Community Detection in Spatial Networks: An Application to Gerrymandering

Alex Book[1,*] and Cole Sturza[1,†]

[1]*Pupil of the ever-wise Dan Larremore*

Gerrymandering is a tactic that has been used throughout modern history to enforce unfair district boundaries in favor of a political party or affiliation. A method to objectively (or as close to objectively as possible) decide district boundaries has long been sought after, as such a discovery would bring parity to many majoritarian political systems, namely in the United States (both federally and locally). Solutions such as the introduction or popularization of entirely proportional voting systems have been suggested, but we seek a solution that could be implemented to help the current political system seen in the US. Modularity calculations are commonplace in network analysis, but it becomes a bit trickier than usual when dealing with spatial networks, as the connections between nodes are limited by physical constraints. Through exploring techniques to effectively modularize given areas and communities, we propose a possible fairer, more objectively-centered method to draw district boundaries, and remove the unfair control that is placed in the hands of those that stand to benefit from the manipulation of such systems.

I. INTRODUCTION AND BACKGROUND

Spatial networks are a special kind of network in which physical constraints impede the formation of connections between areas that exist far away from each other. Therefore, typical modularity calculations don't necessarily work particularly well in predicting/finding the community structure of a given population.

Spatial networks can be found at nearly every turn in everyday life. Mailing letters or packages to friends or family that live far away is more difficult and time-intensive than dropping an item in a neighbor's mailbox. Engaging in a relationship, romantic or otherwise, with someone over a long distance is typically more difficult (and less likely) than doing so with a person who lives nearby. Simply put, physical distance complicates things. To many, this idea is obvious when in comes to political districting, yet the hacky way in which boundaries are drawn continues to be an incredibly relevant issue.

Gerrymandering is the process of constructing districts in such a way to give your political party an advantage in elections. There are two principle tactics in Gerrymandering, Packing and Cracking. Packing is when you create a single district that concentrates a large number of voters of a single party in order to reduce that party's voting power in other districts. Cracking is when you spread out the voting power of a party across many districts in order to dilute their overall voting power in certain districts. The overall goal is to pack opposition voters into some districts, effectively guaranteeing them some small number of government seats, while cracking across other districts in order to raise the chances of one's own party gaining a high number of seats.

In an equitable system, boundaries wouldn't be created by those with strictly partisan interests, but would instead be created in a way that is as close as possible to objective fairness. Fairness that would disallow tactics like Gerrymandering through an automated system that draws boundaries, which can possibly be reviewed if absolutely necessary. As has been discovered over time in countless other fields, humans are rarely the best at most tasks. Districting is no different, and if it can be made fairer by using computation, math, and a network-based approach, then such an approach should be seriously considered.

II. PROJECT AIMS AND/OR RESEARCH QUESTION

Our overall goal for this project is to find one or more objective, non-partisan method(s) to draw district boundary lines, in the pursuit of pushing against the rampant inequality and misrepresentation that is ever-present in US politics (as well in the political machines of other nations).

Specifically, we would like to find an effective approach to partition spatial networks, test those methods on real-life data sets, and brainstorm the details necessary for applying such methods to drawing district boundaries involving real-life communities.

We were largely inspired to pursue such ideas during the events of the politically tumultuous year that 2020 has been thus far. We find it absolutely imperative to bring impartiality to the table in decisions that could affect communities (and on a macro scale, states and countries) for years, or even decades, to come.

III. APPROACH/METHODS

There were two approaches we looked into when considering how to partition spatial networks. The first was looking into whether or not modularity could be maximized and used as an effective means to accurately partition a spatial network into sub-communities. The nice thing about modularity in this sense is that algorithms like greedy agglomerate would allow you to pick the number of merges. In the US, states are split up into a number of districts equal to the number of seats they have in the House of Representatives. So when using

* albo6624@colorado.edu
† cost5824@colorado.edu

something like the greedy agglomerate algorithm, we could simply tell it to stop merging after a certain number (once the proper number of districts has been reached). This would give us the partition with the correct number of districts and a maximized modularity. The second approach we looked into was if there are any other techniques for partitioning spatial networks that have shown some level of success. After various considerations and exploration, we decided to explore an algorithm called Infomap.

A. Modularity

Modularity is a measurement of assortativity and asks the question: "How much more often do attributes match across edges than expected at random?" In order to determine whether some attributes mix with a higher degree of assortativity, we need to write down a null model for mixing at random. This is generally expressed in the following form.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - P_{ij}\right] \delta\left(x_i, x_j\right) \tag{1}$$

where A is the adjacency matrix, m is total the number of edges in the network, x_i is the label of vertex i, and $\delta\left(x_i, x_j\right)$ is the Kronecker delta function, which equals 1 when its arguments are the same and 0 otherwise. P_{ij} is the expected weight of a link between nodes i and j over an ensemble of random networks with certain constraints. These constraints correspond to known information about how the network is organized [1]. Q in (1) is essentially computing

$$\begin{aligned} Q = &(\text{fraction of links within communities}) \\ &- (\text{expected fraction of such links}). \end{aligned} \tag{2}$$

The most popular choice of null model for modularity, proposed by Newman and Girvan (NG) [2] is

$$P_{ij}^{\mathrm{NG}} = k_i k_j / 2m, \quad \text{then } Q = Q_{\mathrm{NG}}. \tag{3}$$

Where k_i is the degree of vertex i. In this model randomized networks preserve the strength of each node [1]. The NG null model uses the basic structural information encoded in the adjacency matrix. This model assumes that the network is well mixed, and that any node can be connected to any other node. The model prioritises connectivity. This is appropriate when no additional information on the nodes is available, but not when additional constraints are known. In spatial networks, this assumption does not hold because there are inherent restrictions that prevent certain nodes from being connected to others. For example, if we had a network of neurons in the brain, it is not possible to connect a neuron from one side of the brain to one on the other side.

There have been several proposals to create null models for modularity that address this concern. They use techniques to model how the connections actually function in spatial networks. One such proposal is known as the gravity model. The gravity model is inspired by the 'gravity model' of human mobility [1, 3]. This model assumes that the interaction between two locations is proportional to their importance (e.g. population) and that it decreases with distance. In the standard gravity model, the interaction between locations i and j with populations N_i and N_j that are a distance d_{ij} apart is

$$P_{ij} = N_i^{\alpha} N_j^{\beta} f\left(d_{ij}\right). \tag{4}$$

Where the "deterrence function" $f(d)$ describes the effect space has on the interaction between two nodes. There are several common choices for the deterrence function. Some of the more popular ones are inverse proportionality to distance (i.e. $f\left(d_{ij}\right) = 1/d_{ij}$), inverse proportionality to squared distance (i.e. $f\left(d_{ij}\right) = 1/d_{ij}^2$), exponential decay (i.e. $f\left(d_{ij}\right) = e^{-d_{ij}}$), and more generally any function of the form $f\left(d_{ij}\right) = d_{ij}^{\kappa}$ [3]. Expert et al. propose another deterrence function of the form

$$f(d) = \frac{\sum_{\{i,j|d_{ij}=d\}} A_{ij}}{\sum_{\{i,j|d_{ij}=d\}} \left(N_i N_j\right)}. \tag{5}$$

This deterrence function is the weighted average of the probability $A_{ij}/\left(N_i N_j\right)$ for a link to exist at distance d. It is a popular choice to use bins instead of an exact distance d because of the variation of distances that could be in a dataset [1]. Combining (4) and (5) and setting $\alpha = \beta = 1$ we get the gravity model of the form

$$P_{ij}^{\mathrm{Grav}} = N_i N_j \frac{\sum_{\{i,j|d_{ij}=d\}} A_{ij}}{\sum_{\{i,j|d_{ij}=d\}} \left(N_i N_j\right)}. \tag{6}$$

In our analysis we chose to use this form with bin sizes of 5, 10, 20, and 100. Also, we experimented with the deterrence functions $f\left(d_{ij}\right) = 1/d_{ij}$, $f\left(d_{ij}\right) = 1/d_{ij}^2$ and $f\left(d_{ij}\right) = e^{-d_{ij}}$. We also used $P_{ij}^{\mathcal{NG}}$ in our analysis too. To maximize the modularity we used the "generalized Louvain" MATLAB code for community detection provided by [4]. This code allows for the user to define a quality function in terms of a generalized-modularity null model.

B. Infomap Algorithm

We also were interested in other techniques for partitioning spatial networks other than modularity. The second method we looked into was the Infomap algorithm. This algorithm has been proposed for partitioning spatial networks by Munoz-Mendez et al. In their paper, they use the algorithm on a network of a bicycle sharing system in London. In this paper we aim to recreate their results as well as determine

how modularity compares to this algorithm. The Infomap algorithm was first proposed by Rosvall & Bergstrom [5]. This algorithm acknowledges that the system structure drives the flow in the system, leading to system-wide interdependencies. The algorithm tries to minimize the cost function

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H\left(\mathcal{P}^i\right). \tag{7}$$

Where the first term is the entropy of the movement between modules, and the second is the entropy of movements within modules. $H(\mathcal{Q})$ is the entropy of the module names. $H\left(\mathcal{P}^i\right)$ is the entropy of the within-module movements, including the exit code for module i. $q_{\curvearrowright}$ is the probability that the random walk switches modules on any given step. The weight $p^i_{\circlearrowright}$ is the fraction of within-module movements that occur in module i, plus the probability of exiting module i [5]. Or in other words $p^i_{\circlearrowright}$ gives the proportion the walker spends in the respective module $\mathcal{P}^i$ [6]. $L(M)$ gives the average number of units per step that it takes to describe an infinite random walk on a network partitioned according to M [5].

This method is particularly interesting for partitioning spatial networks because it retains the information about the directions and the weights of the edges. Also, Infomap acknowledges the interdependence in networks inherently characterized by flows. Traditional modularity disregards valuable information like directed edges and weights. In the case of Gerrymandering, it may be interesting to create partitions using data from different types of flows in civilization. For example, weighting the edges as the number of people that travel from one precinct to another. This could give insight on how the various communities in a state interact. We could potentially use this type of information to build a partition that accurately represents a district.

C. Bicycle Sharing Network in London

The first data set we looked at was a network of a bicycle sharing system in London. Munoz-Mendez et al. used this data set along with the Infomap algorithm to partition the network into various communities [6]. Our aim is to recreate their work and also apply the aforementioned modularity equations. The data consists of 1,469,945 unique shared bicycle trips from June and July 2014. There are 750 bike stations which represent the nodes in the network. Edges are represented by the "flow" of trips from one station to another. Basically a weight that represents the number of trips from one station to another. The locations of the bike sharing stations can be seen in Figure 1.

For the modularity equations we used the degree of the bike station as the importance of a node. The distance between the nodes was measured as the euclidean distance between the longitude and latitude coordinates of the bike stations. We also scaled the distance between the nodes by a factor of a 1000 because of how close the bike stations are to each other.

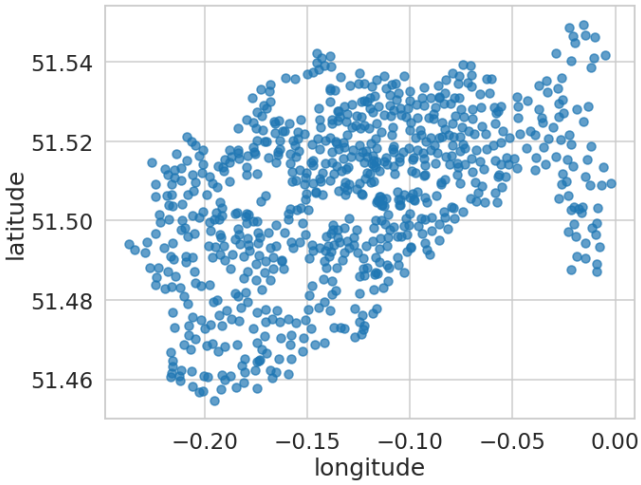

FIG. 1. **Location of bike sharing stations in London.** This is a plot of the physical locations of the bike sharing stations in London.

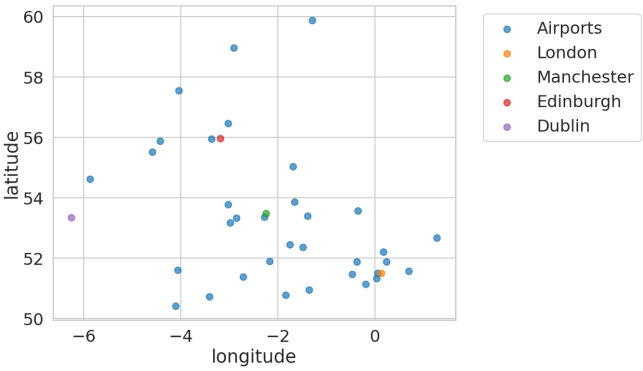

FIG. 2. **Location of airports in UK.** This is a plot of the physical locations of the airports in the UK, partitioned using Infomap.

Also, for the modularity we needed to make the network undirected. Therefore, we averaged together the edge i to j and j to i. This made the adjacency matrix symmetrical and preserved some information about the directed form of the network.

D. UK Airports

The second data set we looked at was a network of the airports in the UK from [7]. The data tracks flights from airport to airport from 1990 to 2003. There are 34 nodes representing the airports, and the edges and weights represent the number of passengers that flew from one airport to another. We used the same procedure as before with the bike sharing network. The locations of the airports can be seen in Figure 2.

Method	Number of Communities
bin size of 5	26
bin size of 10	27
bin size of 20	27
bin size of 100	26
$f(d_{ij}) = 1/d_{ij}$	660
$f(d_{ij}) = 1/d_{ij}^2$	659
$f(d_{ij}) = e^{-d_{ij}}$	27
NG model	110

FIG. 3. **Results of modularity partitioning on bike sharing data set.** The results of maximizing modularity on the bike sharing dataset. The first seven use the gravity model. The first 4 use (6) with different bin sizes.

IV. RESULTS/FINDINGS

The first thing we did was to look at how the modularity equations performed on the bike sharing data set. We ran the aforementioned variations of the gravity model and the NG model on the network and recorded the number of communities found and the maximized modularity. The algorithm used to maximize modularity was the "generalized Louvain" MATLAB code for community detection provided by [4]. The results can be seen in the table in Figure 3.

As one can see, the gravity models using different bin sizes performed about the same on this data set. They all found 26-27 communities. The gravity models using $f(d_{ij}) = 1/d_{ij}$ and $f(d_{ij}) = 1/d_{ij}^2$ were not able to partition the network in to less than 659 communities. This was rather surprising to us. We expected that they would do rather poorly as they are rather simplistic in terms of expressing how space effects the network, but not as poorly as they did. The exponential decay gravity model performed around the same as the gravity models using (6). Lastly, the NG model came in around the middle of the pack, with 110 communities. This was also surprising as we thought that this one would perform the worst.

A select number of the modularity partitions can be seen plotted spatially in Figure 4. As one can see in Figure 4 the communities are scattered all over London. We found that these null models did not work spatially for this data set. There is some semblance of community, particularly in the gravity model using $f(d_{ij}) = e^{-d_{ij}}$. Overall, we were rather surprised to find that none of the modularity models performed well spatially. Something we could do in the future is test these models on different data sets or possibly use a different algorithm to maximize the modularity. We also believe that these methods performed poorly because we removed some information about the directed nature of the network, as well as used the degree as the importance of a bike station. If we had some metric that could accurately represent the population of a bike station, we may have arrived at a different outcome.

We then moved on to looking into using the Infomap algorithm to partition the bike sharing network. We decided to also compare how the algorithm does against two other parti-

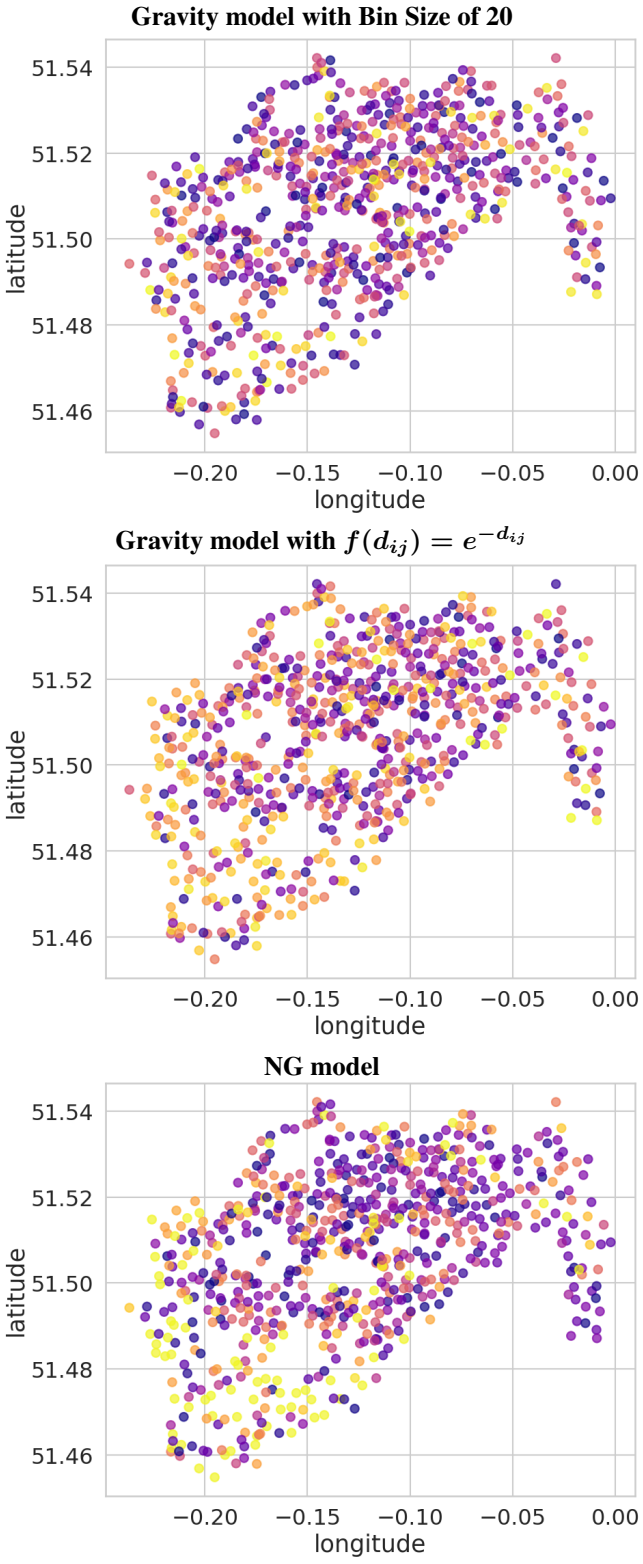

FIG. 4. **Results of modularity partitioning on bike sharing data set plotted spatially.** This is a plot of the physical locations of the bike station communities in London found by maximizing modularity.

tioning algorithms, namely, Walktrap and fast greedy. Walktrap partitions graphs based on random walks and was first introduced by Latapy & Pons. Fast greedy partitions a network based on the greedy optimization of modularity. The partitions created by these algorithms can be seen in Figure 5.

The Infomap algorithm does particularly well at partitioning this network into communities. It was able to partition the network spatially into 5 separate communities. The Walktrap algorithm also did somewhat well too. The fast greedy performed poorly, which was expected based on the previous experiment using the NG model. What is surprising, however, is that using fast greedy allowed it to decrease the number of communities to 3.

We then used the Infomap algorithm to attempt to find an effective partitioning of the UK Airport data set. As per the results, there is no discernible usefulness in partitioning the airports as there was with bike stations. This could be for a number of reasons, namely the lack of regularity when compared to the bike sharing network, in that there are many people that only ever take a certain flight once or twice in their lifetime, whereas many people bike the same route daily (whether it be for work, school, or any other functional or leisurely purpose).

As can be expected, drawing communities out of raw data sets isn't necessarily always going to prove useful. When one is considering locations that can be represented by "neighborhoods", such as bike stations, or neurons in various sections of the brain, or (hopefully) voting districts, useful results can be found. However, data tracking the flow between airports isn't quite useful in the same sense, as airports tend to be much more independent of one another, making it more difficult to group them together in a meaningful way.

V. EXTENSIONS TO GERRYMANDERING

As far as extending this work to fix the problem of gerrymandering in the US, we have a couple of thoughts. In order to effectively create partitions of states, the methods used need to be refined and proven to work before they can be applied. Infomap has shown that it can be effective in partitioning the bike sharing network into separate communities spatially, but the modularity models failed to do so in our testing. So in order for these methods to be applied against real world partitions, we would need to make sure the data we use fits the methods used.

Expert et al. [1] were able to use the modularity equation in (6) to partition a network of Belgium spatially. The data consisted of various towns and cities and their phone calls to other towns/cities. The edges are represented by the frequency of calls and the importance of the nodes by the population of the town/city. They were able to find communities based on language, with Belgium having two large Flemish and French populations. After reading this we have high hopes that being able to partition networks into communities based on the languages they speak could be a useful method

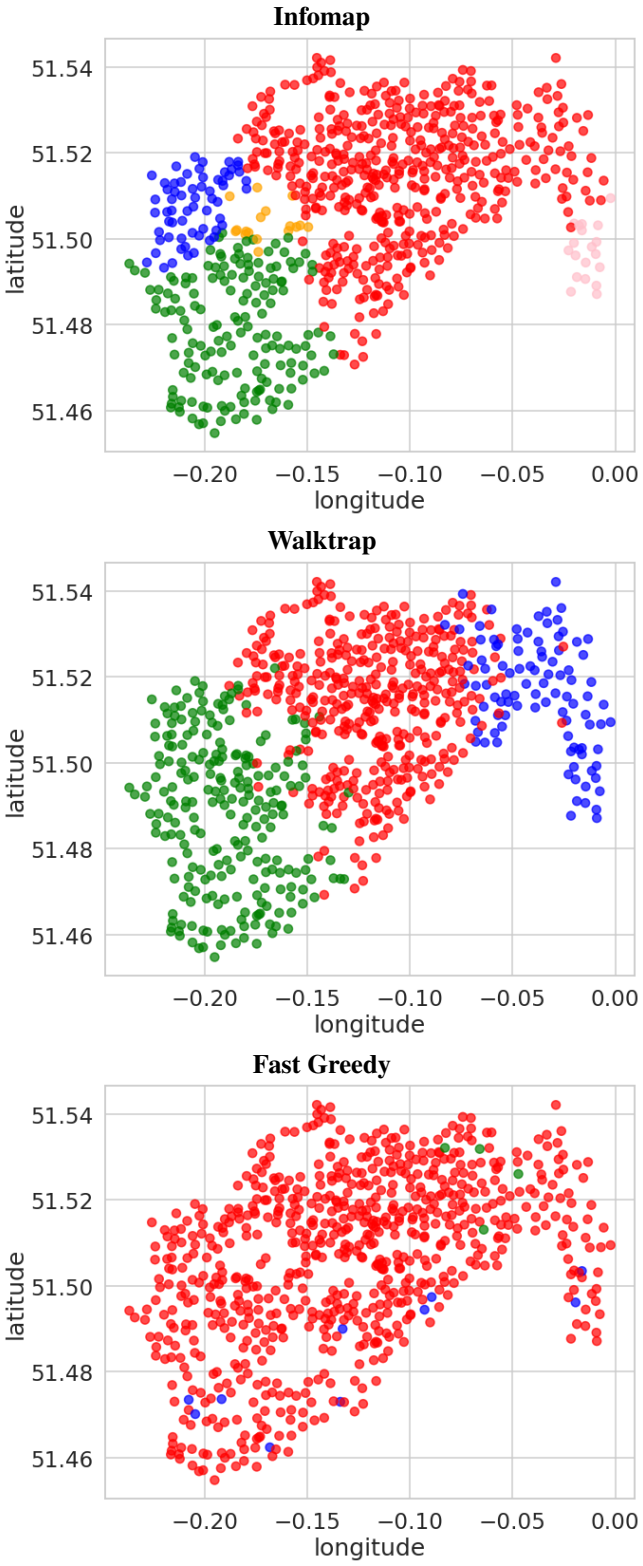

FIG. 5. **Results of Infomap partitioning on bike sharing data set plotted spatially.** This is a plot of the physical locations of the bike station communities in London found by using the Infomap algorithm, the Walktrap algorithm, and the fast greedy algorithm.

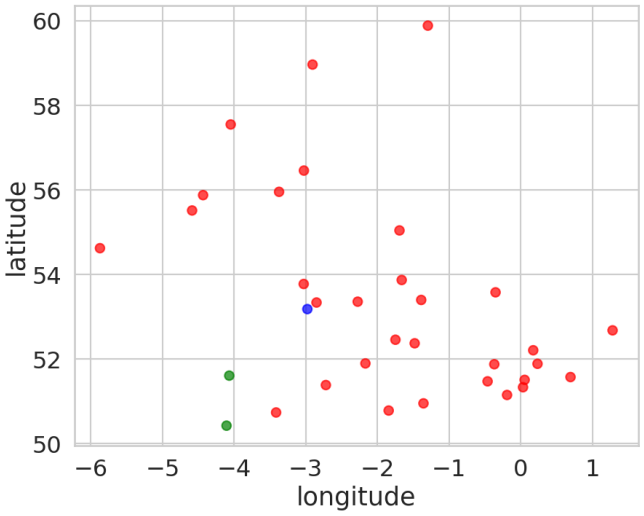

FIG. 6. **Results of Infomap partitioning on UK airports data set plotted spatially.** This is a plot of the physical locations of the airport communities in the UK found by using the Infomap algorithm.

in comparing against districts created by states.

As far as comparing these techniques to the partitions created by states, we believe that the best way to do this would be to create an ensemble of partitions using various data sets and methods discussed in this paper. Then use a metric like normalized mutual information (NMI) or variation of information (NVI) to compare a state's proposed plan against the ensemble. This would give a good estimate on how the state's partition compares with how people actually interact within their communities. Districts should represent strong communities of people, not the geographically "stringy" boundaries that are drawn purely for political advantage.

The networks that form a state would ideally be represented by precincts and how they interact with one another. So the nodes of the network would be the precincts themselves and the edges would then depend on the data being used. So in the case of the Belgium data set, it would be the frequency of calls from one precinct to another, or we could possibly use a data set on the mobility of people from one precinct to another (which precincts do people travel to frequently from their own).

VI. CONCLUSIONS/DISCUSSION

As expected, finding communities in data sets isn't always useful, nor is it necessary. However, based on our results with the bike sharing data, as well as the discussions posed in the previous section, we believe that finding communities that are effective as voting districts is possible. However, there are a couple barriers beyond those that we previously mentioned that would have to be surmounted before such a partitioning process could see success.

The first is the question of data. Data privacy is an important issue, and rightly so. The data necessary for such a titanic undertaking could be seen as an invasion of privacy by many. Whether or not the general public would see the objective worth that fairer voting districts would bring (at least enough worth to put up with movement-tracking data being used) is a toss-up.

The second is the question of general acceptance. There is no shortage of people, especially in the US, that are resistant to change. This expanse of people becomes even larger when technology is brought into the conversation. Despite our best wishes, it's hard to believe that most people would entrust something as important as voting districts entirely to algorithms that all but the most aptly educated would have little or no idea of how they work. Many people would likely be more on board once there exist tangible results, but getting to the point where such results can be seen may end up being quite a challenge.

While there doesn't seem to be a perfect solution, we have thought about various compromises that could be made in order to bring a higher level of parity than currently exists in the drawing of voting districts. As explained in the previous section, politicians would still be able to propose voting districts, but they would be checked against the districts created algorithmically, and must be within a certain threshold of "fairness" (under the assumption that the algorithms used will create proposals that are as close to objective fairness as possible) in order to be signed into official existence/law.

It's plain to see that gerrymandering doesn't have an easy solution. If it did, the problem likely wouldn't exist in the present at all. However, solutions that can, at the very least, help to remedy the issue may be right around the corner (or may already exist, unbeknownst to the general populous). Regardless, coming as close to objective fairness in politics is a goal that is worth striving for, however difficult it may be.

VII. DATA AND CODE

Data for this project can be found at UK flights and London bike sharing. Code can be found at this GitHub repo.

[1] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Proceedings of the National Academy of Sciences **108**, 7663 (2011).

[2] M. E. Newman and M. Girvan, Physical review E **69**, 026113 (2004).
[3] M. Sarzynska, E. A. Leicht, G. Chowell, and M. A. Porter, Journal of Complex Networks **4**, 363 (2016).
[4] I. S. Jutla, L. G. Jeub, and P. J. Mucha, URL http://netwiki. amath. unc. edu/GenLouvain (2011).
[5] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **105**, 1118 (2008).
[6] F. Munoz-Mendez, K. Han, K. Klemmer, and S. Jarvis, in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (2018) pp. 1015–1023.
[7] I. Morer, A. Cardillo, A. Díaz-Guilera, L. Prignano, and S. Lozano, Physical Review E **101**, 042301 (2020).
[8] M. Duchin, T. Needham, and T. Weighill, arXiv preprint arXiv:2007.02390 (2020).
[9] B. Gonçalves, R. Menezes, R. Sinatra, and V. Zlatic, *Complex Networks VIII: Proceedings of the 8th Conference on Complex Networks CompleNet 2017* (Springer, 2017).
[10] A. J. Comber, C. F. Brunsdon, and C. J. Farmer, International Journal of Applied Earth Observation and Geoinformation **18**, 274 (2012).
[11] M. Z. Austwick, O. O'Brien, E. Strano, and M. Viana, PloS one **8**, e74685 (2013).
[12] M. Coscia and R. Hausmann, PloS one **10**, e0145091 (2015).