Delft University of Technology

Departement of Electrical Engineering, Mathematics and Computer Science

TI3150TU

Social Mixing in US Cities From Social Network Data

Final Report

Authors:

Student Name	Student Number
M.M. van Arnhem	4918738
V.W. Chen	5169003
V.K. Kuramae Izioka	4685822
N. Tran	5168767
K.L. Reith	4667247

Supervisors:

Name	Role
T.J. Viering	Responsible Instructor
L.J. Spierenburg	Client
D.V.Q. Nguyen	Teaching Assistant

Date: 23-01-2022

Contents

1 Introduction

In the following section the problem will be described along with the goals of the research. Specification and scope of the project will be given.

1.1 Background and Project Aim

People can be clustered into social groups; clumps of individuals who interact with each other. These social groups are defined by the relatively high interaction levels within the group, but they can still interact with other social groups, which is the case with social mixing. Social mixing is thought to reduce income inequality and increase student performances (McCoy et al. 2014), thus the opposite, which is social isolation, will need to be prevented.

The aim of this research is to find a possible correlation between physical features and social isolation in cities. This research could help cities solve problems around urban design and social mixing. Namely, as earlier research has shown, there is a positive though limited correlation between physical features such as the segregation by physical barriers and the social fragmentation in Hungarian cities (Tóth et al. 2021). This research will however inspect 50 cities in the United States. The data used is from the closed down social media platform Gowalla, which features data from Feb. 2009 - Oct. 2010. The data includes friendship connections of Gowalla users as well as geographical location information on social media posts made by users. The hypothesis is that a city with a lot of spatial division generally includes more social isolation than a city with little spatial division. Spatial division can be defined by physical barriers, for example rail roads or rivers that go through a city, but could also be defined by urban sprawl.

The final product should show possible correlations between physical features and social isolation. Using the data, appropriate metrics have to be found for both the spatial constraints and the social isolation in order to analyze the correlation, which is the main challenge for this project. To tackle this challenge, a correct validation process is essential, since false conclusions are likely to happen when designing metrics. The final product can be beneficial for many stakeholders. The stakeholders for this project are: the machine learning developers, urban designers, users of the app Gowalla, citizens of the place where the model will be implemented, governmental bodies and public administration. While the research can evidently be of use for society, the research will also look into the ethical considerations. A more extensive elaboration on the stakeholders and ethics can be found further on (section 6.2). This project is meant to support a PhD project conducted by the client concerning social isolation in the examined cities. The scope of this project only includes the correlation between the physical constraints and urban features of cities and their social isolation level. Research regarding the characteristics, like income and education, of social groups and their isolation lies not within this project, but is an element of the PhD project.

2 Methods

In this chapter an overview will be provided of the steps taken to produce the results, which will be further elaborated in the sub-chapters. Along with the validation approach that describes the method of our result validation.

In order to produce results a number of steps were taken. Below a short overview of these steps taken can be seen, while the rest of this chapter will go more in depth on each individual steps.

The steps taken are largely based on the predefined requirements (See appendix A), but deviate with respect to the order. The steps are as followed:

1. Define city shapes in terms of geographical coordinates (polygons);
2. Assign users to cities based on these new geographical city shapes;
3. Build a social connection network per city using the connections between individuals;
4. Build a movement network describing visiting patterns per city using the check-ins of individuals;
5. Determine the (social) isolation in a city based on the social network as well as the movement network;

6. Determine the physical constraints present in each city, using several metrics;
7. Analyse relations between the (social) isolation metrics and the physical metrics.

2.1 Validation Approach

A validation approach has been established in order to verify whether the set requirements (appendix A) have been accomplished and to analyze the validity of the results produced in this research.

In order to be able to verify the completion of the requirements, the requirements were made according to the SMART philosophy (Mannion and Keepence 1995). As a result, the requirements are measurable and it is binary whether they are achieved or not. Whether the results produced are correct is not included in the description of the requirements, however this was further discussed with the client. In order to analyze the results they will be compared to the results of a similar research which was performed in Hungary (Tóth et al. 2021). In the Hungarian report, (income) inequality was also part of the research, but this lies outside of the scope for this report.

In conclusion, validation will be done with the use of a Hungarian report (Tóth et al. 2021). In order to correctly complete the requirements, a comparable methodology should be deployed, the results should be compared and correctly explained for significant differences. The main measure for comparing will be the correlation coefficient between the physical and social metrics. This validation approach was discussed and agreed upon with the client.

2.2 Shape of the Cities

An essential part of building a network per city is assigning the users to a city. As this is done using the user's location data we first require to define the boundaries of all cities of interest. A general definition used is the so-called Functional Urban Area (FUA), which defines the geographical shape of a city as an area within people commute towards a shared centre. For our purposes, however, this is not sufficient as this results in large city shapes with large low-density areas. This was improved by implementing a convex hull, drawn around the high density ($> 100\ p/km^2$) parts within the previously mentioned FUA's. See Figure 1.

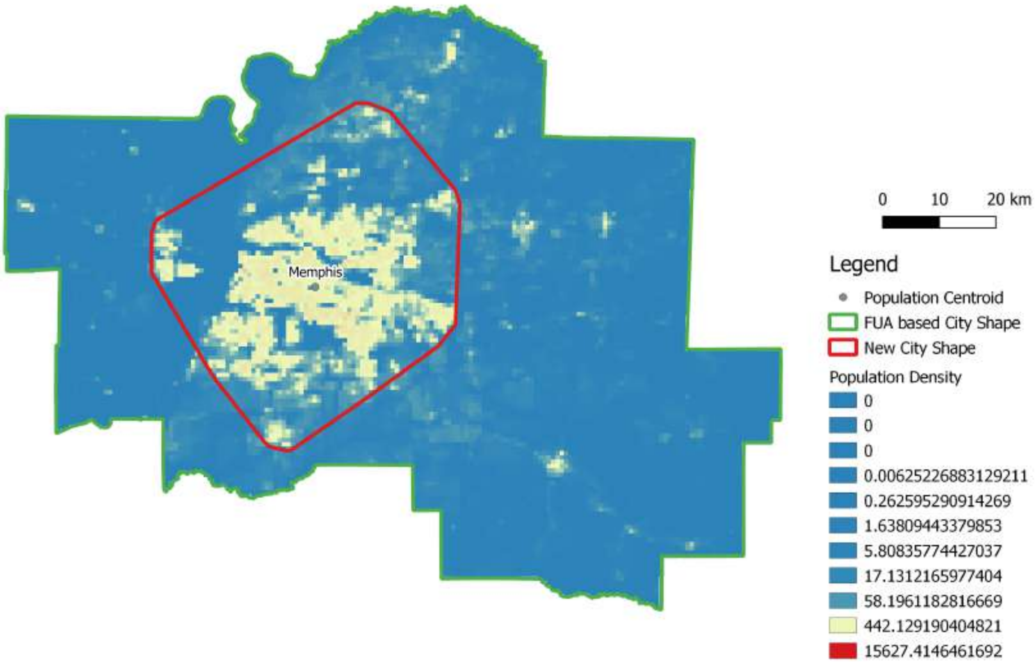


Figure 1: The newly defined shape of Memphis together with The Functional Urban Area based shape.

2.3 Network Analysis

For the social isolation, the dataset of the Gowalla app has been used. This contains a database of 196,590 users in the United States. This database is broken up into two parts:

- The edges: the friendship connections between users. In total there are 1,900,654 edges.
- The check-ins: the locations where the users have shared a social media post. In total there are 6,442,892 check-ins.

This data is then used to build both a social a network, based on the connections between individuals; as well as a movement network, based on the check-in data of individuals. Both networks are explained below.

2.3.1 Social Connection Network

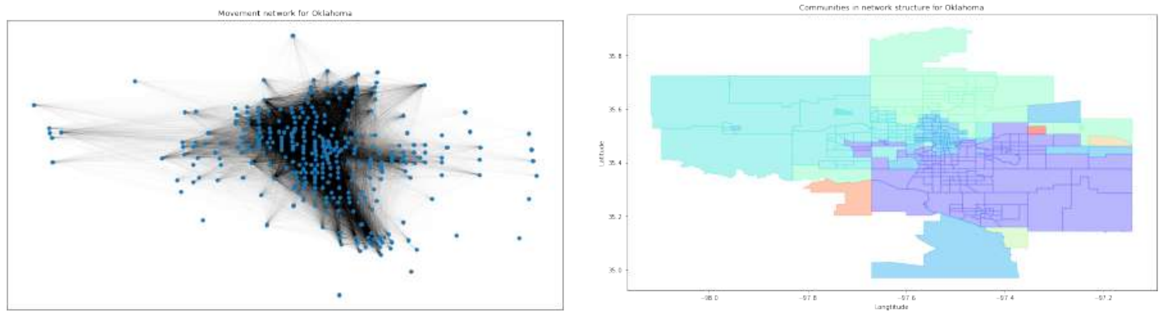
The social connection network is based on the connections (friendships) between individuals. For every city only the connections between users in the city are taken into account to build a network. To assign people in a city, the check-ins of the users are necessary. By linking those check-ins with the city shape files, described in chapter 2.2, the users are linked to every city they have been to according to the Gowalla app. It is possible that users have been assigned to multiple cities. Meaning that users are also assigned to cities when having a few check-ins in that city. This could lead to users not being part of a community in that city or being part of a disconnected community both of which will influence the metrics results. To remove those misclassified users in the city, the communities have to be determined to check if those users are a part of the society in that city, this will be handled by detecting communities.

Another important issue to take into account is about the inactive users of the Gowalla app. But defining criteria for the check-ins and connections of the users would be difficult. And since the subject of this project is about communities, the filtering of the inactive users will happen by detecting communities.

A lot of cities have a few users, this does not give a good reflection on the reality and would affect the results in a negative way. To this extent, only the 50 biggest cities (according to the amount of users of the Gowalla app) will be part of this project.

2.3.2 Movement Network

A big issue in this project is the limitation of the amount of data. One way this could lead to inaccuracies in the results, is due to outliers damaging the accuracy of the community structure. A possible solution that was applied in the project is to zoom out and look from a neighborhood scale rather than an individual scale. Besides that, this network focuses on the mobility of the individuals, looking into the spatial behaviour within the cities. This network is based on the check-ins of the Gowalla app users instead of the friendship connections. Every neighborhood is a node and the edges between those nodes are weighted by the number of people that have visited both neighborhoods. The neighborhoods are defined by the census tracts (USCB 2016a).


Movement network for Oklahoma
Communities in network structure for Oklahoma
35.8
35.6
35.4
35.2
35.0
Latitude
-98.0
-97.8
-97.6
-97.4
-97.2
Longitude

(a) Movement network of Oklahoma with weighted edges

(b) Neighborhoods colored by the communities of the movement network of Oklahoma

Figure 2: Networks of Oklahoma

Table 1: A comparison on modularity, connectedness and loading times for the three considered community detection algorithms

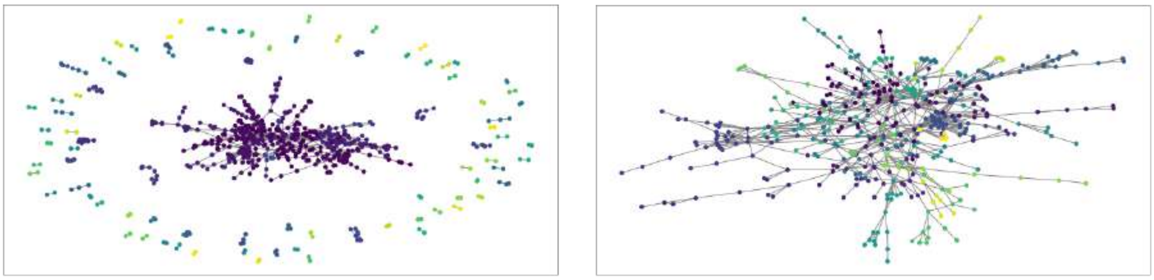
Algorithm	Average Modularity	Disconnected Graphs	Average Loading Time
Clauset-Newman-Moore	0.648890	0	5.877161
Louvain	0.671232	4	0.266733
Leiden	0.677273	0	0.027750

2.3.3 Defining Communities

Three different algorithms were applied for finding a well defined community structure within each of the social/movement networks. The three algorithms used are the Clauset-Newman-Moore algorithm (Clauset et al. 2004), the Louvain algorithm (Blondel et al. 2008) and the Leiden algorithm (Traag et al. 2019). Each of these algorithms work by a greedy, modularity based method for finding the best possible structure. A community structure is better when the nodes interact more with others within their community, rather than with others outside of their community, and the modularity is a quantification for this. The 'greedy' part means that during each stage, the locally optimal choice will be chosen.

There were three characteristics by which the algorithms were ranked in order to decide which one to use. The first one is the modularities it produced, for which higher modularities meant better performance. When the modularity is high, it means the algorithm classified the users in communities with others it interacts with, so the community structure is correct. The second characteristic is the loading time, or cost, of the algorithm and the third one is whether the community structure featured internally disconnected communities. When the community structure features internally disconnected communities, the modularity will become ill defined due to bad community identification (Traag et al. 2019). The Leiden algorithm ranked the highest for each of these characteristics (as seen in figure 1), so it was chosen as the algorithm to be used from hereafter.

For each algorithm, completely isolated communities were identified. To be able to analyze a network correctly it needs to be a connected network, so either edges would have to be added between these isolated communities and the rest of the network, or those communities would have to be deleted. Since the former would need state of the art methods and is not required for the project aim (as discussed with the client), it was decided that deleting the isolated communities would be sufficient, see figure 3.



(a) Network with disconnected communities

(b) Connected network

Figure 3: Community graph for the social connection network in Sacramento before and after removing the disconnected communities

2.4 Social Isolation Metrics

To quantify the social isolation within a city, the following five metrics were found and applied:

- Modularity
- Scaled Modularity
- Betweenness Centrality
- Community Fragmentation

- Ratio Disconnected Users

In the following paragraphs these metrics will be explained and supported. For each of these metrics a high number means more social isolation.

2.4.1 Modularity

The modularity, which can directly be extracted from the community detection method, describes how much users interact within their community rather than outside their community. When building the community structure this number is maximized. However, a high modularity does mean that there is less interaction between communities. In other words, the higher the modularity, the less social mixing occurs, which means that there is more social isolation.

2.4.2 Scaled Modularity

As mentioned in Tóth's report (Tóth et al. 2021), the modularity is heavily influenced by the network size and density. To account for this bias the modularity was divided by the theoretical maximum modularity (Wachs et al. 2019), which resulted in the scaled modularity.

2.4.3 Betweenness Centrality

In order to inspect the quality of the network in terms of connections, the betweenness centrality was calculated for each city. This metric describes the how heavily the connectedness is influenced by single nodes. If the number is high, it means that communities are connected through few nodes, which means they are more segregated.

2.4.4 Community Fragmentation

For the social connection network, another metric option would be the 'community fragmentation'. This method is described by equation 1.

$$community\, fragmentation\, in\, city = \frac{amount\, of\, communities}{amount\, of\, users} \tag{1}$$

If the community fragmentation is equal to 1, every community consists of one user. If the community fragmentation decreases, the total users per community increases. Since the communities are based on the ties between individuals, this metric shows that a larger number means more isolation. If there is just one community in a city, it means that everyone has a connection with everyone and this would give the lowest community fragmentation.

2.4.5 Ratio Disconnected Users

The ratio disconnected users is a metric for the social connection network. This method takes disconnected users into account. Disconnected users are part of a completely isolated community, which were mentioned before. There are two options for the reasoning of this disconnection: inactive users that have a few friends or communities with only internal connections. The second group of users leads to more isolation in a city, because they are isolated from the rest, therefore they are taken into account in this metric. If the ratio is higher, there are a lot of disconnected communities, which means the city is more isolated.

2.5 Physical Constraints Metric

In this chapter the use of different metrics for physical constraints will be explained. As well as the types of physical constraints that have been taken into account. This is part of requirement 4: Must find a metric for physical constraints for all cities provided by the PhD student. And requirement 8: Should find at least one additional metric for the physical constraints in all cities provided by the PhD student. Found in the requirements appendix A.

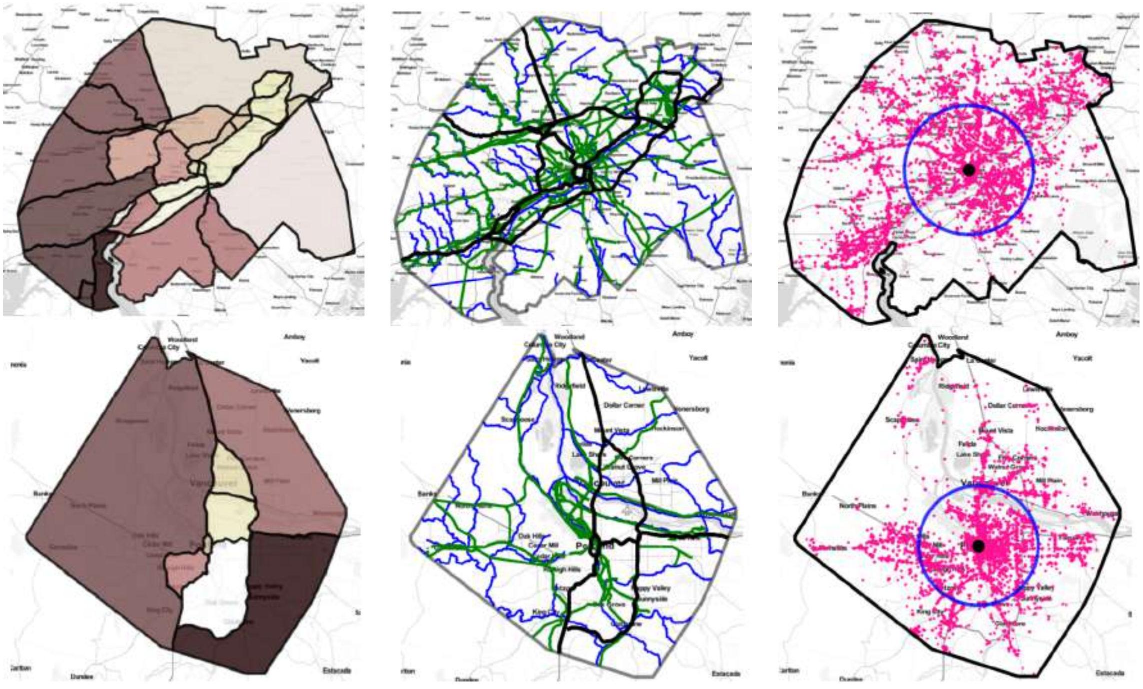
2.5.1 Types of Physical Barriers

There exists fast amount of different types of physical barriers that can be used to analyze segregation. Physical barriers come in different forms and can be differentiated between man-made or naturally occurring. The physical barriers that are most appropriate for isolation of communities would be those that limit or prohibits access and mobility. The most significant physical barriers will be determined by existing literature. Taking too many types into account will most likely make the results and correlation to social isolation too complex. The physical barriers that will be taken into account and the rationale therefore are listed below.

- **Interstate Highway System** was approved under president Eisenhower and congress in 1956 as to improve the mobility between cities. The interstate was only accessible through designated points and were very wide, often destroying housing, schools and facilities in its way (Dottle R. 2021). While interstate connectivity improved, within the cities connectivity between neighborhoods got destroyed. Planning and design documents have shown that the interstate tracts had been designed, with minority isolation in mind. This happened during a time where segregated neighbourhoods were increasing (Rose and Mohl 2012). Considering parts of this infrastructure has ties with segregation, it will be used as one of the barriers.

- **US Railway System** was created in the 19th century and segregation of minority communities is still seen to this day in some cities. Just like the interstate system this was build in a time where social engineering was prevalent (E. and D. 2015; Ananat 2011). Railway can only be crossed at certain intersections and heavily limits the mobility between neighborhoods.

- **Rivers** are one of the few natural barriers which, without infrastructure, completely isolates areas from each other.



(a) Interstate SPB index for Philadelphia (greater): 0.91 and Portland: 0.69. The FUA is subdivided by the interstate network, each color representing a different subdivided area.

(b) Average barrier index for Philadelphia (greater): 0.0028 and Portland: 0.0039. Each color represents a type of physical constraint. In blue the rivers, green the railroads and in black the interstate network.

(c) Average distance to center for Philadelphia (greater): 0.44 and Portland: 0.23. In the figure the FUA outskirt is marked by a black border, the pink dots represent check-ins within the FUA, the black dot the center of gravity and the blue circle the ADC.

Figure 4: Visualization of physical barrier metrics between greater Philadelphia (upper) and Portland (lower).

2.5.2 Segregation of Physical Barriers Index

To determine spatial divide of cities the Segregation of Physical Barrier (SPB) index will be used as one of the metrics. It is a measurement which takes into account how many area's there are within a city, which are divided by given physical constraints. This metric is based on the Railroad Division Index, which is an equation used to quantify to what extend a railroad segregates an area. The functionality of this metric can be expanded by using more than just the railroad network as a physical barrier (Ananat 2011).

The formula used is $SPB_i = 1 - \sum_a (S_a/S_i)^2$ where S_i refers to the size of the city's area and S_a denotes the size of area constraint by physical barriers. The SPB value is between 0 - 1 and can be read as the smaller the SPB index, the less the city is divided by physical barriers. And the larger the SPB index, the more fragmented the city is geographically by physical barriers (Tóth et al. 2021). As seen in figure 5.

2.5.3 Average Barrier Index

To make sure the SPB index gives us a correct indication on the measurement of the physical barriers, multiple indexes will be used to compare it with the SPB. The first comparison will be made with the Average Barrier Index (ABI). This is a simplified version of the final barrier, which works in three spatial class indices, within these three spatial classes the area is calculated for the barriers (Barber et al. 2021). Using the calculated area's, percentages can be gained in order to specify the frequency of physical boundaries in a given city. Scoring each physical boundary will result in frequencies of certain scores given the amount of physical boundaries given an area. With this every city will gain a simplified 'Final barrier' index which will be called the ABI, then comparison with the SPB index will be made. If the scores give a similar relation between the separation of the boundaries, then the SPB index will be a good indicator to use in further research. If they differ however another research might need to be done, to see if anything went wrong or where the variance lies. It might be that both index correlate other aspects of the physical barrier (see Figure 5).

2.5.4 Average Distance to Center

The center is an important location for social ties, as the probability for links to occur decrease further if the distance between the center is large. Thus as the distance grows the more fragmented social networks can become, because of less interactions between the individuals. As the location to facilitate integration is mostly centered in a specific location in the city, this mean that the center will be defined as the most dense populated space of the city. To calculate the effect of the distance to center the following formula will be used: $ADC_i = \sum_p^p \frac{D_{p,c}}{P} / S_i$, where $D_{p,c}$ is the distance from point p with center of gravity c of point P to the defined center. Lastly S_i is the size of the city's area. The ADC value depends on the size of the town and fragmentation of the population. Where a smaller ADC value indicates smaller cities with tight communities. And a larger ADC indicates bigger cities where communities are more spread out (see Figure 6).

2.5.5 Urban Sprawl

Interest in the relation between urban sprawl and isolation follows a similar reasoning as with the previous metrics: a more spread out city is expected to be less socially coherent.

To capture urban sprawl in a quantitative way multiple methods as outlined in OECD 2018 were tried. The one used for analysis is the mean population density (MPD), which is a way to describe how the population is located within a city (See Figure 6). A higher value means less urban sprawl.

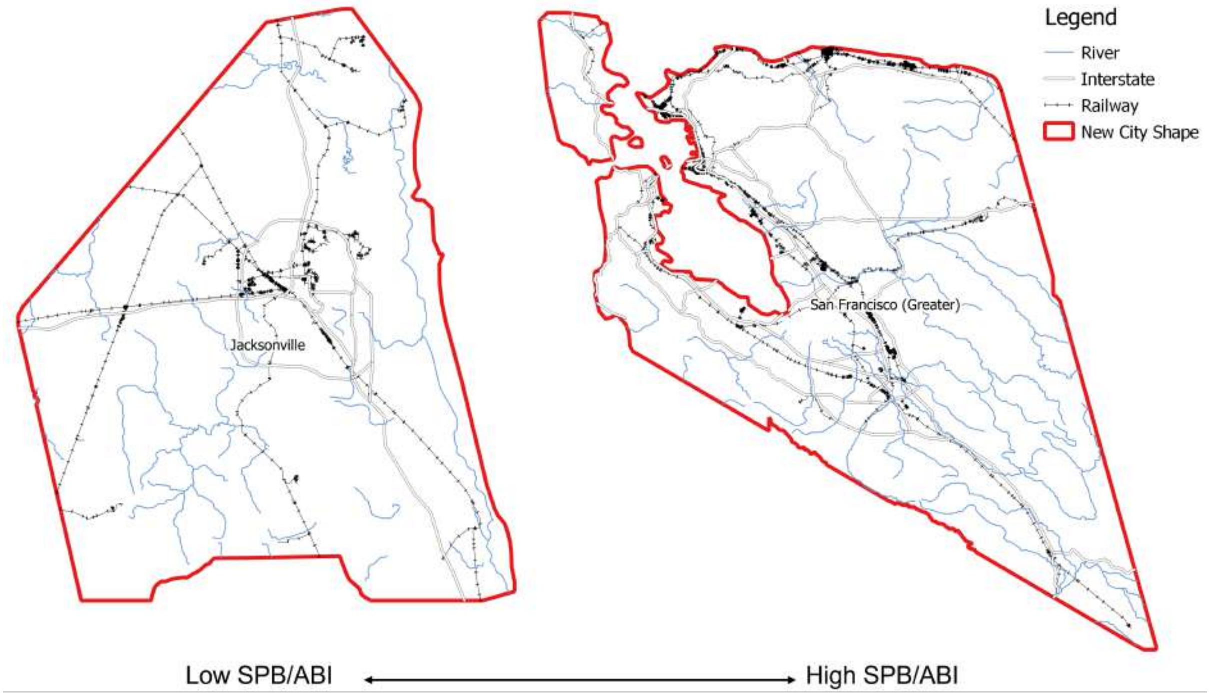


Figure 5: Jacksonville with relatively low values for the SPB and ABI metrics, compared to San Francisco which has relatively high values for the SPB and ABI metrics.

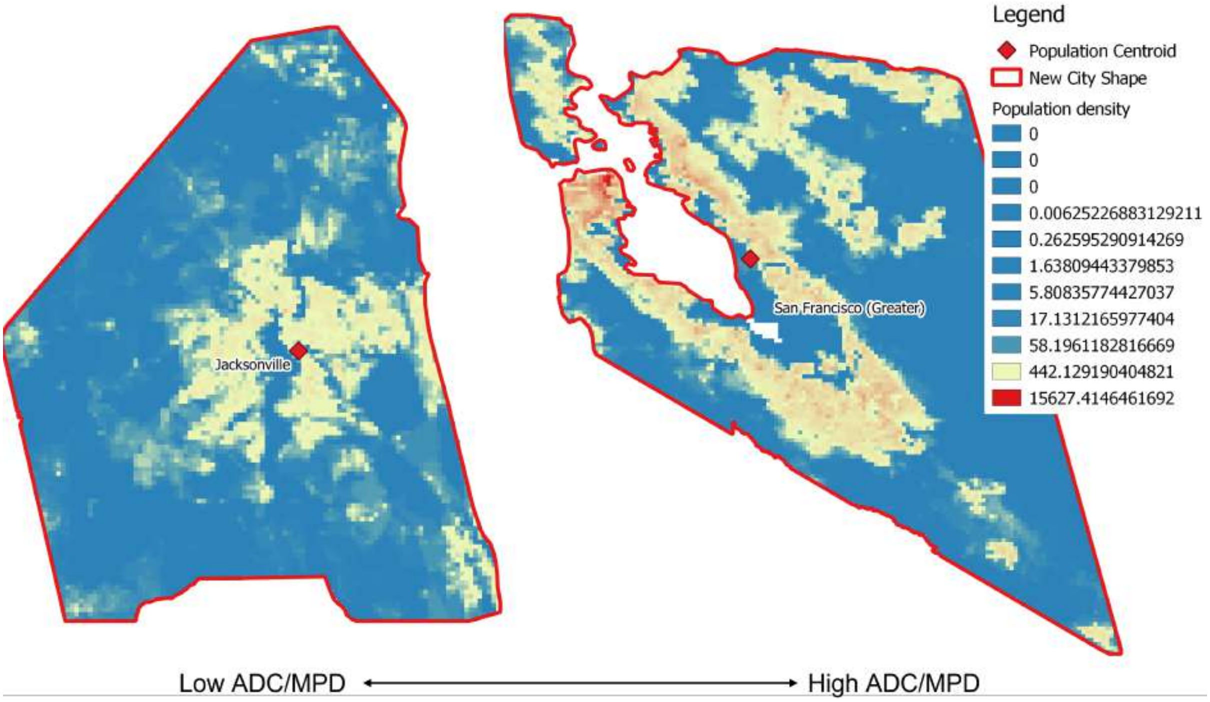
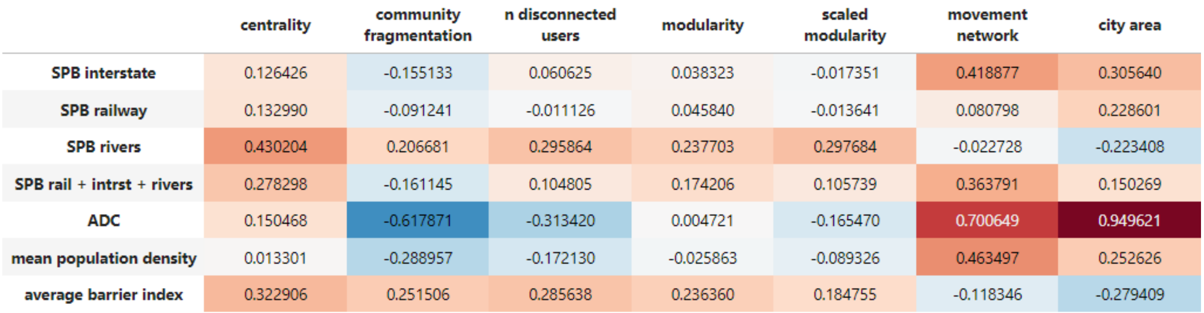


Figure 6: Jacksonville with relatively low values for the ADC and MPD metrics, compared to San Francisco which has relatively high values for the ADC and MPD metrics.

3 Results

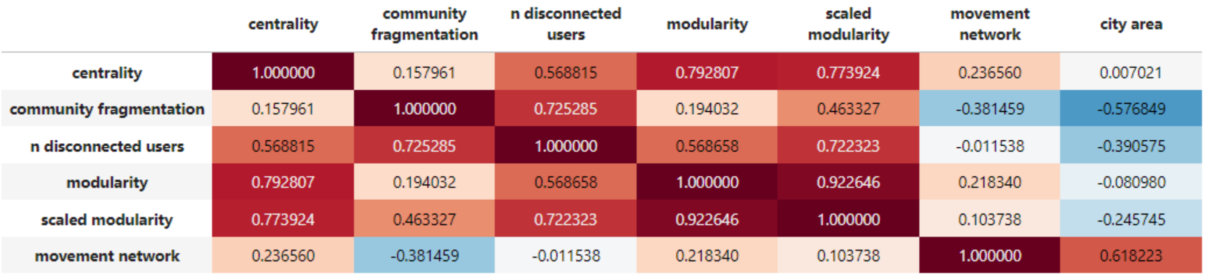
In this section, the most interesting results will be highlighted, which will be discussed further in sections 4 and 5. The purpose of this section is to show the level of correlations, which corresponds to requirements 5 and 9 (see: A).



	centrality	community fragmentation	n disconnected users	modularity	scaled modularity	movement network	city area
SPB interstate	0.126426	-0.155133	0.060625	0.038323	-0.017351	0.418877	0.305640
SPB railway	0.132990	-0.091241	-0.011126	0.045840	-0.013641	0.080798	0.228601
SPB rivers	0.430204	0.206681	0.295864	0.237703	0.297684	-0.022728	-0.223408
SPB rail + intrst + rivers	0.278298	-0.161145	0.104805	0.174206	0.105739	0.363791	0.150269
ADC	0.150468	-0.617871	-0.313420	0.004721	-0.165470	0.700649	0.949621
mean population density	0.013301	-0.288957	-0.172130	-0.025863	-0.089326	0.463497	0.252626
average barrier index	0.322906	0.251506	0.285638	0.236360	0.184755	-0.118346	-0.279409

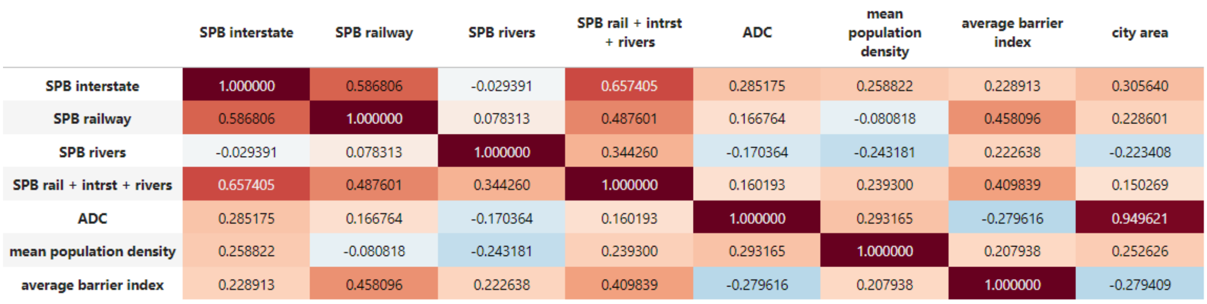
Figure 7: Correlation matrix between isolation metrics and physical constraints metrics. The complete correlation matrix can be found in appendix B figure 12.

By analyzing the correlation coefficients between the isolation metrics and physical constraints metrics separately, it can be determined whether or not there are biases in the chosen metrics.



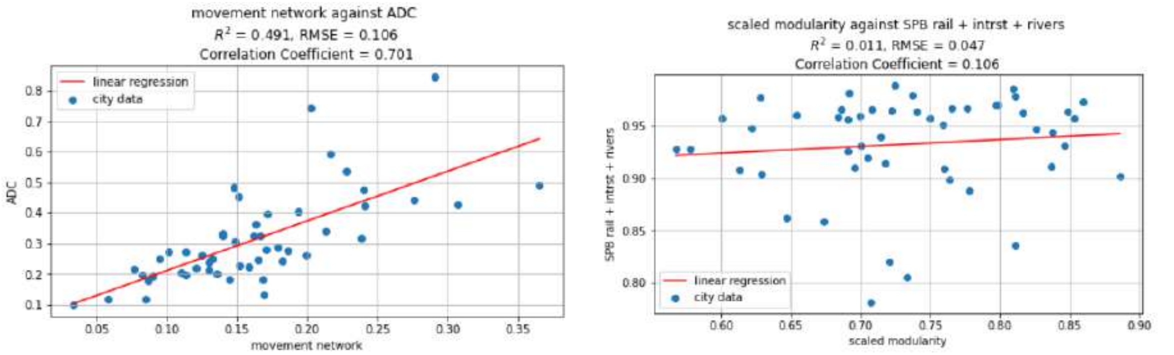
	centrality	community fragmentation	n disconnected users	modularity	scaled modularity	movement network	city area
centrality	1.000000	0.157961	0.568815	0.792807	0.773924	0.236560	0.007021
community fragmentation	0.157961	1.000000	0.725285	0.194032	0.463327	-0.381459	-0.576849
n disconnected users	0.568815	0.725285	1.000000	0.568658	0.722323	-0.011538	-0.390575
modularity	0.792807	0.194032	0.568658	1.000000	0.922646	0.218340	-0.080980
scaled modularity	0.773924	0.463327	0.722323	0.922646	1.000000	0.103738	-0.245745
movement network	0.236560	-0.381459	-0.011538	0.218340	0.103738	1.000000	0.618223

Figure 8: Correlation matrix between isolation metrics



	SPB interstate	SPB railway	SPB rivers	SPB rail + intrst + rivers	ADC	mean population density	average barrier index	city area
SPB interstate	1.000000	0.586806	-0.029391	0.657405	0.285175	0.258822	0.228913	0.305640
SPB railway	0.586806	1.000000	0.078313	0.487601	0.166764	-0.080818	0.458096	0.228601
SPB rivers	-0.029391	0.078313	1.000000	0.344260	-0.170364	-0.243181	0.222638	-0.223408
SPB rail + intrst + rivers	0.657405	0.487601	0.344260	1.000000	0.160193	0.239300	0.409839	0.150269
ADC	0.285175	0.166764	-0.170364	0.160193	1.000000	0.293165	-0.279616	0.949621
mean population density	0.258822	-0.080818	-0.243181	0.239300	0.293165	1.000000	0.207938	0.252626
average barrier index	0.228913	0.458096	0.222638	0.409839	-0.279616	0.207938	1.000000	-0.279409

Figure 9: Correlation matrix between physical constraints metrics



(a) Movement network against the ADC index

(b) Scaled modularity against the SPB index for interstate highway, railways and rivers

Figure 10: Relation between isolation metrics and physical constraints. (a): When looking at figure 7 the highest correlation found is between the movement network and ADC, with a correlation coefficient of 0.701 and r^2 of 0.491. (b): The relation between the scaled modularity and SPB index for highways, railways and rivers

4 Discussion

In this chapter the results will be discussed along with the limitations and recommendations. The following four statements are discussed:

1. The movement network has a positive correlation with the average distance to center, see figure 10(a).
2. The physical constraints metrics have a better correlation with the movement network than with the social connection network, see figure 7.
3. Overall, the physical barriers defined by the rivers influence the isolation of the social connection network the most, see figure 7.
4. The scaled modularity for the social connection network does not have a correlation as good as in the validation report with the segregation physical barriers, see figure 10(b)

4.1 Movement Network and Average Distance to Center

Figure 10(a) confirms a positive correlation between the average distance to the center (ADC) and the modularity of the movement network. This means that if the population of the city is more spread out (higher ADC) (Tóth et al. 2021), neighborhood communities are more cohesive, thus there is little social mixing between neighborhoods. This may be due to travelling time between neighborhoods, in which case social mixing is highly influenced by a city's urban layout.

For this part, figure 2(b) will be looked at. In this figure the community deviation of the movement network is shown, where the communities of neighborhoods are defined by the check-in locations of individuals. By observing these communities it is shown that communities are mostly composed by adjacent neighborhoods. This confirms that people are more more likely to visit places close by, as there is more interaction between adjacent neighborhoods.

To conclude, the results show that cities with a spread out population (high ADC, see figure 6), tend to have less social mixing between distant neighborhoods (high movement network modularity). This is explained by the fact that a spread out city has less opportunities for encounters. In other words, a city with a high ADC value has a fragmented movement network, as there exists little opportunity for social interactions.

4.2 More Impact on Movement Network

The results from figure 7 confirm that physical constraints have more impact on the movement of the inhabitants compared to the social network, if we compare it to figure 8. It states that people are more likely to move within their neighborhoods, if the city contains more physical barriers such as rivers, interstates or railroads see figure 5. Peoples' social networks are less impacted by the physical constraints. This difference in results could be due to people getting a lot of connections through high school, university or their work. So a place that could combine multiple communities in a social network. Since not every neighborhood has their own school, university or job opportunities, physical constraints have less influence on the urban travel to these places. But regular activities such as doing the groceries, grabbing a coffee, doing sports which are available within peoples' own community, physical constraints could be seen as a barrier. Those regular activities give spots to go to (check ins).

4.3 Impact of Rivers

Rivers have more impact on the network since they have a larger area and are more difficult to pass by the limited amount of crossovers (bridges, tunnels or ferries). The limited amount of ways to cross the rivers is due to expensive and difficult infrastructure needed. Building a bridge is more expensive and challenging than creating a road underneath a railroad. Kondolf states in his article that rivers have a big impact on social segregation (Kondolf 2017). And as the river becomes wider, the more impact the river has on social connectivity.

The results of figure 7 show that rivers are the biggest barrier for isolation in the social connection network compared to interstates and railroads. This is shown by the fact that the SPB index for rivers has a higher correlation with each of the social isolation metrics than the SPB index for interstates and railways. This means people have less friendship connections in other communities due to the impact of rivers rather than interstates and railroads. This further confirms the fact that rivers have more impact on the network of social connectivity as described in the article of Kondolf.

4.4 Scaled Modularity and Segregation of Physical Barriers

As said in section 2.1, the main idea of the validation approach was to compare our results with the results of a Hungarian report (Tóth et al. 2021). This report found a coefficient of determination (r-squared) of 0.185 between the scaled modularity and the SPB index. In our research, these two metrics had a r-squared value of 0.106 see Appendix (B).

Even though there is a marginal difference between these two values, we conclude that our research was done correctly. The Hungarian report does feature the same type of data input: the connections between users (edges) and self-reported locations of those users (check-ins). However, this data of course stems from Hungarian users and cities. In 2019, research was done to investigate the level of 'mimicry' found in Eastern European regions compared to typical Western urbanisation and the results show that Eastern European regions are significantly different than Western cities with regard to urbanisation (Taubenböck et al. 2019). Also, the Hungarian report features data of 2.8 million users, while we only had access to data of a little over a hundred thousand users. In conclusion the main features that cause the difference in correlation between our research and the research performed by Tóth et al, are that two different types of urban areas were investigated and that the datasets are of different sizes.

4.5 Limitation and Recommendations

During the research the biggest barrier was the limitation of the data. Building networks of 200,000 users spread over the United States of America does not give a good reflection of the reality. Besides that, there are quite some inactive users, which leads to an even lower number of users. Furthermore the users are a specific group of individuals, so the users are not a good representation of the real world.

In the statement of section 4.2 peoples' social networks are less impacted by the physical constraints. For this statement to hold the difference in network structures, the limitation of the data needs to be included. The social network has been based on the individual level and the movement network has been plotted on the neighborhood scale. Since the database has its limitations, the social network (which is more looking at individuals) might not be the best to trust. Zooming out, to the neighborhood level, these problems are less influencing.

Recommendations on further research: due to the time constraints not all the research was included in the report since it was incomplete. Some of the research we recommend on expanding further on: Looking at the network between cities as the idea is to compare the travel time and network of the cities. A clique analysis, can be made to better understand and detect social fragmentation. Another topic that was mentioned but not included is accessibility, different modes of transport within the cities and see if there is a correlation with social fragmentation. Lastly to limit the scope of the project, social groups' characteristics (income level, education level, etc.) were not included, but it is still relevant to the subject for investigation.

5 Conclusions

The aim of this report was to find correlations between physical features and social isolation in cities. To this end social network data from the (discontinued) app Gowalla was analyzed, using a number of network analysis tools. The resulting metrics were then analyzed to metrics describing physical features, which might contribute to the aforementioned social isolation. From this correlation based analysis the following conclusions can be drawn:

- If a city is more spatially fragmented, people are more likely to just visit nearby neighborhoods;
- The physical constraints present in a city have a larger impact on the mobility of people than on their social interactions;
- The presence of rivers in a city is a larger contributor to social isolation than that of interstates or railways;
- The results for Hungarian cities found by Tóth et al. 2021 could not be reproduced using our data applied to 50 US cities, no correlation could be established between the Segregation of Physical Barriers and the scaled modularity of the social network;
- Further research could:
  - Analyse the network of cities throughout the US;
  - Explore the relation between social isolation and accessibility;
  - Deepen our understanding of social isolation by performing a clique analysis in conjunction to the modularity based approach we followed;
  - Include socioeconomic factors in the analysis.

6 Ethical Considerations

Ethical problems that may arise in the use of data, methodology and results of the research should be taken into consideration, the issues need to be analyzed by a framework. For this project the framework used is value sensitive design, supported by consequentialism.

6.1 Ethical Theory and Framework

Value sensitive design is an ethical framework where in the values of human beings are taken into account throughout the design process. The tripartite methodology subdivides the framework into the conceptual, empirical and technical investigation. In the conceptual investigation the values of the stakeholders are highlighted and how competing values should be handled. Next in the empirical investigation it goes a step further, looking into the human context of the of the technology. Finally the technical investigation focuses on how human values may be impacted (Friedman et al. 2002).

Consequentialism ethics looks at purely the consequences of actions. The general view is that actions that make the future of the world a better place are good, because what has passed can no longer be changed. If an unfavourable act leads to a better outcome, than act is still considered good (Sinnott-Armstrong 2003).

6.2 Conceptual Investigation

For this project every stakeholder will be discussed. Looking through the eyes of a consequentialist an initial problem will be listed for every stakeholder. After that, a design requirement will be described to take that initial problem into account.

Machine learning developers

- Involvement: The machine learning developers are responsible for the code of the method. The code can be used to check if a city will have isolated groups. On the other hand, the code could be misused by other developers.
- Initial problem: Duplicating other ones' designs is stealing and is against the norms of a consequentialist.
- Design requirements: to prevent this, the algorithms and models will be put in a black box, so that it will not be open sourced.
- Counter argument: Certain stakeholders may want to have access to see if the code is not biased in discriminatory ways. Having the algorithm stored in a black box is in that case impossible.

Urban designers

- Involvement: Urban designers are responsible for the layout of an urban area. They could use the model to prevent isolation and segregation of groups in a city by looking at the effects of social barriers. This might be helpful by the designing process.
- Initial problem: The urban designers could misuse the information, for example to isolate groups on purpose.
- Design requirements: Since it is the aim of the project to investigate if there is a correlation between segregation and physical barriers, nothing could be done to prevent this problem.

Users of the app Gowalla

- Involvement: For the algorithm the connections and locations of the users of the app Gowalla will be used. Their information is the base of the method. The data is available for everyone on the internet. Every user has been converted to a number to keep the users 'anonymous', the userID.
- Initial problem: Even though the users have an anonymous name, someone could come up with the identity of the user by knowing some specifications which is deanonymization. This could hurt someones privacy, this is against the norms of a consequentialist.

- Design requirements: Since the data is open for everyone, it is not possible to change particular settings of the existing data set. But for the future, it is better to put rules in the app that ensures the protecting of the users privacy and a plan in case a breach happens to minimize the damage done. It is also possible to only give the locations of the users for one day, instead of all their locations.
- Counter argument: For this project the users of Gowalla have not seen any compensation for the usage of their data in this project. Which is true, but the developers of this project are not responsible for the publicity of the data. Thus compensation will not be available for the users. Also since the data is more than ten years ago, it is unlikely that the data will be misused against the users.

Citizens of the place the model will be implemented

- Involvement: Models created from the results gathered from the research done from Gowalla users will be used for the urban design of these citizens.
- Initial problem: What is not taken into account is whether the Gowalla users and the citizens of where these models will be implemented have the same social behaviour. Since a community from the east coast might have a very different behavioural pattern from the community of the west coast. What may happen is that the model becomes to generalized, this might lead to more harm than good.
- Design requirements: Making sure that the machine learning model is not too general when applying it to different locations. Have a general understanding of the the type of people that have used Gowalla.
- Counter argument: It may be that the type of information is not separable from the geographic location that has been sent to Gowalla by the users. Some citizens may not want to mix with other groups and cultures and this may cause friction between them.

Governmental bodies

- Involvement: While focusing on the segregation of groups in cities, the government could use this method to categorize the cities and see if physical barriers have something to do with the segregation. If so, the government could do something about these geographical features.
- Initial problem: The government could also misuse these information, by for example isolating the groups.
- Design requirements: Since it is the aim of the project to investigate if there is a correlation between segregation and physical barriers, nothing could be done to prevent this problem.

Politicians

- Involvement: Politicians have interest in more power by having a bigger support from communities. Knowing the behavioural patterns of people can be used to the benefit of politicians.
- Initial problem: Politicians are increasingly using more unethical ways to get more votes, in the 2016 USA elections one fifth of the tweets about the election where published by bots. Knowing the location of social groups and how they interact with one another may result in creating another way for politicians to unethically gain more following (Berkowitz 2020).
- Design requirements: In the model that will be created, exclusively the connectivity of the communities will be observed and how barriers can effect it. Making sure to exclude the ethnicity and social class from the model will be beneficial.

Big relevant corporations and institutions, such as banks or insurance companies

- Involvement: These big relevant companies might use this project to categorize individuals in certain classes based on the neighborhood city.
- Initial problem: Knowledge about groups could have a large negative financial impact on the people involved. Individuals could be generalized into a class. Mortgage and insurance rates could be higher for undesirable labeled neighborhoods.
- Design requirements: Set up a goal that the method is only relevant for designing cities or to prevent segregation.

6.3 Values Hierarchy

All ethical issues are taken into account to design a Values hierarchy see figure 11. Here the value integrity is chosen since it compliments consequentialism with the following definition: A conceptualized ideal that is working towards a positive path. Orienting your own being towards good faith. Where participation into good faith means to align yourself with reality in a truthful manner (Peterson 2002). Using this definition we're able to encompass all design requirements with the following norms: transparency, privacy and safety. As these three norms are vital to make sure the usage of our design is used positively and in good faith. This is translated into design requirements which have been written in the sections below. These requirements are now tied to the values hierarchy, rather than to a specific stakeholder.

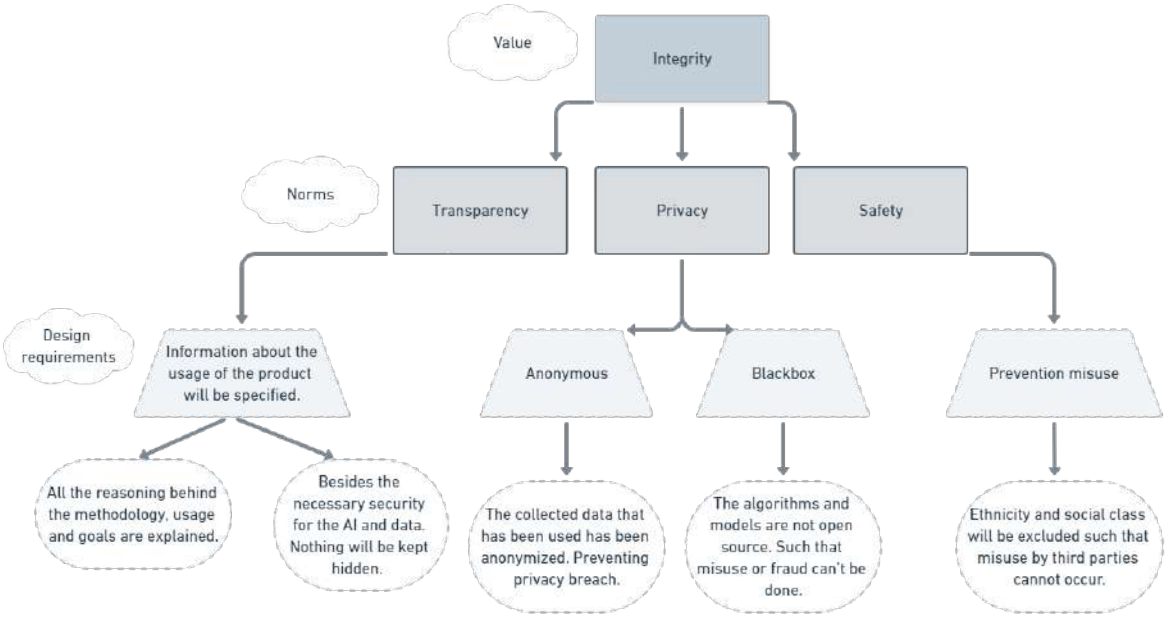

Value
Integrity
Norms
Transparency
Privacy
Safety
Design requirements
Information about the usage of the product will be specified.
Anonymous
Blackbox
Prevention misuse
All the reasoning behind the methodology, usage and goals are explained.
Besides the necessary security for the AI and data. Nothing will be kept hidden.
The collected data that has been used has been anonymized. Preventing privacy breach.
The algorithms and models are not open source. Such that misuse or fraud can't be done.
Ethnicity and social class will be excluded such that misuse by third parties cannot occur.

Figure 11: Values hierarchy tree

6.4 Empirical and Technical Investigation

Since the final product of our design is essentially our report, it is hard to do a proper empirical and technical investigation. Since the product or prototype needs to be tested on the assumptions made in the conceptual stage by involving all stakeholders and eventually test it in a real world setting. This way the iterative design process can be implemented, such that if there exists any flaws they can be fixed. Because of the limitation of our project we'll be looking into how it will impact human values as a replacement of the empirical and technical investigation, this way we will be able to see if our assumptions need some tweaking in case there are still some concerns or problems. The main goal is to justify the usage of the product, without impacting human values negatively. To see if this is properly done we will define the negative impact on human values with Immanuel Kant's formulation of the categorical imperative. There are four formulations that describe the same idea, for this we'll describe the first two formulations that will be used to see the impact on human values. First formulation: 'act only according to that maxim by which you can at the same time will that it should become a universal law without contradiction'. The second formulation: 'Act so that you treat humanity, whether in your own person or in that of another, always as an end, and never as a mere means' (Kant 2008). All the stakeholders will be tested on the categorical imperative, now with the implementation of the design requirements.

Machine learning developers: To uphold the second formulation of the categorical imperative, prevention on misuse and fraud the blackbox has been properly implemented to prevent this. The code is not open sourced and without permission cannot be gained accessed. Without this, other developers can use this with their own benefit in mind without considering our expertise. Thus treating humanity as a mere means rather than an end.

Urban designers/Governmental bodies/Politicians: This report could unfortunately be used to isolate groups by urban designers or public administration to facilitate discrimination. Which would breach the second formulation of the categorical imperative. Since the isolated groups are used as a mere means, without the consideration of the negatively impacted groups. For prevention social group characteristics are not included and the code will be hidden away such that the implementation of misuse will be counter-acted.

Users of the app Gowalla: The data that we use to conduct our research is meant to research isolation to hopefully prevent this from happening. Even though we're using the data of gowalla that is publicly available, which has been anonymized, there is still some risk of privacy breach. But since this has been publicly released this is not something we can solve. Also we're using the data towards an end, the data is used for to hopefully prevent isolation, which is a problem that needs to be addressed. While also implementing the ethics we hope to fullfill the second formulation where we treat humanity as an end and not simply as a mere means.

Citizens of the place the model will be implemented: By making the model not too general, we hope to achieve a product that can be used to prevent isolation. This confirms along with the first formulation of the categorical imperative where we should always use models to prevent isolation. By doing this we hope for a positive contribution towards the citizens of the place where the model will be implemented.

Big relevant corporations and institutions, such as banks or insurance companies: This product can be used to generalize certain people into groups that is undesirable for banks or insurance companies which would negatively impact them. This would breach the first formulation of the categorical imperative. Where we should not always make decisions on simply on their 'group' identity which would be considered discrimination. But by making the data anonymous and setting the goal of segregation prevention, this cannot be misused.

Since the outcome did not breach any of the categorical imperative, we can conclude that our product can be released without any ethical issues. By implementing the value sensitive design we hope to be thorough with our research seeking for every potential bottleneck when it comes to ethics. By stating each part of every stakeholders, we hope to illustrate the potential problems such a product can deliver. Hopefully by giving the solution and testing them with an ethical framework, we have delivered a design while incorporating human and moral values.

7 Reflection

In section 7.1 the requirements will be recalled and their completion will be evaluated. Afterwards in section 7.2, the overall struggles and learning points of the project will be discussed. Finally in section 7.3, the fulfilment of the group made agreements will be discussed.

7.1 Requirement Reflection

The requirements can be found in section A. In this section, each requirement and it's completion will be discussed shortly.

Requirement 1 (must) was completed. More work was done than initially expected, since the borders of each city had to be redefined, but eventually each user was assigned to the city they have visited.

Requirement 2 (must) was completed. This requirement was done using the Clauset-Newman-Moore algorithm, which was a logical choice as this is algorithm is built into NetworkX. Using only the edges that stay within the city (so not the ones between cities) the communities within each city were detected.

Requirement 3 (must) was completed. This requirement only requires one social isolation metric to be found, which was the modularity.

Requirement 4 (must) was completed. This requirement only requires one physical constraints metric to be found, which was the Segregation of Physical Barriers (SPB). There were some difficulties with the shape files. Defining the physical constraints geometry within each FUA geometry using GeoPandas did not succeed. Instead Qgis was used to form these shapefiles.

Requirement 5 (must) was completed. Between all of the found metrics for social isolation and physical constraints the correlations were inspected. The interesting ones made the report, but each (non existing) correlation was investigated.

Requirement 6 (should) was completed. The requirement states the need for only one more community detection method, but the research looked at two more, Louvain and Leiden, as Louvain was not meeting the expectations. The Leiden algorithm was later chosen over both Clauset-Newman-Moore and Louvain.

Requirement 7 (should) was completed. In total, 4 additional social isolation metrics were found; the scaled modularity, betweenness centrality, community fragmentation and the ratio disconnected users.

Requirement 8 (should) was completed. In total, 2 additional physical constraints metrics were found; the average barrier index and the average distance to centre.

Requirement 9 (should) was completed. The patterns in within the social isolation metrics and within the physical constraints metrics were analyzed.

Requirement 10 (could) was not completed. The requirement was started, but was quickly deemed unimportant with regard to the research. The network was built, but no analysis of the network was done.

Requirement 11 (could) was completed. The neighborhood network was built and named the movement network. It shows different numbers than the social network metrics, thus this requirement was successful.

Requirement 12 (won't) was obeyed. As the report shows, no characterization of any of the users was done during the research.

7.2 Project Evaluation

In this section an evaluation will be made on the project. In the evaluation the whole process of the project will be dissected into parts that went well and also what issues were present in the project.

In the early stages of the project everything went quite smoothly; all the deadlines were met and we were even ahead of schedule. Even though this was planned, it was quite surprising that everything went as well as it did. This does not mean we did not encounter any problems, but by using the scrum method, sprints were made on the code and the results were frequently shown to the client. This enabled us to ask for feedback and help whenever we got stuck. This removed the doubt that our findings were negative and prevented us from revisiting our past work for reparations. The early success was also mainly due the diverse skill set the group had, because of the background of multiple disciplines everyone was able to cover each other in different areas. Luckily our client also had past experience with the project in hand, thus most of the struggles that were presented, were fixable with the help of the client, group or TA. In the end, a few of our additional researches seemed interesting, but this unfortunately led to a lot of time and effort wasted. This might've been prevented if we had presented the new research properly elaborated to our client and discussed the feasibility within our time constraint, as it was later revealed we needed more time to properly finish everything we wanted to. We had scheduled a deadline for the additional research and after that was past, we would move on to the final report. This meant we still had plenty of time to finish the report, receive feedback and ask additional questions to the client and TA to properly finish our final product. We would like to thank them for their help and time. The client was very cooperative during our project, guiding us towards the correct path, validating our results and answering every question we had. This made our job very easy as the goals were clearly defined.

7.3 Group Process Evaluation

Before we started with the project, agreements were made in the code of conduct (see Appendix C. And the scrum method was implemented for a successful development throughout the project. For the most part the process throughout the project was excellent, everyone communicated well on time of their absence and tried to adhere the code of conduct. Everyone was quite forgiving when it came to small interventions when coming late, as the reasoning behind it was understandable or justified. The deadlines were mostly met and if someone had a problem, someone else would offer their help for a successful completion of the tasks at hand. There were no real conflicts between the group members as every discussion was a civil discourse, where the goal was to improve the report in mind. This made collaboration between the members easy as everyone took their responsibility seriously and treated each other with respect. This can be seen throughout the whole project as requirements were finished in due time and decisions were made quickly with the same consensus throughout the whole group. Everyone was up to date, since communication was kept throughout the whole project. Having an appointed a chair throughout the meetings everything was structured in a way where efficiency was met. Decisions were made using opinions of every group member, without leaving any opinion out. This way we hope to keep everyone satisfactory and the ability to give their input into the project.

8 Files Location

In this chapter, the agreements and rules on the file management will be described. Files will exclusively be stored on the shared Google Drive folder following the structure as seen below. The most important aspect is that the parts of the project are subdivided well and clear from each other.

- **Main folder** includes folders for each main product.
  - **Code** includes the used and processed data, as well as the code files for all methods used.
    - **Raw data** is where all data can be found that is obtained directly sources, explained in section 9.2.
    - **Processed data** contains the output from the methods implemented.
      - **Graphs**
      - **Tables**
    - **Requirement 1**
    - **Requirement ...** each requirement from the requirements appendix A has it own folder.
    - **Requirement 11**
  - **Design document**
  - **Meetings** where all meeting agenda's and notes are stored.
  - **Planning**
  - **Presentation**

9 Tools Used

Python 3 has been used as the programming language for the code. There are however libraries and other tools that have been used to achieve the desired results. In this chapter information regarding these libraries and tools can be found.

9.1 Libraries & Tools

The following libraries and tools have been used to preform the research needed.

- Pandas: data manipulation and analysis
- GeoPandas: expansion of Pandas that supports geospatial data
- QGIS: software for viewing, editing and analysis of geospatial data
- NetworkX: structure and analysis of networks
- Community package: network structures with Louvain
- Leidenalg package: network structures with Leiden
- Contextility: extended plotting of geographical data

9.2 Dataset

The datasets that have been used are all open sourced and some have been suggested by the client to use. The following datasets below are used:

- **Gowalla users** data collected from users that have checked in their location between February 2009 - October 2010. This data has been sourced from a research done in 2011 about the user movement in location-based social networks. There are two txt files of which "loc-gowalla_edges.txt.gz" contains the friendship network of Gowalla users and "loc-gowalla_totalCheckins.txt.gz" the time and location of the check-ins made (Cho et al. 2011).
  Data location: http://snap.stanford.edu/data/loc-Gowalla.html

- **Functional urban areas by country** data regarding the functional urban areas as defined by the Organisation for Economic Co-operation and Development - European Union (OECD-EU) standard. This standard was created as a global standard so that comparisons can be made for research related work. Each country has two files available. One file contains the general information about the population size, finctional urban area and methodology. The second file is a zip file that contains the country's shape file in different data types. The latest version of these files available is from November 2020 (OECD 2020).
  Data location: https://www.oecd.org/regional/regional-statistics/functional-urban-areas.htm

- **Population distribution** datasets for the spatial distribution of the population in the USA. Zip folder contains shape files of different data types. Data has been collected but the United Stated Census Bureau published on 17 september 2016 (USCB 2016a).
  Data location: https://www2.census.gov/geo/tiger/TIGER2016/TRACT/

- **Geography of rivers** shape file data has been obtained from the National Operational Hydrologic Remote Sensing Center, a governmental body that provides river and flood forecasting (NORA 2016).
  Data location: https://www.nohrsc.noaa.gov/gisdatasets/

- **Geography of interstate network** shape file is obtained from the United Stated Census Bureau (USCB 2016b).
  Data location: https://www2.census.gov/geo/tiger/TIGER2016/PRIMARYROADS/

- **Geography of railroad network** shape file is obtained from the United Stated Census Bureau (USCB 2016c).
  Data location: https://www2.census.gov/geo/tiger/TIGER2016/RAILS/

9.3 Preliminary Analysis of Dataset and Libraries

The tools and libraries that need to be installed separately are QGIS and GeoPandas. Pandas and NetworkX can be imported from Python directly. Documentation on the functionality of these libraries can be found on the internet, linked below. GeoPandas has been tested for installation on Google Colab and QGIS has already been installed. In case no further progress can be made after testing, help can be acquired from the client as he has knowledge with these libraries and tools. Documentation for the libraries and tools can be found below.

- **Python 3** https://docs.python.org/3/
- **Pandas** https://pandas.pydata.org/docs/user_guide/index.html
- **GeoPandas** https://geopandas.org/en/stable/docs.html
- **QGIS** https://www.qgis.org/en/docs/index.html
- **NetworkX** https://networkx.org/documentation/stable/reference/index.html
- **Community package** https://python-louvain.readthedocs.io/en/latest/api.html
- **Leidenalg package** https://leidenalg.readthedocs.io/en/stable/reference.html
- **Contextility** https://contextily.readthedocs.io/en/latest/index.html

References

Ananat, E. O. (2011). The wrong side(s) of the tracks: The causal effects of racial segregation on urban poverty and inequality. *American Economic Journal: Applied Economics*, *3*(2), 34–66. https://doi.org/10.1257/app.3.2.34

Barber, A., Haase, D., & Wolff, M. (2021). Permeability of the city–physical barriers of and in urban green spaces in the city of halle, germany. *Ecological Indicators*, *125*, 107555.

Berkowitz. (2020). The evolving role of artificial intelligence and machine learning in us politics. *Center for Strategic & International Studies*. https://www.csis.org/blogs/technology-policy-blog/evolving-role-artificial-intelligence-and-machine-learning-us-politics

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1082–1090.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Dottle R., R. P., Bliss L. (2021). *What it looks like to reconnect black communities torn apart by highways*. https://www.bloomberg.com/graphics/2021-urban-highways-infrastructure-racism/

E., B., & D., C. (2015). *How railroads, highways and other man-made lines racially divide america's cities*. https://www.washingtonpost.com/news/wonk/wp/2015/07/16/how-railroads-highways-and-other-man-made-lines-racially-divide-americas-cities/

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report*, (2-12).

Kant, I. (2008). *Groundwork for the metaphysics of morals*. Yale University Press.

Kondolf, G. M. (2017). *The social connectivity of urban rivers*. Elsevier.

Mannion, M., & Keepence, B. (1995). Smart requirements. *ACM SIGSOFT Software Engineering Notes*, *20*(2), 42–47.

McCoy, S., Quail, A., & Smyth, E. (2014). The effects of school social mix: Unpacking the differences. *Irish Educational Studies*, *33*(3), 307–330.

NORA. (2016). Gis data sets. https://www.nohrsc.noaa.gov/gisdatasets/

OECD. (2018). *Rethinking urban sprawl*. https://doi.org/https://doi.org/https://doi.org/10.1787/9789264189881-en

OECD. (2020). *Functional urban areas by country*. https://www.oecd.org/regional/regional-statistics/functional-urban-areas.htm

Peterson, J. B. (2002). *Maps of meaning: The architecture of belief*. Routledge.

Rose, M., & Mohl, R. (2012). *Interstate: Highway politics and policy since 1939*. University of Tennessee Press. https://books.google.nl/books?id=akgToyGbE-YC

Sinnott-Armstrong, W. (2003). Consequentialism.

Taubenböck, H., Gerten, C., Rusche, K., Siedentop, S., & Wurm, M. (2019). Patterns of eastern european urbanisation in the mirror of western trends–convergent, unique or hybrid? *Environment and Planning B: Urban Analytics and City Science*, *46*(7), 1206–1225.

Tóth, G., Wachs, J., Di Clemente, R., Jakobi, Á., Ságvári, B., Kertész, J., & Lengyel, B. (2021). Inequality is rising where social network segregation interacts with urban topology. *Nature communications*, *12*(1), 1–9.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

USCB. (2016a). Spatial distribution. https://www2.census.gov/geo/tiger/TIGER2016/TRACT/

USCB. (2016b). United states census bureau. https://www2.census.gov/geo/tiger/TIGER2016/PRIMARYROADS/

USCB. (2016c). United states census bureau. https://www2.census.gov/geo/tiger/TIGER2016/RAILS/

Wachs, J., Yasseri, T., Lengyel, B., & Kertész, J. (2019). Social capital predicts corruption risk in towns. *Royal Society open science*, *6*(4), 182103.

Appendices

A Requirements

In order to answer our research question and define the scope of this project a MoSCoW analysis was performed. Together with our client we defined the following Musts, Shoulds, Coulds and Won't, which we will elaborate upon in this section. Each requirement is numbered for easy reference

1. **Must** assign users of the Gowalla app to cities based on their coordinate data in the data set provided by the PhD student:

   The dataset we will be using is broken up into two parts, primarily it contains the connections between people in the dataset and a number of locations people visited. This first **Must** means we need to link individuals in the dataset to cities, not taking into account possible ambiguities which could arise when a person visited multiple cities. These ambiguities can be resolved by simply linking this individual to each city they visited.

2. **Must** find a method to detect communities in all cities provided by the PhD student:

   A first step in determining isolation is finding communities, which are defined as: a group of individuals in which these individuals tend to interact more with each other than with others outside the community.

3. **Must** find a metric for the social isolation of groups, and determine its value for all cities provided by the PhD student:

   Once we extracted the users and linked them to cities we determine the isolation of the previously found communities within these cities.

4. **Must** find a metric for physical constraints for all cities provided by the PhD student:

   Using the mapping data, a physical constraint metric will be found. An example of such a metric is described by Tóth et al. 2021: Segregation by physical barriers (SPB), which uses the method used in defining Rail Road Division Index (**ananat2011wrong**), but also applied to other barriers such as rivers, major roads, etc. and combines it into a single metric.

5. **Must** compare the physical constraints metric(s) with the found isolation metric(s) in cities, and validate results:

   After finding at least one metric for both the social isolation and the physical constraints, we can inspect whether the metrics correlate or not. Finding more metrics will lead to more possible correlations (n isolation metrics and m physical metrics lead to n x m possible correlations). Produce visualizations and validate whether the results are reasonable.

6. **Should** use at least one additional method to detect communities in all cities provided by the PhD student:

   To properly define communities, an additional metric needs to be used to compare and fixate communities in different levels. Additionally, it is to find the best metric to use for the project.

7. **Should** find at least one additional metric for the isolation in all cities provided by the PhD student:

   To properly define isolation, an additional metric needs to be used to compare and fixate communities in different levels. Additionally, it is to find the best metric to use for the project.

8. **Should** find at least one additional metric for the physical constraints in all cities provided by the PhD student:

   To properly define physical constraints, an additional metric needs to be used to compare and fixate communities in different levels. Additionally, it is to find the best metric to use for the project.

9. **Should** analyse the relations between the metrics found, identify patterns and draw conclusions:

   This should captures the analysis of the previously found relations between the metrics for isolation and physical constraints. We should look at all meaningful combinations of metrics, so also comparing different physical constraint metrics to each other (possibly within clusters of cities with similar levels of isolation). Additionally this requirement contains the drawing of conclusions in the form of pattern identification.

10. **Could** extract the city network based on friendship connections between cities. Analyze those results to find interesting patterns:

    This could was an added requirement, to replace the original requirement, which was: Could find a metric for accessibility using different modes of transport within each of the cities and compare this to the isolation metric in those cities. The idea behind the requirement is that effects of physical barriers, and thus social isolation, might be amplified when looking at such a large scale such as the country scale. Interesting findings might be found when analyzing isolated cities.

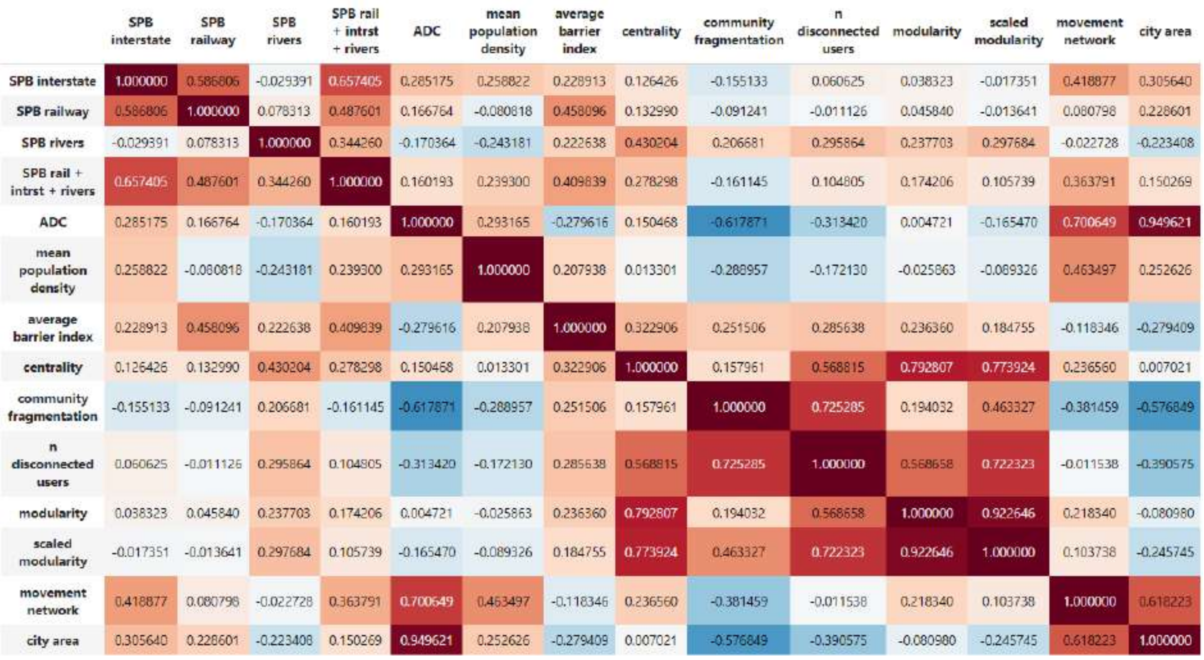
11. **Could** build a new network per city where the neighborhoods (defined by physical boundaries) form the nodes, and the edges between these nodes are weighted according to the visiting pattern of the individuals in the dataset. Next we visualize the results and analyze:

    First the cities will be divided based on the physically defined neighborhood boundaries. These will form the nodes of a new network with the edges weighted according to the visiting pattern of individuals: if a person visits two neighborhoods the weight of the edge between said neighborhood nodes will be incremented with one. This new weighted network will then be visualized and analyzed, possibly providing new insights in the isolation level of certain neighborhoods.

12. **Won't** infer the social groups' characteristics (income level, education level...). The students will focus on the network's topology (identifying the number of groups, and their isolation). They will not characterize the group:

    This mainly serves to limit the scope of the project, looking into the social groups' characteristics is obviously relevant to the subject matter but will be part of further research by the client.

B Correlation Matrix



	SPB interstate	SPB railway	SPB rivers	SPB rail + intrst + rivers	ADC	mean population density	average barrier index	centrality	community fragmentation	n disconnected users	modularity	scaled modularity	movement network	city area
SPB interstate	1.000000	0.586806	-0.029391	0.657405	0.285175	0.258822	0.228913	0.126426	-0.155133	0.060625	0.038323	-0.017351	0.418877	0.305640
SPB railway	0.586806	1.000000	0.078313	0.487601	0.166764	-0.080818	0.458096	0.132990	-0.091241	-0.011126	0.045840	-0.013641	0.080798	0.228601
SPB rivers	-0.029391	0.078313	1.000000	0.344260	-0.170364	-0.243181	0.222638	0.430204	0.206681	0.295864	0.237703	0.297684	-0.022728	-0.223408
SPB rail + intrst + rivers	0.657405	0.487601	0.344260	1.000000	0.160193	0.239300	0.409839	0.278298	-0.161145	0.104805	0.174206	0.105739	0.363791	0.150269
ADC	0.285175	0.166764	-0.170364	0.160193	1.000000	0.293165	-0.279616	0.150468	-0.617871	-0.313420	0.004721	-0.165470	0.700649	0.949621
mean population density	0.258822	-0.080818	-0.243181	0.239300	0.293165	1.000000	0.207938	0.013301	-0.288957	-0.172130	-0.025863	-0.089326	0.463497	0.252626
average barrier index	0.228913	0.458096	0.222638	0.409839	-0.279616	0.207938	1.000000	0.322906	0.251506	0.285638	0.236360	0.184755	-0.118346	-0.279409
centrality	0.126426	0.132990	0.430204	0.278298	0.150468	0.013301	0.322906	1.000000	0.157961	0.568815	0.792807	0.773924	0.236560	0.007021
community fragmentation	-0.155133	-0.091241	0.206681	-0.161145	-0.617871	-0.288957	0.251506	0.157961	1.000000	0.725285	0.194032	0.463327	-0.381459	-0.576849
n disconnected users	0.060625	-0.011126	0.295864	0.104805	-0.313420	-0.172130	0.285638	0.568815	0.725285	1.000000	0.568658	0.722323	-0.011538	-0.390575
modularity	0.038323	0.045840	0.237703	0.174206	0.004721	-0.025863	0.236360	0.792807	0.194032	0.568658	1.000000	0.922646	0.218340	-0.080980
scaled modularity	-0.017351	-0.013641	0.297684	0.105739	-0.165470	-0.089326	0.184755	0.773924	0.463327	0.722323	0.922646	1.000000	0.103738	-0.245745
movement network	0.418877	0.080798	-0.022728	0.363791	0.700649	0.463497	-0.118346	0.236560	-0.381459	-0.011538	0.218340	0.103738	1.000000	0.618223
city area	0.305640	0.228601	-0.223408	0.150269	0.949621	0.252626	-0.279409	0.007021	-0.576849	-0.390575	-0.080980	-0.245745	0.618223	1.000000

Figure 12: Correlation matrix between all metrics

C Code of Conduct

Communication and respect

- Be considerate to your colleagues.
- Decisions that influence the group need a majority vote (3+ people).
- Always give others a chance to voice their opinion in meetings.
- Do not interrupt when your colleagues are speaking in a meeting.
- In case of a conflict, first attempt to solve it with the person in question. If the problem is still not solved, discuss it with the group at a Monday morning meeting. If it still is not fixed, the group will contact the TA.
- In case you cannot meet a deadline, let the group know 2 working days beforehand.
- Work communication (including communication with the TA) takes place through Mattermost and small things can be communicated through Whatsapp.
- Starting from week 2.5, communicate your reachability through your Mattermost status.

Meetings

- Every Wednesday 9-10, we hold an update meeting with the client.
- Every Monday 14-15, we hold a Q&A meeting with just the client.
- We have a Monday morning meeting where we discuss and update on weekly tasks and discuss whether tasks need to be done on campus or not.
- We hold a small daily meeting every morning to discuss daily activities and goals.
- At the end of every Monday, we discuss and prepare the Wednesday meeting.
- If a member cannot attend a meeting, they need to inform the group at least a day before, but everyone is expected to attend all meetings otherwise.
- For every 5 minutes that a group member is late to a meeting, they will receive a turfje.

Work and deadlines

- We use Scrum to organize the software development process.
- Tasks must be well defined and achievable.
- If a deadline is not met, the reasoning is discussed within the group and solutions are discussed. Without a valid reason, that person will receive 5 turfjes. When this occurs regularly, this is discussed with the TA.
- When a task is completed sooner than expected, the group member(s) has/have to discuss this with the group.
- Use comments in your code for easy compiling.
- Do not hard code.
- Provide your code to the others before Friday, so compiling can be done on Friday and fixing can be done during the weekend, so a working code file is achieved each Monday.

Efforts and goals

- We want to achieve an 8 as a result and are expecting each group member to work proportionally.
- Each member is expected to work an average of 36 hrs a week.
- Efforts are based upon hours working on the project, not on the number of deadlines met.

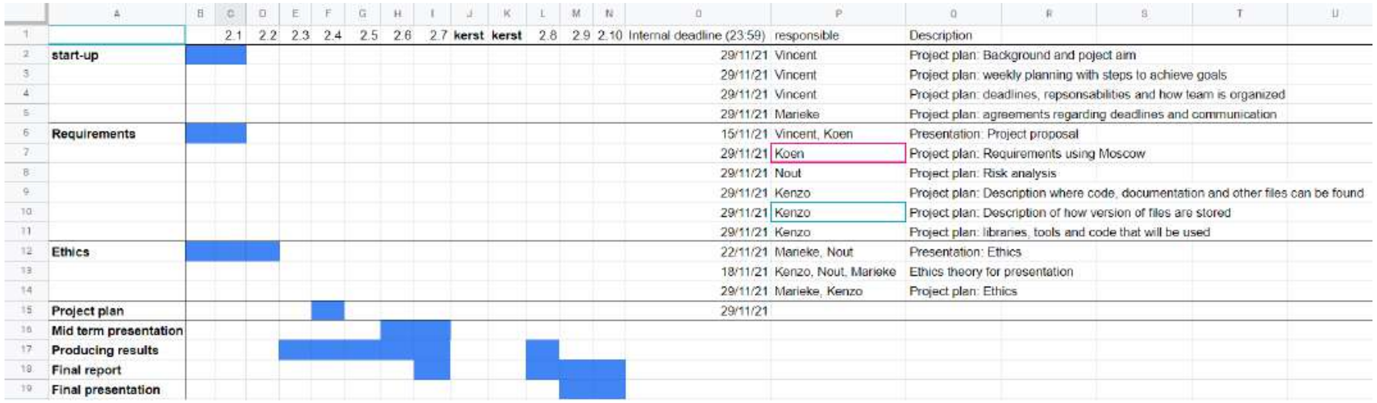
Responsibilities

- Koen will be the group leader, which makes him the chair of each meeting and makes him responsible for the communication with the TA and client.
- Marieke is responsible for compiling the code.
- A different person is assigned every week to take notes during the meetings.
- Two duo's are assigned as presenters, Nout & Marieke and Koen & Vincent.
- If a script does not work how it is supposed to in the compiled code file, Marieke will contact the group member and they will try to find the problem. If the problem is within the code, the group member should fix it themselves.

Corona

- Before each in-person meeting, each group member does a self-test.
- In case of an infection or an infected housemate, that group member attends the meeting online.

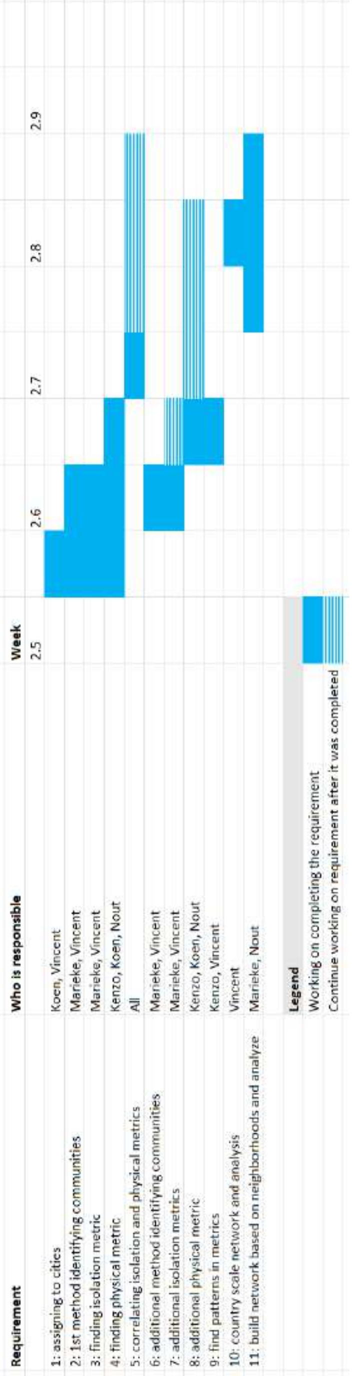
D GANNT Chart Phase 1



		2.1	2.2	2.3	2.4	2.5	2.6	2.7	**kerst**	**kerst**	2.8	2.9	2.10	Internal deadline (23:59)	responsible	Description
start-up	🟦	🟦												29/11/21	Vincent	Project plan: Background and poject aim
														29/11/21	Vincent	Project plan: weekly planning with steps to achieve goals
														29/11/21	Vincent	Project plan: deadlines, repsonsabilities and how team is organized
														29/11/21	Marieke	Project plan: agreements regarding deadlines and communication
Requirements	🟦	🟦												15/11/21	Vincent, Koen	Presentation: Project proposal
														29/11/21	Koen	Project plan: Requirements using Moscow
														29/11/21	Nout	Project plan: Risk analysis
														29/11/21	Kenzo	Project plan: Description where code, documentation and other files can be found
														29/11/21	Kenzo	Project plan: Description of how version of files are stored
														29/11/21	Kenzo	Project plan: libraries, tools and code that will be used
Ethics	🟦	🟦	🟦											22/11/21	Marieke, Nout	Presentation: Ethics
														18/11/21	Kenzo, Nout, Marieke	Ethics theory for presentation
														29/11/21	Marieke, Kenzo	Project plan: Ethics
Project plan					🟦									29/11/21		
Mid term presentation							🟦	🟦								
Producing results				🟦	🟦	🟦	🟦	🟦			🟦					
Final report								🟦			🟦	🟦	🟦			
Final presentation												🟦	🟦			

Figure 13: GANTT chart phase 1

E GANNT Chart Phase 2



Requirement	Who is responsible
1: assigning to cities	Koen, Vincent
2: 1st method identifying communities	Marieke, Vincent
3: finding isolation metric	Marieke, Vincent
4: finding physical metric	Kenzo, Koen, Nout
5: correlating isolation and physical metrics	All
6: additional method identifying communities	Marieke, Vincent
7: additional isolation metrics	Marieke, Vincent
8: additional physical metric	Kenzo, Koen, Nout
9: find patterns in metrics	Kenzo, Vincent
10: country scale network and analysis	Vincent
11: build network based on neighborhoods and analyze	Marieke, Nout

Legend

Working on completing the requirement

Continue working on requirement after it was completed

Figure 14: GANTT chart phase 2