Community detection in spatial networks: Examining human mobility

[redacted]

1 Introduction

Geographic online social networks such as Foursquare, where users voluntarily signal their current location, enable detailed studies on human movement. In this paper we analyse a data set of transitions of Foursquare users between locations in three different US cities: Washington, New York and San Francisco. To analyse how those transitions are structured, we examine the aggregated network with the help of different community detection methods. They differ in the null-model they are based on: the configuration model, the gravity model and the degree-constrained gravity model. They therefore reveal communities that are influenced by different spatial and non-spatial factors. We use these three methods to examine how mobility changes from day to night time in the three cities.

In Section 2 we first give some background on current paradigms in the analysis of human mobility. In Section 3, we then introduce our methodology by giving an overview about spatial networks and explain community detection with the three different null models. We also introduce the Louvain algorithm, which we use in an adapted form for the computations of the communities in the actual data set. Subsequently, we describe the Foursquare data set in Section 4 and give details about the aggregation of the spatial networks and how we divide into night and daytime transitions. We then do the analysis on those networks and describe the results in Section 5. In this section we both compare methodology and then try to make qualitative conclusions about the three cities, comparing day and nighttime transitions. We then end with a conclusion about our result and further research possibilities in Section 6.

2 Human mobility in cities

The mobility of people and their movement in space has been fascinating for researchers from different disciplines. First works on this topic date back to the works of Ravenstein in 1885 who investigated the laws of migration.

A common assumption that one can take with regard to human mobility is that it is hindered by geographical distance. How exactly this relationship is to be mathematically formulated is an open question that is highly relevant to applications such as urban planning and social studies such as the investigation of the influence of wealth on mobility [14, 13].

In the literature there can generally be found two different approaches to modelling the dependence of mobility and distance:

The first group of models, referred to as *gravity models*, are inspired by Newton's law of gravity and argue that mobility is directly limited by the cost of physical distance [3]. This has the outcome that flow of people decreases with increasing distance between places.

The second approach distances itself from the assumption that physical distance is an intervening factor by itself but states instead that it is a surrogate for the effect of intervening opportunities [2]. This theory assumes that the distance covered by humans is determined by the number of opportunities (for example places of interest) within that distance, and not by the distance itself. People attempt to satisfy certain needs by their journeys and therefore their mobility is determined by the number and ranking of opportunities closer than their destination. In contrast to the gravity models in which displacements are only limited by distance, displacements are here driven by the spatial distribution of places of interest and thus by the response to opportunities [12].

Generally, gravity models are easier to compute and therefore more frequently used. Nevertheless, even if the nature of intervening opportunities is hard to capture, many studies have shown that the second camp of models relying on the theory of intervening opportunities has high explanatory power [4]. Noulas et al. have shown in [12], that variations in human movement in different cities are predominantly due to heterogeneous distribution of places across different urban environments which are a proxy for different densities of intervening opportunities.

3 Community detection in spatially embedded networks

As network models represent connections between nodes and not their spatial relation this type of model represents just the topology of relations. Nevertheless, most complex networks are spatially embedded and their connections cannot simply be explained by their topology. Examples of this include transport networks, mobility networks or even social networks. It is important to consider the importance of space for those spatial networks, as space has a high influence on analysing both structure and processes on spatial networks. In this work we will mainly focus on the aspect of community detection. We therefore first want to introduce the concept and some properties of a spatial network following [9] that we will use later in our methodology for community detection in a spatially embedded

network derived from Foursquare user transition data in three US cities.

3.1 Spatial networks

A spatial network is a network for which the nodes are located in a space equipped with a metric. Usually we consider Euclidean distance and a two-dimensional space, but this is not the only possibility. The introduced characteristic just mimics that the probability of observing an edge between vertices depends on the distance between them. This is true because there is a cost associated with the length of an edge. This can also be true for social networks that display friendship relations. The probability of those relationships also decreases with growing spatial distance, without that being exactly encoded in the network. Nevertheless, we will focus on networks in this work that have an exact location in two dimensional space associated with each node, as our data set consists of coordinates of Foursquare venues.

Due to the embedding in a metric space the network is equipped with a distance function $d(u,v)$ which measures the distance between nodes u and v. In our case this is the geographic distance between two locations, but other options like transportation time instead of physical distance are possible.

3.2 Community detection

Community detection is an important part of complex network analysis. It aims to divide the network into sets of nodes that have more connectivity among themselves than with the rest of the nodes, the so called modules. It is not always clear what the best partition into communities is. This is due to the fact that in contrast to other graph partitioning methods, community detection aims to uncover the mesoscale organisation of a network in an automated way and therefore the total number and size of communities is not a priori known [11]. There are various methods that try to achieve this goal and most of them aim to maximise a mathematical definition of the quality of a partition, the so called modularity. It measures if links are more likely to be present within a community than would be expected:

$$Q = (\text{fraction of links within a community}) - (\text{expected fraction of such links}) \quad (1)$$

For the mathematical definition of modularity it is important to define what the null hypothesis for the model is, meaning how many links we would expect in a community. This is represented by the matrix P_{ij} of a so called *null-model*. Its entries represent the expected

weights of a connection between i and j over an ensemble of random matrices with certain constraints [11]. The constraints are chosen depending on which structural properties of the original network we want to conserve in the null-model to be able to assess the relevance of certain partitions. Two basic considerations hold true in general for a null-model for a weighted network, represented by its adjacency matrix W, whereas W_{ij} is the weight of the link between i and j:

1. If W is symmetric, we choose P to be symmetric.
2. The total edge-weight w of the network is conserved: $\sum_{ij} W_{ij} = \sum_{ij} P_{ij} = 2w$.

Mathematically we can define the generic modularity according to a specific null-model represented by P^{NM} as introduced in [15]:

Definition 3.1 (Generic Modularity, from [15])**.** For a weighted static network with adjacency matrix W, modularity is

$$Q = \frac{1}{2w} \sum_{ij} (W_{ij} - P_{ij}^{NM}) \delta(c_i, c_j) \tag{2}$$

where $w = \frac{1}{2} \sum_{ij} W_{ij}$ is the total edge-weight of the network, c_i denotes the community that contains node i, the Kronecker delta δ is 1 if $c_i = c_j$ and 0 if $c_i \neq c_j$, and P_{ij} is the ij-th element of the null-model matrix.

$Q = 0$ is achieved if the number of within-community edges is no better than random, $Q = 1$ if the network has strong community structure [5].

In the following we want to introduce three null-models with different constraints that we will use in this work to detect communities in our spatial data set to examine how different spatial constraints change the nature of modules that are being detected. We follow the work of [11, 16].

3.2.1 Configuration model

The most common null-model that that is used for the definition of modularity is the configuration model which was introduced by Newman and Girvan (NG) in [6]. The configuration model proposes to locate links at random in the network while keeping the degrees of the nodes. In the case of an undirected, weighted network the expected weight of a node is

$$P_{ij}^{NG} = \frac{k_i k_j}{2w} \tag{3}$$

with k_i the strength of node i, which is defined as $k_i = \sum_j W_{ij}$ and $2w = \sum_{ij} W_{ij}$ is the sum of all edge weights in the network as before. In this case we have the Newman-Girven modularity given as

$$Q^{NG} = \frac{1}{2w} \sum_{ij} (W_{ij} - \frac{k_i k_j}{2w}) \delta(c_i, c_j) \tag{4}$$

The NG-modularity only constraints the node strength and only takes structural information provided by the adjacency matrix into account. This is based on the assumption that the given network is well-mixed and only connectivity of nodes matters for the probability that two random nodes are connected. It does not take spatial effects into account. This is a useful choice if we do not have additional information about the network available. It is interesting that such null-models in community detection methods yield spatially connected regions, and some regions coincide with administrative units rather well. For example , Cazabet, Borgnat, and Jensen reported in [16] that the communities obtained from bicycle sharing data in Lyon match rather well the administrative borders of the city.

3.2.2 Gravity-based null-model

In spatial networks, higher distance between nodes strongly decreases the probability of them being connected. The configuration model does not take any further information like spatial distribution of nodes into account and therefore overestimates the probability of a connection between two very distant nodes. To include the available spatial information, we use the null-model introduced in [11] that is inspired by gravity models that are frequently used in the transportation domain to repartition trips between different cities and areas [16]. A general version defines the null-model matrix as

$$P_{ij}^{Gra} = n_i n_j f(d_{ij}) \tag{5}$$

whereas n_i is the intrinsic strength of node i, d_{ij} the distance between node i and j and $f(d)$ any deterrence function. The intrinsic strength measures the importance of node i an can vary depending on the application and available information (e.g. population, number of jobs, degree of node etc.). In settings when this is unknown the degree of the node is used as a proxy for the intrinsic strength of a node [16]. In the traditional form of the gravity model, the deterrence function is a priori defined as

$$f(d_{ij}) = d_{ij}^{-\gamma} \tag{6}$$

whereas γ is an optional parameter, usually tuned by regression analysis [17]. However, as shown in [11], the deterrence function does not have to be defined beforehand but can directly be measured from the data:

$$f(d) = \frac{\sum_{i,j|d_{ij}=d} W_{ij}}{\sum_{i,j|d_{ij}=d} n_i n_j} \tag{7}$$

This is the weighted average of the probability $\frac{W_{ij}}{n_i n_j}$ for a link to exist at distance d. Using this deterrence function, the null-model takes into account the constraint that the total weights between nodes at a certain distance is preserved:

$$\sum_{i,j|d_{ij}=d} P_{ij}^{Gra} = \sum_{i,j|d_{ij}=d} W_{ij}. \tag{8}$$

The conservation of the total weight of the network is therefore also given. In the following we will denote the modularity using the gravity null-model as Q_{ij}^{Gra} and use the deterrence function learned from the data itself as defined in Eq. (7). This modularity, additionally to topological structure of the graph, takes physical location of nodes into account and therefore favours communities made of nodes that are more connected than expected for their distance. Therefore, it is expected that it reveals communities that are formed by nonspatial factors [11].

3.2.3 Connection between configuration and gravity-based null-models

If we use the degree of the nodes as a proxy for their intrinsic strength, i.e. $n_i = k_i$, and we assume that distance does not play a role i.e. the system is well mixed ($f(d)$ is independent of d), using $k_i = \sum_j W_{ij}$ and $\sum_{i,j} W_{ij} = 2w$ we derive the NG null-model:

$$P_{ij}^{Gra} = k_i k_j f(d_{ij}) = k_i k_j \frac{\sum_{i,j} W_{ij}}{\sum_{i,j} k_i k_j} = k_i k_j \underbrace{\frac{\sum_{i,j} W_{ij}}{\sum_{i,j} (\sum_j W_{ij})(\sum_i W_{ij})}}_{\frac{1}{\sum_{i,j} W_{ij}}} = \frac{k_i k_j}{2w} = P_{ij}^{NG} \tag{9}$$

Let us finally emphasise, that, as suggested in [11], there is the possibility to introduce a mixing parameter ξ and interpolate between those two null-models to balance the importance of spatial and topological effects according to the relevant application. The interpolated null-model then reads

$$P_{ij}(\xi) = \xi P_{ij}^{NG} + (1 - \xi) P_{ij}^{Spa} \tag{10}$$

Taking a weighted model into account for tuning the influence of distance is a possibility for future investigation that we will not take into account in this work.

3.2.4 Degree-constrained gravity-based null-model

One draw-back of the gravity-based null-model is that there is no simple relation between the intrinsic strength of a node n_i and its actual strength $k_i = \sum_j W_{ij}$. Therefore, if we do not have information about the intrinsic strength of the nodes given by data about population sizes or similar quantities, we have to take the strength of a node as a proxy for its intrinsic strength. Then, the gravity null-model will not preserve degrees of nodes. Because the observed strength of a node in a network generated according to the gravity null-model depends both on its intrinsic strength and on its distance to other nodes, this model usually systematically underestimates the intrinsic strength of nodes with few nodes around and overestimate the strength of those located in the centre. This highly depends on the data set we are working with and we will later see when looking at the Foursquare data sets that the assumption about nodes in the periphery having generally lower degree is not always true and highly depends on the functionality of the different localities.

We can measure the connection between degree bias and spatial distribution as suggested in [16] by correlating the spatial eccentricity and the degree bias of the data set. In our context, those two measures are defined as follows:

Definition 3.2 (Spatial eccentricity and degree bias)**.** Given a network G with N nodes $i = 1, \cdots, N$ and distances d_{ij} between node i and j, the spatial eccentricity of node i is defined as the average distance to all other nodes

$$ecc(i) = \frac{1}{N} \sum_{j=1}^{N} d_{ij} \tag{11}$$

The degree bias $db(i)$ for in and out degrees of node i respectively is defined as :

$$db(i) = \frac{deg^{GM}(i)}{deg^{D}(i)} \tag{12}$$

with deg^{GM} the degree according to the gravity model and deg^{D} the degree observed in original data.

To eliminate this bias, a degree constrained gravity-based model was proposed in [16]. It is derived from the doubly constrained gravity model [20] and can be applied to both undirected or directed networks. We will derive it here following [16].

The idea of the model is to find values for the *Incoming estimated intrinsic strength* n^{Oeis} and the *Outgoing estimated intrinsic strength* n^{Ieis} that would best explain the observed degrees. For undirected networks we have $n^{Oeis} = n^{Ieis}$.

This leads to an iterative method that estimates new values for those strengths while satisfying the observed in-degrees deg^{in} and out-degrees deg^{out} in each step. The estimated intrinsic strengths are initialised with the in- and out-degree and then computed as

$$n^{Ieis} = \frac{deg^{out}(i)}{\sum_i n^{Oeis} f(d_{ij})} \qquad n^{Oeis} = \frac{deg^{in}(i)}{\sum_i n^{Ieis} f(d_{ij})} \tag{13}$$

After computing those values, we can use them to calculate the matrix of the corresponding degree-constrained gravity null-model that is defined as

$$P_{ij}^{DCGra} = n^{Oeis} n^{Ieis} f(d_{ij}). \tag{14}$$

The modularity that is defined by this null-model will be referred to as Q^{DGGra}. We use the same deterrence function as the gravity model as defined in Eq. (7). Nevertheless, it has to be noted that the deterrence function also depends on the intrinsic strength of the nodes. Therefore an approximation from the data using the degrees of the nodes as a proxy leads to a biased approximation. We therefore recompute the function in each iteration to correct this bias with the newly estimated intrinsic strengths.

This is iteratively done until the degrees in the Degree Constrained gravity model are close to the target network. This procedure is known to converge [20], but as suggested in [16], we use a fixed number of 5 iterations.

3.3 Community detection algorithm

One of the most widely used algorithms for community detection is the Louvain algorithm [7]. This algorithm optimizes the defined modularity in a greedy fashion. It is therefore adaptable to all of the notions of modularity that we want to consider. The algorithm is divided into two phases that are repeated iteratively [7].

The **first phase** consists of the following steps:

1. Start with a weighted network of N nodes.
2. Assign a different community to each node.
3. For each node i, compute the gain of modularity if it is put in the same community with each of its neighbours j. Put i in the community that gives maximal gain. If there is no positive gain, i stays in its community.
4. Repeat this process sequentially for all nodes until no further gain possible.

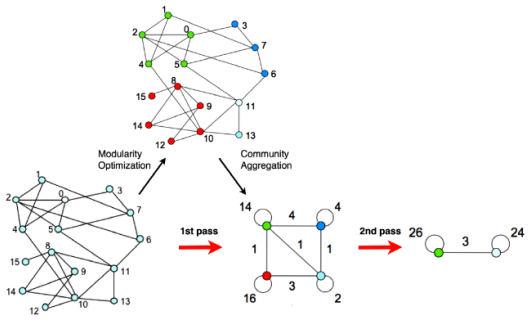


Figure 1: Visualisation of two passes of the Louvain algorithm. In each step, nodes are first being assigned to modules to optimise the modularity and then a new network is created whose nodes are the communities from before. From [7]

In the **second phase** a new network is build whose nodes are the communities found in the first phase:

1. Weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities.
2. Links between nodes of the same community lead to self-loops for this community in the new network.

These two phase, collectively called a *pass*, can be reapplied until the maximum of modularity is attained. This builds a hierarchy of communities. The algorithm is visualized in Fig. 1. In our work, we use the python-based implementation of [16], that adapts the Louvain algorithm to take in different definitions of null-models and adapt this implementation for our purposes

4 Description of data set

In the following section we want to apply the three community detection methods introduced before to a data set of Foursquare user transitions between venues in three US cities. The data set we are working with consists of 4-year-long data from Foursquare describing movements between places in New York, San Francisco and Washington. For each Foursquare venue in a city, the data set contains the unique venue ID, the geometric location (latitude and longitude), the general Foursquare category (e.g. nightlife spot), the specific Foursquare category (e.g. bar), the total number of check-ins and the total number of unique visitors. Additionally to the general venue information the data set contains all

transitions of customers in the four-year time period. A transition is defined to be a pair of check-ins by a single user to two different venues less than 3h apart in time [19]. Nevertheless, the transition data does not contain information about the identity of the user. For each transition, we have start time and end time and source and destination venue ID.

As proposed in [19], we want to exclude all the venues that are added mistakenly or maliciously. We set our cutoff value a bit higher then in this paper, due to computational limitations of our implemented community detection algorithms and available hardware and therefore excluded all locations that have less than 3000 total check-ins for Washington and San Francisco and less than 10000 for New York city, as the transition data set is a lot richer. We choose those cut off values differently for the different cities to achieve a number of locations that is comparable. We visualise the venues including the 5 places with the most check-ins in the three cities in Fig. 2.

4.1 Aggregation to spatial networks

We aggregated the transitions for the relevant locations to an undirected, weighted network for each city in the following way:

The network consists of nodes that represent the Foursquare venues. Node i and j are connected by a weighted edge w_{ij} if there were in total w_{ij} observed transitions between i and j in the four year period. To transform this into a spatial network we have to take to following processing steps: We have data available about latitude and longitude of the different locations. To make use of this spatial information for the computation of distances between locations in metres (e.g. nodes in the spatial network) we have to project this into a local coordinate reference system. Because of the cities location in Northern America we project into UTM zone 18N.

4.1.1 Division in day and nighttime transitions

Because we are interested in the difference of formed communities at day and night, we divide each transition network into two networks: The set of daytime transitions, that start between 7 am and 7 pm, and the set of nighttime transitions, which consist of all the others. It must be noted here that due to the definition of a transition to be a check-in at two different places within a time frame of 3 hours, a transition that ends shortly before 10 pm is still counted as a daytime transition here if it started before 7 pm. This includes e.g. main

Washington	Total number of locations: 730
Check-ins	Location
299410	'Ronald Reagan National Airport'
178748	'Union Station'
70935	'Nationals Park'
60116	'Verizon Center'
57673	'Gallery Pl-Chinatown Metro Station'

New York	Total number of locations: 634
Check-ins	Location
543442	'New York Penn Station'
465646	'John F. Kennedy International Airport'
410435	'LaGuardia Airport LGA'
327828	'Grand Central Terminal'
179738	'Times Square'

San Francisco	Total number of locations: 936
Check-ins	Location
107392	'ATT Park'
101098	'Montgomery St. BART Station'
91834	'San Francisco-Oakland Bay Bridge'
66008	'Golden Gate Bridge'
58461	'Westfield San Francisco Centre'

Figure 2: City statistics: 5 most checked in places and total number of locations in data set after a cutoff value of 3000 total check-ins (San Francisco and Washington) and 10000 check-ins (New York).

rush hour times for work in all three cities into the daytime data set

In the next two sections we first want to apply the three methods introduced before to this data set.

5 Evaluation of the three methods on Foursquare data set

In this section, we want to investigate the performance of the three different community detection methods on the transition networks derived from the data sets we introduced in

the last section. In our analysis, we want to focus on three aspects

1. Examine the relationship between degree bias and spatial eccentricity for the three cities.
2. Examine how the deterrence function that is learned from the data depends on binning distance and on outliers in the different data sets.
3. Compare how communities change from day to night time in the different cities.

5.0.1 Degree bias and spatial eccentricity

Even if we have information about the intrinsic strength of locations based on e.g. the total number of check-ins ore unique users in this data set, we will assume that our model is uninformed for this work and use the degree of the nodes as a proxy for their intrinsic strength for Q^{Gra}. By doing so, we want to evaluate the different performance of this null-model compared to the degree-constrained gravity model. As explained in Section 3, in most data sets the intrinsic strength of nodes in the periphery is systematically underestimated when using node degrees as proxies as they have fewer nodes around then the nodes in the center. To compute this dependence we introduced the notion of spatial eccentricity and the degree bias in Definition 3.2. We now want to compute the correlation between those two values for our three cities to see if we can observe such a dependency in our data sets as well.

This is not necessarily true for all data sets, as we can see when we compare the correlations between degree bias and spatial eccentricity in our three data sets, split up in day and night time. For that we compute these values as defined in Definition 3.2 for all the nodes and then compute the Pearson correlation coefficient r_{XY} that measures the linear relationship between two data sets X and Y, whereas $r_{XY} = 0$ implies no correlation and correlations of -1 or +1 imply an exact linear relationship. We can see in Table 1 that the correlation of the spatial eccentricity and the degree bias highly varies for the three cities. For San Francisco, there is a strong positive correlation which accounts for high spatial eccentricity going along with high values of degree bias $db(i) = \frac{deg^{GM}(i)}{deg^{D}(i)}$, meaning that the degrees in the periphery are getting significantly underestimated. This effect is less strong for New York and hardly there for Washington DC.

We now want to look for possible explanations for the different correlations. If we look at the most visited places in the three cities, displayed in Fig. 2, San Francisco has its main attractions and most visited localities in the center of the clustering of all locations. in contrast to this, New York's JFK airport is a spatial outlier with high importance looking at

	Washington	New York	San Francisco
day	0.1461	0.4291	0.8223
night	0.1949	0.5132	0.7523

Table 1: Correlation of Spatial Excentricity and Degree Bias

the total number of check-ins, even though it is in the periphery. The same is true for the Washington DC airport. This relates to Stouffer's law of intervening opportunities in the sense that some places with a specific attraction, e.g. special sights or an airport, are not as easily replaceable as e.g. restaurants and therefore also suffer less from spatial effects as the number of intervening opportunities is very low.

Even if this is an observation limited to the most visited places in the city and more analysis would be necessary for a detailed analysis of spatial outliers with high node degrees in the different data sets, it motivates the following observation: Nodes in the periphery often have lower degrees because of less nodes around them if the data set consists of locations that are homogeneous in functionality and therefore replaceable, such as bike stations of a bike sharing network as in [16]. As our data set is functionally not homogeneous but includes all different kinds of venues ranging from transportation hubs to restaurants, the function has to be considered in the intrinsic strength of a node. A node that has a very unique functionality such as a transport hub or a specific tourist attraction therefore might lead to the effect that we have high node degrees even if the node is distantly very remote. Summing up we can say, that correlation of degree bias and spatial eccentricity may differ in a data set with functionally inhomogeneous nodes, because some locations are not easily reparable by closer ones. This agrees with Stouffer's law of intervening opportunities.

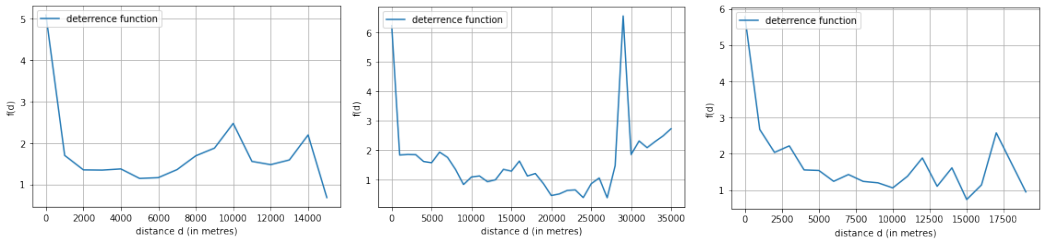
In any case, this motivates to use a degree constrained model for community detection as it is taking this effect into account by constraining the degrees of nodes and therefore systematically eliminating the degree bias, as explained in Section 3.

5.0.2 Deterrence function

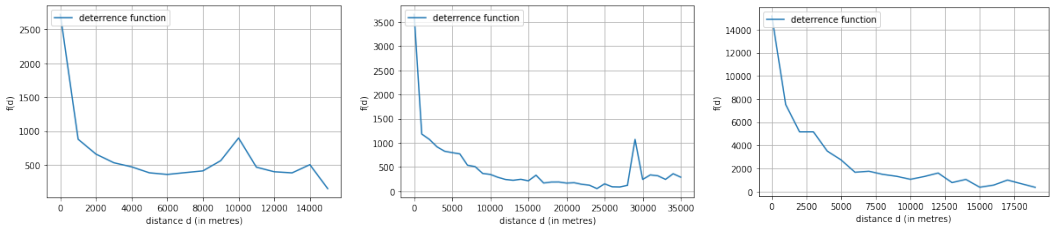
In this section, we want to make some observations about the deterrence function

$$f(d) = \frac{\sum_{i,j|d_{ij}=d} W_{ij}}{\sum_{i,j|d_{ij}=d} n_i n_j} \tag{15}$$

that is learned directly from the data. This deterrence function is used for both Q^{Gra} and Q^{DCGra}, however the gravity null model uses the node degrees as intrinsic strengths, $n_i = k_i$,



(a) Washington Q^{Gra} **(b)** New York Q^{Gra} **(c)** San Francisco Q^{Gra}



(d) Washington Q^{DGGra} **(e)** New York Q^{DGGra} **(f)** San Francisco Q^{DGGra}

Figure 3: Deterrence functions for the gravity model (left) and the degree-constrained gravity model (right), binning distance 1km.

and the degree corrected null model the iteratively constrained values $n^{eis} = \frac{deg(i)}{\sum_i n^{eis} f(d_{ij})}$, as described above. The two deterrence functions are displayed in Fig. 3. Looking at those functions, we can observe the outlier effect we have discovered in the previous section. We can see clear spikes for long distances in both the functions for New York and Washington DC, whereas the deterrence function of San Francisco is more similar to a classical power law of spatial dependency. This accounts for the fact that there are important locations with high node degrees in a spatially distant position. One can also observe that the degree constrained model partly eliminates those effects.

5.0.3 Dependence on binning distance

As highlighted in [11, 18], the evaluation of the deterrence function

$$f(d) = \frac{\sum_{i,j|d_{ij}=d} W_{ij}}{\sum_{i,j|d_{ij}=d} n_i n_j} \tag{16}$$

when measured from data is highly dependent on the binning distance d. The choice of the binning of d influences the precision of the deterrence function. When working with Foursquare data we are dealing with an incredibly fine granularity of the location data, with GPS accuracy down to 10 meters. Nevertheless, as we evaluate the data set it might not make sense to work with that high precision as the deterrence function does not bring clear

distinguishing power for too high precision due to a lot of noise. A fitting binning distance might vary for different city sizes and distribution of places. Therefore we compute the deterrence function for binning distances of $s = 10, 100$ and 1000 metres for all three cities and show the results in the appendix in Fig. 9. We can see from the results that a binning distance of 1km makes sense for our purposes as the noise in more precise bins is negligible but the deterrence function preserves the main properties. We will use this distance for the community detection below.

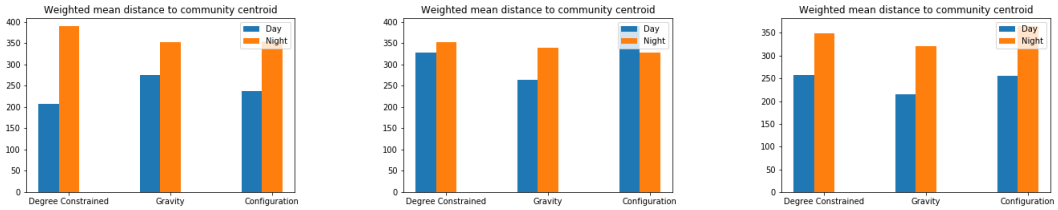
5.0.4 Qualitative Analysis of the cities at day and night

We now want to apply the methods to the data sets and compare the derived communities for night and daytime transitions. As explained in Section 3, the three different null models take into account different properties of the networks and therefore discover different communities in the data set. We visualise those communities in Fig. 6, Fig. 7 and Fig. 8. It is already visible from these maps that the communities differ both for the three different methods as well as for day and night time. From theory we would expect, that communities for the configuration model are highly influenced by spatial factors and therefore spatially more clustered than the communities formed by the gravity and the degree-constrained gravity model. The latter ones are both governed by non-spatial factors whereas the degree biases in the data set are eliminated for the degree-constrained model. Nevertheless, we cannot make any conclusions about what is influencing the formation of the different communities for the latter models as we do not have enough information about underlying possible influences such as transportation routes or significance of certain nodes. It would be a matter of further research to try to relate the formed communities to possible influences in the different cities such as common work permutations etc. Nevertheless, we make a first attempt here to measure the spatial extents of the communities under the different models and times of day. For this, we compute the radius of gyration of the different communities, that is defined as follows:

Definition 5.1 (Radius of gyration, from [8])**.** For a set consisting of n nodes $p_1, \cdots, p_n$ with centroid $p_{cen} = \frac{1}{n}\sum_{i=1}^{n} p_i$ the radius of gyration is defined by

$$r_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(p_j - p_{cen})^2} \tag{17}$$

The radius of gyration represents the standard deviation of distances between points of a trajectory and the center of mass of these points. A low radius of gyration indicates that



Washington New York San Francisco

Figure 4: Weighted mean radius of gyration

transitions in a certain community are mostly locally while a high value of this metric generally stands for predominantly long distance transitions [8, 10]. The results displayed in Fig. 4 show a weighted mean of the radii. We weight the mean by the size of the communities to not overestimate the influence of small communities. A more detailed analysis of the radii of gyration for the individual communities can be found in the appendix in Fig. 10. Generally we can see from these quantities that communities generally consist of more locally bounded transitions at day than at nighttime. This might be due to the fact that people work in the day and therefore generally do not make a lot of long distance transitions. Further research is necessary to investigate this further.

If we want to compare the different community structures for night and daytime, we can also look at the modularity values that were achieved by the displayed partitions. Hereby it must be considered that a direct comparison of values for Q^{NG}, Q^{Gra} and Q^{DGGra} is not possible because modularity is a way to compare different partitions of the same graph and so its absolute value is meaningless [11]. In general, the modularity is expected to be lower when its null model is closer to the real structure of the data, as it is the case for Q^{Gra} and even more for Q^{DGGra}, because more constraints on the null model are added. This result is observed in the computed modularities displayed in Fig. 5 (The precise values can be found in the appendix in Table 3, Table 2 and Table 4). Nevertheless, the different cities can be compared with each other for each modularity value individually. It can be clearly seen in the tables, that the modularity values are higher for daytime than for nighttime for all three methods. It must be noted that this seems to align with the previous result that communities are more spatially extended during the night. This connection can just partly be made, because communities in the two gravity-based models form mainly due to non-spatial factors and therefore a higher modularity value for daytime in those models is not expected to go along with higher spatial clustering. Nevertheless, those results are interesting as they suggest that the higher modularity values at daytime might come from somehow more structured transition patterns during the day. We cannot draw conclusions

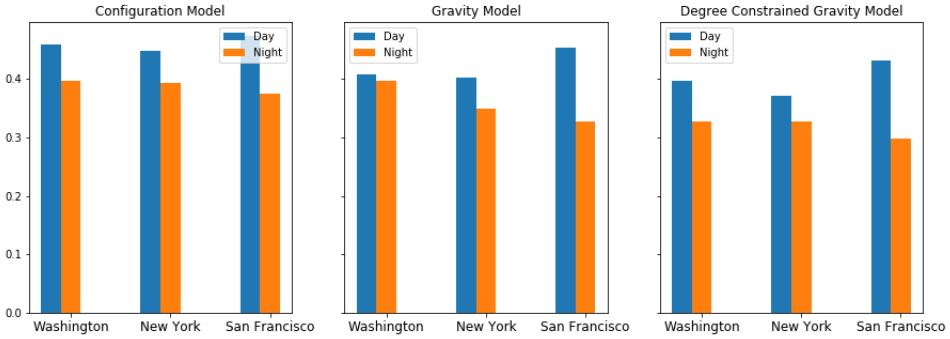


Figure 5: Modularity values of derived communities for the three different null models

about the reason for this and further investigation would be necessary if it is for example due to work transitions vs. leisure activities.

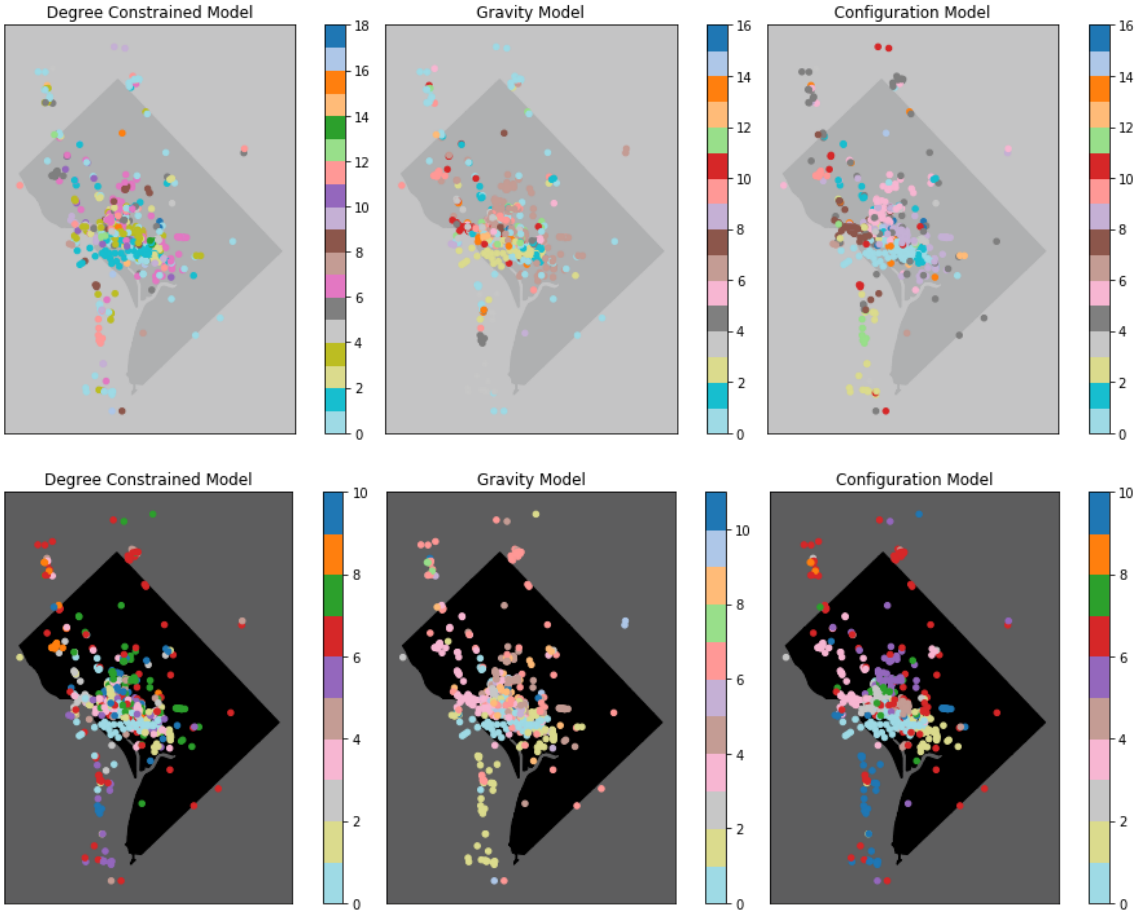


Figure 6: Communities for Washington DC for daytime (top, 725 nodes, 14645 edges and 31017.0 trips) and nighttime (bottom, 730 nodes, 27982 edges and 52594.0 trips)

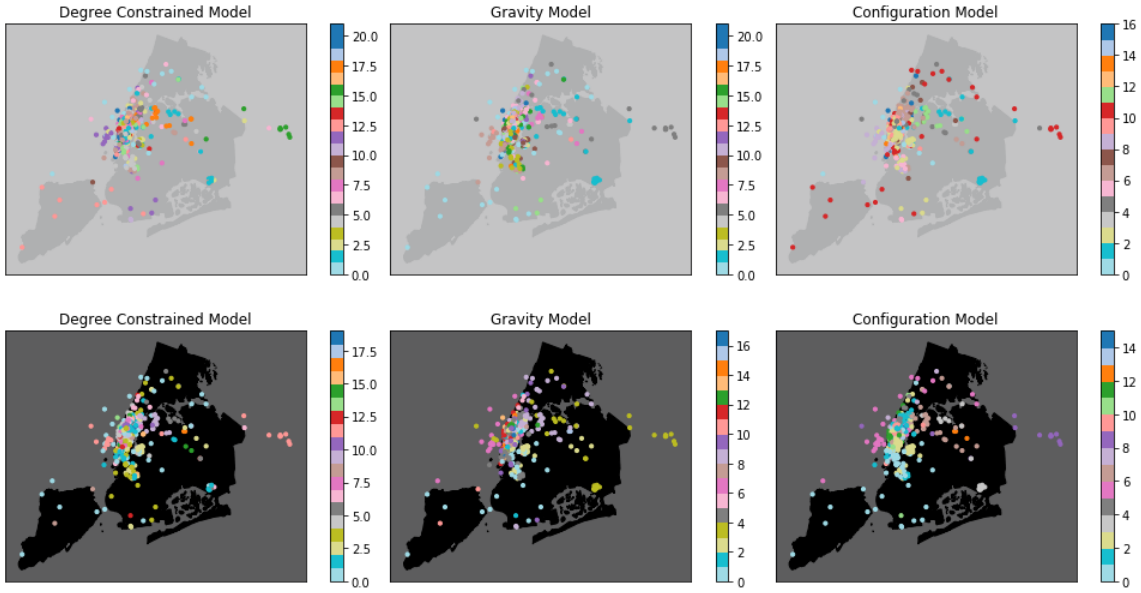


Figure 7: Communities for New York City for daytime (top, 633 nodes, 23883 edges and 62569.0 trips) and nighttime (bottom, 634 nodes, 38934 edges and 90256.0 trips)

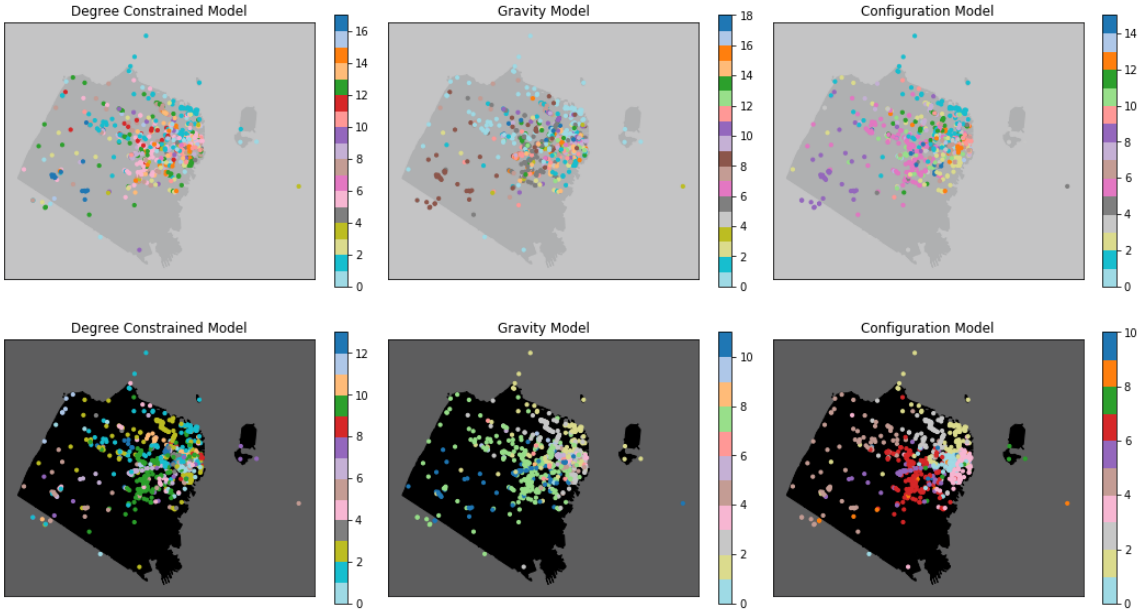


Figure 8: Communities for San Francisco for daytime (top, 936 nodes, 10628 edges and 17708.0 trips) and nighttime (bottom, 939 nodes, 37093 edges and 60957.0 trips)

6 Conclusion

In this work we have presented three different null models for community detection and have applied them to a data set of transitions between different localities in three US cities. We have shown, that there are significant degree biases in the data set that come from outliers with unexpected high node degrees that have to be considered in the analysis and therefore the application of a degree constrained model is senseful when using node degrees as proxys for their intrinsic strength. Our new observation was, that this does not necessarily have to come from lower node degrees in the periphery but can also be biased by the functional heterogeinity of our data set, in which outliers like airports or main sites of a city highly influence the commutation of people, disregarding the spatial distance. We also saw that the modularity values for all three methods are generally higher for daytime than for nighttime. We cannot draw any general conclusion from this yet as the underlying influences on the community formation when spatial effects are eliminated are unclear and vary for different cities and importance of localities for different day and nightimes. Furthermore, the spatial extend of the communities, measured by the radii of gyration, is higher at night than at day for all the models. In further research this can be explored further with regard to possible explanations of those patterns. It could be explored if more connectivity is caused between certain boroughs due to work flows. It would be also interesting to examine how the probability of transitions between various types of establishments changes as function of day vs. night. Maybe nighttime communities are more facilitated by social interactions (so clubs to bars, etc.) vs. daytime communities (facilitated by workplaces to cafes, workplaces to workplaces). This could be a first step towards an explanation for the different spatial clustering that we observed in this work.

7 Acknowledgements

We would like to thank Anastasios Noulas (Center for Data Science, New York University) for kindly providing us with the Foursquare data set and providing useful advice.

References

[1] E. G. Ravenstein. “The Laws of Migration”. In: *Journal of the Statistical Society of London* 48.2 (1885), pp. 167–235.

[2] Samuel A. Stouffer. “Intervening Opportunities: A Theory Relating Mobility and Distance”. In: *American Sociological Review* 5.6 (1940), pp. 845–867.

[3] Gerald A. P. Carrothers. “An Historical Bedew of the Gravity and Potential Concepts of Human Interaction”. In: *Journal of the American Institute of Planners* 22.2 (June 30, 1956), pp. 94–102.

[4] Edward Miller. “A Note on the Role of Distance in Migration: Costs of Mobility Versus Intervening Opportunities*”. In: *Journal of Regional Science* 12.3 (1972), pp. 475–478.

[5] M. E. J. Newman and M. Girvan. “Finding and evaluating community structure in networks”. In: *Physical Review E* 69.2 (Feb. 26, 2004), p. 026113.

[6] M. E. J. Newman. “Modularity and community structure in networks”. In: *Proceedings of the National Academy of Sciences* 103.23 (June 6, 2006), pp. 8577–8582.

[7] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 9, 2008).

[8] Marta C. Gonzalez, Cesar Hidalgo, and Albert-Laszlo Barabasi. “Understanding Individual Human Mobility Patterns”. In: *Nature* 453 (July 2008), pp. 779–82.

[9] Marc Barthelemy. “Spatial Networks”. In: *Physics Reports* 499.1 (Feb. 2011), pp. 1–101.

[10] Zhiyuan Cheng et al. “Exploring millions of footprints in location sharing services”. In: *Aaai Icwsm*. 2011.

[11] P. Expert et al. “Uncovering space-independent communities in spatial networks”. In: *Proceedings of the National Academy of Sciences* 108.19 (May 10, 2011), pp. 7663–7668.

[12] Anastasios Noulas et al. “A Tale of Many Cities: Universal Patterns in Human Urban Mobility”. In: *PLOS ONE* 7.5 (May 29, 2012), e37027.

[13] Yu Liu et al. “Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data”. In: *PLoS ONE* 9.1 (Jan. 17, 2014). Ed. by Peter Csermely, e86026.

[14] Laura Lotero et al. “Several Multiplexes in the same City: The role of wealth differences in urban mobility”. In: *arXiv:1408.2484 [physics]* (2016).

[15] Marta Sarzynska et al. “Null models for community detection in spatially embedded, temporal networks”. In: *Journal of Complex Networks* 4.3 (Sept. 2016), pp. 363–406.

[16] Remy Cazabet, Pierre Borgnat, and Pablo Jensen. “Enhancing Space-Aware Community Detection Using Degree Constrained Spatial Null Model”. In: Springer Proceedings in Complexity. Feb. 23, 2017.

[17] Anes Bendimerad et al. “Contextual Subgraph Discovery with Mobility Models”. In: *Complex Networks & Their Applications VI*. Ed. by Chantal Cherifi et al. Vol. 689. Cham: Springer International Publishing, 2018, pp. 477–489.

[18] Remy Cazabet, Pierre Borgnat, and Pablo Jensen. “Using Degree Constrained Gravity Null-Models to understand the structure of journeys’ networks in Bicycle Sharing Systems.” In: (), p. 6.

[19] Daggitt Matthew L. et al. “Tracking urban activity growth globally with big location data”. In: *Royal Society Open Science* 3.4 (), p. 150688.

[20] Benjamin G Heydecker. “On the calibration of the gravity model”. In: (), p. 16.

	Washington	New York	San Francisco
day	0.4591	0.4473	0.4734
night	0.3964	0.3929	0.3758

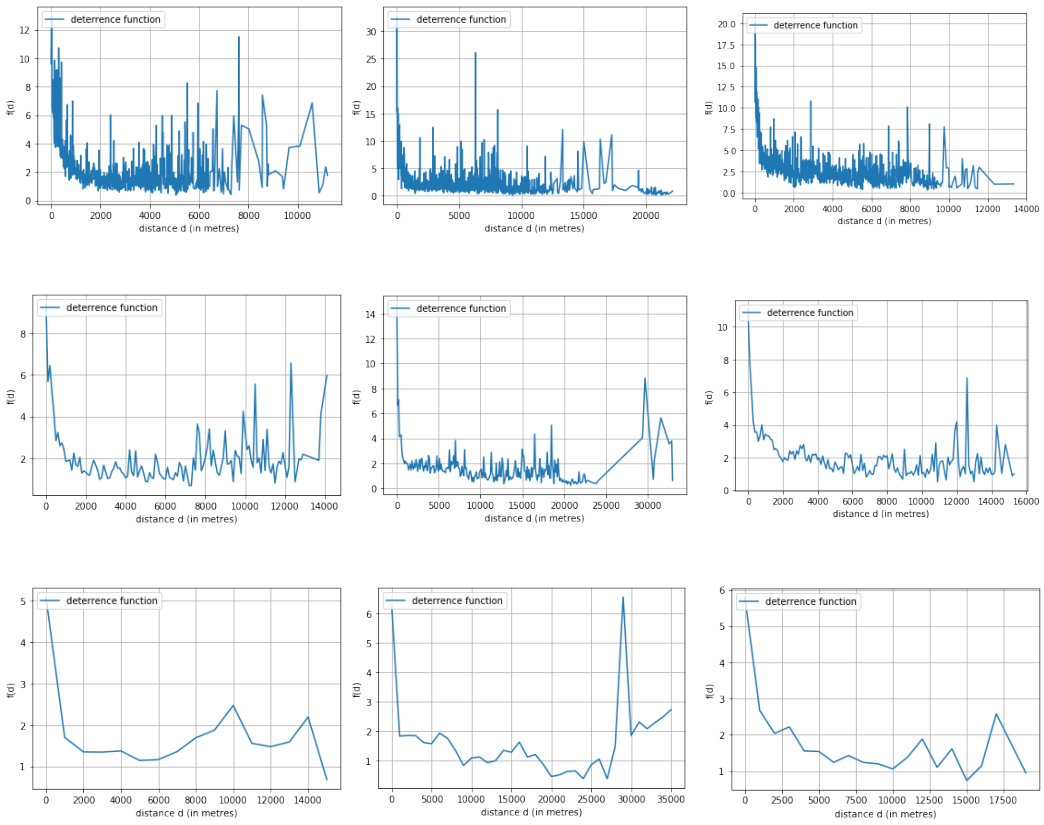
Table 2: Modularity Configuration Model

	Washington	New York	San Francisco
day	0.4087	0.4022	0.4536
night	0.3964	0.3495	0.3279

Table 3: Modularity Gravity Model

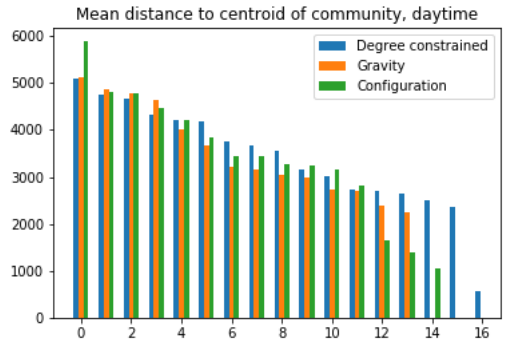
	Washington	New York	San Francisco
day	0.3975	0.3705	0.4309
night	0.327	0.3272	0.2989

Table 4: Modularity Degree constrained Gravity Model

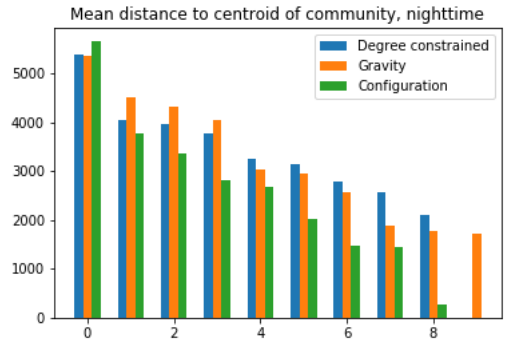


(a) Washington **(b)** New York **(c)** San Francisco

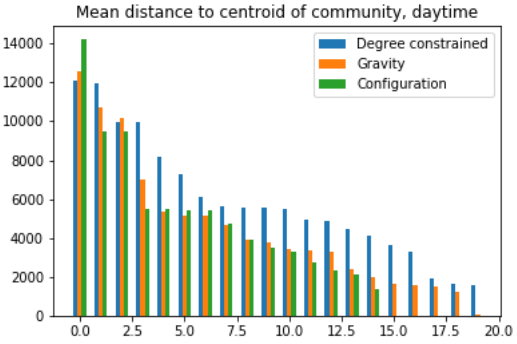
Figure 9: Binning distances for the deterrence function learned from the data sets of the three cities for bins of (from top to bottom) 10m,100m,1000m



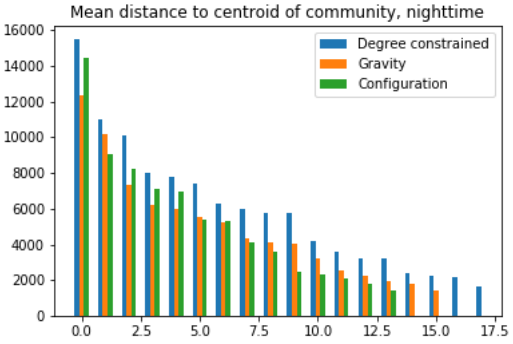
Washington day



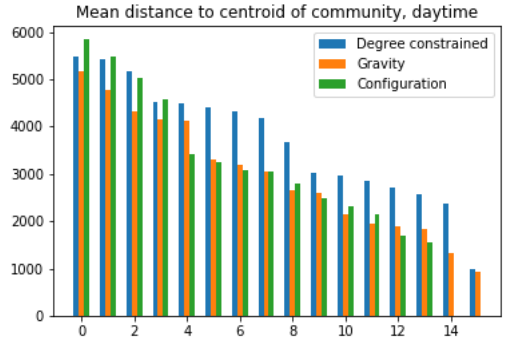
Washington night



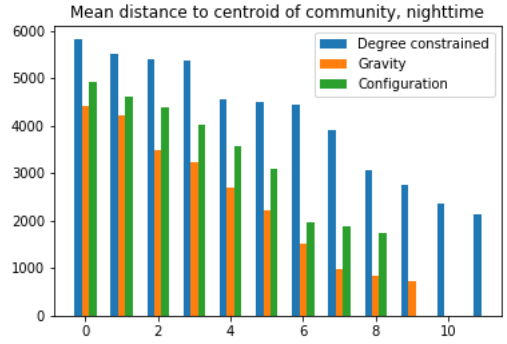
New York day



New York night



San Francisco day



San Francisco night

Figure 10: Radius of gyration for all the communities