividing each cell by its respective column sum, setting all the columns sums to one. This removes the disturbing influence of the different size destinations. The elements show, where the in-migrants to a certain state come from. They can be interpreted as the probability that a migrant ending up in a certain state comes from another state. We denote this matrix as *mi*, the migration matrix standardized to in-migration.

Because of the different sizes of origin and destination state, the two matrices *mo* and *mi* usually differ. Each standardization removes one part of the problem, but not all of it. The matrix elements will add up to one either over the rows or the columns, but usually not over both dimensions at the same time. Bi-proportional adjustment is a technique that can generate such a matrix. It applies the two adjustment procedures repeatedly to the matrix until the matrix shows the desired row and column sums sufficiently well. This technique is also called RAS-procedure and is "one of the most popular methods for adjusting input-output, social accounting, and demographic matrices" (Okuyama, et al., 1998). In the migration context it can be applied to standardize row and column sums to identical values, usually one. This standardization removes the scale difference at the origin and destination side of the migration flows simultaneously. We denote the resulting matrix by *moi*, the migration matrix standardized simultaneously to in- and out-migration. It shows the spatial structure of the migration. We will use the matrices, *M*, and *moi* in section 4 of the paper.

An alternative to bi-proportional adjustment would be the use of a saturated log-linear model. It allows us to split up the migration matrix into four parts: a scale factor, an origin factor, a destination factor, and an interaction factor. The first one captures the overall scale of the information. The origin effect captures the origin specific component in the migration matrix, the destination factor the destination specific one. The interaction effect finally contains the interrelation between the states net of the other three effects. So, in order to look at the spatial structure of the migration matrix, we would concentrate on the matrix representing the interaction effect.

3. Social Network Analysis – a brief overview

Social network analysis (SNA) is a technique for visualizing, describing and analyzing a system of relations. The main features of SNA include the analysis of empirical data, interactions of social actors, visualizations and mathematical models (Freeman, 2004, p. 3). Those characteristics have developed since the dawn of this discipline in the 1920s. The developments in computer hardware and software – particularly graphical computing – since the 1970s have enabled researchers to make significant progress. Software has been developed to cope with large matrices and algorithms have been designed to visualize graphs in an efficient way. The program UCINET that will be used in our analysis dates back to 1983 and has been constantly improved (Freeman, 2004, p 139-140). This rise in the application of ICT also stimulated the application of the technique.

SNA is solidly rooted in graph theory. Frequently, networks are displayed in the form of graphs with nodes representing the actors or units of investigation and links representing the relations between them. The nodes may represent people (in a network of friendship relations, for example), firms (network of innovation flows), states (trade or migration flows), journals or articles (citation network), etc., to give a few examples. The links in these examples may represent sympathy, transfers of innovation, trade flows or number of migrants, and citations. SNA is most developed for simple, non-directional, dichotomous networks. These networks represent only one relationship and show, whether a relationship exists between a given pair of nodes or not.

A basic form of use of SNA is to display the network visually. Major advances have been made in this area in recent years due to the developments of graphical systems. The software places the nodes as geometrical figures on screen and characterizes the relations between them by lines. Shape, size, and color of nodes and relations can be adjusted to reflect attributes of actors or a specific relation. This visualization by itself can add a lot to the understanding of some relationship under investigation. Different algorithms are in use for locating the nodes somewhat “optimally” on screen. Criteria for optimization of graph would include lines of similar length, minimum number of crossing lines, angles that are not too small and vertexes that are not too close to the line. The purpose of those characteristics is to generate a graph that is easy to read, where overlap does not exist.

SNA has developed a number of indicators to describe a network and its nodes. All these indicators focus on the relationships between the nodes. Simple indicators, for example, are the degree of nodes, i.e., the number of direct relations a node has with other nodes in the network. In the graphical representation of a network the degree is equal to the number of lines connected to the respective node. This measure can be extended to directed data. In this case two types of degree, indegree and outdegree, can be calculated. Those measures indicate how often a given node is adjacent to or adjacent from other nodes (Wasserman/Faust, 1999, p. 125). A normalized measure for degree exists that adjusts the results according to network size. The degree is divided by the maximum possible degree and is expressed as a percentage. This normalization procedure should only be used for binary data, since might not be meaningful for valued data (Borgatti et al., 2002). The density of a network is the number of links in a network divided by the maximum number of links possible in this network. Such indicators can be used to answer questions like “what is the most central node in the network” or “how well connected is the network”.

An important aspect of SNA is the connectedness of a network. Two nodes are connected, when we can follow an uninterrupted series of links from one node to the other. This series of links is called a path; the number of links we need to travel is called the distance between the two nodes. When there are nodes in a network that are not connected, the network can be divided into components. A component is a part of a network where every node is connected to every other node, but not connected to any node outside the component. Identifying components is an important step in SNA. Distances between nodes in the same component are finite; distances between nodes in different components are infinite or undefined.

Also within a component we can identify different forms of subgraphs through SNA. Cliques are particularly important subgraphs. According to the definition of Wasserman and Faust (1999) “a *clique … is a maximal complete subgraph of three or more nodes*”. Maximal complete means, that every actor in the clique is connected to all other actors in the clique. For a clique with 3 actors there are 3 links, for 4 actors 6. More generally, a clique with n nodes has $n(n-1)/2$ links. Cliques represent robust structures, since there is a great deal of redundancy of links. Therefore it might be useful to identify cliques to uncover the “core” structures of a graph. Cliques are one of the strictest forms of cohesive subgroups that can be identified, since the removal of a single link will result in the destruction of the clique.

Hierarchical clustering can be used to assign actors to specific subgroups. According to the chosen criterion clusters can be formed following the interpretation of the researcher. In the case of the analysis of this paper similarity has been chosen as the appropriate criterion.

Once subgroups are identified we can identify different roles that actors play in the network. According to Gould and Fernandez (1989) five roles are possible. Coordinators broker within their own group. Consultants indirectly link actors that belong to the same other group, but cannot reach each other directly. Gatekeepers have incoming relations from another group and outgoing relations to members of their own group. In this respect they are gatekeepers, since they’re the ones who may filter inflows. Representatives have the opposite role. They represent their own group to another group. Liaison actors broker ties where all actors are members of different groups. The table below shows the roles with the corresponding positions involved. The relevant actor is always the second one (printed in bold)

Role	Positions
Coordinator	A → **A** → A
Consultant	B → **A** → B
Gatekeeper	B → **A** → A
Representative	A → **A** → B
Liaison	B → **A** → C

Another interesting approach to measure the connections among groups is to measure the E-I index. This index measures the ratio of ties within and between groups. It may range from -1 to +1. A value of -1 indicates that there are only ties within the respective group. A value of +1 means that only external ties exist. A value of zero indicates a balance between internal and external ties. The index is calculated for individual actors and the groups. The results for the group can be displayed in a density matrix that can be visualized to show the ties between groups instead of single actors (Borgatti et al. 2002).

4. Applying SNA to the analysis of migration

In section 3 we have discussed various versions of migration matrix that could be used in the analysis. We will begin by analyzing the raw migration matrix M the elements M_{ij} of which give the number of migrants from state i to state j. This part is mainly included because it shows the common knowledge of interstate migration in the US. The more detailed analysis will use *moi*, the matrix simultaneously standardized to in- and out-migration.

4.1. Using the raw migration matrix

Figure 1 shows the network generated from the raw migration matrix M. The figure demonstrates some of the visualization capabilities of contemporary SNA software (the figure is generated by NetDraw version 2.4). The width of the lines is proportional to the number of migrants, the size of the nodes to the absolute value of net-migration. The color of the node reflects whether this state is gaining (yellow) or losing (red) population because of internal migration. In addition to size and color we could also set the shape to the node indicator according to another attribute.

We see clearly from this figure that New York and California are major sources of migration, and that Florida is a major destination. The largest migration flow is from New York to Florida with over 308,000 migrants, represented in the figure by the thick line between these two states.

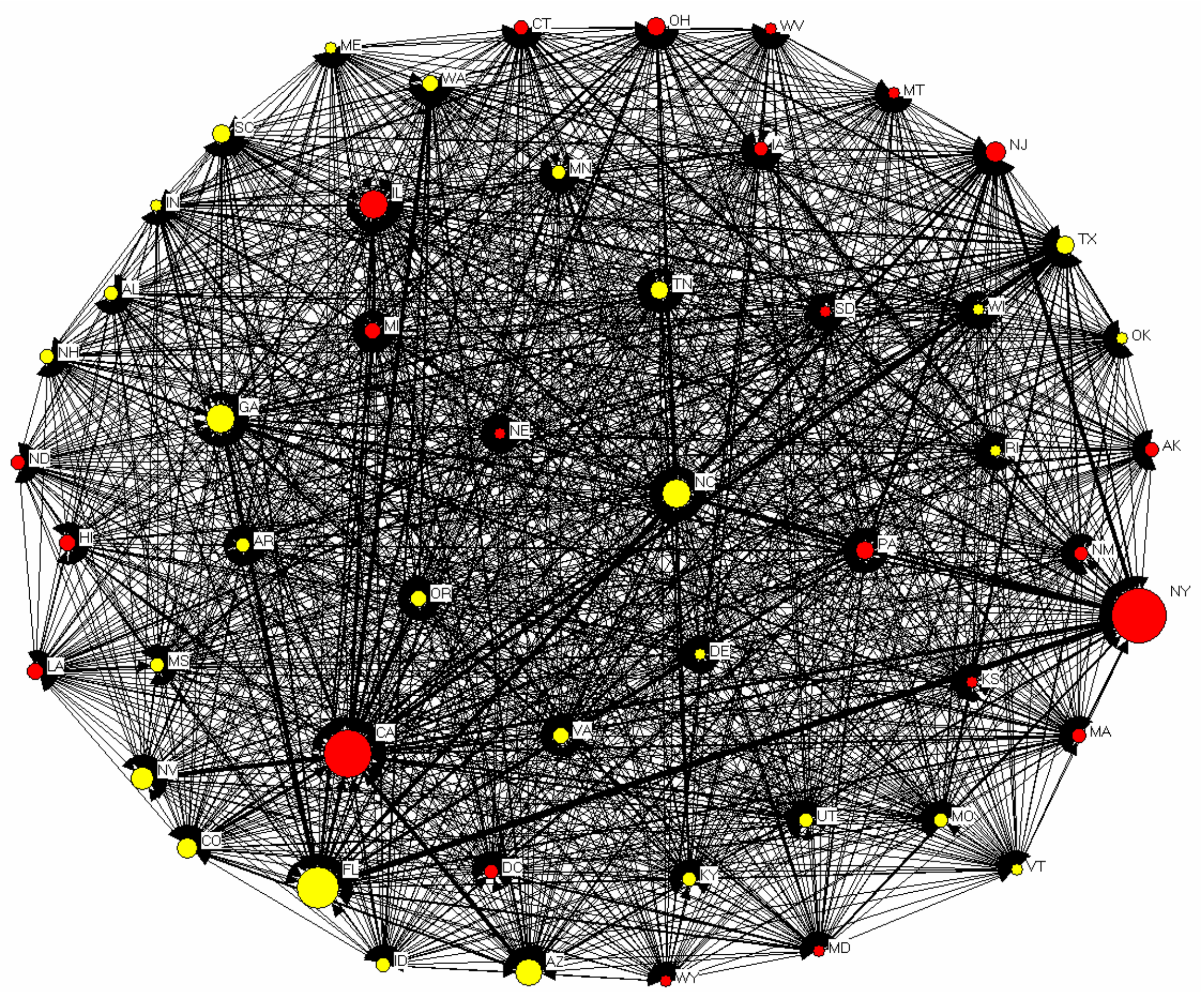

Figure 1:

One conceptual problem with the application of SNA to migration also becomes apparent in this figure. SNA is typically applied at a much more disaggregate level than US-states, showing whether a relationship between two actors exists of not. In the case of interstate migration, the actors are states and at this level migration flows usually exist between most states. In the migration matrix for 1995-2000 of the 2550 cells only one, that representing migration from Rhode Island to North Dakota, is empty. Therefore, the density of the resulting network is very high (0.9996), producing a picture with too many lines to show some structure.

To get more visual information, we have to reduce the density of the network. We can do this by selecting a threshold level and suppressing all links representing migration flows below it. When we lower the threshold level step by step, we can see how the network in figure 1 is built up from the migration links with declining importance.

When we apply this approach to the migration matrix M, we first see two components growing; one on the East-coast and one on the West-coast. In the West California is the major source of migrants. A significant counter flow only exists with Texas. The West-coast component clearly has the shape of a star graph with California being the central node. On the East-coast, the structure of the component is slightly more complex. The Northern states (New York, New Jersey, Pennsylvania) are losing population due to migration, the Southern states (Florida, Georgia, North Carolina) are gaining. Above the threshold level, however, Pennsylvania and North Carolina are only receiving migration (only arrows pointing toward them). The component has no clear source of migration (only arrows pointing from the node) as New York and New Jersey exchange migrants in both directions. Although there are fewer states involved, the component on the East-coast has a more complex structure than the one on the West-coast.

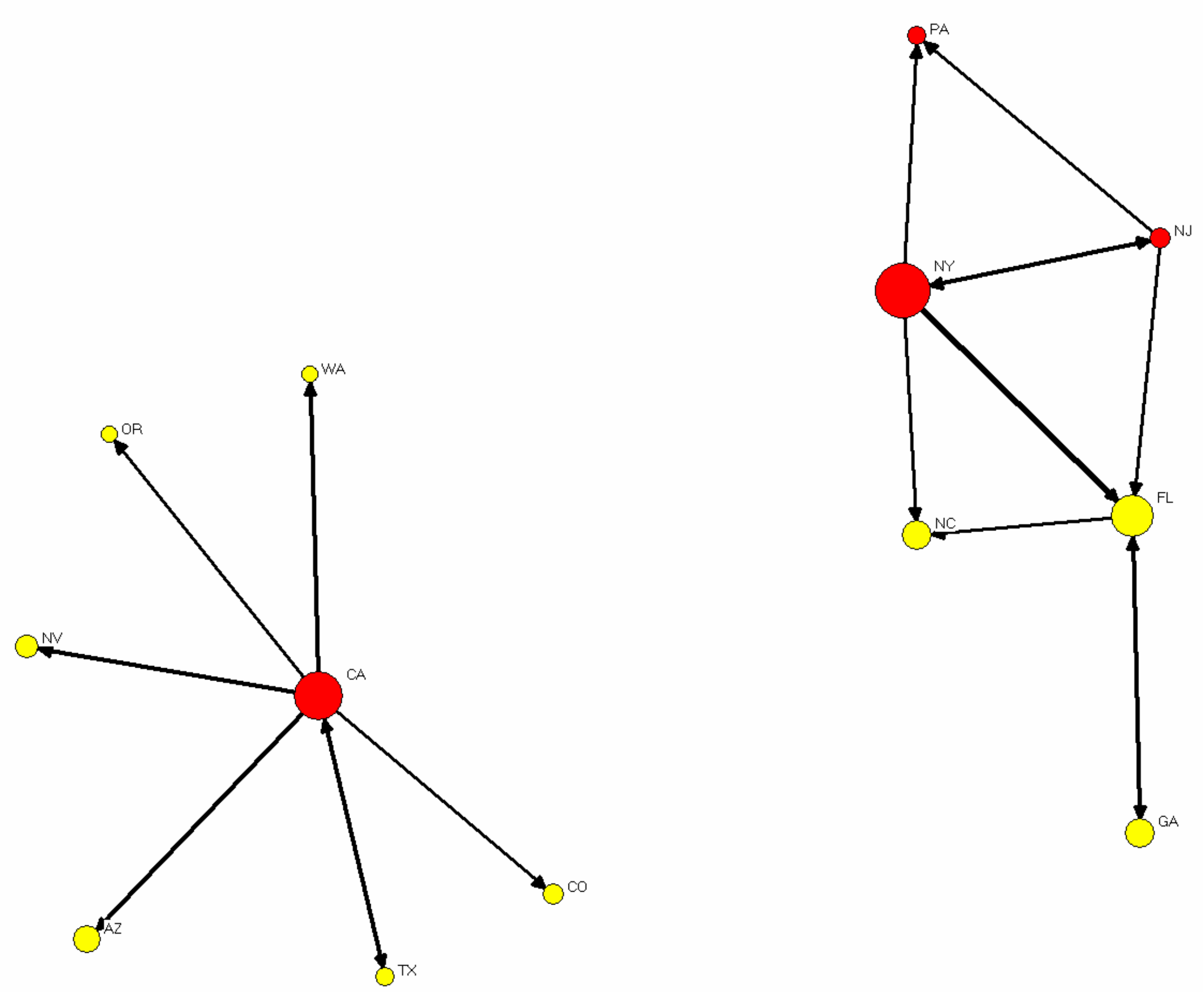

Figure 2:

When we lower the threshold, the two components become connected through a migration flow from New York to California, then through a flow from California to Florida. This triad – California, Florida, New York – forms the central element that links the East-cost and the West-coast sub-networks together for many threshold levels while the two sub-networks become more differentiated. Figure 3 shows the network at a threshold level of 81.000 migrants.

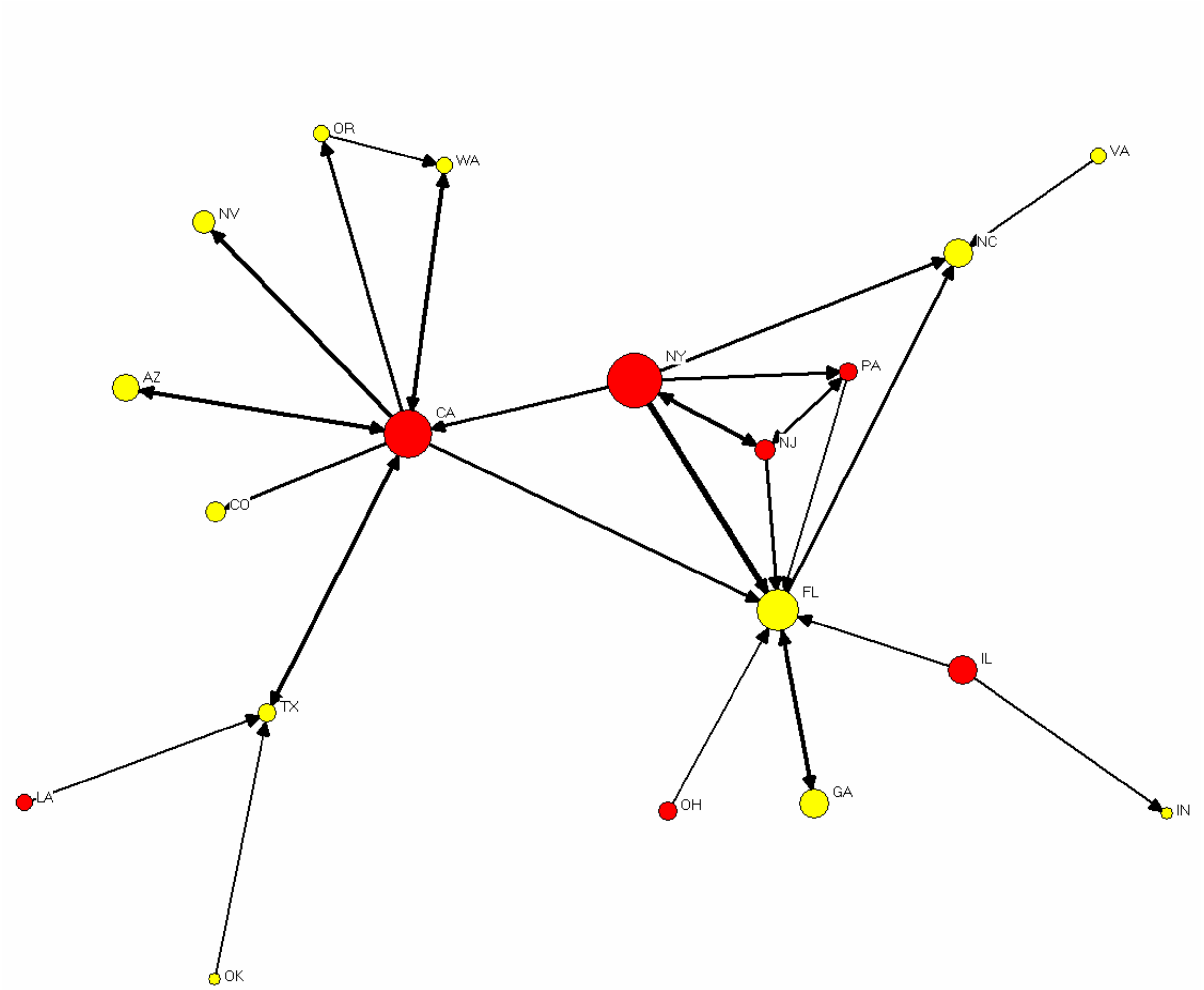

Figure 3:

At later stages Texas and Illinois add to the tissue that links East and West. Figure 4 shows the network at a threshold level of 71,000. Qualitatively, however, little changes. The original division between East- and West-coast is still clearly visible.

It makes sense to look at the degrees of the nodes at this level. Table 1 gives the out- and in-degrees of the states in the network formed at this level of the threshold. The most important sources of migration are New York and California with out-degrees of 8 and 7, respectively, the most important destinations are Florida and California with out-degrees of 10 and 5, respectively. Looking at states with more than one in- or out-going connections, we see that Illinois is a pure source (only outgoing connections), and North Carolina a pure sink (only incoming connections). The absolute difference between in- and out-degree is largest for Florida (with much higher in-degree) and New York (with much higher out-degree).

Table 1: In-degrees and out-degrees

	OutDegree	InDegree
AZ	1	1
CA	7	5
CO	0	1
CT	0	1
FL	3	10
GA	1	1
IL	4	0
IN	0	1
LA	1	0
MD	1	0

MA	0	1
MI	1	0
NV	0	1
NJ	3	2
NY	8	1
NC	0	3
OH	1	0
OK	1	1
OR	1	2
PA	2	2
TX	3	4
VA	2	2
WA	2	2
WI	0	1

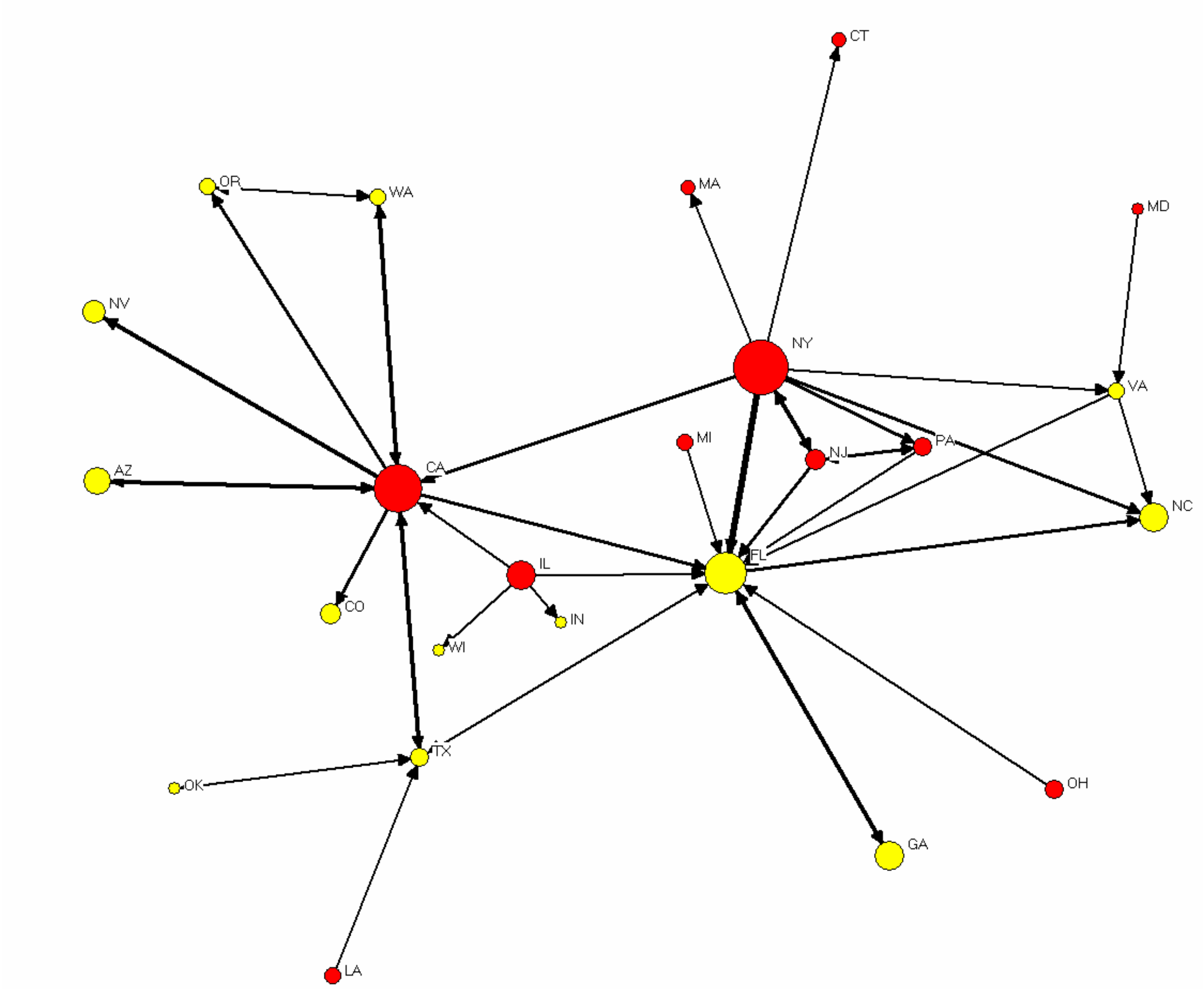

Figure 4:

4.2. Using the standardized migration matrix

As mentioned in section 2, one of the problems with using the raw migration matrix is that the states differ substantially in size. To remove these size effects, we have standardized the migration matrix simultaneously to in- and out-migration by use of bi-proportional adjustment. This yields migration matrix *moi*. We will deal with this standardized migration matrix in the rest of the section.

This standardization basically scales down those links that start or end in states with large numbers of in- or out-migrants. When the link is between a state with a large number of out-migration and one with a large number of in-migration, this think will be down-scaled particularly strongly. We can expect this to be the case for the links between New York and Florida, and between New York and California, for example. Moreover, since we remove the scale effect on both sides, the resulting migration matrix is expected to reflect more strongly the spatial structure of migration. Therefore, we also generated the network of US-states based on adjacency. This network is shown in Figure 5. A line between two states is drawn, when they have a common border. Of course, Alaska and Hawaii are not connected to the component of the contiguous US. The only place where two lines intersect is at the four-states-corner in the Southwest.

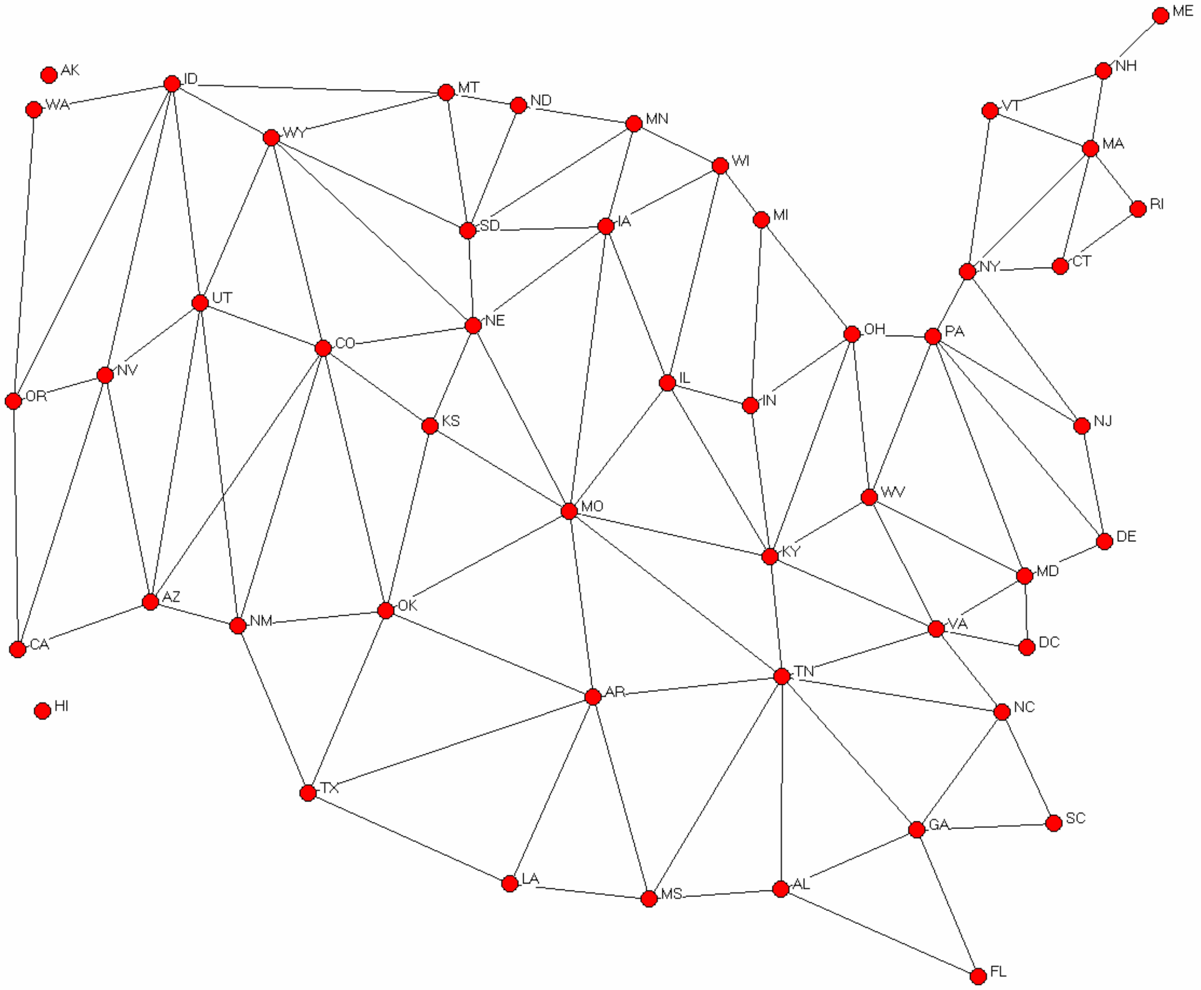

Figure 5:

By far the largest relationship in the standardized matrix is that from DC to Maryland with a value of 0.366. The values in all the other cells are smaller than 0.232. When we apply the same approach as above, the network starts off with a number of small components. At a threshold level of 0.13 we see the picture of Figure 6. The isolated nodes on the left hand side of the figure are those states that are not yet connected to any other state at this threshold level. We see that states that were very important before, like California, Florida, and Texas now remain isolated. On the other hand, the components (with more than one state) are formed from neighboring states. The larger components represent specific regions of the US: New England (Connecticut, Rhode Island, Massachusetts, New Hampshire, Vermont, and Maine), "Greater DC area" (DC, Maryland, Delaware, and Pennsylvania), "Northern Prairie" (Wyoming, Montana, North and South Dakota, Minnesota, and Wisconsin), and "Deep South" (Alabama, Mississippi, Louisiana, Tennessee). Of all the links displayed in this figure, only one (Delaware, DC) connects states that are not adjacent.

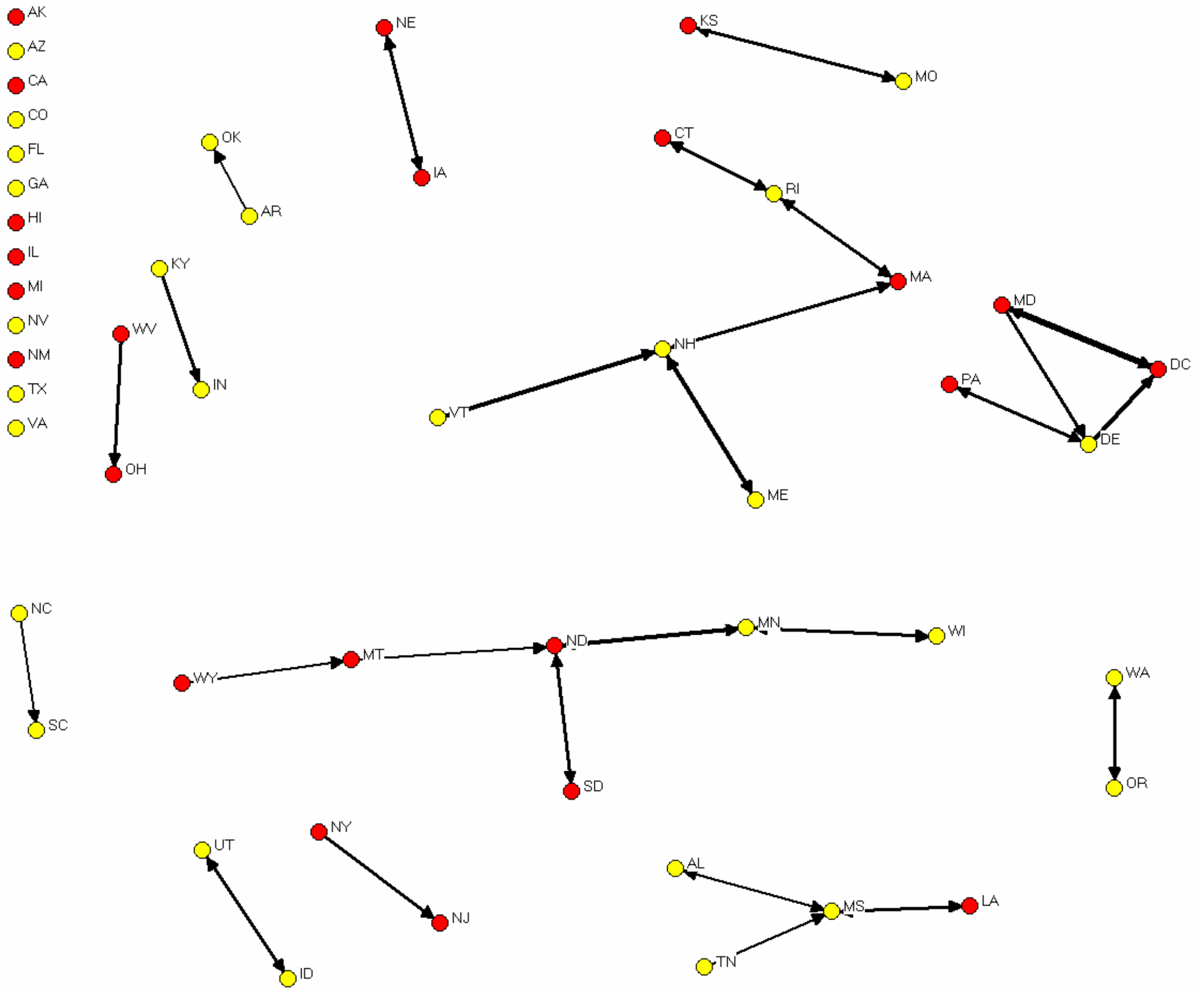

Figure 6:

When we lower the threshold level further, the components get connected and previously isolated states get added, all forming long chains. At the threshold level of 0.092 we get Figure 7. Note that Florida and Texas are still isolated. Besides the isolates (in addition to Florida and Texas: Alaska, Colorado, Hawaii, and Michigan), there are only three components remaining: Arizona, New Mexico; Missouri, Arkansas, Oklahoma, Kansas; and the large component including all the other states. This component forms almost a perfect line graph. It deviates from this structure only in New England, on the East-coast around DC and Pennsylvania, in the Northern Prairies and in the Northwest. The South is connected via a side-arm of the network. Consequently, the density of the network displayed in Figure 7 is very low (0.033). Most of the states have very low in- as well as out-degrees. The largest out-degree is found for Idaho (4), the largest in-degree for South Dakota (5). In average, the states connected to the three components only have 1.89 connections each. Consequently, the index of network centralization is also very low: 4.76% for out-degree, and 6.8% for in-degree.

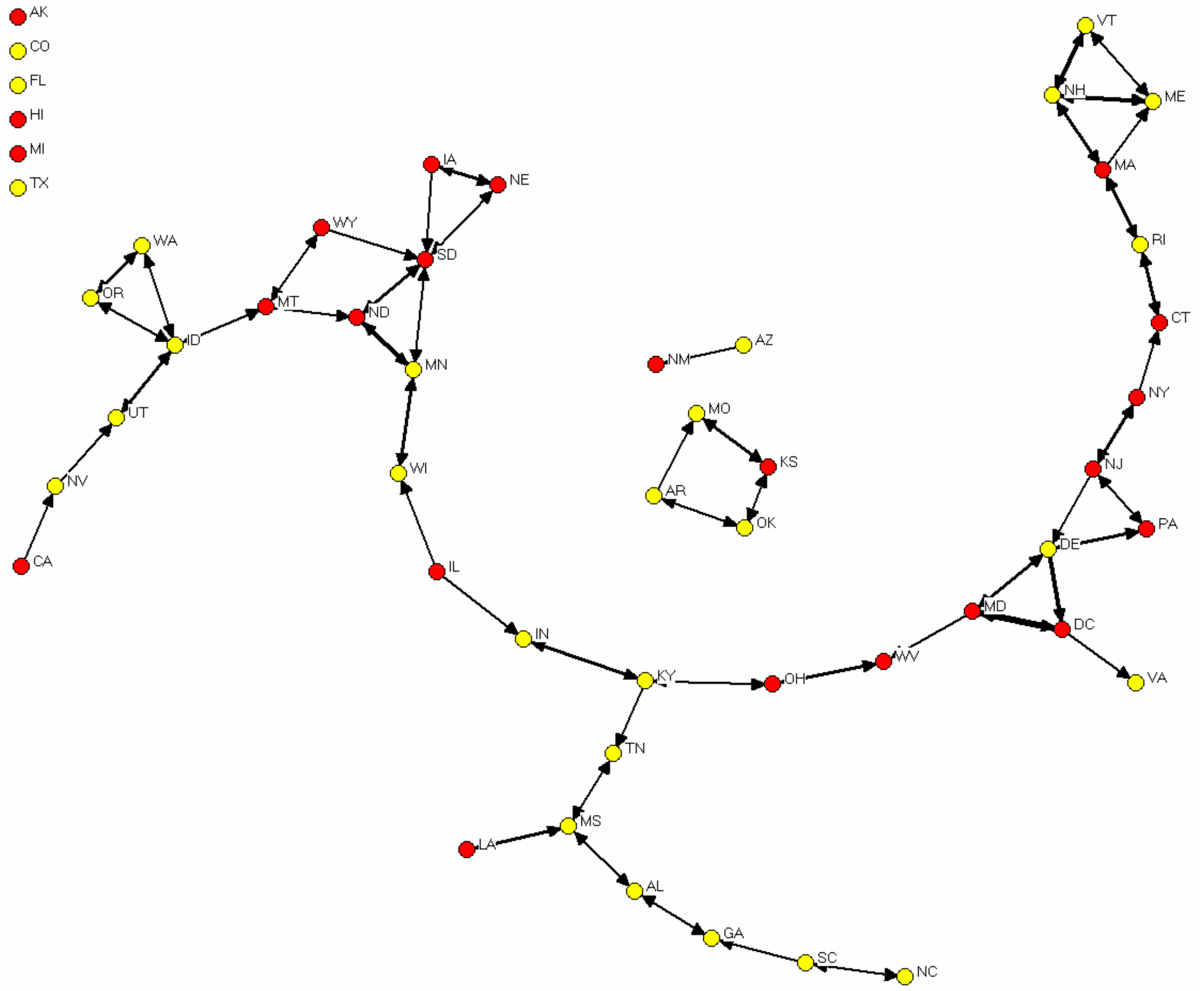

Figure 7:

Florida, the most important destination in the raw migration matrix is the last state to be connected to the network. This happens at a threshold level of 0.066. The network at this stage is shown in Figure 8. Note that New York, New Jersey, Delaware, and West Virginia are so called "cut points" in the network, nodes that if removed would disconnect the network. Additionally, the link between New York and New Jersey has the same property and is called a bridge[2].

From visual inspection we seem to see various groups of states. On the one hand the New England states, on the other hand the states in the West seem to be closely connected among each other and only weakly connected to other states. The area around New York and DC seems to form a group, just as the Northern Prairie states. Visual inspection, however, is not a very reliable method for identifying groups and in our case the picture of the network also depends upon the threshold value we apply. Therefore, we apply hierarchical clustering (Johnson, 1967) to the full *moi* matrix. This technique uses the average of the two sides of each migration relation and treats the resulting values as indicators of similarity. It then groups the most similar states together step by step until at the end all states are merged into one group. The choice one has to make when applying this method is about the number of groups. We have decided to use 8 groups. The clustering sequence is shown in Figure 9, our threshold resulting in the 8 groups is indicated by the dotted vertical line. Figure 10 shows the resulting regions on a map.

[2] Georgia and New Mexico and their respective links to Florida and Arizona have the same properties. Their removal, however, would only isolate one state in each case.

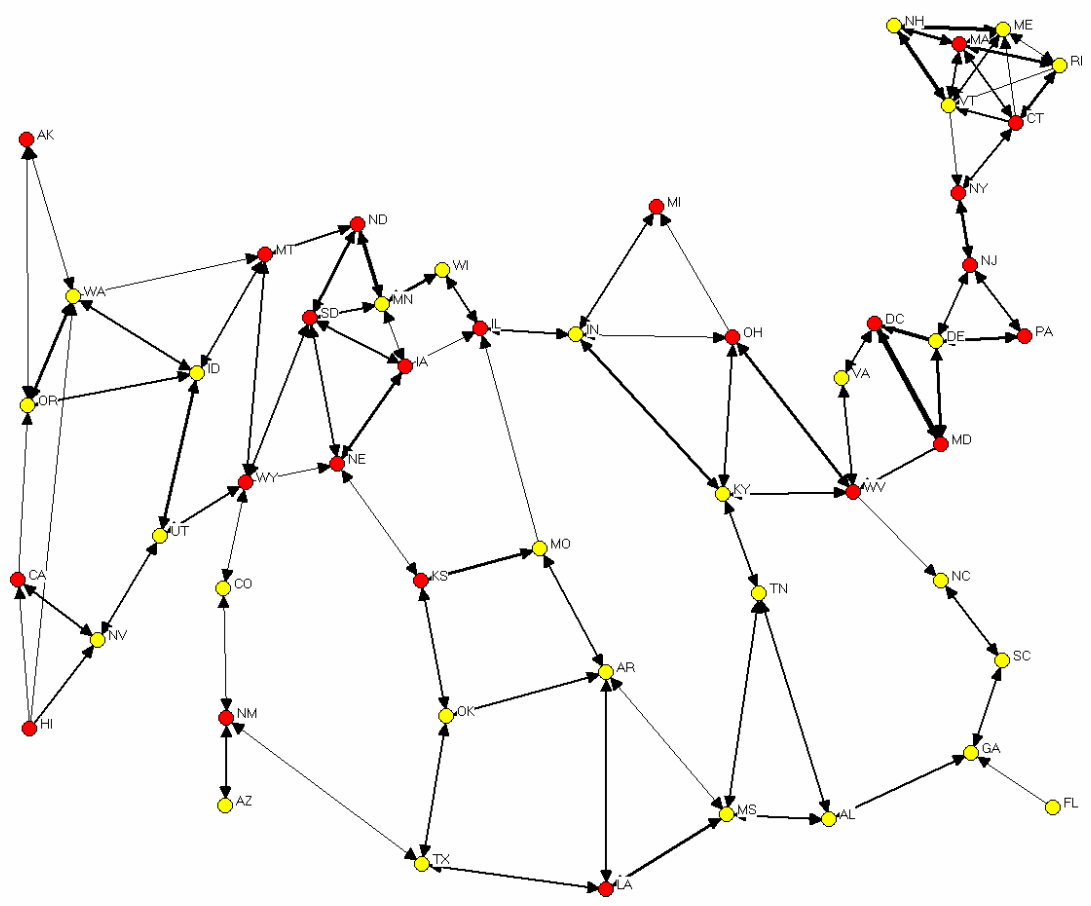

Figure 8:

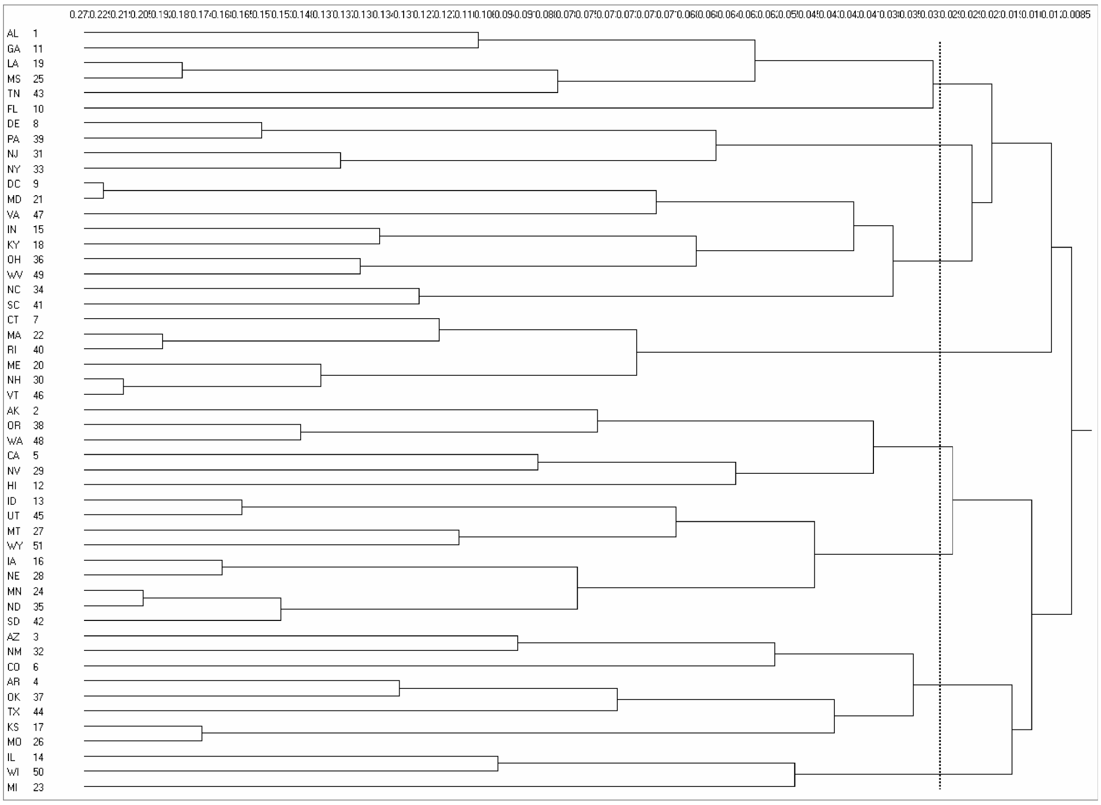

Figure 9:

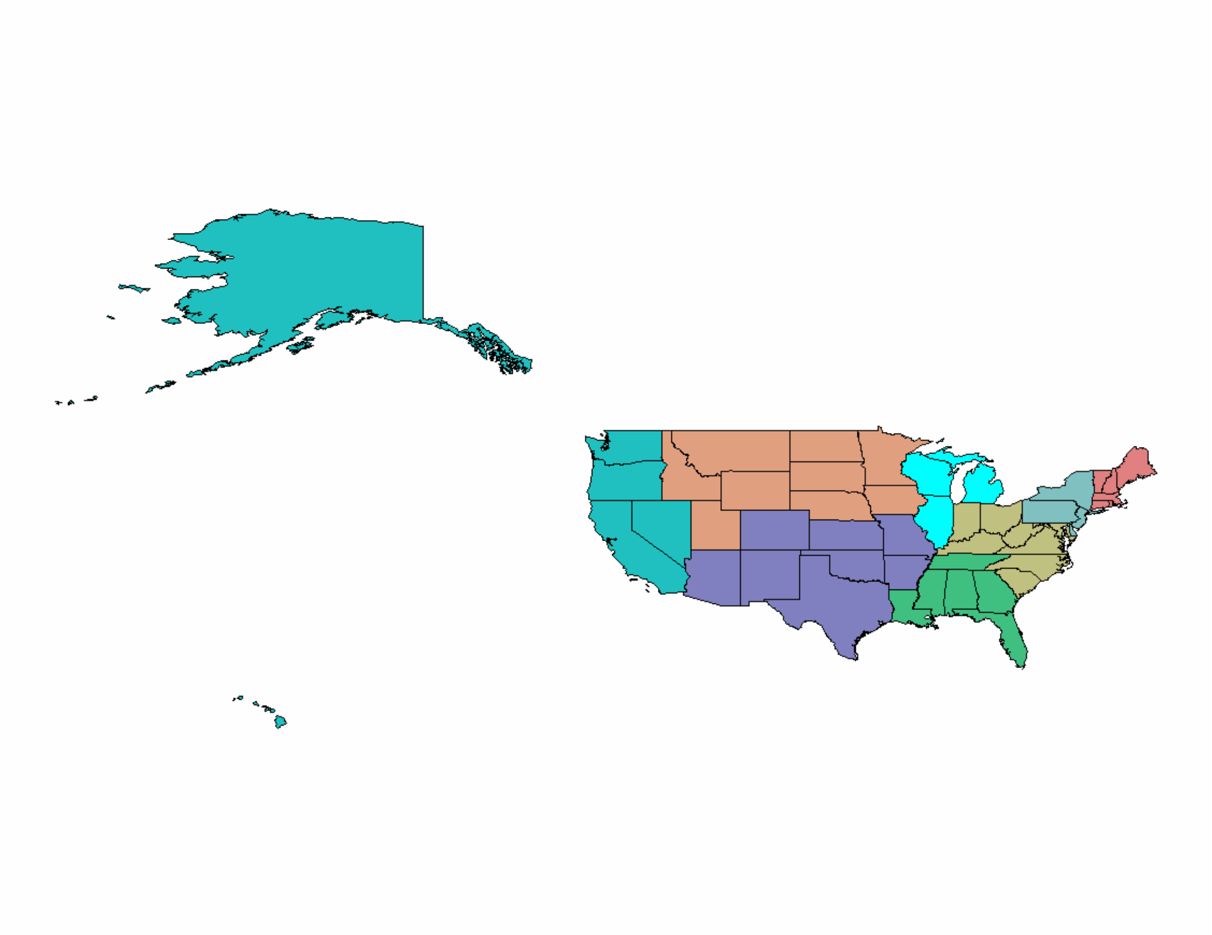

Figure 10:

Figure 11 shows the migration matrix *moi* at a threshold level of 0.05 with the colors of the nodes representing their regional assignment. We see that the regional boundaries are usually running through the less dense parts of the network and that the densely connected states are grouped together into one region. To further check the validity of this regionalization we compute the E-I index that was described in section 3. This index shows to which extent groups are connected internally and externally. The E-I index for the network shown in Figure 11 is -0.355, with an expected value of 0.758. The difference is significant at the 5% level. At the threshold level of 0.05 the states have significantly more ties to states within their groups than to states in other groups. This result also holds for six of the eight groups individually. Only the group with Illinois, Wisconsin, and Michigan has a positive E-I index (more external than internal ties), the group with New York, New Jersey, Delaware, and Pennsylvania has the same number of external and internal ties (E-I index: 0). This probably also results from the fact that these are the two smallest groups.

Figure 11 also shows different roles the states have in this migration relation. The shape of the nodes indicates the main role. We distinguish the following four roles:

- coordinator (circle)
- representative (triangle pointing up)
- gatekeeper (triangle pointing down)
- gatekeeper and representative (square)

The size of the node indicates the number of different roles the state has.

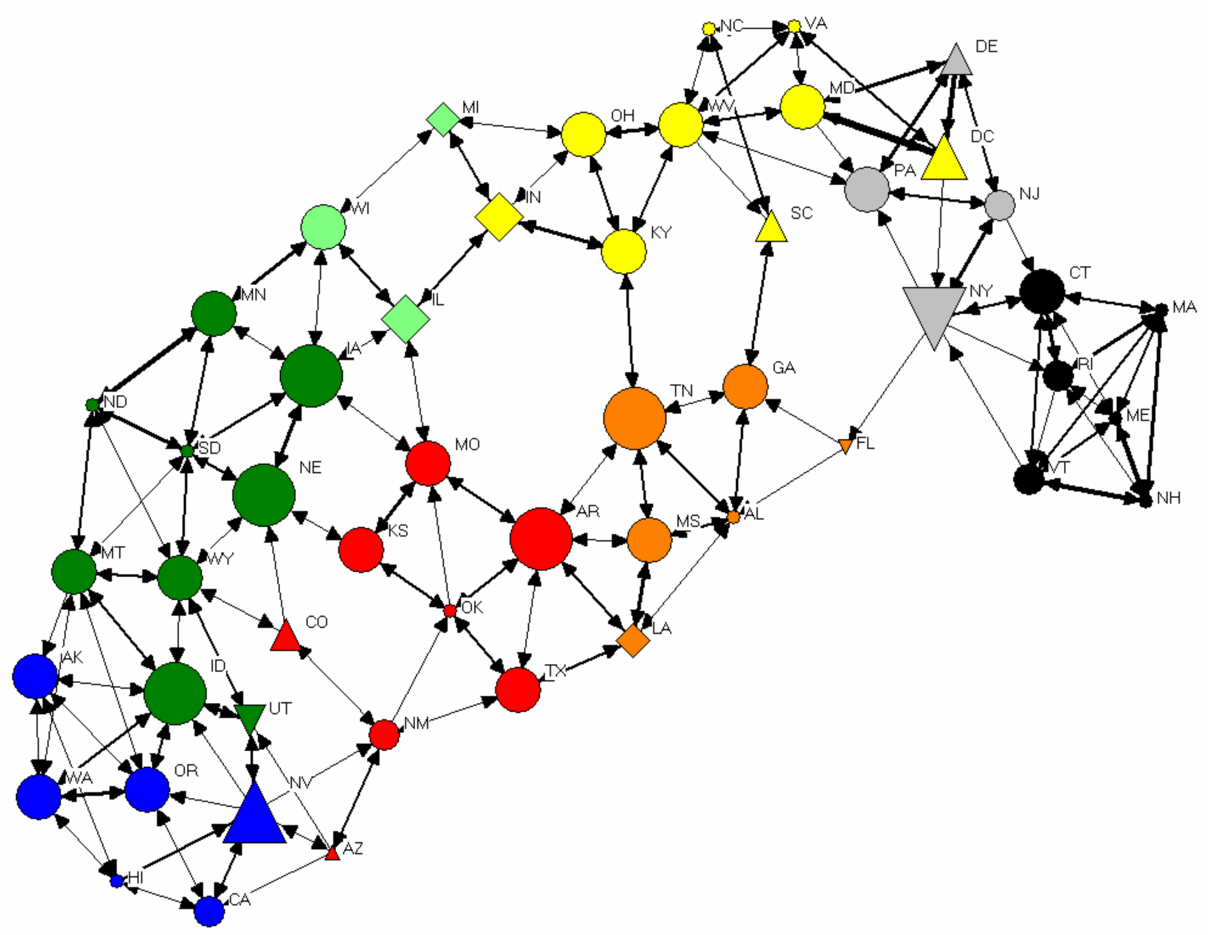

Figure 11:

5. Summary and conclusions

This paper is not finished yet; not even as a conference paper. For a more elaborate and final version, please check http://wuw-sre.wu-wien.ac.at/dpprice.html .

Therefore, we cannot draw any major conclusions yet. The analysis so far, however, indicates to us that Social Network Analysis can be applied to the analysis of migration and can generate some new insights. Some of the tools and techniques need to be explored more thoroughly in this context. Also, their possible relationship to more conventional tools of migration analysis need to be studied more carefully.

References

Borgatti, S.P., M.G. Everett, and L.C. Freeman (2002). Ucinet 6 for Windows – Software for social network analysis. Analytic Technologies, Harvard.

Freeman, Linton (2004). *The Development of Social Network Analysis – A study in the sociology of sciene*. Empirical Press, Vancouver.

Gould, J. and Fernandez, J. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. Sociological Methodology :89-126.

Wasserman Stanley, Katherine Faust (1999). *Social Network Analysis – Methods and applications*. Cambridge University Press, Cambridge.

Yasuhide Okuyama, Geoffrey J.D. Hewings, Michael Sonis and Philip R. Israilevich, An econometric analysis of BiProportional properties in an econometric input-output system, REAL 98-T-12, Rutgers University (http://policy.rutgers.edu/cupr/iioa/OkuyamaHewingsSonis&Israilevitch.pdf)

Johnson, S C (1967). 'Hierarchical clustering schemes'. Psychometrika, 32, 241-253.