Demarcating Geographic Regions using Community Detection in Commuting Networks

Mark He *, Joseph Glasser 1, Nathaniel Pritchard 2, Shankar Bhamidi PhD †, and Nikhil Kaza PhD ‡

1 Statistics & Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
2 Statistics, University of Wisconsin at Madison, Madison, WI, 53706, USA
3 City & Regional Planning, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

Abstract

We find agglomerations of U.S. counties that are partitioned by commuting patterns by representing inter-county commuting patterns as a weighted network. To do so, we develop a community detection method based on the configuration model to identify significant clusters of nodes in a weighted network that prominently feature self-loops which represent same-county commuting. Application of this method to county level commuting data from 2010 yielded regions that are significantly different from existing delineations such as Metropolitan Statistical Areas and Megaregions. Our method identifies regions with varying sizes as well as highly overlapping regions. Some counties may singularly define a region, while others may be part of multiple regions. Our results offer an alternative way of categorizing economic regions from existing methods and suggest that geographical delineations should be rethought.

1 Introduction

Delineations of agglomerations of geographic objects are usually determined by their analytic and political purposes (Jones and Paasi, 2013; Paasi, 2013). For example, watershed boundaries are determined by hydrological flows, councils of governments are based on political connections among jurisdictions, and labor market areas are based on matching employment opportunities with worker residences. On the other hand, megaregions are identified by the Regional Planning Association (RPA) based on linkages between economic systems, infrastructure patterns, and projected population growth (Hagler, 2009). These agglomerations are regions, which are fundamental objects in regional science (Pike, 2013). In this paper we present a method to delineate regions using community detection methods from network science. One of the key innovations of this approach is to identify multiple overlapping regions of different types and at different scales in the same statistical inference framework. We also extend community detection methods to include self-loops that have typically been ignored in other community detection work. However, self-loops are of great importance in commuting networks and the more general context of spatial networks that characterize collective behaviors of agglomerations of social or physical particles.

Commuting networks have been used in many different settings (see e.g. Manley, 2014; Tonev et al., 2017; Nelson and Rae, 2016). For example, the U.S. Office of Management and Budget (OMB) uses commuting networks to define core based statistical areas (CBSA). Surrounding counties are added to

*MH was supported by an NSDEG fellowship.
†SB was supported in part by NSF grants DMS-1613072, DMS-1606839 and ARO grant W911NF-17-1-0010.
‡Corresponding Author: Corresponding email at nkaza@unc.edu
Figure 1: Conceptual diagrams representing a) Overlapping non-nodal communities b) Nodal communities (trees) c) monads. The different colors represent different clusters/regions

A CBSA core county (population thresholds of 10,000 or 50,000 in urban areas for micropolitan (µSA) and metropolitan areas (MSA)), if social and economic ties (as defined by commuting to the core county) are more than 25% of the total ties. Adjacent CBSAs are merged into a single CBSA when the central county or counties of one CBSA qualify as an outlying county or counties to the other CBSAs. The county to CBSA relationship is a one-to-one relationship; i.e. a county belongs to only one CBSA. If a county satisfies threshold requirements for multiple CBSAs, then it is allocated to the CBSA that has the strongest ties unless it is a core county. Furthermore, CBSAs are necessarily geographically contiguous (Office of Management and Budget) 2010).

The core-based approach to identifying urban regions has a long history (since 1950s). Key to this approach is the identification of a core and its connections with the hinterland (nodal communities) (Nystuen and Dacey 1961). Research has focused on the core as the central subject of study rather than acknowledging the polycentric regional forms that have come to dominate regions around the world (Fowler et al. 2018). Since the 1980s, research has shown that the majority of the commuting flows in urban systems are lateral (i.e. between different parts of suburbs and hinterland) rather than core-centered (Plane 1981). Methods have been proposed to account for these peripheral commuting patterns and use them to delineate regions (e.g. Tong and Plane 2014).

In addition to this issue of core focus, another problem of unique membership persists in standard delineations. While objects can logically belong to multiple agglomerations, many of the applications tend to assign them to one by breaking membership ties. Many counties have large numbers of commuters to different cities that are relatively close to one another, especially in densely urbanised regions (Han and Goetz; Han and Goetz; Kim et al. 2017). In such instances, it is useful to relax the unique relationship condition between an object and the agglomeration it belongs to.

The other major issue that has received less attention in the literature (both in the context of regional demarcation as well as in the area of network science) is the idea of self-connection. Many agglomeration delineations in both network science and regional science have focused on the dyadic connections between two nodes/counties. However, commuting networks have significant self-loops (i.e. commuters within a county). In fact, 56% of the total commuters in 2010 in the US commuted within the same county they resided in. Ignoring this large commuting pattern skews the results of agglomerations. Since some nodes may be preferentially attached to themselves (measured by the weight of the self-loop), they should be treated as their own agglomerations known as monads (section 2.5). To this end, we propose a methodology that accounts for these three different types of regions that account for multiple memberships, core-based nodal communities and monads (see figure 1).

Networks are used to model the relational structures between individual units of an observed system. A multitude of data structures may be conceived of as networks from the biological, physical, and social sciences. Over the last few years, owing to the explosion in data from a host of areas including social
networks, information networks such as the Internet and biochemical networks such as gene regulatory networks, there has been a concerted inter-disciplinary approach to understand this data, see (Newman 2018, 2003; Boccaletti et al. 2006; Durrett 2007; Van Der Hofstad 2009). Due to the inherently relational nature of the commuting data, network methods are a naturally fitting approach to identify clusters.

In this paper, we present a method to identify clusters in a weighted network with self-loops. Using this method, we analyze the commuting between counties in the United States in 2010 to identify the clusters of counties that are significantly connected with themselves and with others. We proceed by describing the methodology and theoretical basis. We then present and discuss the results of regional delineation. We conclude with the limitations and further research directions.

2 Methodology

Community detection is an approach to divide a set of nodes in a given network into clusters that are strongly connected within clusters as opposed to between clusters. Many techniques have been proposed for unweighted or binary data ranging from modularity optimization (Girvan and Newman 2002; Clauset et al. 2005), structured stochastic blockmodels (Holland et al. 1983; Nowicki and Snijders 2001; Peixoto 2018; Yan et al. 2014), and extraction methods (Zhao et al. 2012; Lancichinetti et al. 2011). Most of these methods are based on a null network model. Typically, one constructs a network model that preserves some aspects of the observed network (in the context of unweighted networks, this is most often the degree distribution of the network, see Section 2.2 for details) but creates a network with no inherent clustering tendency. One then compares the observed network with this null model to extract subsets which seem more densely connected within the subset as compared to the null model.

The Configuration Model on a simple graph, first introduced by Bollobas (1980) and Bender and Canfield (1978), is a probability measure on a family of multigraphs that preserves the degree sequence. The input to the model is an observed graph from which we extract the degree sequence namely a list consisting of the vertices and their corresponding degrees. The model then constructs a random graph as follows: start with the degrees of nodes with \(d_u\) denoting the degree of vertex \(u\); associate every vertex \(u\) with \(d_u\) “stubs”. One then performs a uniform matching on these stubs to form full edges, thus resulting in a random graph with the prescribed degree sequence but without any other inherent clustering tendency. The relative proclivity of each node to form ties is determined purely upon its degree.

Many of the aforementioned community detection methods utilize the configuration model as the null (Lancichinetti et al. 2011; Palowitch et al. 2018; Fosdick et al. 2018). We significantly extend the methodology developed by a subset of the authors (Palowitch et al. 2018) for weighted network data by developing a new null model for weighted networks with self-loops. The outcome of the methodology reveals both significantly connected communities, monads, as well as nodal communities (see figure 1).

In the following sections, we elaborate on our methodology by first defining the terminology in section 2.1. We then describe the null model in section 2.2 paying close attention to the adjustments for weighted self-loops. In section 2.3 we discuss the variance parameters of the latent variable associated with the null model. We describe the community detection steps (initialization, update, filtering) in section 2.4 and describe the post-processing step to separate the nodal communities from non-nodal communities in section 2.6. In addition to communities, since monads are important, we describe their identification in section 2.5. We finally describe methods to benchmark the results with other delineations in section 2.7.

2.1 Notation

Let an undirected weighted network on \(n\) nodes be denoted by the triple \(G = ([n], \mathbf{A}, \mathbf{W})\), where \([n] := \{1, 2, \ldots, n\}\) is the set of \(n\) labeled nodes; \(\mathbf{A} = (A_{uv})\) is an \(n \times n\) square, symmetric adjacency matrix with
A_{uv} = 1 if there is an edge between u and v, and A_{uv} = 0 otherwise. Since we are interested in networks with self-loops, we assume A_{uu} ≡ 1 for all u ∈ [n]. W = (W_{uv}) is another symmetric matrix representing (non-negative) weights of edges between vertices with W_{uv} denoting the weight between u, v ∈ [n] with W_{uv} ≡ 0 if there is no edge between u and v.

We let d_u = ∑_{v∈[n],v≠u} A_{uv} denote the degree of a vertex, which specifically is the total edges connecting to u but not u itself. The total strength of a node u is defined as: s_u = ∑_{v∈[n]} W_{uv}. We let d_T = ∑_{u∈[n]} d_u and s_T = ∑_{u∈[n]} s_u denote the total degree and weight of the network respectively. We define ϱ_u to be the propensity of the node to connect to itself by

\[ ϱ_u := \frac{W_{uu}}{s_u}, \quad u ∈ [n]. \]

We let the baseline propensity of the self-loop ratio for the entire network be:

\[ p = \frac{∑_{u∈[n]} W_{uu}}{s_T}. \]  \hspace{1cm} (1)

We let d = (d_1,...,d_n) and s = (s_1,...,s_n) denote the degrees and strengths of nodes in [n] respectively.

2.2 Null model for weighted networks with self-loops

A null model in the context of community detection is a random network model which preserves some aspects of an observed network but without any explicit community structure. The most common null model used in the context of unweighted networks is the configuration model as described in the previous section. Once one has a null model for the network, communities based on the null model are groups of nodes that deviate from the notion of a baseline by being more connected to each other than expected under the null.

2.2.1 Related Null Models

Various functionals are used to measure the deviation of a set or a partition of the entire node set from the null model, the most popular among these being the modularity score [Girvan and Newman (2002); Clauset et al. (2005)]. One can then try to optimize such scores to find the best partitions. [Fosdick et al. (2018)] introduced a framework for configuration models that accounts for self loops and used a modularity optimization approach for community detection as an application, but do not focus on weighted networks whose most prominent edges are self loops. Another null-model based approach is to assess the statistical significance of the deviation of subsets from what one would expect under the null, correcting these estimates for multiple testing, and then extracting communities that appear to be more significantly connected than under the null. Several approaches have implicitly utilized the notion of deviation against null partitions such as likelihood ratios (Yan et al., 2014) or Bayes factors (Peixoto, 2018).

Significance-based testing was directly pursued in the context of unweighted networks in [Wilson et al., 2014] and weighted networks in [Palowitch et al., 2018], but do not account for networks with strong self-looping characteristics. We extend the significance testing based approach in which we now extend both in scope and application by adjusting for self-loops in a dataset with prominent weighted self-loops.
2.2.2 Model Construction

We start by describing the null model for weighted networks with self-loops that will serve as a comparative model for an observed weighted network \( G = ([n], A, W) \). The model is indexed by a family of parameters \( \theta = (d, s, \kappa_{SL}, \kappa_{nSL}, a, b) \) where \( d, s \) are as before the degree and weight sequence of the observed network respectively, \( \kappa_{SL}, \kappa_{nSL} > 0 \) are parameters that control the variance of self-loop and non self-loop edge distribution in the null model and \( a, b > 0 \) are parameters constrained by the relation \( a/(a + b) = p \) where \( p \) is as in (1) is the global self-looping tendency of the observed graph. Implicitly we will fix two distributions \( F_{SL} \) and \( F_{nSL} \) on \( \mathbb{R}^+ \) with mean one and variance \( \kappa_{SL} \) and \( \kappa_{nSL} \) respectively. Using the above ingredients we construct a random weighted graph \( \hat{G} = ([n], \hat{A}, \hat{W}) \) as follows:

(i) **Network topology:** By design \( \hat{A}_{uu} = 1 \) for all \( u \in [n] \). For all \( u \neq v \) we let

\[
\mathbb{P}(\hat{A}_{uv} = 1) = \frac{d_u d_v}{d_T},
\]

(ii) **Self-loop edges:** For self-loop edges, we generate edge strengths as follows: First for each vertex \( u \in [n] \) (independently across vertices), we generate its self-loop propensity \( \hat{\varrho}_u \sim \text{Beta}(a, b) \) (i.e. a Beta distribution). Next we generate \( \xi_{uu} \) from distribution \( F_{SL} \) (independent of \( \hat{\varrho}_u \)). Hence, we model \( \hat{W}_{uu} := \hat{\varrho}_u s_u \xi_{uu} \).

(iii) **Non self-loop edges:** For \( u \neq v \) generate edge strengths as follows: first if \( \hat{A}_{uv} = 0 \) from step (i) then let \( \hat{W}_{uv} = 0 \). If \( \hat{A}_{uv} = 1 \) then let \( \xi_{uv} \sim F_{nSL} \) and let

\[
\hat{W}_{uv} = (1 - \hat{\varrho}_u) q_{uv} \xi_{uv},
\]

where each \( q_{uv} \) represents the following ratio of strengths and degrees of \( u \) and \( v \):

\[
q_{uv} = \frac{s_u s_v}{d_u d_v},
\]

Writing \( D(u) \) and \( S(u) \) for the degree and strength of vertex \( u \) in the random graph it is easy to check that

\[
\mathbb{E}(D(u)) = d_u, \quad \mathbb{E}(S(u)) = s_u, \quad \mathbb{E}(\hat{W}_{uu}) = ps_u.
\]

Thus the model preserves (on average) the degrees and strengths of the observed graph without any other specific notion of clustering. Further in the null model, each vertex has no particular preferential notion of self-loopiness, rather on the average the self-loopiness density of each vertex is the same as the observed graph.

2.3 Parameter specifications

Palowitch et al. [2018] use a method-of-moments estimator to specify parameters. We extend this assumption and posit two variance parameters to capture the variation in interconnectivity as well as self-looping tendencies of the null model. We specify two types of variables to describe the uncertainty arising out of the strengths of the nodes’ connection propensities. Recall that we denote \( \kappa_{SL} \) as the variance of self-looping edge weight distribution (with distribution \( F_{SL} \)) and \( \kappa_{nSL} \) as the variance of the non-self-looping edge weight distribution (with distribution \( F_{nSL} \)). Both of these variables have mean one to ensure the preservation of the strengths and degrees for the configuration null model.

5
2.3.1 Variance Parameters

The expressions of $\kappa_{SL}$ and $\kappa_{nSL}$ are as follows:

\[
\hat{\kappa}_{SL} = \frac{\sum_{u \in [n]} (W_{uu} - ps_u)^2 - \sum_{u \in [n]} s_u^2 \hat{\sigma}_p^2}{\sum_{u \in [n]} s_u^2 (\hat{\sigma}_p^2 + p)}
\]

\[
\hat{\kappa}_{nSL} = \frac{\sum_{u \in [n]} \sum_{v \neq u} (W_{uv} - (1 - p)q_{uv})^2 - \hat{\sigma}_p^2 \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2}{(\hat{\sigma}_p^2 + (1 - p)^2) \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2}
\]

where $\hat{\sigma}_p^2$ represents the estimated variance of $\varrho_u$ using empirical method-of-moments.

\[
\hat{\sigma}_p^2 := \text{Var}(\hat{\varrho}_u) = \frac{1}{n-1} \sum_{u \in [n]} \left( \frac{W_{uu}}{s_u} - p \right)^2.
\]

$\hat{\kappa}_{nSL}$ represents the variation in relative weights between two edges when we know that the strengths (total sum of weights) are the same, and $\hat{\kappa}_{SL}$ represents the variation within a single self-directed edge. Details on the derivations of these parameters can be found in the Appendix A.

2.3.2 Beta Random Variable to Model Self Looping Propensity

We specify $\varrho_u$ as adhering to a Beta($a,b$) distribution with mean $p$ and with variance equal to $\sigma_p^2$, the sample variance of $p$. The beta distribution is supported on $[0,1]$. We are considering the proportion of self-looping commuters in each node as being made up of the averages of decisions to either commute in-county or out-of-county by a host of commuters, and have found that the beta distribution is a suitable description of such data. We have verified that the empirical distribution of $\hat{\varrho}_u$ closely match the simulated values of $\hat{\varrho}_u$, except for a few nodes that are 0% or 100% self-commuting (Appendix B).

\[
E(\varrho_u) = \frac{a}{a+b} = p,
\]

\[
\text{Var}(\varrho_u) = \frac{ab}{(a+b)^2(a+b+1)} := \sigma_p^2
\]

The variance of $\varrho_u$ is approximated using its sample variance $\hat{\sigma}_p^2$, which is calculated with a method-of-moments estimator, as in equation 8.

2.4 Community Detection Algorithm

The Continuous Configuration Model Extraction (CCME) algorithm used here can be split into three general phases: initialization, update, and filtering. These steps compose the general procedure of iterative testing. Significant communities as imputed by the iterative testing algorithms on networks are those with cross-edges that deviate considerably from the expected edges under a null model. Significant communities are determined by repeatedly applying an iterative search algorithm that starts with a seed set $B_0$, finds all nodes with edges connecting to the seed set, then summing up the edge-weights as a test statistic. The test statistic is then evaluated against the expected values of the sums of the weights in the set under the null model (described in appendix A.4) with respect to each node in the starting seed set $B_0$, and imputes a p-value for each node.
Each p-value from the whole set is then rejected after being filtered by the Benjamini-Hochberg correction. The nodes with p-values that are rejected after the correction are extracted as the significant nodes upon the first iteration and are used as the initial seed sets for the next iteration step. The final set \( B \) is extracted when the node-set becomes stable: when at some iteration step \( k, B_k = B_{k+1} \). Nodes in the final set have stronger affiliation with each other and have less edge connections with all other nodes outside the set.

2.4.1 Initialization

We initialize (step 1) sets \( B_0 \) in our application by setting counties with high commuting volume as seed nodes/counties. We select counties that have above 20,000 self-commuters as seed counties. For each seed county, we then find all counties which have commuters connected to the seed county. The seed county/node in addition to counties connected to it are used as the initializing sets \( B_0 \). The seeds are largely irrelevant to the final outcome: the final outcomes are essentially same regardless of what the initial counties selected are, so long as a majority of the high-volume counties are included.

2.4.2 Update

Stable communities are found using iterative node-set updating scheme based on the p-value of the connectivity between a single node \( u \in [n] \), a candidate (testing) set \( B \in \mathcal{G} \). We denote \( S(u, B, \mathcal{G}) \) as the connectivity of a single node (in this case, county) to the set of nodes which is hypothesized to be a community:

\[
S(u, B, \mathcal{G}) = \sum_{v \in B} W_{uv}.
\]

When the observed value of \( S(u, B, \mathcal{G}) \) significantly exceeds the expected sum of weights under the null model, then there is evidence to support that there is some additional structure undergirding the set of nodes than that which is posited by the null model, which attributes connectivity between sets of nodes as only dependent on the strengths and degrees of the aggregations of the nodes themselves.

The p-value representing the significance of node-set is given by:

\[
p(u, B, \mathcal{G}) = \mathbb{P}(S(u, B, \mathcal{G}) > S(u, B, G)).
\]

In the above equation \( G \) is fixed but \( \mathcal{G} \) is random with distribution given by the null model \( \mathbb{P} \) and each \( B \) represents the candidate set to be tested. When the observed value of \( S(u, B, \mathcal{G}) \) is much larger than the expected value, expressed as \( S(u, B, G) \), the p-value is low. Low p-values are rejected in an iterative fashion as to allow the formation of node-sets with edges that are consistently significantly connected to each other. We define these sets as communities.

The iterative method is described as follows: for each \( u \in [n] \), given a set \( B \) (or denoted by \( B_k \) at \( k^{th} \) iteration if each iteration step is indexed), we find all counties that are connected to the present set, then p-values are imputed for members of the candidate set \( B \) and repeatedly tested until the set becomes stable upon sequential iterations:

(i) calculate p-values \( p = p(u, B, \mathcal{G}) \). P-value are calculated using a normal approximation of the values of \( S(u, B, \mathcal{G}) \). Details on this part of the procedure are given in appendix A.4.

(ii) obtain threshold \( \tau(p) \) using a Benjamini-Hochberg multiple testing procedure. The procedure is used for sets of p-values that are obtained through multiple hypothesis testing. The rejection method ensures that the expected number of falsely rejected hypotheses divided by the total number of rejected hypotheses (false discovery rate, or FDR) has a maximum percentage of \( \alpha \) (Benjamini and Hochberg, 1995). A false discovery rate threshold \( \alpha \) of 5\% is common in many applications.
however, for community detection, however, we find empirically that such a threshold should be lower to avoid excess overlaps.

(iii) the next set reached by the iteration is defined as \( B' = \{ u : p(u, B, G) \leq \tau(p) \} \)

The above steps are iterated with \( B' \) replacing \( B \) until we reach a fixed point. We set \( \alpha = .01 \) at each step of iterative testing to perform community detection. The threshold can be made higher or lower, and such adjustments do not change the results drastically (see supplement for details), but the threshold of .01 appears to be optimal for maximizing coverage and minimizing overlaps.

2.4.3 Filtering

After obtaining \( J \) initial communities, with the \( j \)th labeled \( B_j \), we filter communities first by redundancy, then we differentiate between types of clusters using the clustering coefficients of nodes. Redundant sets that have a high Jaccard overlap similarity are filtered and removed. A Jaccard similarity index of two sets is defined as the ratio of the size of common elements between the two sets over the total distinct elements, or concisely expressed as \( J(A, B) = |A \cap B|/|A \cup B| \) for two sets \( A, B \) [Jaccard, 1901]. Such an index for measuring overlaps is utilized in the same way for filtering clusters in [Palowitch et al., 2018].

For each pair of found communities, labeled \( B_i, B_j \), we evaluate their Jaccard similarity \( J(B_i, B_j) \). If it is above a given pre-set threshold \( \tau \), then the clusters are redundant and we select a preferred cluster by calculating the average weight per connection between nodes. We use a simple formula for a given node-set \( B \):

\[
K(B) = \frac{\sum_{v \in B, v \neq u} \sum_{u \in B} W_{uv}}{|B|}.
\]

\( K(B) \) roughly measures the average sum of weights among cross-edges per node within a community \( B \). A higher \( K(B_i) \) compared to \( K(B_j) \) signifies more interconnectivity between nodes in \( B_i \), and thus makes \( B_i \) redundant compared to \( B_j \) if there is ample overlap between the two sets. In this application, when comparing \( B_i, B_j \), we choose whichever \( K(B_i) \) or \( K(B_j) \) is higher. In the implementation of CCME, we set the \( \tau \) parameter to be 0.80, meaning that for two communities with more than 80% overlap, the one with higher average connectivity is used, while the other is filtered out.

2.5 Detection of Monads

One unique feature of geographical commuting networks is that many prominent nodes that have large strengths are not found in communities because most of the residents stay within their counties to commute. We define the degree of monadicity of a given node as the following:

\[
I_u := W_{uu} - ps_u.
\]

Assuming the observed graph originated from the null model, the variable \( I_u \) measures two things. Firstly, \( I_u \) measures how much more \( \varrho_u \) is for a given \( u \) than the global mean self-looping tendency \( p \). Secondly, \( I_u \) measures (assuming the network did originate from the null model) how much more the latent \( \xi_{uu} \) component of \( \hat{W}_{uu} \) is than its expected value of one (recall that self-loop weights are modeled as \( \hat{W}_{uu} = \hat{\varrho}_u \xi_{uu} s_u \) from equation [3]).

The exact form of the variance of \( I_u \) is difficult to calculate, but we use a simulation method to approximate a p-value for \( I_u \). We first determine estimates for \( a, b \) from the mean and variance of \( \varrho_u \) under the null model. Given \( p \), and the sample standard deviation \( \hat{\sigma}^2_p \), we find estimates for \( a, b \) as follows, by solving for the concentration parameters of the beta distribution from equation [3].
\[
\hat{a} = -\frac{p(p^2 - p + \hat{\sigma}_p^2)}{\hat{\sigma}_p^2},
\]
\[
\hat{b} = \frac{(p - 1)(p^2 - p + \hat{\sigma}_p^2)}{\hat{\sigma}_p^2}.
\]

We use these estimated parameters to simulate a measurement of how extreme \( \hat{I}_u \) of a given node is, compared to what would be expected assuming random generation from a beta(\( \hat{a}, \hat{b} \)) distribution. If the actual value \( I_u \) is large, as measured by whether it is above the \( \alpha \)-th quantile of the simulated values, then the node is deemed significant because it consistently exceeds what is to be expected from if \( \varphi_u \) was randomly generated from a distribution of the same parameters. Such a simulation-based method to approximate p-values is commonly used in the sciences (Ewens 2003).

The process to find significantly monadic nodes may be concisely described by the set of the following procedures. First, we simulate \( \tilde{\varphi}_u \) from beta(\( \hat{a}, \hat{b} \)) for every node (county). We then obtain the empirical distributions of how monadic a given node is by computing the empirical distribution of \( \hat{I}_u = W_{uu} - \tilde{\varphi}_u s_u \). Following this step, we consider \( I_u = W_{uu} - ps_u \). If \( I_u \) is in the \( 1 - \alpha \)-th tail of the distribution \( \hat{I}_u \) then the node is a monad at this instance of simulation. We repeat the procedure 10,000 times and the nodes that have been identified 10,000 times are conclusively categorized as monads. In practice, we set \( \alpha \) equal to .05 for this test. The resultant group of nodes are significantly monadic at the 5% significance level.

### 2.6 Differentiating Nodal Communities from Non-nodal communities

In the post-processing phase, we identify nodal communities within the statistically significant communities detected by using the local clustering coefficient as defined by Opsahl and Panzarasa (2009) for weighted networks in the post-processing phase. For an unweighted network, the local clustering coefficient of a node \( u \) is based on fraction of number of present ties over the total number of possible ties between the node’s neighbours. Thus root node of a tree-like structure, where there is a single core node that is connected to all its peripheral nodes, but none of the peripheral nodes connect to each other, will have low coefficient, while a node in a complete graph has a coefficient of 1.

For a weighted network, the minimum clustering coefficient of a particular node \( u \) in a set \( J \) is defined as

\[
C_{\text{min}}(u, J) = \frac{\sum \sum \sum_{v,w} C_{\text{min}}(u,v,w) \in J \min(W_{uv}, W_{vw})}{\sum \sum \sum_{v,w} \delta(v,u,w) \in J \min(W_{uv}, W_{vw})}
\]

A triplet defined as \( \delta(v,u,w) \) is any three nodes that have at least one edge to at least one other node in the triplet. A closed triangle \( \Delta(u,v,w) \) is a set of three nodes that have edges which are all connected with each other. The ratio \( C_{\text{min}}(u, J) \) is the ratio of sum of all closed triangles incident on \( u \). If the sum of minimum values of triangles are low compared to those of all triplets within a cluster, then that implies the presence of dominant node that projects high edge weights across many peripheral nodes. As such, a low clustering coefficient signifies that communities are bound together by a common node, while a high \( C_{\text{min}}(u, J) \) signifies that the nodes are all connected to each other.

We determine the overall clustering coefficient of a community by calculating an average of \( 1 - C_{\text{min}}(u, J) \) within set \( J \) weighted by the proportion of the sum of weights of each node \( u \) divided by the sum of total strengths in \( J \):

\[
C_{\text{min, total}}(J) = \sum_{v \in J} \frac{(1 - C_{\text{min}}(v, J))s_v(J)}{\sum_{u \in J} s_u(J)}.
\]
This metric is constructed to capture how tree-like the highest-strength nodes in a given cluster is. The lower $C_{min}(u, J)$ is, the more tree-like the node is. $1 - C_{min}(u, J)$ weighted by the strengths of nodes in a given cluster assigns a value of how tree-like and strong a node is. Summing these values gives an overall measure of how monocentric a cluster is, as the strongest nodes tend to have the smallest $C_{min}(u, J)$. The empirical distribution of the above coefficient is bimodal (see Appendix fig B) with a split around 0.4. We use this to identify nodal communities.

2.7 Methods to compare communities with other delineations

We compare our results with other existing delineations (in particular OMB’s Metropolitan areas) by using the Fuzzy Rand Index (FRI) \cite{Chakraborty et al. 2017}. The Fuzzy Rand Index is a metric that measures the similarity of two covers, $C_1$ and $C_2$ are. A cover $C$ is the assignment of vertices of a graph into $K$ groups. Our model allows our vertices, counties, to be assigned to multiple groups, or communities. Counties that are not assigned to any community will belong to their own group, a community of counties that do not belong to any other group. Let $V$ represent the set of vertices and $P$ represent the fuzzy partition of $V$, a cover. Each element $v \in V$ is characterized by its membership vector $P(v) = \{P_1(v), P_2(v), ..., P_k(v)\} \in [0, 1]^k$, where $P_i(v)$ is the degree of membership of $v$ in the $i^{th}$ community $P_i$. The norm of $P(u) - P(v)$ in a cover with $N$ communities will be defined to be $||P(u) - P(v)|| = \sum_{i=1}^{N} |P_i(u) - P_i(v)|/2$. For some cover $C$, the similarity measure between two nodes $u, v$ is defined as $E_C(u, v) = 1 - ||P(u) - P(v)||$. The distance measure between two covers, $C_1$ and $C_2$ of a network $V$ is defined as:

$$d(C_1, C_2) = \frac{\sum_{u,v \in V} |E_{C_1}(u,v) - E_{C_2}(u,v)|}{n(n-1)/2}.$$

Likewise, the similarity measure between two covers is $1 - d(C_1, C_2)$.

If a county, $u$, belongs to $n$ communities, then its membership vector will be $1/n$ for each group that it belongs to and 0 for all other groups. The MSAs are disjoint so the membership vectors for counties in this cover will be 1 for the group it belongs to and 0 otherwise.

3 Data Description

We downloaded the data from the Local Origin Destination Employment Statistics (LODES) by US Census. This data contains commuter data between census tracts for all of the continental United States in the year 2010, which are then aggregated to the county level. The data is stored as an undirected and weighted network with self-loops such that each node represents a county and each edge represents the number of commuters between the connected counties. Edges with fewer than 100 commuters are removed from the network. Commuters who travel more than 100km (distance between population weighted centroids) are also ignored to remove the effect of telecommuters or supercommuters. Such pre-processing was also done in \cite{Nelson and Rae 2016}. The resulting network has 3,091 nodes and 17,222 edges.

4 Results

We use the term ‘clusters’ to refer to all groups of counties or single counties that were imputed from our detection procedures, ‘communities’ to refer to clusters that contain more than one county, and ‘nodal communities’ to refer to clusters that have strong nodal centers with little lateral commuting and non-nodal communities to refer to communities that have strong peripheral commuting characteristics. We refer to ‘monads’ as those counties that are strongly connected to themselves.
Table 1: Size distribution of the identified regions. Note that the number of counties are the total number captured by these regions with duplicates removed, to avoid double counting within each subgroup.

From the 3,091 total number of counties, we found a total of 182 significant communities. Of these clusters, 14 are nodal communities, 78 are non-nodal communities, and 90 are monads. Together they cover 90.3 % of the population and commuters (93 % within-county, 87 % between-county.). The method simultaneously delineates both small (such as singletons and dyads) and large clusters consisting of hundreds of counties (see table [1]). For example, Santa Barbara and San Luis Obispo counties in California are strongly connected to one another and are separate from other clusters. Of the regions identified, 99 are dyads or singleton counties. 68 of these clusters are medium sized, comprising between 2 and 50 counties. In 20 other instances, our method identifies clusters consisting of 50 or more counties that span several states. These tend to be polycentric regions centered around multiple large cities such as Philadelphia, Washington DC, and Baltimore (see figure [2]). Out of the 182 total clusters, the average size of a cluster is 24 counties, with a standard deviation of 59. The median size, however, is only 2, signifying that the majority of imputed clusters are monads. When only considering communities, the average size is 49, with a standard deviation of 78. The median size of communities is 15, suggesting that the distribution of community sizes is right-skewed (see table 1). In general, many of the counties belong to overlapping clusters, with some belonging to as many as six different ones (see figure 3).

The parameters associated with the null model are as follows: $p$, roughly the mean of within-county commuters (though weights between cross-edges are double counted, the proportion of same-county commuters is 56%) is .46. $\hat{\sigma}^2_p$, the sample standard deviation of $G_u$ is .04, signifying that around half of commuting volume in the US are within the same county, and that the dispersion about the mean is fairly tight. The estimates for concentration parameters $a$ and $b$ of the beta variable $G_u$ derived from $p$ and $\hat{\sigma}^2_p$ are respectively 3.86 and 4.50, signifying that the self-looping tendencies of counties is not sparse. $\hat{k}_{SL}$ and $\hat{k}_{nSL}$ are .08 and .79 respectively, signalling a much higher variance for the latent variable undergirding between-county commute volume than for same-county commuting.

4.1 Benchmarking with other delineations

In general, the clusters found by our proposed method are much larger than existing delineations and account for much more of the inter-county commuting activity. We define the coverage rate as the number of commuters between all edges of the given sets of clusters divided by the sum of the total edges. The criterion captures how much of the total commuting activity is ‘captured’ by different modes of aggregation. The rate of coverage for all inter-county commuting was 86 % for all clusters, compared to only 48 % from OMB delineated MSAs and 77 % from megaregions. The coverage rate for all within county, or same-county, commuting was 92 % from clusters, compared to 86 % from MSAs and 74 % from megaregions.

Similarities between clusters and MSAs are assessed by using the Fuzzy Rand Index (Section 2.7) and by comparing size and volume characteristics in table 2. Clusters and MSAs are most similar in West North Central and Mountain regions and most different in East South central. In the Mountain and Pacific regions, the number of clusters are nearly the same, and mean commuter volumes are more comparable than in the eastern states.
Communities are most consistent with MSA delineations in large urban areas (figure 5). The size differences between the detected clusters and other delineations are muted in the regions around population centers such as Chicago, New York and Texas cities. The more populous and denser a core urban area is, the more concentrated a community becomes.

In Texas, for example, the economic regions of major cities Austin, Houston, Dallas, and San Antonio are all detected by our method, but regions are larger and overlapping. The Austin region is the only Texas community is drastically differs from MSA, perhaps indicating greater economic influence of the region that demarcated by the MSA. The Houston region can be split into coastal and inland commuting zones, which overlap at the core Houston area. The overlapping counties between Austin and San Antonio areas include Bastrop and Hays counties and the communities identify the dual memberships of these counties in larger cities’ economic spheres of influence. Each of the four cities are also quantified as monads (not shown), signifying the relationship between large urban centers and their associated suburbs and hinterlands [Nystuen and Dacey 1961]. However, the larger, more diffuse nature of the communities signals emergent lateral commuting behavior between these peripheral counties [Plane 1981].

The communities associated with New York all include the five boroughs; and two of the commu-
Figure 3: Heatmap of frequencies of each county to appear in any cluster (community, nodal cluster, or monad)

Communities are similar in size with the MSA, although the communities do not incorporate as much of New Jersey as the MSA. One community in particular stretches North to upstate NY and part of New Hampshire. In California, an imputed community in the Bay area region is similar to the San Francisco-Oakland-Fremont MSA, but the community amalgamates counties from San Jose–Sunnyvale–Santa Clara, CA MSA (see figure 4). The inter-regional commuting among these counties warrants this amalgamation. Thus in the Bay Area case, the community and MSAs are alike in the core area, but the community includes more peripheral counties. The same could be said for many other urban-centric areas, such as the community surrounding Minneapolis, which encompasses the MSA but incorporates more peripheral counties (figure 5).

Megaregions, in general, are larger than communities or nodal clusters, though with some exceptions. The largest megaregion is in the Midwest connecting Minneapolis from the west to Western Pennsylvania in the east, comprising 388 total counties. The largest community yields a similar number of counties at 356 but is located in the Georgia-Tennessee region, centered around Atlanta (see table 1). The largest communities, though similar in size to megaregions, appear to be in the southern central part of the US, where megaregions cover the least amount of ground (fig 2). Though megaregions tend to be larger than the average community, the megaregions in the South and central part of the US still do not take into account the counties with smaller commuter volumes. The coverage rate of commuting volumes is lower than communities for both within-county and between-county cases. The CCME algorithm tends to find larger communities in counties that are more tightly connected by commuting, while megaregions tend to identify clusters of large population centers that are relatively proximal. The Piedmont megaregion stretches from Atlanta, GA to central NC, and only comprises around 150 counties while most communities around that region are more concentrated in shape and spread towards smaller, more
Figure 4: Example of a tightly connected community in the Bay Area in Northern California

| Census Division | FRI | # Clusters | Mean Commuters | Mean Size | Coverage |
|-----------------|-----|------------|----------------|----------|---------|
|                 | Clus/MSA | Clus | MSA | Clus | MSA | Clus | MSA | Cliques | Clus | MSA |
| Northeast       | .66 | 8 | 15 | 703,315 | 133,223 | 10 | 2 | 86% | 92% | 60% |
| Midatlantic     | .64 | 18 | 31 | 2,023,890 | 442,430 | 16 | 3 | 97% | 97% | 81% |
| E N Central     | .63 | 24 | 70 | 1,557,092 | 179,574 | 22 | 2 | 81% | 83% | 68% |
| W N Central     | .76 | 15 | 34 | 796,990 | 163,363 | 36 | 4 | 61% | 68% | 63% |
| S Atlantic      | .60 | 33 | 80 | 1,225,996 | 195,596 | 32 | 4 | 85% | 90% | 69% |
| E S Central     | .58 | 15 | 34 | 1,062,264 | 107,392 | 59 | 3 | 97% | 97% | 56% |
| W S Central     | .66 | 33 | 43 | 661,861 | 230,651 | 19 | 3 | 76% | 85% | 70% |
| Mountain        | .80 | 33 | 35 | 265,302 | 173,790 | 5 | 2 | 51% | 82% | 71% |
| Pacific         | .70 | 42 | 43 | 706,892 | 321,860 | 5 | 2 | 47% | 79% | 74% |

Table 2: Characteristics of clusters (monads, nodal and non-nodal communities) compared to MSA for each Census Division. Cliques (non-nodal communities) are also included in the comparison of coverage.

...
Tulsa counties are members of 5 different communities. North Carolina is covered by three large, heavily overlapping communities. Nodal clusters are regions that have strong central attractors but little peripheral commuting. Most nodal clusters are found in the Mountain West, where the counties are on average much larger and more sparsely settled. Monadic regions are much more common along the western US because the larger size of counties. Most of the major cities are identified as monads, including Cook, Harris, and Los Angeles. Notably, New York County (Manhattan) and King County (Brooklyn) are not monads because the counties that comprise New York City are all they are highly interconnected to each other (see figure 2).

5 Discussion

Nearly all U.S. major urban areas are grouped into clusters, though not with the same extent as what other delineations would suggest. Counties like St Louis, MO and Fulton, GA (Atlanta), connect to many such smaller, less populated counties. The cities that surrounded by such counties yield much larger clusters than more urban areas such as New York or Harris (Houston.) On the other end, counties in Southern California is not grouped into a community at all but are identified as individually distinct monadic regions because counties in Southern California are on average have much higher rates of same-county commuting. In Northern California, however, one community shows cities like Sacramento, Stockton,
are all grouped with the Greater Bay Area (figure 3). The clusters are heavily overlapping; similarly, some counties in the Bay Area are also part of communities associated with the Sacramento metropolitan area.

New York City is linked to upstate New York in some communities, Philadelphia is linked to Baltimore and Washington DC, and Baltimore/Washington DC are more linked to each other and to other urban centers along the eastern seaboard, stretching down to North Carolina. However, the Northeast and Midatlantic regions yield smaller communities on average compared to the South and Midwest. Because the commuting volumes in the Northeast are higher and counties are smaller, the counties in this region tend to cluster around each other in medium-sized communities of usually between 20-30 counties, which are larger than those in the West, but smaller compared to those in the South and Midwest.

Communities have some similarities to megaregions in the Midatlantic. New York, Philadelphia, Washington DC and Baltimore are all grouped in the same megaregion, like several communities (figure 2). However, megaregions also do not account for the interconnected nature of Baltimore and Washington DC to Eastern North Carolina. That the clusters from the CCME method allows for overlaps allows one to say that both Eastern North Carolina and New York are connected to Washington DC significantly, but without being connected to each other, allows different urban areas to be flexibility classified in different regions.

The wide range of communities’ sizes is due to the different area sizes of counties as well as the differences in population along different regions of the United States. A county in California or Colorado tends to be at least a few times larger than a county in a state like Georgia. Small communities are mostly found in the Western part of the Contiguous US, while the largest clusters are found in the South and Central US.

The imputed clusters are larger than MSAs except for in populous regions. CCME naturally finds larger communities when inter-county commuting is relatively high compared to total commuting and smaller communities when large population centers are included. Communities have large variation in size partly because they account for the power-law distribution of human populations. Compared to MSAs, which necessarily must be at most a few counties, (excepting some metro regions like the Atlanta MSA) and megaregions, which (excepting regions in the West like Southern California) necessarily must be large, community sizes are much more variable. That a single county can be compared to a massive region is a natural adjustment to the highly uneven dispersion of population in the US; this is a natural feature of results from an unsupervised method such as CCME. Los Angeles county is the most populous county in the US with over 10 million inhabitants and 4 million commuters, which more commuters than all counties combined in most states. LA County is identified as monadic and is within the same class of comparison with the 300-county communities stretching across several states in the South. Such a comparison is not tenable in considering only small or large regions with each other as constrained by the MSAs or megaregion delineations.

Communities found by our proposed method account for much more of the total cross-county commuting than MSAs. The most significant improvement that the CCME method has over MSA demarcation is in accounting for all the inter-county commuting in the Southeast, Northeast, and Midwest parts of the US. These regions are where communities and MSAs are most dissimilar according to the FRI and exhibit the largest relative discrepancies in the size, volume, and number of delineations (table 2). MSAs in these regions capture between 50-70 % compared to over 90 % from communities. Counties in these regions are most interconnected by commuting and their corresponding MSAs may be classifying counties in too fragmented a manner to effectively describe the highly interconnected nature of the regions (table 2). In the Central US and the South, where the clusters are massive, commuting patterns within counties are more spread out across counties.

The CCME method produces clusters with high variation in size. The most populous MSAs house similar counties as communities (see figure 5). However, the sizes of clusters can be as large as megaregions, although the clusters tend to cover different types of regions. Megaregions capture cities with large
overlapping commuting volumes that are closely located to each other, but large communities capture regions that are closely linked by commuting. The larger the community, the smaller the absolute strength of each member tends to be, but the higher the relative strength each linkage is. In the case of the South and Central US, counties tend to be small and rural and commuting behavior tend to be across counties, thereby more counties tend to be more interconnected in clusters. Such aggregations have not been depicted in existing official regional delineations may be a novel contribution of the method in this study.

Though detection of singletons may be antithetical to typical notions of community in network theory, they are very important in this particular application of spatial commuting networks with strong self-looping tendencies. The of monads is closely tied to the prevalence of self-looping behavior, and this study calls into attention the role of self loops in community detection of networks modelling human mobility with spatial constraints. As described in Clauset et al. (2009), most network data arising from nature adhere to power laws, and commuting flows, which are reflective of human populations, are no exception. This study shows that power laws in spatial settings are intrinsically linked to self-referential behavior. Many studies have described human populations adhering to heavy-tailed distributions such as the Zipf Law (Newman, 2005; Barabási and Albert, 1999), but this study is among the first to indicate how might regional delineations arise to adjust for such phenomena.

Strongly self-commuting counties that are identified as monads can also be classified as clusters are they are found using tests of similar hypotheses in evaluating how unlikely the size of the total commuting activity in a given geographical unit is compared to a scenario where commuting activity was generated at random subject to constraints arising from the weighted configuration model. The null hypothesis for monadicity parallels the null hypothesis for each community connectivity in evaluating whether a test node is significantly more strongly connected to other nodes, or itself, than expected.

Monads tend to be parts of communities and tend to have multiple community memberships: nodes with high strengths extend influence both within county and between county. This may be explained by viewing high volumes of within-county commuting as a driving force for outward expansion, first becoming nodal clusters, then to tightly interconnected communities. One may conceive of a sequential toy model that offers a compromise between the concentric and lateral models. Population concentrates at first in a core area, then economic activity spreads out nodally. However, the once-peripheral areas begin to commute amongst themselves after a certain point (Plane, 1981). Identifying certain nodes as simultaneously monads and members in a cluster gives more insight to the type of effect that the self-looping behavior of single node exerts on connectivity patterns around the node.

6 Limitations and Further Research

Several limitations accompany applying CCME to demarcate regions using commuting networks. Though almost entirely unsupervised, there are several tuning parameters that must be specified, such as the threshold of commuting volume cutoff, the overlap parameter, and the p-value. However, sensitivity analysis (supplementary materials) over a range of p-values and overlap thresholds demonstrates that results do not change much when tuning parameter are tweaked. The post processing step of determining the nodal communities does not capture the extent of monocentricity that arise from areas that we expect to exhibit these properties such as Multnomah County (Portland) and Hennepin County (Minneapolis).

Several fruitful directions of further research may emerge from this study of commuting regions. This study documents the influence of self loops in spatial networks describing complex patterns arising from the collective behavior of many individuals. As network data in these realms become more and more available, we expect emergence of more methods accounting for self-looping behavior in networks to arise. Another extension will be to extend the methodology proposed in this study to directed networks as to model the orientation of commuter flows. A further extension of the method described in this study is to use a temporal model to measure change in communities across time. One such application is to look at several time points and to track the evolution of clusters imputed by CCME. Another avenue
is to investigate the characteristics of power-law distribution of populations that are embedded within the commuting networks and how distributional characteristics of power laws may play in community detection.

Much of the three types of communities hints at a relationship between monads, nodal clusters, and communities. Future studies may explore the dynamics of a core-periphery evolutionary model in which initial cores (monads) branch out into nodal clusters, which become further interconnected into communities, across time. Further research may also necessitate a more theoretically justified way of quantifying nodal clusters.

7 Conclusions

Traditional delineations of geographic regions have relied on agglomerations of smaller geographies, historical and political boundaries, separating edges and central foci. The boundary characterization is important not only for scientific purposes of tracking and tracing historical evolution of urban systems but also for administrative purposes of allocating infrastructure investments and formulating economic development strategies. For example, boundaries of metropolitan areas in the United States are artifacts of delineation definitions, yet are central to tracking demographic and economic changes, funding allocations, determination of fair market rents, housing subsidies that depend on area median income and a host of other federal and state programs, even when the agencies caution their use for non-statistical purposes. These delineations are central but invisible to the lives of many. In this paper, we provide a robust method for accounting for multiple regions that a place can belong to.

The main methodological contribution of this paper is in augmenting CCME to account for self loops. This paper also is one of the first forays of applied network analysis that makes much use of self-looping characteristics of spatial networks. The application of the model suggests that regional clusters in commuting networks are far larger than other existing delineations. It also suggests that accounting for within-county commuting patterns dramatically changes our understanding of the metropolitan structure of the United States. Furthermore, allowing for regions to overlap allows us to create institutional structures and policies that are tailored not only to singular geographical entities, but to multitudinous, interacting identities that space and place assume.
Acknowledgements
The authors thank John Palowitch and Eric Friedlander for helpful advice.

Conflict of interest
The authors declare no conflict of interest.

References
Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. Science 286(5439), 509–512.
Bender, E. and A. Canfield (1978). The asymptotic number of labeled graphs with given degree sequences. Journal of Combinatorial Theory, Series A 24(3), 296–307.
Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57(1), 289–300.
Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang (2006). Complex networks: Structure and dynamics. Physics reports 424(4-5), 175–308.
Bollobas, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. European Journal of Combinatorics 1(2), 311–316.
Chakraborty, T., A. Dalmia, A. Mukherjee, and N. Ganguly (2017, August). Metrics for community analysis: A survey. ACM Comput. Surv. 50(4), 54:1–54:37.
Clauset, A., M. E J Newman, and C. Moore (2005, 01). Finding community structure in very large networks. Physical review. E, Statistical, nonlinear, and soft matter physics 70, 066111.
Clauset, A., C. R. Shalizi, and M. E. Newman (2009). Power-law distributions in empirical data. SIAM Review 51(4), 661–703.
Durrett, R. (2007). Random graph dynamics, Volume 200. Cambridge university press Cambridge.
Ewens, W. J. (2003). On estimating p values by the monte carlo method. American Journal of Human Genetics 72(2), 496–498.
Fosdick, B. K., D. B. Larremore, J. Nishimura, and J. Ugander (2018). Configuring random graph models with fixed degree sequences. SIAM Review 60, 315–355.
Fowler, S., C., L. Jensen, and D. Rhubart (2018). Assessing U.S. Labor Market Delineations for Containment, Economic Core, and Wage Correlation.
Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences 99(12), 7821–7826.
Hagler, Y. (2009). Defining US megaregions. Technical report, Regional Planning Association, New York.
Han, Y. and S. J. Goetz. Overlapping labour market areas based on link communities. Papers in Regional Science 0(0).
Holland, P. W., K. B. Laskey, and S. Leinhardt (1983, June). Stochastic blockmodels: First steps. Social Networks 5(2), 109–137.
Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, 241–272.

Jones, M. and A. Paasi (2013, January). Regional World(s): Advancing the Geography of Regions. *Regional Studies* 47(1), 1–5.

Kim, K., Y. Chun, and H. Kim (2017). p-Functional Clusters Location Problem for Detecting Spatial Clusters with Covering Approach. *Geographical Analysis* 49(1), 101–121.

Lancichinetti, A., F. Radicchi, J. J. Ramasco, and S. Fortunato (2011, 04). Finding statistically significant communities in networks. *PLOS ONE* 6, 1–18.

Manley, E. (2014, January). Identifying functional urban regions within traffic flow. *Regional Studies, Regional Science* 1(1), 40–42.

Nelson, G. D. and A. Rae (2016, November). An Economic Geography of the United States: From Commutes to Megaregions. *PLOS ONE* 11(11), e0166083.

Newman, M. (2005). Power laws, pareto distributions and zipf’s law. *Contemporary Physics* 46(5), 323–351.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review* 45(2), 167–256.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association* 96(455), 1077–1087.

Nystuen, J. D. and M. F. Dacey (1961, December). A graph theory interpretation of nodal regions. *Papers of the Regional Science Association* 7(1), 29–42.

{Office of Management and Budget} (2010, June). 2010 Standards for Delineating Metropolitan and Micropolitan Statistical Areas: Notice. Federal Register 123, Washington, D. C.

Opsahl, T. and P. Panzarasa (2009, May). Clustering in weighted networks. *Social Networks* 31(2), 155–163.

Paasi, A. (2013, September). Regional Planning and the Mobilization of ‘Regional Identity’: From Bounded Spaces to Relational Complexity. *Regional Studies* 47(8), 1206–1219.

Palowitch, J., S. Bhamidi, and A. B. Nobel (2018). The Continuous Configuration Model: A Null for Community Detection on Weighted Networks. *Journal of Machine Learning Research* 18, 1–48.

Peixoto, T. P. (2018, Jan). Nonparametric weighted stochastic block models. *Phys. Rev. E* 97, 012306.

Pike, A. (Ed.) (2013, September). ‘Whither regional studies?’ New York, NY: Routledge.

Plane, D. A. (1981). The geography of urban commuting fields: some empirical evidence from New England. *The Professional Geographer* 33(2), 182–188. bibtex[publisher=Taylor & Francis].

Tonev, P., Z. Dvořák, P. Šašínka, J. Kunc, M. Chaloupková, and Z. Šilhan (2017, March). Different approaches to defining metropolitan areas (Case study: cities of Brno and Ostrava, Czech Republic). *Geographia Technica* 12(1), 108–120.

Tong, D. and D. A. Plane (2014). A New Spatial Optimization Perspective on the Delineation of Metropolitan and Micropolitan Statistical Areas. *Geographical Analysis* 46(3), 230–249.

Van Der Hofstad, R. (2009). Random graphs and complex networks. *Available on http://www.win.tue.nl/rhofstad/NotesRGCN. pdf* 11.
Wilson, J. D., S. Wang, P. J. Mucha, S. Bhamidi, and A. B. Nobel (2014). A testing based extraction algorithm for identifying significant communities in networks. *Annals of Applied Statistics* 8(1), 1853–1891.

Yan, X., C. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborova, P. Zhang, and Y. Zhu. (2014). Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment* 5, 05–07.

Zhao, Y., E. Levina, and J. Zhu (2012, August). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* 40(4), 2266–2292.

## Appendices

### A Probability computations for parameters

#### A.1 Variance of $\xi_{uu}$: $\kappa_{SL}$

Note that

$$\text{Var}(\xi_{uu} q_u) = \text{Var}(\xi_{uu}) \text{Var}(q_u) + \text{Var}(\xi_{uu}) \mathbb{E}[\mathbb{E}[q_u]^2] + (\mathbb{E}[\xi_{uu}])^2 \text{Var}(q_u)$$

$$= \kappa_{SL} (\sigma_p^2 + p^2) + \sigma_p^2$$

as the two variables are assumed to be independent. The sample variance of $W_{uu}$ may be decomposed in the following way:

$$\frac{1}{n} \sum_{u \in [n]} \text{Var}(W_{uu}) = \frac{1}{n} \sum_{u \in [n]} s_u^2 \text{Var}(\xi_{uu} q_u)$$

$$= \frac{1}{n} \sum_{u \in [n]} s_u^2 \left( \kappa_{SL} (\sigma_p^2 + p^2) + \sigma_p^2 \right)$$

$$= \frac{1}{n} \sum_{u \in [n]} s_u^2 \kappa_{SL} (\sigma_p^2 + p^2) + \frac{1}{n} \sum_{u \in [n]} s_u^2 \sigma_p^2$$

We derive another calculation of the sample standard deviation of self looping weights using a method of moments estimator.

$$\frac{1}{n} \sum_{u \in [n]} \text{Var}(W_{uu}) \approx \frac{1}{n} \sum_{u \in [n]} (W_{uu} - p s_u)^2$$

From the above two equations, we derive the following approximation

$$\frac{1}{n} \sum_{u \in [n]} (W_{uu} - p s_u)^2 \approx \kappa_{SL} \frac{1}{n} \sum_{u \in [n]} s_u^2 \left( \sigma_p^2 + p^2 \right) + \frac{1}{n} \sum_{u \in [n]} s_u^2 \sigma_p^2$$

$$\implies \kappa_{SL} \sum_{u \in [n]} s_u^2 \left( \sigma_p^2 + p^2 \right) \approx \sum_{u \in [n]} (W_{uu} - p s_u)^2 - \sum_{u \in [n]} s_u^2 \sigma_p^2$$

---

21
Rearranging the above equation and replacing unknown parameters by their estimates, we derive the estimate for $\hat{\kappa}_{SL}$

$$\hat{\kappa}_{SL} = \frac{\sum_{u\in[n]}(W_{uv} - ps_u)^2 - \sum_{u\in[n]}s_u^2\sigma_p^2}{\sum_{u\in[n]}s_u^2(\hat{\sigma}_p^2 + \hat{p}^2)}$$

A.2 Variance of $W_{uv}$

The properties of the weighted configuration model stipulates that the expectation of an edge weight given that there exist an edge, by equation 4, is:

$$E[W_{uv} | 1(A_{uv})] = (1 - p)q_{uv}$$

$q_{uv}$ in the above equation is defined in the main body of the paper in equation 5. We calculate the variance of $W_{uv}$ by decomposing it into two terms $I$ and $II$ using the following identity

$$\text{Var}(W_{uv}) = E[\text{Var}(W_{uv} | A_{uv})] + \text{Var}(E[W_{uv} | A_{uv}])$$

$$= I + II \quad (12)$$

To calculate $I$, we note that,

$$\text{Var}(W_{uv} | A_{uv}) = q_{uv}^2 \text{Var}((1 - \theta_u)\kappa_{uv} | A_{uv})$$

Therefore, calculating $I$ first necessitates calculating the conditional variance of $(1 - \theta_u)\kappa_{uv}$, given that there exists an edge, under an expectation. As shorthand, we define the operation $\text{Var}_A(\cdot) := \text{Var}(\cdot | A_{uv})$ and $E_A(\cdot) := E[A | A_{uv}]$.

$$\frac{1}{q_{uv}^2}E[\text{Var}(W_{uv} | A_{uv})] = E[\text{Var}((1 - \theta_u)\kappa_{uv} | A_{uv})]$$

$$= E[\text{Var}_A(1 - \theta_u)\kappa_{uv}]$$

$$+ \text{Var}_A(1 - \theta_u)E_A[\kappa_{uv}]^2 + \text{Var}_A(1 - \kappa_{uv})E_A[1 - \theta_u]^2$$

$$= E[\text{Var}(\kappa_{SL}) + \text{Var}(\theta_u) + \kappa_{SL}(1 - p)^2]$$

$$= E[(\sigma_p^2\kappa_{SL} + \sigma_p^2 + \kappa_{SL}(1 - p)^2)]$$

$$= E[(\sigma_p^2 + (1 - p)^2)\kappa_{SL} + \sigma_p^2]$$

$$= (1 - p)^2q_{uv}^2 \kappa_{SL} + \sigma_p^2) \quad (13)$$

Since relation $13$ holds for all $A_{uv}$, it becomes apparent that

$$\text{Var}(W_{uv} | A_{uv}) = q_{uv}^2((\sigma_p^2 + (1 - p)^2)\kappa_{SL} + \sigma_p^2) \quad (14)$$

Using equation $13$ the calculation of $I$ becomes straightforward:

$$I = E[\text{Var}(W_{uv} | A_{uv})]$$

$$= \frac{1}{q_{uv}^2}[\text{Var}_A(1 - \theta_u)\kappa_{uv}]$$

$$+ \text{Var}_A(1 - \theta_u)E_A[\kappa_{uv}]^2 + \text{Var}_A(1 - \kappa_{uv})E_A[1 - \theta_u]^2$$

$$= \frac{1}{q_{uv}^2}(\sigma_p^2 + (1 - p)^2)\kappa_{SL} + \sigma_p^2)$$

$$= q_{uv}^2(\sigma_p^2 + (1 - p)^2)\kappa_{SL} + \sigma_p^2) \cdot P(A_{uv}) \quad (15)$$

The calculation of $II$, similarly, is as follows:

$$II = \text{Var}(E[W_{uv} | A_{uv}])$$

$$= \text{Var}(q_{uv}(1 - p)1(A_{uv}))$$

$$= (1 - p)^2q_{uv}^2\text{Var}(1(A_{uv}))$$

$$= (1 - p)^2q_{uv}^2P(A_{uv})(1 - P(A_{uv})) \quad (16)$$
Putting the two equations 15 and 16 together, we are able to solve for the variance of \( W_{uv} \) from equation 12 and substituting the expression for \( \mathbb{P}(A_{uv}) \) from equation 2.

\[
\text{Var}(W_{uv}) = \mathbb{E}[\text{Var}(W_{uv}|A_{uv})] + \text{Var}(\mathbb{E}[W_{uv}|A_{uv}])
\]

\[
= q_{uv}^2 \cdot (\sigma_p^2 + (1 - p)^2) \cdot \kappa_{nSL} + \sigma_p^2 \cdot \mathbb{P}(A_{uv}) + q_{uv}^2 \cdot \mathbb{P}(A_{uv})(1 - \mathbb{P}(A_{uv}))
\]

\[
= q_{uv}^2 \mathbb{P}(A_{uv})((\sigma_p^2 + (1 - p)^2) \cdot \kappa_{nSL} + \sigma_p^2 + 1 - \frac{d_u d_v}{d_T})
\]

**A.3 Variance of \( \xi_{uv} \): \( \kappa_{nSL} \)**

Now we construct a similar method of moments estimator for \( \kappa_{nSL} \) as was done for \( \kappa_{SL} \). However, we also make use of the expression for the conditional variance of \( W_{uv} \) given the existence of an edge in equation 14.

\[
\sum_{u \in [n]} \sum_{v \neq u} \mathbb{E}[(W_{uv} - \mathbb{E}[W_{uv}])^2|A_{uv}] \approx \sum_{u \in [n]} \sum_{v \neq u} \text{Var}(W_{uv}|A_{uv})
\]

\[
= \sum_{u \in [n]} \sum_{v \neq u} ((\sigma_p^2 + (1 - p)^2) \kappa_{nSL} + \sigma_p^2 q_{uv}^2)
\]

\[
= (\sigma_p^2 + (1 - p)^2) \kappa_{nSL} \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2 + \sigma_p^2 \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2
\]

After solving for \( \kappa_{nSL} \) in the above equation and substituting unknown variables with their estimates, then changing the approximation to an equation, we derive the estimate \( \hat{\kappa}_{nSL} \) for \( \kappa_{nSL} \), thus obtaining the estimate as given in equation 6.

\[
\hat{\kappa}_{nSL} = \frac{\sum_{u \in [n]} \sum_{v \neq u} (W_{uv} - (1 - p)q_{uv})^2 - \sigma_p^2 \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2}{(\sigma_p^2 + (1 - p)^2) \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2}
\]

**A.4 Central Limit Theorem for \( S(u, B, \mathcal{G}) \) in set \( B \)**

In this section we detail the calculation of the expectation and variance of the relative strength \( S(u, B, \mathcal{G}) \) of a given node-set \( B \) used in iterative testing, described in the body of the text in Section 2.4.2.

\[
S(u, B, \mathcal{G}) = \sum_{v \neq u, v \in B} W_{uv}
\]

\[
= \sum_{v \neq u, v \in B} (1 - \vartheta_u) q_{uv} \xi_{uv}
\]

\[
= \sum_{v \neq u, v \in B} (1 - \vartheta_u) \frac{\vartheta_u \kappa_{nSL}}{d_T} \xi_{uv}
\]
Taking the expectation of each $\xi_{uv}$ gives the following expression for the strength of the node set

$$E[S(u, B, G)] = (1 - p) \sum_{v \neq u, v \in B} d_u d_v \frac{s_u s_v}{d_T}$$

$$= (1 - p) \sum_{v \neq u, v \in B} s_u s_v$$

$$= s_u \left( (1 - p) \sum_{v \neq u, v \in B} \frac{s_v}{s_T} \right)$$

We have found, in section A.2, that the variance of a given $W_{uv}$ is expressed as:

$$\text{Var}(W_{uv}) = r_{uv} \left( (\sigma_p^2 + (1 - p)^2) \kappa_{nSL} + \sigma_p^2 + 1 - \frac{d_u d_v}{d_T} \right)$$

Adding the variance terms together in set $B$ yields:

$$\text{Var}(S(u, B, G)) = \sum_{u \neq v, u \in B} \text{Var}(W_{uv})$$

$$= \sum_{u \in B} r_{uv} \left( (\sigma_p^2 + (1 - p)^2) \kappa_{nSL} + \sigma_p^2 + 1 - \frac{d_u d_v}{d_T} \right)$$

Then given that $B$ is ‘typical’ and that $d_u$ and $B$ are sufficiently large, that $S(u, B, G)$ is approximately normal. For $\mu(u, B) = E[S(u, B, G)]$ and $\sigma(u, B)^2 = \text{Var}(S(u, B, G))$

$$\frac{S(u, B, G) - \mu(u, B)}{\sigma(u, B)} \implies \mathcal{N}(0, 1)$$

Hence in each step of iterative testing in the update step of CCME, the normal p-value is used to iteratively reject insignificant nodes in a candidate community. Assumptions for and proofs of this Central Limit Theorem can be found in [Palowitch et al., 2018].

B Supplementary Figures
Figure B.6: Figure B.6: Histogram of comparison of $\varrho_u$ with simulated data generated from Beta($\hat{a}$, $\hat{b}$). In practice, $\hat{a}$ and $\hat{b}$ are respectively 3.86 and 4.50 for the commuting dataset in this study.

Figure B.7: Figure B.7: Histogram of weight clustering coefficients as described in equation [11].