Migratory Determinants Between Member-States of the OECD

Table of Contents

## Executive Summary

In this report we investigated the dynamics of international migration from the network perspective, between the member countries of the Organisation for Economic Co-operation and Development. We modelled international migration as a directed and valued network where the nodes represent countries, and the directed edges represent the number of countries migrating from one country to another in the year 2019. We tested whether this network exhibited clustering, and found an exceptionally high, statistically significant level of clustering. We investigated the factors that drive migration using a complex generalized exponential random graph model, and, contrary to our expectations based on the literature on the determinants of migration, found no significant effects of any of the covariates we included in the analysis. The only predictors of migration were the endogenous popularity and sociality effects, while reciprocity was not present. Future work from the network perspective is needed to determine the reasons behind the lack of significant covariate effects found.

## Introduction

### Determinants of international migration

International migration is an important feature of global economy, continuously shaping the demographics of our globalized world. It is predicted that international migration is likely to grow and become more complex in the coming decades, due to factors such as an aging population, growing economic inequalities and climate change (Klugman 2009). As it has an impact on receiving as well as sending countries, researching why people migrate, how they choose their destination countries, and how this migration changes over time are important to better understand global trends and inform policy decisions.

One of the classical methods for investigating the determinants of migration are gravity models (Boyle, Keith, and others 2014), their main components are population size and geographical proximity. There are more factors that influence migration according to the available literature, and these are typically grouped in two categories: push and pull factors (Langley et al. 2016). Push factors are those associated with the country of origin of the migrant and are related to their decision to migrate, examples of these are natural disasters, violence, poverty, and unemployment. Pull factors are associated with the destination country of the migrant, these can be economic prosperity and employment opportunities, favourable climate, and political stability. Migration movements can also be encouraged by homophily in the dominant religion and language, which lower the expected costs of acculturation (Belot and Ederveen 2012).

### Network perspective

The flows of international migration naturally constitute a directed and valued network where nodes represent countries. The outbound edges are the number of citizens each country sends to every other country in the network, this way the inbound edges of a node represent the number of citizens received from each country. Characteristics describing the nodes and the edges between them can be included as covariates.

Statistical network analysis captures the interconnectedness of countries and allows for the influence of various factors on migration to be measured without assuming independence that does not hold in migration data (Desmarais and Cranmer 2012). Despite network analysis being suitable for modelling such phenomena, to this day surprisingly few studies have used this perspective to better understand international migration Windzio (2018).

An example of a study that looks at international migration from a network perspective was conducted by Windzio (2018). Their results from a global dataset were in line with the gravity models and theories of global inequality when examining the determinants of migration using temporal and cross-sectional exponential random graph models (ERGMs). ERGM is a method for statistical inference on networks, which, given a number of statistics that capture the essential generative structures of the network, estimates the probability of the observed network (Cranmer and Desmarais 2011). This method takes into account the interdependent nature of migration and allows for modelling both endogenous dependencies and exogenous covariates of the network. Instead of flows of migrants, however, Windzio (2018) used data on stocks of persons born in a different country, and instead of looking at the number of migrants, they used quartiles of the measure of the migration distribution in order to reduce the data into a binary network. While ERGMs are a ubiquitous means of network analysis, their inability to model networks with valued edges is a strong limitation for modelling migration networks, as thresholding weighted edges into binary ones leads to a considerable loss of information on the vastly different magnitudes of migration that different edges of the network represent.

The generalized exponential random graph model (GERGM), developed by Desmarais and Cranmer (2012), is a method that generalizes ERGMS to be used for value-edged networks, making it suitable for studying migration networks. Desmarais and Cranmer (2012) used GERGMs to model interstate migration flows within the United States, and K. E. Abramski (2018) used the method to study refugee migration flows from four countries experiencing violent conflicts, finding that the GERGM is a better way for modelling this specific problem compared to the ERGM.

### Our research

In this report, we aim to contribute to the literature on international migration by examining it from a network perspective using generalized random graph models. Using a dataset of migration flows within the member countries of the Organization for Economic Co-operation and Development (OECD) in 2019, we look at the properties of this network and set up a model on the determinants of international migration.

Using the network perspective, it becomes possible to quantify the network properties of the international migration network. It is generally assumed that migration networks exhibit a high level of clustering (Fagiolo and Mastrorillo 2013), but to statistically test this proposition, we apply a conditional uniform graph (CUG) test. We focus our attention to clustering only, as this is a network-level measure for which a generalization has been proposed and implemented which is suitable for weighted networks (Opsahl and Panzarasa 2009).

*Research question 1: Does the international migration network of the OECD countries exhibit a high degree of clustering?*

*Hypothesis 1: The international migration network of the OECD countries exhibits a high degree of clustering.*

We examine the effects of several driving factors of migration suggested by earlier literature: population, population density, shared borders, shared language, shared religion, GDP, crime index, freedom index, and unemployment rate Windzio (2018). Our contribution is examining this subject from the network perspective and employing GERGM, a model that is highly suitable for this kind of data but has not been used in the research of international migration before. This model allows for the inclusion of endogenous effects, regarding which we hypothesise that there is no significant reciprocity in the network, similarly to the interstate migration network within the United States studied by Desmarais and Cranmer (2012).

*Research question 2: What factors drive migration within the OECD countries?*

*Hypothesis 2: Large population, high population density, low GDP, high crime, low freedom, and high unemployment rate are push factors of international migration within OECD countries.*

*Hypothesis 3: Large population, high population density, high GDP, low crime, high freedom, and low unemployment rate are pull factors of international migration within OECD countries.*

*Hypothesis 4: Sharing the same dominant language, religion, as well as sharing borders increases migration between OECD countries.*

*Hypothesis 5: The international migration network of the OECD countries does not exhibit reciprocity.*

The structure of the rest of the report is as follows. The method section explains the data used in this study, and the methods we use to run our analysis. Subsequently, the results section presents and discusses the outcomes of the analysis for each research question respectively. Finally, the report is closed with a short conclusion section.

## Methodology

### Dataset

The statistical analysis in this project focuses on the member countries of the Organisation for Economic Co-operation and Development (OECD). The OECD maintains a collaboration environment for various countries via the OECD council, which is set out in the OECD convention. There are currently 38 OECD countries.

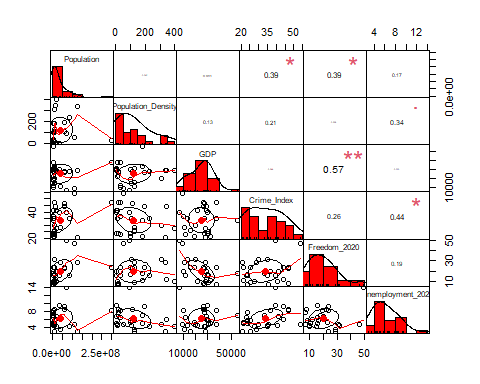
The essential network that is used in this report is obtained from the International Migration Database of the OECD (to access this, and other data sources we used, see the Table in the Appendix). Available information is inflow and outflow of foreign population by nationality for each country. Furthermore, the data is divided into one-year blocks. For this project, the migration data from the year 2019 is used, this being the last year on which data is currently available. In the Migration Database, every country has its own data frame for inflow and outflow. Each data frame has countries as rows and years as columns. Values that correspond to these rows and columns are the inflow/outflow of migration by/to the foreign population. However, some parts of the dataset were missing or corrupted so the authors had to remove some countries from the dataset in order to maintain an organized dataset. Therefore, the number of countries included in our analysis decreased from 38 to 27, the exact list can be found in the Appendix.

As covariates, we focused on borders, gross domestic product (GDP), crime rate, main religion, main language, unemployment rates, and freedom attributes for each country. For this purpose, datasets have been collected from various internet resources (see the list of links in the Appendix). A dataset on the country borders was used to create a network with the same nodes as the migration network, where undirected edges connect any two countries that share a border. The GDP and population data for this project were obtained from a Kaggle dataset, as well as the crime index. Information on the main religions and main languages were collected from Wikipedia. Unemployment rates were retrieved the website of The World Bank, and finally, the freedom index is a rating from Reporters Without Borders.

This array of different, and often not official, sources of data is indisputably a strong limitation of the present study. The language and religion data are particularly lacking, since for ease of computation, each country was assigned only one language and one religion, which severely neglect the diversity on these facets. For example, in a country such as Switzerland, where four languages are official and widely used, only one had to be selected, the one spoken by the highest percentage of the population. In the present project it was out of scope to work with more sophisticated data, however, for reliable conclusions, future work is needed to supplement our results.

Descriptive statistics of the covariates are presented in the table below.

covariates <- read.table("../data\_final/covariates.csv", row.names = 1, sep = ";", header = TRUE)  
  
covariates <- subset(covariates, select = -c(Area, Safety\_Index))  
  
is\_numeric <- sapply(covariates, is.numeric)  
  
bootcamp2021::betterpairs(covariates[, is\_numeric])



### Data analysis (Research Rationale)

Research question 1 investigates the degree of clustering within the network of internal migration among OECD countries. Clustering in a network is a form of graph level index or network descriptive statistic (Butts 2008), (Cranmer, Desmarais, and Morgan 2020). These types of descriptive statistics are useful for understanding attributes of a network, such as clustering, centralization, transitivity or reciprocity (Cranmer, Desmarais, and Morgan 2020). The network generation process by itself does not trivially translate into a clear answer, such as “there is a high degree of clustering,” which is where the conditional uniform graph (CUG) method is useful. It works by constructing a null hypothesis, or benchmark, which does not include a tendency towards the process we wish to study (meaning that the null distribution generated would not involve a high degree of clustering). This null distribution creation is often created through the use of Monte Carlo distributions, and ensures this benchmark that can be tested against (Butts 2008).

Research question 2 investigates which factors drive migration between the OECD countries. Several papers have been published using GERGMs to explain migratory patterns, but these focus on refugees (K. Abramski, Katenka, and Hutchison 2019) or intra-national migration (Aksoy and Yıldırım 2020). On the other hand, it is more common in literature to use ERGMs to study both international and intra-national migration (Windzio, Teney, and Lenkewitz 2021), (Nourpanah et al. 2018), (Belkhiria et al. 2019). At the time of writing, the Google Scholar search results for “GERGM migration” number 23, whereas “ERGM migration” numbers 739. While this is not completely accurate because it is an unsupervised search result, it does provide an indication. It is our wish to expand on literature and test the robustness of the available findings with the GERGM model. The ERGM and GERGM both evaluate endogenous and exogenous factors, but the ERGM does not allow for valued edges, whereas the GERGM does. Patterns of migration among countries are not binary, and investigating how the values of migration between countries affects the overall migration network could be beneficial in understanding migratory patterns, and shed light upon factors not previously found.

## Results

### Model 1: CUG

*Clustering in weighted networks*

In order to understand the attributes of the network of internal migration among OECD countries, we investigate the degree of clustering within the network. The global clustering coefficient – often called transitivity - is a network-level measure that in its original form cannot be applied to weighted networks. A generalization has been proposed by Opsahl and Panzarasa (2009), and this is the algorithm we are using in this report.

The method investigates the degree to which nodes appear to be clustered together and it can be applied to both directed and undirected networks. Evidence indicates that in real-world interpretations of networks, particularly in social networks, nodes appear to create “tightly knit groups” that act by a “relatively high density of ties” (Opsahl and Panzarasa 2009). Real-world networks, this probability appears to be above the average likelihood of a tie that is randomly formed between two nodes. The global clustering coefficients approach focuses on triplets of nodes, three nodes that are connected by two or three undirected ties is called a triplet, and their value can be calculated in one of four ways: arithmetic mean, geometric mean, the maximum or the minimum of the edges constituting the triplet (**odopsahl2009clustering?**).

In our work, global clustering coefficients were examined for the network of international migration among OECD countries. We chose to report on results with the method of geometric mean value, as this is the least sensitive to extreme values. Nonetheless, the analyses have been conducted, and all the results hold with all other methods of calculating the global clustering coefficient in weighted networks as well. The observed value for the migration network is 0.989, whose significance was then tested using a CUG test.

*CUG test*

It is generally beneficial to compare observed index values in contrast to those which would be accessed by a baseline model with acknowledged substantive properties for the evaluation of graph-level indices (Butts 2008). Within the consideration of the extent and the direction of derivation for each index from its baseline distributions, the presence of structural biases can be detected, which may provide convenient information about the mechanisms underlying the data (Butts 2008). The Conditional Uniform Graph (CUG) distribution is one of the important baseline models for network data (Butts 2008). One of the most important uses of this method is the CUG test procedure. The CUG test is a method that is applied to a hypothesis for an observed statistic. In our case, the observed statistics were the results of the clustering process. The hypothesis is that there is a high degree of clustering in the data. In order to investigate this hypothesis, we performed a CUG test for the clustering coefficient with 1000 iterations.

Since the implementation of the clustering coefficient for weighted graphs by (**odopsahl2009clustering?**) operates on edge-lists with integer-named nodes, this function was not suitable for calling within the function implementing the CUG test in the R package SNA, as this latter operates on network objects or adjacency matrices. Therefore, our implementation starts with defining a function that converts an adjacency matrix into an edge list, calls the clustering function on this matrix, and returns the clusteredness score.

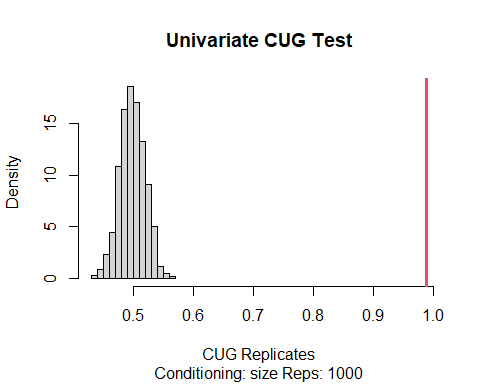
# Load data  
  
migration <- as.matrix(read.table("../data\_final/migration\_adjacency\_matrix.csv", sep = ";", row.names = 1, header = TRUE))

# Function for creating edge list from adjacency matrix and calculating the clustering coefficient  
  
clustering\_function <- function(dat) {  
   
 edge\_list <- matrix(0, (nrow(dat) \* nrow(dat)), 3)  
  
 for(row in 1:NROW(dat))  
 {  
 for(column in 1:NROW(dat))  
 {  
   
 edge\_list[(row-1)\*nrow(dat) + column, 1] <- row  
 edge\_list[(row-1)\*nrow(dat) + column, 2] <- column  
 edge\_list[(row-1)\*nrow(dat) + column, 3] <- dat[row, column]  
 }  
 }  
   
 # Find clustering coefficient  
  
 cc <- tnet::clustering\_w(edge\_list, measure=c("bi", "am", "gm", "ma", "mi"))  
  
 # Here I'm returning the geometric mean value  
   
 cc[[3]]  
}

# Run the CUG test  
  
suppressWarnings(  
   
   
cug <- sna::cug.test(dat=migration,   
 FUN=clustering\_function,   
 mode="digraph",  
 ignore.eval=FALSE,  
 reps=1000) # default is 1000  
  
)  
  
print(cug)

##   
## Univariate Conditional Uniform Graph Test  
##   
## Conditioning Method: size   
## Graph Type: digraph   
## Diagonal Used: FALSE   
## Replications: 1000   
##   
## Observed Value: 0.9887157   
## Pr(X>=Obs): 0   
## Pr(X<=Obs): 1

# Plot the CUG test  
  
plot(cug)



The figure above presents the outcome of the CUG test. The results are in line with Hypothesis 1, showing that there is a high degree of clustering in this network, more than what is expected of a network with similar measures. The measured clustering coefficient is much higher than any of the coefficients in any of the simulated 1000 networks. This indicates that the test resulted with the confirmation of the hypothesis.

### Model 2: GERGM

As discussed earlier in the report, we use a GERGM, which allows for accommodating weights on the edges in the network. The following covariates are considered: population, population density, GDP, crime index, freedom index, unemployment index, language, and borders. While originally we also aimed to include religion as is has been shown to be a key factor related to migration(Belot and Ederveen 2012), our data for the countries remaining in the analysis did not have enough variance on this factor to include it in the model. The covariates showed high correlations as shown in the Data Section of this report, but including them simultaneously did not appear to negatively affect model fits. Methods for calculating variance inflation factors to detect multicollinearity have been implemented for ERGMs (Duxbury 2021) but not yet for GERGMs, these are therefore unknown.

Since GERGMs allow for including endogenous dependencies in the model, we opted for including a number of these as well. Our CUG test results show that the network exhibits a high degree of clustering, suggesting that triads are an interesting metrics to examine. Including transitive or cyclic triads, however, made the model unable to converge, even when utilizing alpha weighting to avoid degeneracy in the model. Our final model included only the remaining three endogenous dependencies implemented in the GERGM package: mutual, which reflects transitivity in the network; in2stars, which reflect popularity of nodes; and out2stars, reflecting sociality of nodes, as their likelihood of sending multiple outgoing edges.

The model uses directed edge values between nodes as the number of migrations between the countries. Since the distribution of edges is highly skewed, weights are estimated with a Cauchy distribution.The estimations are done using 500,000 iterations of Markov chain Monte Carlo with the Metropolis-Hastings algorithm, with a burn-in of 10,000 iterations.

covariates <- read.table("../data\_final/covariates.csv", row.names = 1, sep = ";", header = TRUE)  
borders <- as.matrix(read.table("../data\_final/borders\_adjacency\_matrix.csv",sep = ";", row.names = 1, header = TRUE))  
migration <- as.matrix(read.table("../data\_final/migration\_adjacency\_matrix.csv", sep = ";", row.names = 1, header = TRUE))  
  
  
formula <- migration ~   
 edges +   
 sender("Population") +  
 receiver("Population") +  
 sender("Population\_Density") +  
 receiver("Population\_Density") +  
 sender("GDP") +  
 receiver("GDP") +  
 sender("Crime\_Index") +  
 receiver("Crime\_Index") +  
 sender("Freedom\_2020") +  
 receiver("Freedom\_2020") +  
 sender("Unemployment\_2020") +  
 receiver("Unemployment\_2020") +  
 nodematch("Language") +  
 netcov(borders) +  
 in2stars(alpha = 0.8) +  
 out2stars(alpha = 0.8) +  
 mutual(alpha = 0.8)  
  
  
model <- GERGM::gergm(formula,  
 covariate\_data = covariates,  
 number\_of\_networks\_to\_simulate = 500000,  
 thin = 1/100,  
 proposal\_variance = 0.05,  
 MCMC\_burnin = 10000,  
 seed = 456,  
 convergence\_tolerance = 0.1,  
 transformation\_type = "Cauchy",  
 verbose = FALSE,  
 generate\_plots = FALSE)

## Warning in if (class(net) != "matrix") {: the condition has length > 1 and only  
## the first element will be used

## Warning in if (class(return\_list$network\_matrix\_object) != "matrix") {: the  
## condition has length > 1 and only the first element will be used

## Warning in if (class(net) != "matrix") {: the condition has length > 1 and only  
## the first element will be used

## Warning in if (class(return\_list$network\_matrix\_object) != "matrix") {: the  
## condition has length > 1 and only the first element will be used

## Warning in if (class(net) != "matrix") {: the condition has length > 1 and only  
## the first element will be used

## Warning in if (class(return\_list$network\_matrix\_object) != "matrix") {: the  
## condition has length > 1 and only the first element will be used

## You have specified 13 node level covariate effects.  
## You have provided 1 network covariates.

## Warning in if (class(net) != "matrix") {: the condition has length > 1 and only  
## the first element will be used  
  
## Warning in if (class(net) != "matrix") {: the condition has length > 1 and only  
## the first element will be used

## [1] 1 1 1 1 1 1 1  
## Updating Estimates -- Iteration: 1   
## Starting thetas: 0.1810114 -0.1939442 0.4033215   
## Current iteration of convex hull initialization: 1   
## Average log probability of accepting a proposal: -4.009004 .  
## Standard deviation of log probability of accepting proposal: 2.83832   
## Average (constrained) simulated network density: 0.5017897   
## Distance between observed and mean simulated statistics: 60.8016   
## Proportion of simulated stats that are further away: 0.6532   
## New thetas: 0.2882519 -0.3268649 0.9409202   
## Current iteration of convex hull initialization: 2   
## Average log probability of accepting a proposal: -4.764842 .  
## Standard deviation of log probability of accepting proposal: 3.121519   
## Average (constrained) simulated network density: 0.4993941   
## Distance between observed and mean simulated statistics: 42.7281   
## Proportion of simulated stats that are further away: 0.778   
## New thetas: 0.2934359 -0.3189284 0.5236075   
## Current iteration of convex hull initialization: 3   
## Average log probability of accepting a proposal: -4.736439 .  
## Standard deviation of log probability of accepting proposal: 3.102079   
## Average (constrained) simulated network density: 0.4953101   
## Distance between observed and mean simulated statistics: 7.893459   
## Proportion of simulated stats that are further away: 0.987   
## Converged!  
## Average log probability of accepting a proposal: -4.736439 .  
## Standard deviation of log probability of accepting proposal: 3.102079   
## Average (constrained) simulated network density: 0.4953101   
## Simulated (averages) and observed network statistics...  
## Observed Simulated  
## out2stars 2145.27100 2140.42593  
## in2stars 2202.83472 2196.62746  
## ctriads 687.24179 710.85505  
## mutual 86.91812 87.46758  
## ttriads 2202.68467 2163.80063  
## edges 344.31261 347.70424  
## diagonal 25.35597 25.35597  
## MCMC convergence Geweke test statistic: -0.5931545   
## (If the absolute value is greater than 1.7, increase MCMC\_burnin)  
## Updating Estimates -- Iteration: 2   
## Starting thetas: 0.2975147 -0.3100899 0.2176585   
## Current iteration of convex hull initialization: 1   
## Average log probability of accepting a proposal: -4.721101 .  
## Standard deviation of log probability of accepting proposal: 3.130121   
## Average (constrained) simulated network density: 0.4959685   
## Distance between observed and mean simulated statistics: 2.755432   
## Proportion of simulated stats that are further away: 0.9992   
## Converged!  
## Average log probability of accepting a proposal: -4.721101 .  
## Standard deviation of log probability of accepting proposal: 3.130121   
## Average (constrained) simulated network density: 0.4959685   
## Simulated (averages) and observed network statistics...  
## Observed Simulated  
## out2stars 2145.27100 2145.88971  
## in2stars 2202.83472 2200.15019  
## ctriads 687.24179 712.96932  
## mutual 86.91812 86.86447  
## ttriads 2202.68467 2169.70181  
## edges 344.31261 348.13275  
## diagonal 25.35597 25.35597  
## MCMC convergence Geweke test statistic: -0.02840413   
## (If the absolute value is greater than 1.7, increase MCMC\_burnin)

## Lambda parameter estimates have converged...

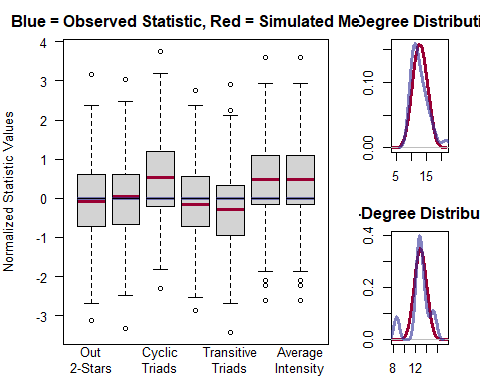
## Average log probability of accepting a proposal: -4.85419 .  
## Standard deviation of log probability of accepting proposal: 3.159985   
## Average (constrained) simulated network density: 0.4948502   
## Metropolis Hastings Acceptance Rate (target = 0.25 ): 0.1221647  
## Average Q-Ratio: -4.096624 Average P-Ratio: -0.757566   
## Mean difference between proposed and current network densities: 9.918775e-05   
## Thinning statistics to correct for autocorrelation in calculating fit diagnostics...  
## Statistics were thinned by a factor of 24, resulting in 209 samples.  
## Statistics of observed network and networks simulated from final theta parameter estimates:  
##   
## t-test p-values for statistics of observed network and networks simulated from final theta parameter estimates:  
##   
## p\_values  
## out2stars 0.143  
## in2stars 0.505  
## ctriads 0.000  
## mutual 0.077  
## ttriads 0.000  
## edges 0.000  
## diagonal NA

## Parameter estimates simulate networks that are statistically indistinguishable from observed network on the statistics specified by the user.

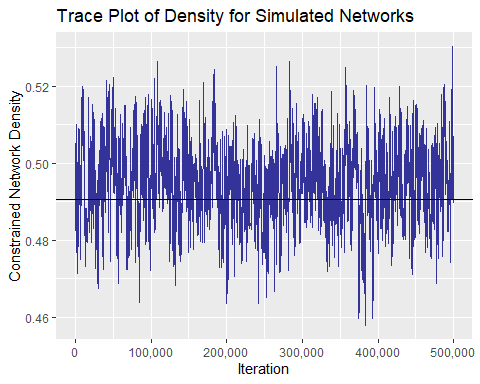
## Transforming networks simulated via MCMC as part of the fit diagnostics back on to the scale of observed network. You can access these networks through the '@MCMC\_output$Networks' field returned by this function...  
## 10% complete...  
## 20% complete...  
## 30% complete...  
## 40% complete...  
## 50% complete...  
## 60% complete...  
## 70% complete...  
## 80% complete...  
## 90% complete...  
## 100% complete...  
## Estimation Complete at: 2021-12-12 22:52:53   
## Elapsed time (seconds): 44.88993

gof <- GERGM::GOF(model, return\_GERGM\_Object = TRUE)

## Thinning statistics to correct for autocorrelation in calculating fit diagnostics...  
## Statistics were thinned by a factor of 24, resulting in 209 samples.  
##   
## Statistics used in comparison:  
## out2stars in2stars ctriads mutual ttriads   
## 2.145271e+03 2.202835e+03 6.872418e+02 8.691812e+01 2.202685e+03   
## edges intensity   
## 3.443126e+02 -2.448166e-17   
##   
## Mean statistic values from simulated networks:  
## out2stars in2stars ctriads mutual ttriads edges   
## 2.136482e+03 2.198768e+03 7.088829e+02 8.647690e+01 2.163371e+03 3.474360e+02   
## intensity   
## 4.449317e-03   
##   
## t-statistics for difference of simulated mean from observed statistic...  
## out2stars in2stars ctriads mutual ttriads edges intensity   
## -1.4686124 -0.6678551 7.2217480 -1.7781064 -4.3900105 6.4424302 6.4424302   
##   
## t-statistics for test of whether observed statistic is an outlier with respect to simulated statistics...  
## out2stars in2stars ctriads mutual ttriads edges intensity   
## 0.1015860 0.0461965 -0.4995388 0.1229942 0.3036634 -0.4456322 -0.4456322

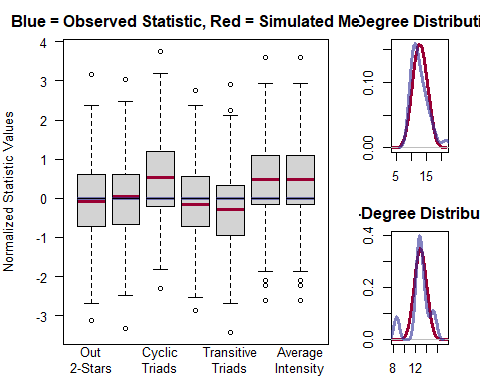


# Generate Trace Plot  
GERGM::Trace\_Plot(model)



# Generate GOF Plot  
GERGM::GOF(model)

## Thinning statistics to correct for autocorrelation in calculating fit diagnostics...  
## Statistics were thinned by a factor of 24, resulting in 209 samples.  
##   
## Statistics used in comparison:  
## out2stars in2stars ctriads mutual ttriads   
## 2.145271e+03 2.202835e+03 6.872418e+02 8.691812e+01 2.202685e+03   
## edges intensity   
## 3.443126e+02 -2.448166e-17   
##   
## Mean statistic values from simulated networks:  
## out2stars in2stars ctriads mutual ttriads edges   
## 2.136482e+03 2.198768e+03 7.088829e+02 8.647690e+01 2.163371e+03 3.474360e+02   
## intensity   
## 4.449317e-03   
##   
## t-statistics for difference of simulated mean from observed statistic...  
## out2stars in2stars ctriads mutual ttriads edges intensity   
## -1.4686124 -0.6678551 7.2217480 -1.7781064 -4.3900105 6.4424302 6.4424302   
##   
## t-statistics for test of whether observed statistic is an outlier with respect to simulated statistics...  
## out2stars in2stars ctriads mutual ttriads edges intensity   
## 0.1015860 0.0461965 -0.4995388 0.1229942 0.3036634 -0.4456322 -0.4456322



# Check for significance  
(EstSE <- rbind(t(attributes(model)$theta.coef), t(attributes(model)$lambda.coef)))

## est se  
## in2stars 3.063624e-01 5.155919e-02  
## out2stars -3.199722e-01 5.695183e-02  
## mutual 2.418823e-01 6.407403e-01  
## Population\_sender 1.381035e-05 2.400005e-05  
## Population\_receiver -1.271428e-02 9.090378e+01  
## Population\_Density\_sender -5.857925e-02 5.358775e+01  
## Population\_Density\_receiver 4.672990e-02 8.141951e+01  
## GDP\_sender 1.170035e-01 6.397814e+01  
## GDP\_receiver 3.552406e-02 1.004950e+02  
## Crime\_Index\_sender 1.321900e-01 6.878762e+01  
## Crime\_Index\_receiver -4.838491e-02 9.840295e+01  
## Freedom\_2020\_sender 9.973170e-02 8.134899e+01  
## Freedom\_2020\_receiver -6.619565e-02 1.092652e+02  
## Unemployment\_2020\_sender 4.585433e-02 7.178460e+01  
## Unemployment\_2020\_receiver -3.118521e-02 8.668115e+01  
## Language\_nodematch 6.847082e-02 9.342942e+01  
## intercept 1.657978e+03 3.630626e+02  
## borders\_netcov 5.805997e-02 7.459879e+01  
## dispersion 7.400137e+00 1.023300e-01

EstSE\_df = as.data.frame(EstSE)  
  
EstSE\_df$LOWER = EstSE\_df$est - EstSE\_df$se\*(-qnorm((1-0.95)/2))  
EstSE\_df$UPPER = EstSE\_df$est + EstSE\_df$se\*(-qnorm((1-0.95)/2))  
  
# If LOWER and UPPER have the same sign (positive/negative) then there's significance  
EstSE\_df$Significant = (EstSE\_df$LOWER > 0 & EstSE\_df$UPPER > 0) | (EstSE\_df$LOWER < 0 & EstSE\_df$UPPER < 0)  
  
EstSE\_df <- subset(EstSE\_df, select = -c(LOWER, UPPER))  
  
knitr::kable(EstSE\_df)

|  |  |  |  |
| --- | --- | --- | --- |
|  | est | se | Significant |
| in2stars | 0.3063624 | 0.0515592 | TRUE |
| out2stars | -0.3199722 | 0.0569518 | TRUE |
| mutual | 0.2418823 | 0.6407403 | FALSE |
| Population\_sender | 0.0000138 | 0.0000240 | FALSE |
| Population\_receiver | -0.0127143 | 90.9037824 | FALSE |
| Population\_Density\_sender | -0.0585793 | 53.5877496 | FALSE |
| Population\_Density\_receiver | 0.0467299 | 81.4195071 | FALSE |
| GDP\_sender | 0.1170035 | 63.9781377 | FALSE |
| GDP\_receiver | 0.0355241 | 100.4949663 | FALSE |
| Crime\_Index\_sender | 0.1321900 | 68.7876236 | FALSE |
| Crime\_Index\_receiver | -0.0483849 | 98.4029504 | FALSE |
| Freedom\_2020\_sender | 0.0997317 | 81.3489925 | FALSE |
| Freedom\_2020\_receiver | -0.0661957 | 109.2652421 | FALSE |
| Unemployment\_2020\_sender | 0.0458543 | 71.7846003 | FALSE |
| Unemployment\_2020\_receiver | -0.0311852 | 86.6811542 | FALSE |
| Language\_nodematch | 0.0684708 | 93.4294161 | FALSE |
| intercept | 1657.9775776 | 363.0625667 | TRUE |
| borders\_netcov | 0.0580600 | 74.5987911 | FALSE |
| dispersion | 7.4001366 | 0.1023300 | TRUE |

The trace plot shows an acceptable level of convergence, and the goodness of fit plots display that the resulting model has the observed and simulated statistics as quite similar to each other. However, as displayed in the table above, none of the attributes other than in2stars, out2stars, the intercept and dispersion are significant. This implies that Hypotheses 2, 3, 4 do not hold, and Hypothesis 5 does. This does not fit with the earlier research results of Belot and Ederveen (2012), Langley et al. (2016) and Windzio (2018), who find that each of these factors is significant.

In this model, in2stars and out2stars are significant, which indicates that there are unmodelled features of countries that influence migration flows to and from each country. Popularity and sociality exist to a significant degree in the network; e.g., if a country is sending emigrants to one country, it is likely to send more to other countries too; and if a country receives many immigrants from one country, it likely receives many from other countries too. In line with our Hypothesis 5 based on earlier studies, reciprocity is not significant, meaning that a country receiving many immigrants from another country does not predict this other country receiving a high number of migrants back from that country.

These results suggest that within the OECD countries, population size, population density, GDP, crime, freedom, unemployment are not push-pull factors in migration (Hypotheses 2, 3). The OECD countries are countries that are relatively well-off and generally viewed as the “first-world”; modern, industrial countries. Previous research has focused on migration primarily in the form of refugees, or where one side of the migration is better off than the other side in aspects such as GDP, freedom, or crime, or others investigated. The range within the OECD countries on these attributes do not go to the extremes, which could be an explanation for why these factors are not significant motivators of migration within the network of observed countries.

Language, and borders are not significant within OECD countries either (Hypothesis 4). This might be related to the fact that most of the countries investigated do not border each other or share a common language. Having a shared language or border does not appear to significantly affect migration, there are therefore other factors which encourage migration more. Perhaps these factors are personal and relatively unique to individuals who migrate within OECD countries, such as job opportunities which are not readily available within their own countries, or the wish to live in a more popular country.

It is quite remarkable that none of the investigated factors contribute significantly to migration. The significant popularity and sociality effects indicate that there are trends in the migration patterns, perhaps further detailing the covariates included in the analysis and controlling for more factors could result in better insights into the determinants of migration. However, since the outgoing edges of migrants were not scaled to the population of the sending country, it is possible that these effects occur partially only due to the imbalanced population sizes of the countries in the network. Our study was limited by a small sample size of only a subset of the OECD countries, and only looking at one specific year. Our results might be due to these limitations, or to generally different dynamics of migration amongst OECD countries than the groups of countries previously studied. Another way to explain our differing findings is the new and different method we used for testing our hypotheses. Since GERGMs are suitable for modelling interconnected and interdependent countries, it is possible that earlier results have been biased because of methods that oversimplified the problem, or methods whose assumptions were violated by migration datasets. Since, however, many factors that we failed to find have been quite robustly established in the literature, a number of further studies would be necessary to confirm these claims.

Our specific way of modeling the data might have also affected the outcomes negatively. The migration data is highly skewed, and while the Cauchy distribution is able to handle somewhat skewed distributions, in some studies the authors opt for log-transforming all edge values, see K. E. Abramski (2018) for an example. In our data, there is also a large variance in the ranges and variances of the covariates. Log-transforming these variables or using z-scores are also frequently taken measures to make the data easier to handle for the models. However, through trying many combinations of the above-mentioned transformations, most versions of the data did not lead to converging models. In addition, while some of these models did converge and confirmed our hypotheses based on the significance values, their diagnostic and goodness-of-fit plots still showed that they were inferior to our model presented above, see an example of such models in the Appendix.

## Conclusion

The aim of our report was to investigate the migration network of the OECD countries. We modelled international migration as a directed and valued network where the nodes represent countries, and the directed edges represent the number of countries migrating from one country to another in the year 2019. We tested whether this network exhibited clustering, and found an exceptionally high, statistically significant level of clustering. We investigated the factors that drive migration through a complex GERGM and found no significant effect of any of the covariates we included in the analysis, the only predictors of migration were the endogenous popularity and sociality effects.

Being unable to confirm our hypotheses about the push and pull factors of international migration might have resulted form our limited data, uncommon patterns of migration amongst OECD countries, or the strictness of the GERGM method. We observed that while migration flows between countries naturally constitute a network, when examining a problem like this, there are still decisions to be made on how to represent the data, especially relating to the scaling of the covariates and the edge weights.

Our contribution has been to examine the topic of international migration from the network perspective. We would like to promote the use of GERGMs for future use in this field, as it provides a suitable method, accounting for the unique characteristics of such data. Future work is called to focus on the factors determining migration that have been missed or misrepresented due to incomplete data in this study, and to examine whether the dynamics of international migration differ between the OECD countries and other sets of countries.

# References

Abramski, Katherine E. 2018. *A Network Based Analysis of the International Refugee Crisis Using GERGMs*. University of Rhode Island.

Abramski, Katherine, Natallia Katenka, and Marc Hutchison. 2019. “A Network-Based Analysis of International Refugee Migration Patterns Using GERGMs.” In *International Conference on Complex Networks and Their Applications*, 387–400. Springer.

Aksoy, Ozan, and Sinan Yıldırım. 2020. “A Model of Dynamic Migration Networks: Explaining Turkey’s Inter-Provincial Migration Flows.”

Belkhiria, Jaber, Modou Moustapha Lo, Fafa Sow, Beatriz Martı́nez-López, and Veronique Chevalier. 2019. “Application of Exponential Random Graph Models to Determine Nomadic Herders’ Movements in Senegal.” *Transboundary and Emerging Diseases* 66 (4): 1642–52.

Belot, Michèle, and Sjef Ederveen. 2012. “Cultural Barriers in Migration Between OECD Countries.” *Journal of Population Economics* 25 (3): 1077–1105.

Boyle, Paul, Halfacree Keith, and others. 2014. *Exploring Contemporary Migration*. Routledge.

Butts, Carter T. 2008. “Social Network Analysis: A Methodological Introduction.” *Asian Journal of Social Psychology* 11 (1): 13–41.

Cranmer, Skyler J, and Bruce A Desmarais. 2011. “Inferential Network Analysis with Exponential Random Graph Models.” *Political Analysis* 19 (1): 66–86.

Cranmer, Skyler J, Bruce A Desmarais, and Jason W Morgan. 2020. *Inferential Network Analysis*. Cambridge University Press.

Desmarais, Bruce A, and Skyler J Cranmer. 2012. “Statistical Inference for Valued-Edge Networks: The Generalized Exponential Random Graph Model.” *PloS One* 7 (1): e30136.

Duxbury, Scott W. 2021. “Diagnosing Multicollinearity in Exponential Random Graph Models.” *Sociological Methods & Research* 50 (2): 491–530.

Fagiolo, Giorgio, and Marina Mastrorillo. 2013. “International Migration Network: Topology and Modeling.” *Physical Review E* 88 (1): 012812.

Klugman, Jeni. 2009. “Human Development Report 2009. Overcoming Barriers: Human Mobility and Development.” *Overcoming Barriers: Human Mobility and Development (October 5, 2009). UNDP-HDRO Human Development Reports*.

Langley, S, M Vanore, M Siegel, I Roosen, M Rango, I Leonardelli, and F Laczko. 2016. “The Push and Pull Factors of Asylum Related Migration: A Literature Review.” *European Asylum Support Office*.

Nourpanah, Shiva, Ivy Bourgeault, Lois Jackson, Sheri Price, Pauline Gardiner Barber, and Michael P Leiter. 2018. “Intersecting Policy Contexts of Employment-Related Geographical Mobility of Healthcare Workers: The Case of Nova Scotia, Canada.” *Healthcare Policy* 14 (2): 12.

Opsahl, Tore, and Pietro Panzarasa. 2009. “Clustering in Weighted Networks.” *Social Networks* 31 (2): 155–63.

Tranos, Emmanouil, Masood Gheasi, and Peter Nijkamp. 2015. “International Migration: A Global Complex Network.” *Environment and Planning B: Planning and Design* 42 (1): 4–22.

Windzio, Michael. 2018. “The Network of Global Migration 1990–2013: Using ERGMs to Test Theories of Migration Between Countries.” *Social Networks* 53: 20–29.

Windzio, Michael, Céline Teney, and Sven Lenkewitz. 2021. “A Network Analysis of Intra-EU Migration Flows: How Regulatory Policies, Economic Inequalities and the Network-Topology Shape the Intra-EU Migration Space.” *Journal of Ethnic and Migration Studies* 47 (5): 951–69.

# Appendix

*List of countries included in the analysis*

knitr::kable(rownames(migration))

|  |
| --- |
| x |
| Australia |
| Austria |
| Belgium |
| Canada |
| Czech\_Republic |
| Denmark |
| Estonia |
| Finland |
| France |
| Germany |
| Hungary |
| Iceland |
| Israel |
| Italy |
| Japan |
| Latvia |
| Luxembourg |
| Mexico |
| Netherlands |
| New\_Zealand |
| Norway |
| Poland |
| Slovenia |
| Sweden |
| Switzerland |
| Turkey |
| United\_States |

*Sources of data*

data\_sources <- read.table("data\_sources.csv", sep = ",", row.names = 1, header = TRUE)  
knitr::kable(data\_sources)

|  |  |
| --- | --- |
|  | Source |
| Migration | <https://stats.oecd.org/Index.aspx?DataSetCode=MIG> |
| Country Borders | <https://github.com/geodatasource/country-borders> |
| Gross Domestic Product(GDP) | <https://www.kaggle.com/fernandol/countries-of-the-world> |
| Crime Index | <https://www.kaggle.com/dumbgeek/countries-dataset-2020?select=Crime+index+by+countries+2020.csv> |
| Religion | <https://www.nationmaster.com/country-info/stats/Religion/Religions/All> |
| Language | <https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory> |
| Unemployment | <https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS> |
| Freedom | <https://rsf.org/en/ranking_table> |

*Second GERGM with log-trasformed edge weights, population and GDP*

# covariates\_logged <- data.frame(covariates)  
# covariates\_logged$Population <- log(covariates\_logged$Population)  
# covariates\_logged$GDP <- log(covariates\_logged$GDP)  
# migration\_logged <- log(migration)  
# migration\_logged[migration\_logged == -Inf] <- 0.01  
#   
# formula\_logged <- migration\_logged ~   
# edges +   
# sender("Population") +   
# receiver("Population") +  
# sender("Population\_Density") +  
# receiver("Population\_Density") +  
# sender("GDP") +  
# receiver("GDP") +  
# sender("Crime\_Index") +  
# receiver("Crime\_Index") +  
# sender("Freedom\_2020") +  
# receiver("Freedom\_2020") +  
# sender("Unemployment\_2020") +  
# receiver("Unemployment\_2020") +  
# nodematch("Religion") +  
# netcov(borders) +  
# in2stars(alpha = 0.8) +  
# out2stars(alpha = 0.8) +  
# mutual(alpha = 0.8)  
#   
# model\_logged <- GERGM::gergm(formula\_logged,  
# covariate\_data = covariates\_logged,  
# number\_of\_networks\_to\_simulate = 10,  
# thin = 1/100,  
# proposal\_variance = 0.05,  
# MCMC\_burnin = 10000,  
# transformation\_type = "LogCauchy",  
# seed = 456,  
# convergence\_tolerance = 0.1)  
#   
# gof\_logged <- GERGM::GOF(model\_logged, return\_GERGM\_Object = TRUE)

# Generate Trace Plot  
# GERGM::Trace\_Plot(model\_logged)

# Generate GOF Plot  
# GERGM::GOF(model\_logged)

# (EstSE\_log <- rbind(t(attributes(model\_logged)$theta.coef), t(attributes(model\_logged)$lambda.coef)))  
#   
# EstSE\_df\_log = as.data.frame(EstSE\_log)  
#   
# EstSE\_df\_log$LOWER = EstSE\_df\_log$est - EstSE\_df\_log$se\*(-qnorm((1-0.95)/2))  
# EstSE\_df\_log$UPPER = EstSE\_df\_log$est + EstSE\_df\_log$se\*(-qnorm((1-0.95)/2))  
#   
# # If LOWER and UPPER have the same sign (positive/negative) then there's significance  
# EstSE\_df\_log$Significant = (EstSE\_df\_log$LOWER > 0 & EstSE\_df\_log$UPPER > 0) | (EstSE\_df\_log$LOWER < 0 & EstSE\_df\_log$UPPER < 0)  
#   
# EstSE\_df <- subset(EstSE\_df, select = -c(LOWER, UPPER))  
#   
# knitr::kable(EstSE\_df)