# Modeling Economical Segregation in US County Network Structure

Sandhya Gopchandani, Atena Farhangian

December 10, 2018

## 1 Abstract

Social Segregation model by Thomas C. Schelling is a simple mathematical model to understand how local rules can produce macro-behaviors. In this report, we have used the idea of Schelling model in network structure of US counties to understand the economical segregation on a micro as well as a macro level.

## 2 Introduction

In 1971 Thomas C. Schelling published an article about how the dynamics of biased choices individuals make can result in segregation [1]. The model mainly discusses how agents with different ethnicity move in the city so they are close to people of their own group [1]. It is designed in a way that the system eventually comes to a complete integration or segregation. By integration here we mean random distribution of individuals. However, experimental studies showed such a subject is not as simple as just two final conditions of total segregation or integration.[1, 2] Schelling model describes two kinds of neighborhoods on a grid of individuals and it is based on the principles of Cellular Automata. The rules of relocation are simple: There is a factor of tolerance threshold (F) which is a fixed number and shows how adapting the individual is to the presence of strangers in its vicinity. The larger the F, the more eager the agent to be close to people of its own group. For each cell on the grid being at the center of a 5 by 5 square, there is a fraction of friends in that neighborhood like the individual itself and if this fraction (f) is less than threshold (F), this particular agent is going to move to a neighborhood where f is equal or greater than F [1]

This project, though inspired by the Schelling model, has a fundamentally different and more robust structure to study the economical segregation in communities. Our model is a network of counties as opposed to square grid of households. The rules of relocation are based on financial status rather than race which we think play an even more significant role in decision making. The model we developed also considers the interaction between three groups rather than two in order to simulate a more realistic system. Our main goal in this project is to study the dynamics of residential interactions in 3140 counties of United States. The idea behind this project is to simulate the segregation dynamics of counties in United States, based on the level of income of households. We decided to use the basic principles of Schelling model and develop a model simulating how different classes of a society based on their income level would make various choices in terms of the neighborhood they want to live in, how they decide to move and what choices of destination they have. Finally we want to show the interactions involved in this dynamic until each individual gets to the point that they are satisfied with their relocation.

Doing this project there are some questions we would like to find answers to:

1. Is the system going to show segregation patterns after T interactions governed by defined rules?
2. If there is any segregation, how would the system look like at both Micro-level (counties) and Macro level (whole network)?
3. How is the population density in the network going to change as a function of time?

# 3 Model

## 3.1 Data

We have a real network data set of USA counties where each node is a county and an edge defines a border between two given counties. We also have a separate data points of each county like education index, level of income, unemployment rate, life expectancy, and also index of obesity but we are not using that information for this project for simplicity sake.

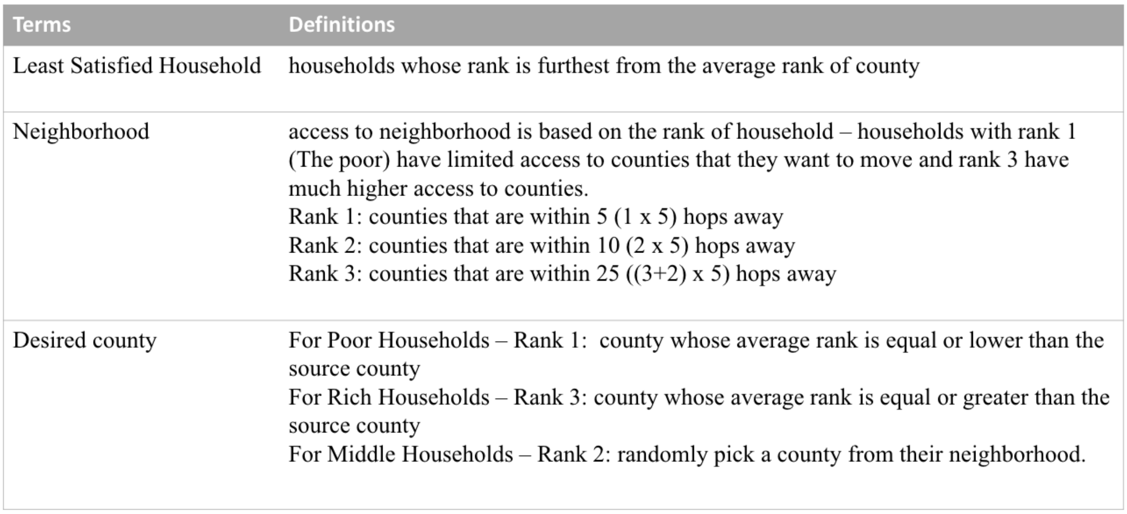
## 3.2 Model Setup

Our model is stochastic model that follows the dynamics of Voter Model in the network structure. The way we have setup this project is as follows: We defined population of each county (Node) where each household in the population belongs to one of the three groups: The Poor (1), The Middle class (2) and The Rich (3). The numbers 1, 2 and 3 also represents the rank of the household. Each node has specific properties defining the state of county (node) at each time step t. These properties include an array of population comprising of P elements that we here know as households in each county. This array with the length of P at the beginning includes households of class 1, 2 or 3 randomly distributed. The second property for each node is the average rank for the node. This rank is determined by the average of the all P ranks of households in each node.

For each household, the distance from the average rank of the county it is in, shows the level of unhappiness/dissatisfaction. The higher this number is, the unhappier and more dissatisfied the household is and more likely to move to a new county that has a rank closer to its own rank. The rank of household also defines the number of resources available to them. For instance, a household with income rank of 1 does not have as much resources for relocating as someone from the level 3. To reflect this condition in the model, we decided to use how many edges away can a household move to, based on its rank. A household of class one can only move to the counties that have closer average rank but are only five edges away. This number for classes 2 and 3 is ten and twenty five edges away respectively. This is where the number of neighbor property of each node comes to a great use. By this definition it is expected that for each household there is a different radius through which they they can relocate. This range is the largest for the Rich.

## 3.3 Model Vocabulary

Here is the list of terms along with their definition that we are using in our model to convey a clear picture.

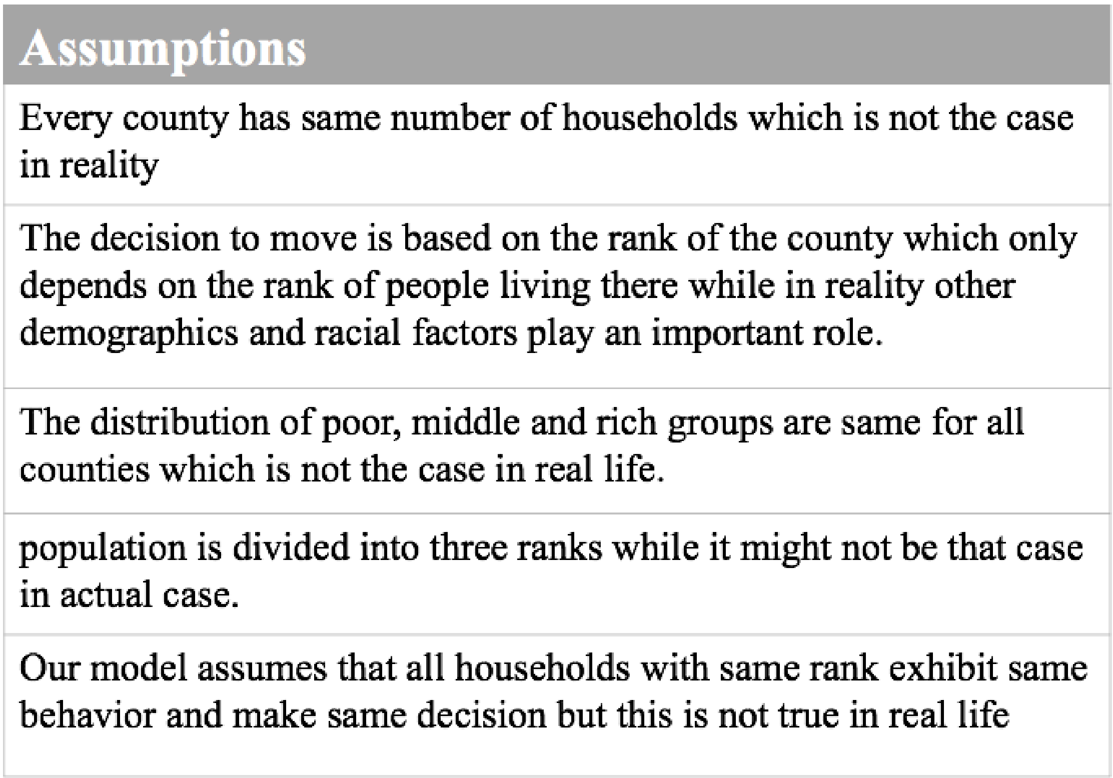


| Terms | Definitions |
|---|---|
| Least Satisfied Household | households whose rank is furthest from the average rank of county |
| Neighborhood | access to neighborhood is based on the rank of household – households with rank 1 (The poor) have limited access to counties that they want to move and rank 3 have much higher access to counties.<br>Rank 1: counties that are within 5 (1 x 5) hops away<br>Rank 2: counties that are within 10 (2 x 5) hops away<br>Rank 3: counties that are within 25 ((3+2) x 5) hops away |
| Desired county | For Poor Households – Rank 1: county whose average rank is equal or lower than the source county<br>For Rich Households – Rank 3: county whose average rank is equal or greater than the source county<br>For Middle Households – Rank 2: randomly pick a county from their neighborhood. |

Figure 1: model vocabulary and definition

### 3.4 Assumptions

In developing this model, we have made some assumptions for simplification. It is assumed that people move to another place just based on the financial status and the desire to live in an area where there are more people around them of their own type while in reality, many other factors might play a role in this decision. Second, our model assumes that there is equal population in all of the counties at initial time-step t while in reality the distribution of people with different levels of income is not the same for every place. Moreover, our model assumes three groups based on their income and also it considers all people in single group make the same decision and exhibit similar behavior while in reality that is not true.



| Assumptions |
|---|
| Every county has same number of households which is not the case in reality |
| The decision to move is based on the rank of the county which only depends on the rank of people living there while in reality other demographics and racial factors play an important role. |
| The distribution of poor, middle and rich groups are same for all counties which is not the case in real life. |
| population is divided into three ranks while it might not be that case in actual case. |
| Our model assumes that all households with same rank exhibit same behavior and make same decision but this is not true in real life |

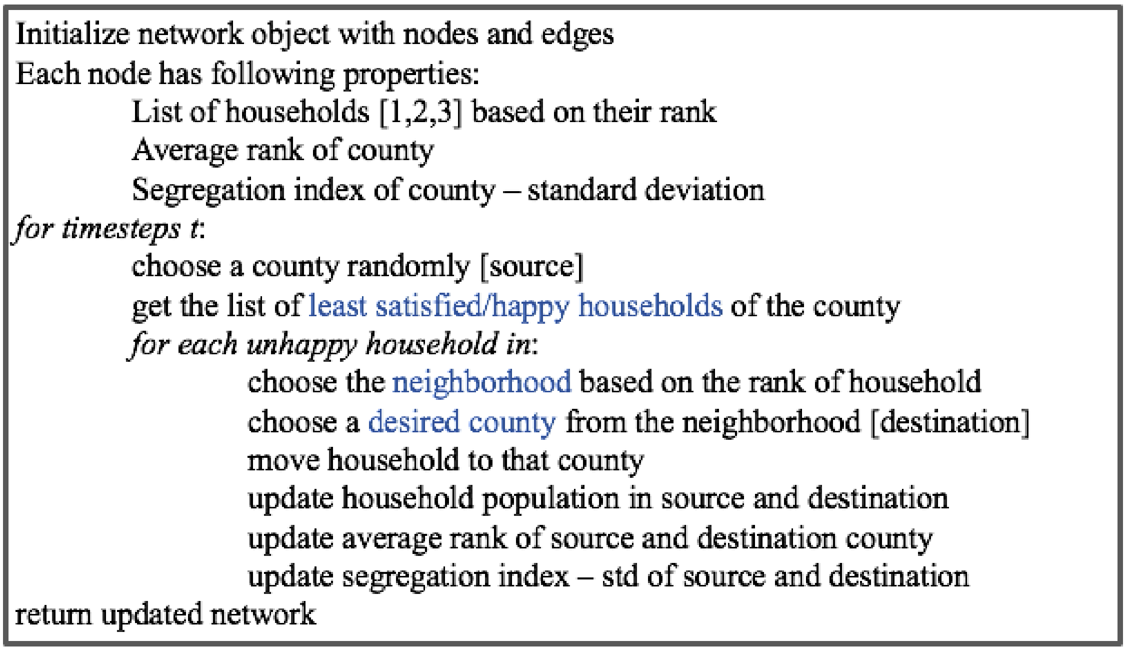
Figure 2: Model Assumptions

### 3.5 Algorithm and Rules

In this section, we will briefly describe the algorithm and rules to model the system. We used Python to build and simulate the model.

The first step is to introduce the network including nodes and how they are connected. The next step parameters were set up and the initial state of the network was defined. Neither the nodes nor the edges are going to change during the process. What gets changed is the attributes defined for each county (node), the array of population composing of P randomly distribution of 1, 2 and 3 and average rank of the node. In terms of the population array, the length of this array varies every time one household moves to another county based on the defined rule. In result, average rank, population count, standard deviation is going to vary as the system runs.

The following figure defines the overall algorithm and how the model is setup. The terms highlighted in blue are explained in the table above in model vocabulary section.



```
Initialize network object with nodes and edges
Each node has following properties:
        List of households [1,2,3] based on their rank
        Average rank of county
        Segregation index of county – standard deviation
for timesteps t:
        choose a county randomly [source]
        get the list of least satisfied/happy households of the county
        for each unhappy household in:
                choose the neighborhood based on the rank of household
                choose a desired county from the neighborhood [destination]
                move household to that county
                update household population in source and destination
                update average rank of source and destination county
                update segregation index – std of source and destination
return updated network
```

Figure 3: Algorithm to simulate the model

## 4 Results

In this section we will show the results that we got from running the model for 3000 time steps. The figures are setup in a way that the top three plots within an image display the initial state of the system and bottom three pictures display the state of system after 3000 time-steps and with population count of 50 households in each county. The Only difference in these three figures is the initial population distribution of the nodes which seems to play an important role in reaching macro and micro level segregation.

Plots a,b,c in the figures 4,5,6 are the measure of segregation, Dominance and dissatisfaction respectively at the initial state of the system, Plots d,e and f in these figures are the same measures after running the model for 3000 time-steps.

### 4.1 Measure of Segregation

standard deviation is one of the values that we have used to measure the segregation in micro level (county) and macro level (overall network). s standard deviation closer to 0 means that all of the households in the node have the same rank in it implying the county is segregated at micro level and the standard deviation closer to 1 means a county has households of all ranks uniformly distributed, implying that there is no segregation. The plots (a&d) in figures (4,5&6) show the measure of segregation. The darker the color is, the segregated the counties are.

### 4.2 Measure of Dominance

Measure of segregation although capable of demonstrating whether all households in the county have the same rank or not, it does not give any information about what rank is dominant in the county. So, we defined measure of dominance as an indicator of the average rank of a county. The idea is if there are more households with rank 3 then average rank of the county should be closer to 3 and the same for 1 and 2. The plots (b&e) in figures (4,5&6) show the measures of dominance.

### 4.3 Measure of Dissatisfaction

To measure how the dissatisfaction level of households with different ranks changes as the model runs we introduced this parameter. Dissatisfaction for a household defines as the difference between a household's rank and the average rank of the county it is in. The bigger the difference, the more dissatisfied a household is. The maximum value of dissatisfaction is 2 while the minimum value is 0. In a complete equilibrium and fair system, the dissatisfaction measure for households of all ranks should be closer to 0 after MAX TIME T but in reality, households with higher rank have higher resources so they should be less dissatisfied with other two on average. The plots (c&f) in figures (4,5&6) shows the dissatisfaction distribution for three different ranks.

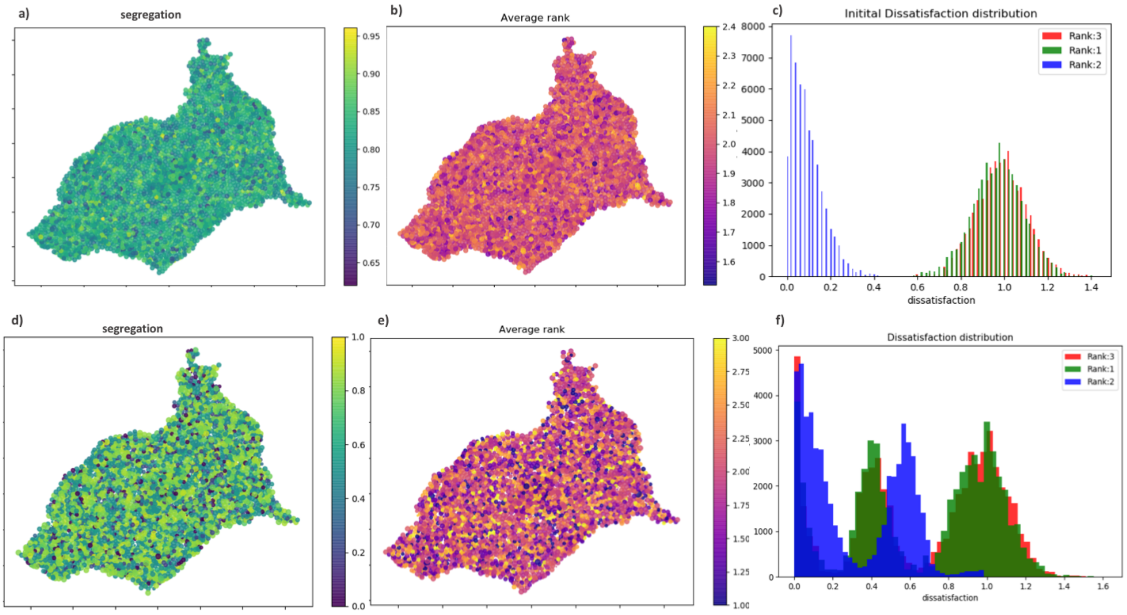


Figure 4: Model Simulation with Pop. Distribution: (1,2,3) = (34%,33%,33%)

Figure 4 shows the results of the model when households with different levels of income were uniformly distributed in counties. As mentioned earlier plots (a,b,c) show the model state at time t0. We can see that there is no segregation (a) as most counties have standard deviation closer to 1. Most counties have average rank closer to 2 (b) as population rank is uniformly distributed. Since average rank of counties is closer to 2, households with rank 2 are the least dissatisfied while households with rank 1 and 3 are most dissatisfied (c). The plots (d,e,f) show the changes in the network after 3000 time-steps. We can see that there is micro level segregation in plot d: counties with standard deviation closer to 0 and darker in color. Though there is not apparent macro level segregation where counties sharing edges would all be segregated combined but it is interesting to see how segregated counties seem to form chains. The overall rank of counties in network seem to go down to being closer to 1 (e) implying that household with rank 1 to move a lot and form segregated counties. We can also see that dissatisfaction level for rank 1 and rank 3 has gone down (f) as dissatisfaction distribution is skewed towards 0.

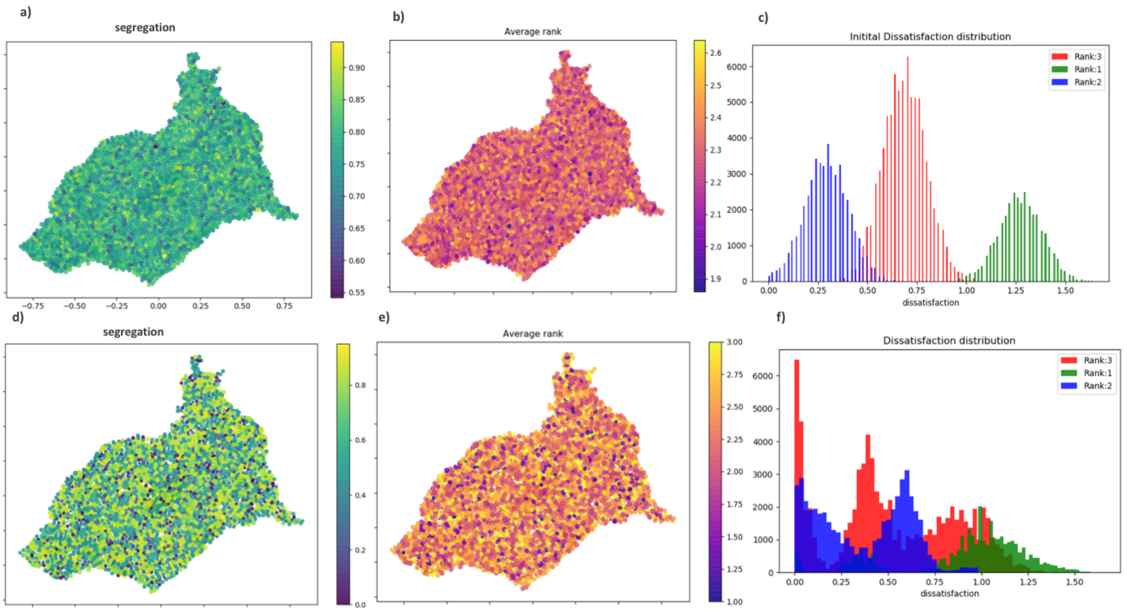


Figure 5: Model Simulation with Pop. Distribution: (1,2,3) = (20%,30%,50%)

Figure 5 shows the results of same experiment but with the change in initial population rank distribution. population of different classes is distributed in a way that there are 50% of household with rank 3, 30% of households with rank 2 and 20% of household with rank 1. Initially, the model exhibits a behavior as it should: there is no apparent segregation (a), the overall average rank of model seems to lean towards lighter colors implying the overall network rank is higher (b) and households with rank 1 are most dissatisfied (c) that's because most counties have higher rank than that closer to rank 1.

The plots (d,e,f) show the changes in the network after 3000 time-steps. Similar to figure 5, We can see the signs of segregation at the county level (d): counties with standard deviation closer to 0 and ones in darker in color. The overall rank of counties in network seem to go up to being closer to 3 (e). It is interesting to note that the counties with lower average rank are surrounded by the counties with higher average rank. We wonder if it is because there is higher number of high ranked households in the model or it is showing the underlying mechanism. Moreover, we can also see that dissatisfaction level for all ranks has seen a decrease implying that relocation based on financial status might be a good indicator.

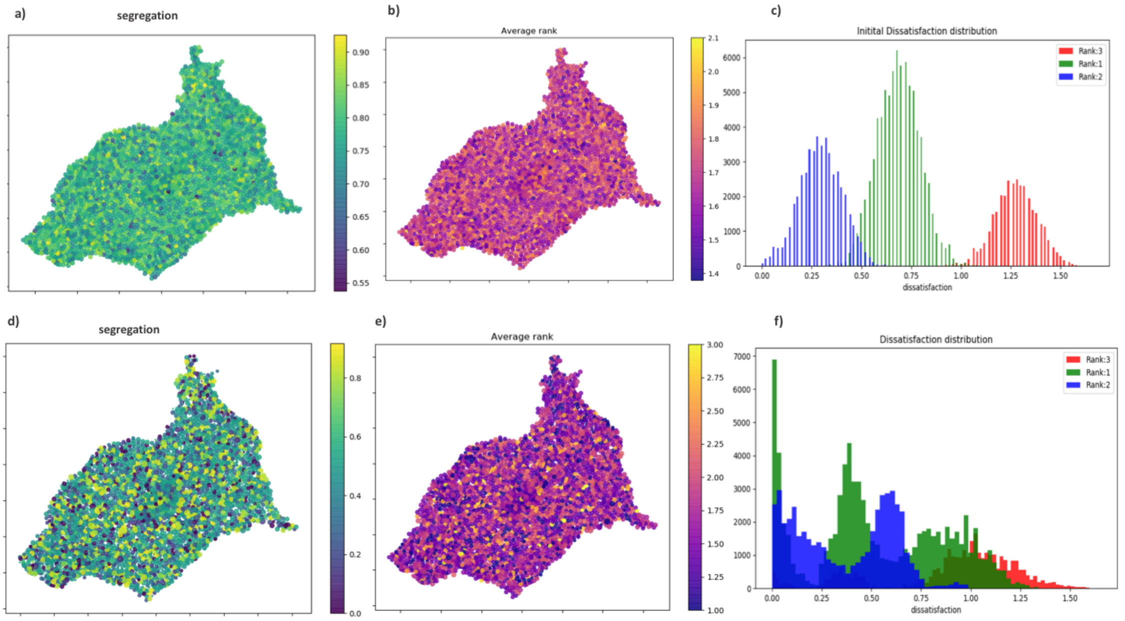


Figure 6: Model Simulation with Pop. Distribution: (1,2,3) = (50%,30%,20%)

Figure 6 shows the results of similar experiment but with the change in initial population rank distribution. Ranks in population is distributed in a way that there are 50% of household with rank 1, 30% of households with rank 2 and 20% of household with rank 3. At initial state, there is no apparent segregation exhibited (a), the overall average rank of model seems to lean towards darker colors implying the overall network rank is lower (b) and households with rank 3 are most dissatisfied (c) that's because most counties have lower rank than highest rank. After 3000 time-steps, we can see that the network is much more segregated towards darker side, There are some counties with complete segregation but there is no apparent macro level segregation. The overall rank of counties in network seem to go down to being closer to 1 (e). It is interesting to note that the counties with higher average rank are surrounded by the counties with lower average rank. We can also see interesting pattern in the dissatisfaction distribution. Though distribution is skewed towards 0 implying there are less dissatisfied people overall but some of the rich - households with rank 3 are most dissatisfied. This may be because how the system started with most households with lower rank and so rank 3 households did not have any place to move to of their interest.

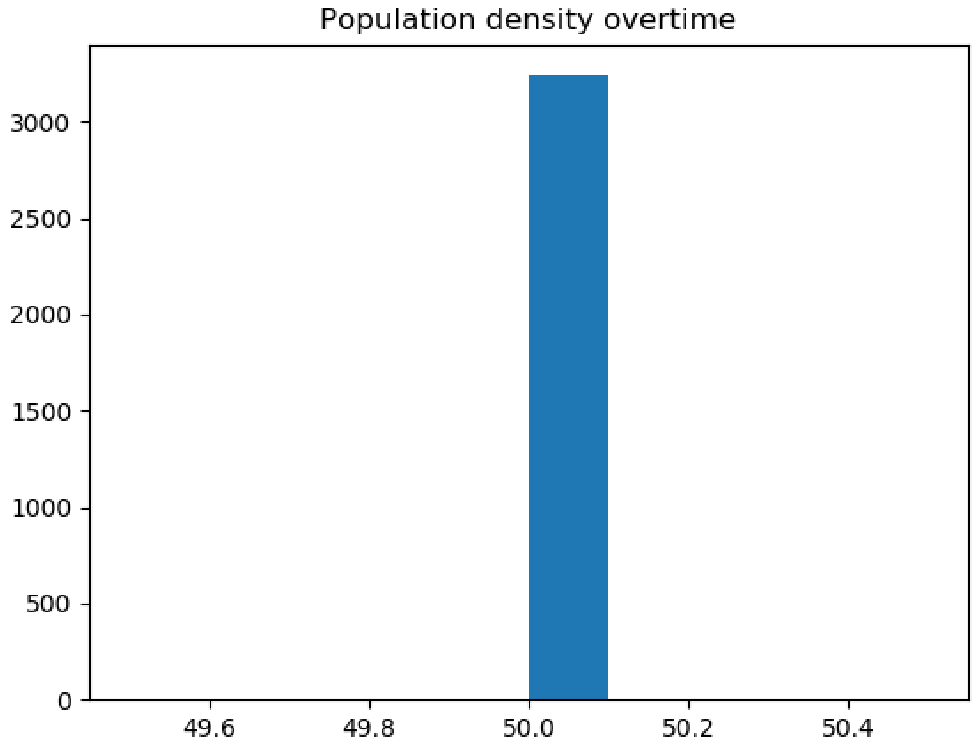


Figure 7: Initial Population Density

Figure 7 shows the population density in the network. Since each county is initiated with same number of population (p=50), every county lies in bin 50.

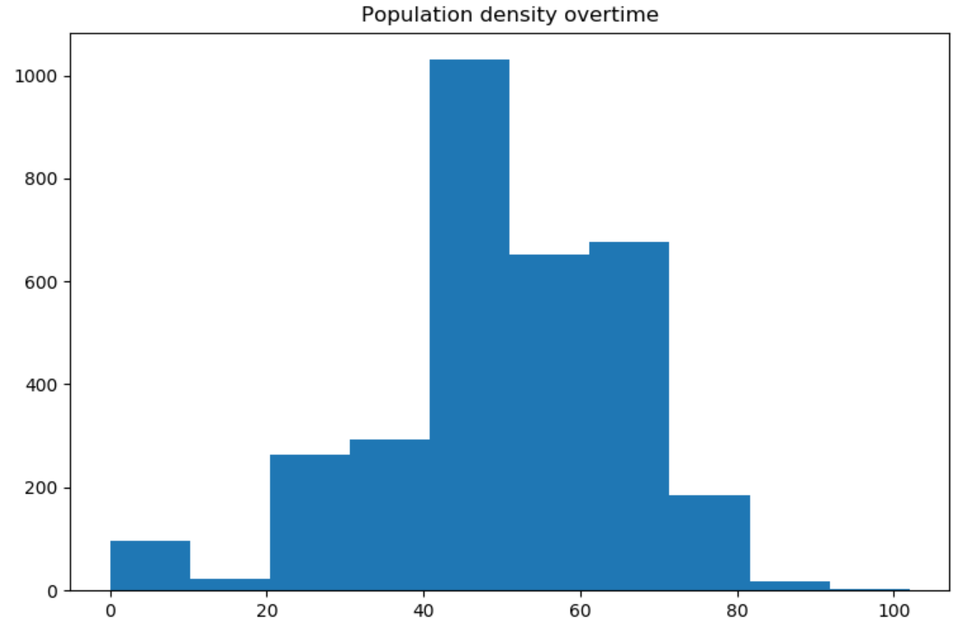


Figure 8: Population Density after 3000 time steps

Figure 8 shows the population density in the network after 3000 time-steps. We can see how population density is changed in the network after relocation of household. A similar distribution takes place regardless of population distribution at the start. We can see that about 1000 counties have population count as 50 implying that 3000 time-steps might not be enough iterations. We think that running the simulation for longer time-steps would result in interesting results.

### 4.4 Discussion

In order to quantify the segregation in overall network, we need a quantitative criterion capable of comparing levels of segregation in different conditions of the system [3]. In literature, measurement of segregation has been defined as the difference between the number of cross-group ties expected by chance and the number observed measured segregation [3, 4, 5]. Over the course of time, different methods have been introduced to do so. Duncan and Duncan suggested that the index of dissimilarity (DSI) can be the most desirable measure of segregation due to its simplicity of calculation. There is also another index being frequently used in analyzing segregation known as Freeman Segregation Index (FSI). The limitation with these two measures is that they just can be beneficial if the system has two states, while our model has three states. As a result, DSI and FSI are not good enough representatives of the segregation in our system. But we think Entropy Index is capable of quantifying segregation and distribution of multiple groups in the network structure [3].

In our system there are 3140 nodes (tracts) and each of these nodes has a specific distribution of itself. At time-step=0 all of the nodes have 50 households. As the system runs the number of households in each county can differ but the whole number of the nodes will remain the same. It is worth mentioning that the sum of the households is also fixed, they are just moving between counties. The entropy index $\boldsymbol{h}$ for each node $\boldsymbol{i}$ (tract) is defined as follows [6, 7]:

$$h_i = -\sum_{j=1}^{k} P_{ij} * ln(P_{ij})$$

where:

- hi = segregation entropy index
- K: Number of income levels

- pij: Proportion of population of jth income level in tract i (nij/ni)
- nij: Number of population of jth income level in tract i (nij/ni)
- ni: Total number of population in tract i

At maximum, ***h*** equals ***Ln(k)*** which would be ***Ln(3)*** for our system since there are three groups of households in the system. The minimum value is 0. The way this index can give us an overview of the segregation in the system is that the higher the h index for a tract, the more diverse the counties are, By this formula then it is very easy to see the course of change of the h index for each and every one of the counties. Having the progress of diversity in nodes, we also liked to see how the segregation in the whole system has gone through change during for instance 3000 iterations. To achieve this, we tried to apply the same intuition to the overall network instead of one single node. As a result, we decided to substitute nodes with the whole network and households in each node with the nodes in the system. And to calculate the Pij, we considered average rank of the node instead of rank of each county. This value is assigned by rounding up the average calculated from household incomes in each node. By assigning this, now we can calculate the h index not only for each node at all time-steps, but also for the 3140 nodes as a whole network. The h index being closer to ln(3)=1.1 would mean network is not segregated at all while h index closer to 0 would imply the signs of segregation in the network

### 4.5 Conclusion

Our model of economical segregation is able to capture the micro level segregation in counties based on their economic rank. It shows that the overall behavior of the system changes depending on how the population is distributed at the initial state in the system. It is also able to capture the fact that higher ranked households having more resources and options to relocation are more satisfied regardless of population distribution at initial state. This makes sense because they have more power in their hands to change their circumstances and relocate where they want to. Though the model shows promising results, there can be many improvements made in the model that would give more realistic results. Some of the shortcomings in this model are as follows:

We believe that model should be run for much more time steps to see steady state where there is apparent macro level segregation. Large network size would need more update to see significant changes. We were not able to do it because it takes a lot of time to run it.

We also think that it would be more realistic to have varying population size to different counties based on some real parameter because the size and distribution of the population would result in different results.

## 5 Code and Data File

We have attached the code and network data file with our submission. The code is commented and will help you get the reproducible results.

## References

[1] Schelling, Thomas C. "Dynamic models of segregation." Journal of mathematical sociology 1.2 (1971): 143-186.

[2] Hatna, Erez, and Itzhak Benenson. "The Schelling model of ethnic residential dynamics: Beyond the integrated-segregated dichotomy of patterns." Journal of Artificial Societies and Social Simulation 15.1 (2012): 6.

[3] Cortez, Vasco, and Sergio Rica. "Dynamics of the Schelling social segregation model in networks." Procedia Computer Science 61 (2015): 60-65.

[4] Freeman, Linton C. "Segregation in social networks." Sociological Methods and Research 6.4 (1978): 411-429.

[5] Moody, James. "Peer influence groups: identifying dense clusters in large networks." Social Networks 23.4 (2001): 261-283.

[6] White, Michael J. "Segregation and diversity measures in population distribution." Population index (1986): 198-221.

[7] White, Michael J. "The measurement of spatial segregation." American journal of sociology 88.5 (1983): 1008-1018.