Community Detection in Relation to the Spread of Epidemics

Candidate Number
562793

Oxford University
Mathematics Part C

Abstract

There have been many recent developments in the study of networks, familiar examples of which include the World Wide Web and social networks. The main focus of my dissertation is *community structure*; this refers to the study of densely connected groups within a network. More specifically, I employ two different algorithms to optimise a quality function known as *modularity* which describes how well a chosen partition divides the network. The two methods I use are a greedy algorithm proposed by Blondel *et al* and Newman's spectral optimisation algorithm. Using time series data that includes the weekly number of reported cases of dengue fever in Peru by province, I define a dynamic network using matrix representations, before using the community detection methods to investigate naturally occurring divisions in the network and their evolution in time. I find that there is a correlation between modularity and epidemic outbreak during the 2000 - 2001 epidemic. Finally, I discuss possible further explorations of the data that could give us further insight into the evolution of the community structure and its relevance to the spread of disease.

Acknowledgements

Thanks to Gerardo Chowell of the Chowell Group at Arizona State University and the Peruvian Ministry of Health for providing the data without which this project would not have been possible. All images and MATLAB code were either created for this project or used and adapted with permission. I gratefully acknowledge Jukka-Pekka Onella for allowing the use of the multislice network image. I attribute the maps of Peru to ErickAgain under the licensing agreement CC-BY-SA and the GNU Free Documentation License, and all community detection code to the Netwiki "software library" collaborators. Finally I would like to express my gratitude to my supervisor, Mason Porter, both for his guidance regarding this project and for his encouragement and advice throughout my years as an undergraduate student.

Contents

Chapter 1

The Biological Problem

Introduction

In this chapter I elucidate the background and motivation behind the project. In the first section, I outline the nature of the problem, before giving details of the data set. I then give examples of methodology that has been used in the past in relation to other epidemiological network problems and finally I introduce community detection, giving two examples from previous studies.

1.1 Dengue Fever in Peru

Dengue fever infects an estimated 100 million people each year, with approximately between 250,000 and 500,000 cases progressing to a more severe disease: dengue hemorrhagic fever or dengue shock syndrome.[4] It is mostly transmitted via the *Aedes aegypti* species of mosquito, though only pregnant adult female mosquitoes feed on human blood, and the time between infecting a human and the time at which symptoms are observed can be up to two months.

There are currently no known vaccines for dengue fever, and mosquito control is the main epidemic prevention method used.[10,12] It is thus especially important to explore the causes of transmission in order to develop further ways of controlling the spread of the disease. Dengue fever is a significant problem in Peru and study of the disease is of particular significance at the moment as, at the time of writing, Peru is amid a dengue outbreak that has already affected over 10,000 people.[15]

1.2 The Data

The data was collected by the Chowell group at Arizona State University in collaboration with the Peruvian Ministry of Health. It includes the number of reported cases of dengue fever and was collected on a weekly basis in 79 different provinces in Peru over a total time period of 15 years - from the beginning of 1994 to the end of 2008.

Using this data, we hope to use methods from network theory to shed light on the transmission of the disease over time, and find the factors that influence the spread of the disease in different parts of the country.

We also have data regarding the population sizes of the provinces for different age groups, which was collected on a yearly basis over the fifteen year period. One would expect to find a correlation between the population size of a province and the average size of an outbreak in that province, and we would eventually like to use this information in the investigation of the spread of the disease. I discuss possible ways to incorporate the population data in Chapter 5.

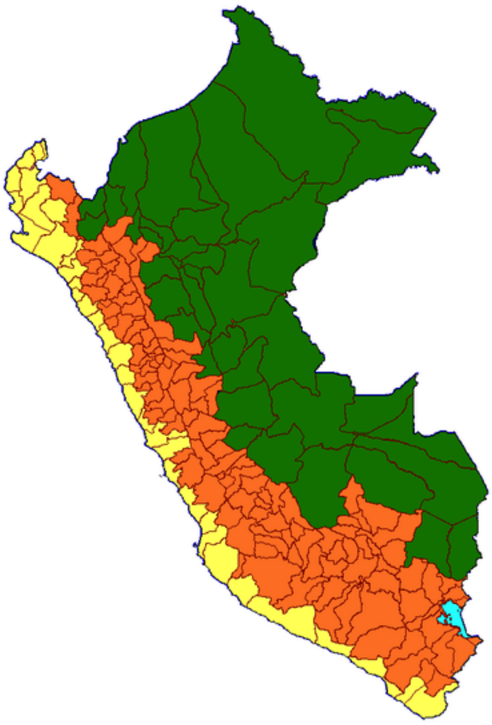


Figure 1.1: This geographic map of Peru illustrates the division of coastal, mountain and jungle regions shown in yellow, orange, and green respectively.

One important characteristic of the data that should be noted is that the provinces are divided into three different geographic regions: jungle, mountain and coast. Mosquitoes thrive in areas of stagnant water, so one would expect to find higher rates of infection in the coastal or jungle regions. The distribution of these regions is very distinctive, as can be seen in Figure 1.1, and we find that the transmission depends not only on the geographic type of regions, but also the relative location of the province within the country.

1.3 Epidemiological Networks

The study of networks is very applicable to epidemiology. If links between nodes of defined in terms of the distance between them, then finding the shortest path between nodes might help one predict the way in which a disease spreads. Calculating centralities (roughly speaking, the relative importance of a particular node in relation to its local or global neighbourhood) might help one identify the nodes to target with vaccinations.

A recent example of an application of network theory in epidemiology is in work by Meyers *et al.*[18] Percolation theory, which is a study of the connectivity of a network, is used to model the spread of *Mycoplasma pneumoniae*,[17] which is spread by droplet contact transmission and is the cause of the infectious disease mycoplasma pneumonia.[5] In their paper, a healthcare institution network is considered in which groups of patients, caregivers and wards are interconnected and statistical properties of the network structure are used to predict the scale of future outbreaks. By reorganising the structure of the network and analysing the predicted outcome in terms of the outbreak size, the study gives an indication of the implications of restricting interactions between those groups. This area of research is applicable to our data, although we need to start by finding the community structure before we can try these methods.

1.4 Why Community Detection?

Many of the epidemiological models used in recent studies are based on compartmental susceptible-infected-recovered SIR models,[19] but our aim is to employ 'community detection' techniques. The purpose of community detection algorithms is to uncover partitions between densely connected groups of nodes, perhaps characterised by properties specific to those groups, and is useful for recognising structural features of the network on different spatial scales.[8,23] Importantly, community detection methods do not assume *a priori* the number or size of the groups. Instead these methods search for naturally occurring communities, which can help to find structural patterns that were previously unknown.

The Zachary Karate Club[26] is frequently used as a pedagogical illustration of the application of community detection algorithms. Zachary was studying the social interactions of a karate club before it divided into two smaller clubs following a conflict between its members. He found that given only the details of the interactions between members of the former club, he would have been able to predict the allocation of members to the two subgroups.

Traud *et al.* used the same community detection method as the one used in this study.[25] In contrast, however, the network under consideration was a social network rather than epidemiological one. Using data from the online social networking site Facebook, they constructed networks based on the online activity of students at five American universities in order to model their offline relationships. Based upon the community structure information that they discovered, Traud *et al.* conclude that undergraduate students at Caltech tend to establish friendships dependent upon their "House affiliation", an arrangement much like the collegiate system that we are familiar with here in Oxford, whereas undergraduate students at the other four universities tended to build relationships within their year group.

These examples demonstrate the way in which results from community detection algorithms can help us understand the structure of a network. The aim of this study is to determine whether these techniques could give us any insightful information into the evolution of epidemics, by analysing the problem of dengue fever in Peru.

Chapter 2

Creating a Network

Introduction

This chapter will be primarily focussed on constructing the network in a comprehensible mathematical form, as an *adjacency matrix*. I start by giving a formal definition of a *graph* and I proceed by presenting some preliminary properties of networks. After careful consideration of characteristics of the data set, I construct a dynamic network which I represent as a sequence of time-dependent adjacency matrices.

2.1 Formal Definition

A network is often represented as a graph, which is made up of a finite set of discrete elements which we call nodes; pairs of nodes are connected by links, or edges. I will give some formal definitions from graph theory, using notation by Boccaletti *et al.*[2]

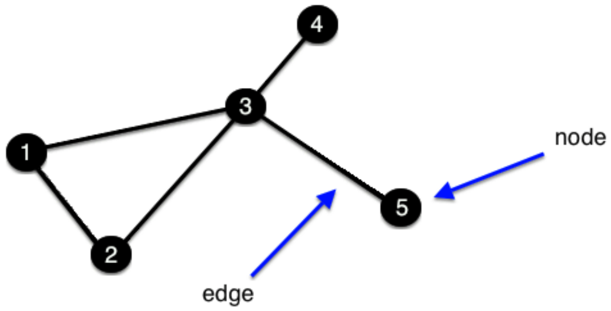


Figure 2.1: An example of a network.

Definition A *node* is an element of the set $\mathcal{N} \cong \{1, ..., n\}$ such that $\mathcal{N} \neq \emptyset$.

In this essay I construct the most intuitive network from the data by allocating each node to a Peruvian province. To aid calculations I number each of the nodes from 1 to 79 so that $\mathcal{N} = \{1, ..., 79\}$; the corresponding provinces are given in Appendix A.

Definition A *link*, or *edge*, is an element of the set $\mathcal{L} \equiv \{\langle i, j \rangle \mid i \in \mathcal{N}, j \in \mathcal{N}\}$.

The edges describe some relationship between nodes. Figure 2.1 is an illustration of a network with 5 nodes and 5 edges. No edge directions are specified, so the network is *undirected* (otherwise, the network is called *directed*). Also note that this network does not contain any *self-edges*, these are edges that connect nodes to themselves.

It is also possible to attach a *weight* to the edge, to indicate its strength. We usually say that the *degree* of a node is the sum of the weights of the edges attached to that node.

Definition A *graph*, or *network*, $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ consists of both a set $\mathcal{N}$ and a set $\mathcal{L}$.

2.2 Adjacency Matrix

An adjacency matrix is a convenient way to represent the network, and is easy to use in computations. There are many choices for constructing an adjacency matrix for time series. We typically denote the matrix by $\mathbf{A}$, where the entries A_{ij} are values corresponding to the edge(s) between nodes i and j.

In this study, I denote the numbers of reported cases of dengue fever in province i at time t by $b_i(t)$. Because the data is collected on a weekly basis, for simplicity I take t to be number of the week in which the data was collected, so that $t = 1$ is the first week of 1994 and $t = 780$ is the final week of 2008. I then define

$$\langle b_i(\tau) b_j(\tau) \rangle = \sum_{\tau=t}^{t+\Delta t} b_i(\tau) b_j(\tau).$$

One possible way to define the entries of the adjacency matrix is thus

$$A_{ij}(t) = \frac{\langle b_i(\tau) b_j(\tau) \rangle}{\sqrt{\langle b_i(\tau)^2 \rangle \langle b_j(\tau)^2 \rangle}} - \delta_{ij}, \tag{2.1}$$

where i, j represent two provinces in Peru and δ_{ij} is the Kronecker delta.

The resulting matrix is normalised (so that A_{ij} takes a value between 0 and 1) and is symmetric (which represents an undirected network). We would expect epidemic networks that model infectious diseases to be directed, and I discuss ways in which we can take this into account in Section 5.2.2. We can see that there is at most one edge between each pair of nodes, and that the Kronecker delta is used to remove self-edges as we would gain no insight by including them.

This appears to be a good choice for the adjacency matrix, as the resulting weighted edges, A_{ij}, denote time-dependent correlations between the nodes. This type of connection has been useful in previous studies, for example in a network constructed for the purpose of modelling the foreign exchange market.[7] I therefore use the definition in (2.1) to construct a network from the epidemic data.

2.3 Time Dependancy

Having defined the adjacency matrix as in (2.1), different values of t and Δt can be introduced to see how the disease evolves over the fifteen years. This should give a sequence of adjacency matrices, with each matrix representing the network at a different time aggregation, or time window. Figure 2.2 is an illustration of a dynamic network like the one that I have defined. The network contains the same nodes at each interval, but the edges connecting them can change over time.

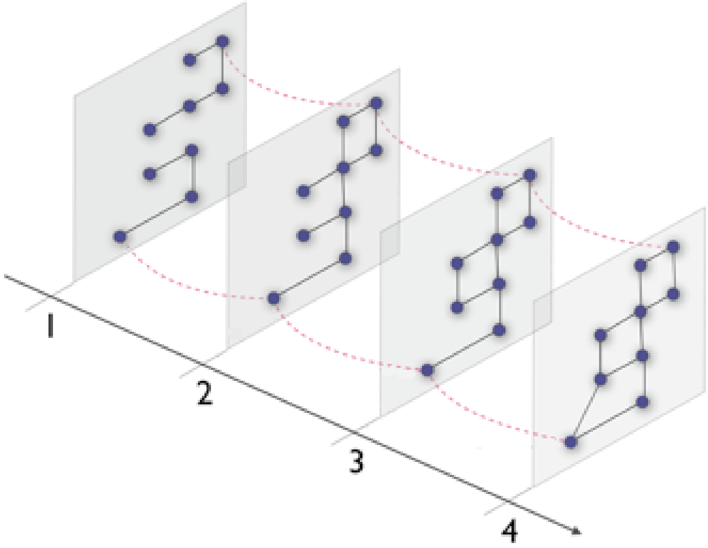


Figure 2.2: A schematic of a dynamic network at different intervals.

In this study, I consider two different choices for the time windows. To begin with, I take $t = 1$ and $\Delta t = 779$. This results in a network that is represented by only one adjacency matrix that is aggregated over the entire fifteen year period. I do this in order to easily compare the community detection algorithms discussed in Chapter 3.

For the second choice, I take $t = 1, 5, 9, ..., 729$ and $\Delta t = 52$ (in other words, the time windows are a year long and are displaces by approximately one month at a time). This seems to be the best compromise as we want to avoid not only unnecessarily noisy data but also over-smoothing the data.[7] A closer look at the epidemic data revealed that the outbreaks typically tend to occur over a time period that varies from between one and six months (see Figure 4.1). This suggests that these time windows are a reasonable choice, and this also makes our life easier computationally, as we will have only 182 adjacency matrices to work with (as opposed to, say, taking two week time windows that move along by one weeks at a time which would result in a total of 779 matrices).

Chapter 3

Community Detection

Introduction

In this chapter I introduce the quality function, *modularity*, which used for community detection. I then give accounts of two different modularity optimisation methods, which seek a partition of the network that will result in the greatest value of Q. The first method is a greedy algorithm called the Louvain method, developed by Blondel *et al.*[1] and the second is a spectral optimisation method given by Newman.[20] There are potential practical issues with the optimisation of Q, as given in a discussion by Good *et al.*[11] (see the discussion in section 3.4).

3.1 Modularity

Having constructed a network, I will now introduce a useful diagnostic used in community detection: the modularity Q. Modularity is a function of a partition of the network into one or more groups, or communities, and the value Q is often used to determine how well the partition divides the network.[8,23] I will use the formulation of modularity given by

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij})\, \delta(c_i, c_j), \tag{3.1}$$

where A_{ij} is the adjacency matrix as given in (2.1). Taking $P_{ij} = \frac{k_i k_j}{2m}$ as defined by Newman and Girvan,[21,22] we get

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{3.2}$$

where k_i denotes the weighted degree of node i and $\delta(c_i, c_j)$ is the Kronecker delta and c_i denotes the community to which node i belongs. Note that the modularity is normalised, as we have divided through by $2m$, where m is the sum of the weighted edges. Hence we see that Q will lie between -1 and $+1$, where a higher value of Q demonstrates a stronger division and a lower value of Q demonstrates a weaker division of the network into smaller groups. The reason for this is that the null model, $P_{ij} = \frac{k_i k_j}{2m}$, gives the expected edge weight between nodes i and j. We know that A_{ij} gives the actual edge weight, so summing over the difference when we allocate the nodes to different communities gives us an indication of how well the partition divides them.

We can thus consider community detection as a *modularity optimisation* problem, as we will want to use an algorithm that will find a partition of the network such that the partition gives us the best approximation to the maximum global modularity.

3.2 A Greedy Algorithm

There are two main stages in the Louvain method presented by Blondel *et al.*[1]

We begin by allocating each node to its own community (in our case, we would start with 79 communities with one node in each). At each iteration, we take a node from one community and add it to another community to find the new set of communities that gives the greatest increase in modularity. If ΔQ is positive, then we continue with further iterations until no further improvements can be made simply by moving a node from one community to another community.

In the second stage of the algorithm, we start by creating a new network. This new network is defined by taking the nodes to be the communities that were found at the end of the first stage, and the new weighted edges between our new nodes are the sums of the weighted edges between nodes in the corresponding first stage communities.

We can then repeat the process, applying the first and second stages in turn until we can find no further alterations that cause a change in modularity.

3.3 Spectral Optimisation Method

Spectral methods usually utilise the matrix formulations by considering their eigenvectors and eigenvalues. The basic idea is to find a matrix whose eigenvectors can be used to find modularity (this is called a *modularity matrix*).[20]

Suppose, for simplicity, that we begin with a single division of the network into two separate communities. Let

$$s_i = \begin{cases} +1 & \text{if node } i \text{ is in the first community,} \\ -1 & \text{if node } i \text{ is in the second community.} \end{cases}$$

Then the Kronecker delta in (3.1) can be rewritten as

$$\delta(c_i, c_j) = \frac{1}{2}\left(s_i s_j + 1\right).$$

Define

$$B_{ij} = A_{ij} - P_{ij} = A_{ij} - \frac{k_i k_j}{2m},$$

so (3.1) becomes

$$Q = \frac{1}{4m} \sum_{ij} B_{ij} s_i s_j = \frac{1}{4m} \mathbf{s}^{\mathbf{T}} \mathbf{B} \mathbf{s}, \tag{3.3}$$

where $\mathbf{B}$ is the modularity matrix. Note that we can write $\mathbf{s}$ in terms of a basis of the eigenvectors of $\mathbf{B}$, with

$$\mathbf{s} = \sum_i a_i u_i$$

where $a_i = \mathbf{u}_i^{\mathbf{T}} \mathbf{s}$ and $\mathbf{u}_i$ $(i = 1, ..., n)$ gives the basis of eigenvectors.

Now (3.3) can be written as

$$Q = \frac{1}{4m} \sum_i a_i u_i^{\mathbf{T}} \mathbf{B} \sum_j a_j u_j = \sum_i a_i^2 \beta_i \tag{3.4}$$

where β_i is the eigenvalue that corresponds to eigenvector $\mathbf{u}_i$. All we have to do now is find the combination of a_i and β_i that gives us the largest value for Q. Finding the associated eigenvectors allows us to find the vector $\mathbf{s}$ and we can divide the network accordingly. Repeating this process recursively should finally give us a good approximation of a division that will optimise modularity.

We should also note that the Kernighan-Lin algorithm can be used in conjunction with either of the methods to further improve the approximation.[8,23] The gist of the method involves node-swapping, moving nodes into different communities to find the greatest change in Q. Used on its own, the algorithm is not only very effective, however if used alongside a divisive technique such as the spectral method, the results can be significantly improved, even for small network sizes.[21]

3.4 Issues With Modularity Optimisation

There are several problems that one encounters when investigating modularity optimisation and when applying the methods to real-world data.[9,11] The first point to make is that modularity optimisation problems have been proven to be NP-hard.[3] This tells us that computing the global maximum modularity would take significantly longer than finding local maxima. We would therefore need to settle for finding the best approximation by using the methods that optimise local maxima, as we have been doing so far. However we have no way of checking how close our approximation is to the actual global maximum value, and in many cases we will not have found the best possible partition of the network into communities.

Having computed an adjacency matrix from the dengue fever data, I tried applying three versions of the two methods given above. The Louvain method, a spectral optimisation method and finally a spectral optimisation method coupled with

Kernighan-Lin node swaps (the results are given in Section 4.1.1). I found that all three algorithms gave similar modularity values, but the distribution of communities was very different. This degeneracy problem was recently illuminated by Good *et al.* We can easily see that even if we could find a unique global maximum modularity, there are often multiple partitions that would give us similar values. This is a problem because we wanted to find naturally occurring communities; our goal was not to find the greatest approximation to the global maximum modularity. We can, however, look at clusters of nodes that frequently occur in the same groups across multiple methods, as these may represent more robust groupings and thus still give us the information that we were seeking.

Chapter 4

Results

Introduction

In this chapter I outline some insightful results of the study. I examine the evolution of the communities over the fifteen year period by first representing the network as a single adjacency matrix and then by creating a sequence of adjacency matrices that represent the network at discrete time intervals. Applying modularity optimisation algorithms to each matrix results in a series of partitions with associated modularities, which I plot as a function of time. I also consider the structure of communities and the role of individual nodes within those communities.

4.1 Modularity Optimisation

As I mentioned in Section 2.3, I considered two choices for the time aggregations. In the first case, I chose $t = 1$ and $\Delta t = 779$, which allowed me to compare the different algorithms outlined in Chapter 3. In the second case I chose $t = 1, 5, 9, ..., 729$ and $\Delta t = 52$, so that I could apply a modularity optimisation algorithm to the adjacency matrix corresponding to each time aggregation with the aim of investigating the evolution of community structure.

4.1.1 First Case Analysis

Having chosen to take $t = 1$ and $\Delta t = 779$ in the first case, the network is represented as a single adjacency matrix. I used this to calculate the modularity by applying the Louvain method (LM), the spectral optimisation algorithm (SO), and the spectral optimisation algorithm with Kernighan-Lin node swaps (KL); output values are given in Table 4.1. These modularity values are very reasonable given the size of the network,[11] but the partition corresponding to these values differs greatly between the algorithms.

The Louvain method and the spectral optimisation algorithm with Kernighan-Lin node swaps divided the network into five different communities, whereas the spectral optimisation algorithm used on its own divided the network into only three communities. It is possible, however, to identify certain groups of nodes that occur in the same community across all three algorithms. For example, nodes 1, 7 and 67 were allocated to the same community by all three methods (see Appendix A for the corresponding provinces and a map of their locations). Nodes 77, 78 and 79, three provinces in the Ucayali region, were also placed in the same community and the same occurred with the eleven nodes corresponding to the provinces in the Piura and Tumbes regions. These patterns suggests that the geographic location of nodes might have some influence on the spread of the disease (this is a sensible result, given the nature of dengue fever[24]), but we would need to perform further statistical analyses before claiming causality.

Algorithm	LM	SO	KL
Modularity	0.2418	0.2218	0.2202

Table 4.1. Community Detection Results

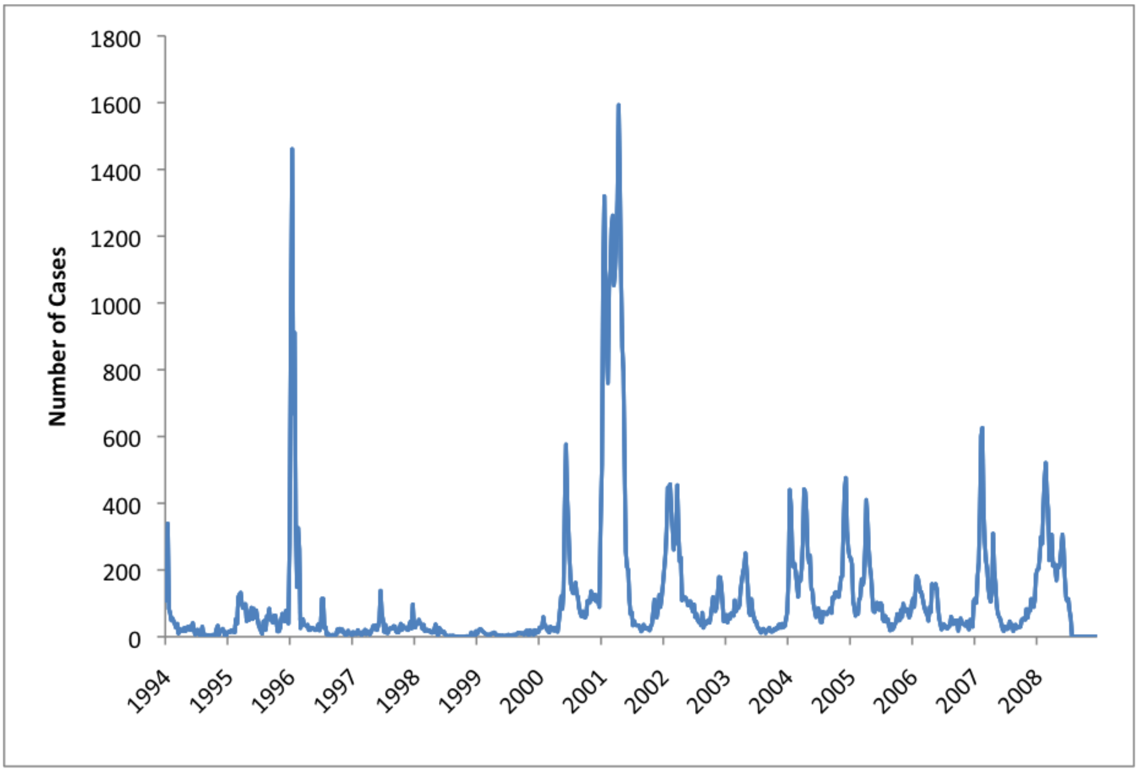
In Figure 1.1, we saw that Peru has interesting geographical characteristics, and I speculated that dividing the network into three communities based upon those characteristics might be an intuitive way to partition the network. Having calculated the modularity when dividing the network in this way, I found $Q = 0.0568$. However, when I allocated all nodes to one single community, the modularity was $Q = -1.2217 \times 10^{-15}$. This suggests that geographical characteristics have some bearing on the community structure, but that there are also other influencing factors involved.

As we know from Section 3.4, modularity optimisation is not an exact science. Most community detection techniques start with some randomised or arbitrary division that is often irrelevant to the real-world situation. We could try starting with a division that we speculate as having some significance within the context of the data. For instance, we could have tried starting with a division of the network into three communities; mountain, jungle and coastal, before applying the division or node-swapping techniques in Section 3.3.

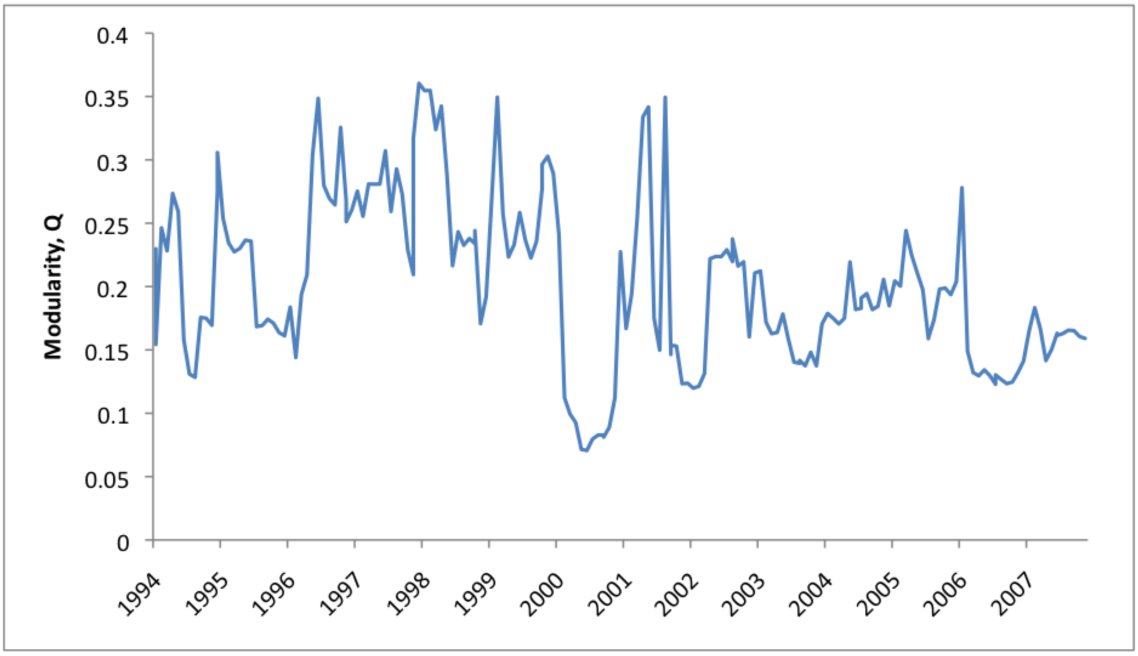
4.1.2 Second Case Analysis

In the second case, I applied the spectral optimisation algorithm to each of the 182 adjacency matrices representing the network at each time aggregation. This resulted in a sequence of modularity matrices, which were used to find the division of the network that will gave the greatest value of modularity.

I have chosen to use the spectral optimisation algorithm with the Kernighan-Lin node swaps in the second case, as it has been found, in general, to give greater modularity



(a) Total Number of Reported Cases of Dengue Fever in Peru per 100,000 People



(b) Modularity of Division of the Network using a Spectral Optimisation Algorithm

Figure 4.1: Graphs showing (a) the total number of reported cases per 100,000 people of dengue fever and (b) the maximum modularity over time.

values than greedy algorithms, even in smaller network sizes.[21] The greedy algorithm is faster computationally and might be useful for networks with size $O(10^6)$, though this is not an issue here as we are dealing with a small network with only 79 nodes.

The plot in Figure 4.1 shows how the modularity changes over time. The most striking result is that we can immediately see a sudden decline in modularity during the 2000 - 2001 epidemic. This suggests that there is a correlation between community structure and epidemic outbreaks, perhaps due to a sudden shift in the structure of the network.

This correlation suggests that we should look at the structure of the network during the time period over which the major outbreaks occur.

4.2 A Closer Inspection of Nodes During Outbreaks

Figure 4.2 shows us that there is a small number of nodes for which the number of reported cases of dengue fever is very high. I have identified these as nodes 7, 14, 33, 45, 61, 74 and 77 (see Appendix A).

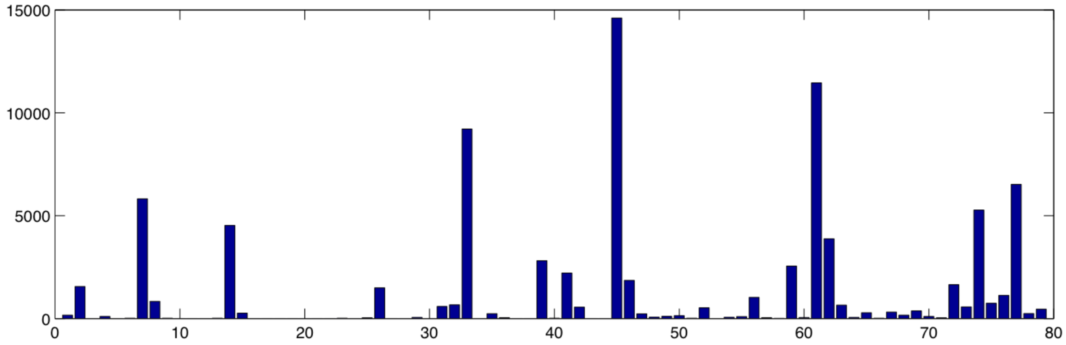
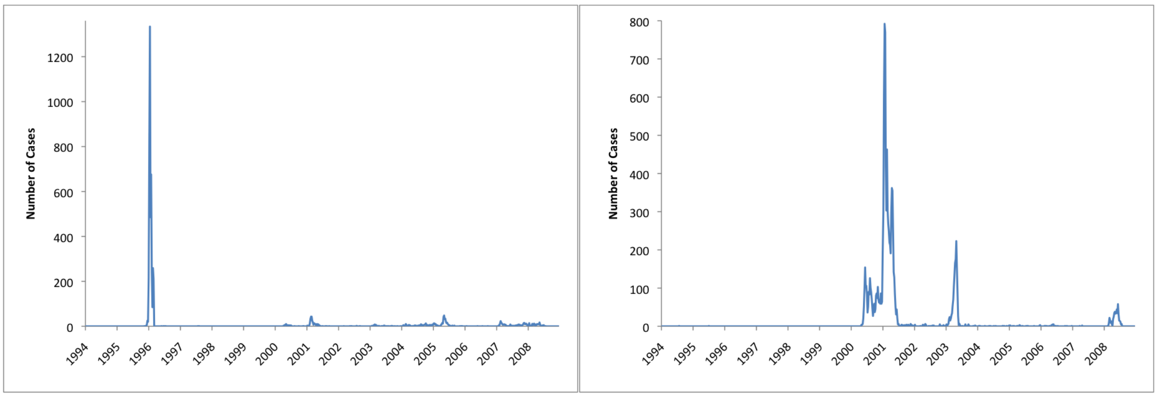


Figure 4.2: Total number of reported cases of dengue fever for each of the nodes over the fifteen year period.

The plots in Figure 4.3 shows us that nodes 7 and 61 have such a large total number of cases of dengue fever because large scale outbreaks occur in those provinces during the nationwide epidemics. Closer analysis of the data during the 1996 outbreak reveals that one node can be distinguished as having particular importance, node 7 (the Utcubamba province). In just one week, the number of reported cases in Utcubamba was 1334; the number of reported cases in other provinces during the epidemic was no more than 129. Similarly, provinces corresponding to nodes 33 and 61 report a maximum of approximately 800 cases of dengue fever in one week during the 2000 - 2001 epidemic. Of the 77 other provinces, only six report over 100 cases in one week during this time, of which the maximum is 415 cases.

Given this information, it seems sensible to compute measures such as the local weighted degree to find the relative importance of each node within its community. This might help us distinguish the hubs, for example, which might have some significant influence on outbreaks in other provinces. In epidemiology this would usually help to identify the regions that could be targeted for vaccinations, or it might suggest restricting travel between certain regions to limit the spread of the disease.[24]

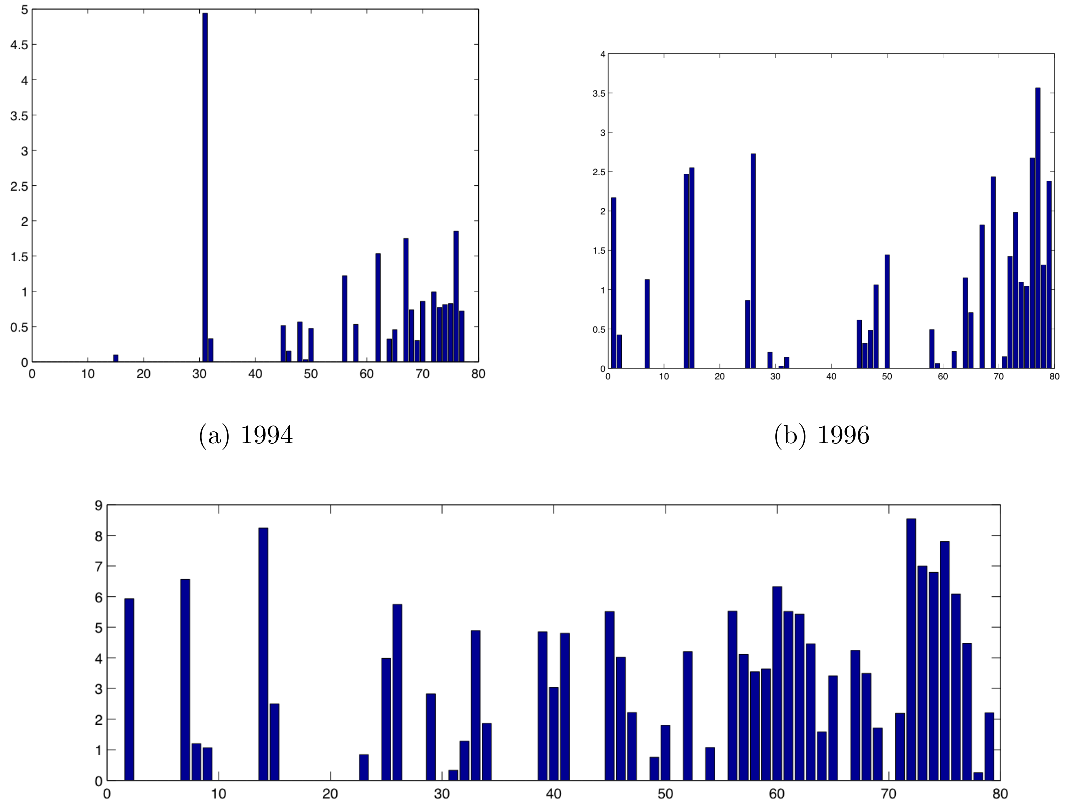


(a) Node 7

(b) Node 61

Figure 4.3: Total number of reported cases of dengue fever in the (a) Utcubamba and (b) Sechura provinces over the fifteen year period.

I calculated the local weighted degree of each node during three different year-long time windows, see Figure 4.4. I took the *local weighted degree* of a node i to be the sum of the edge weights between i and the nodes in the community that i belongs to. It is difficult to infer information from these plots without carrying out other statistical calculations, but due to time constraints these calculations were not possible.



(a) 1994

(b) 1996

(c) 2000

Figure 4.4: Local degree distribution for each of the nodes during the years (a) 1994, (b) 1996, (c) 2000.

Chapter 5

Discussion

5.1 Conclusions

I have used methods in network theory to try to model the spread of dengue fever in Peru from 1994 to 2008. I defined a time-dependent network to model the data and represented it in the form of a sequence of adjacency matrices. Constructing the network in the form of an adjacency matrix made it easier to use modularity optimisation techniques to find the best division of the network into communities. I began by aggregating the data over the entire fifteen year period in order to compare the uncovered partitions that result from applying different modularity optimisation algorithms. Using time aggregations representing yearly windows that shift by four weeks at a time, I also constructed a sequence of adjacency matrices and from this I was able to calculate and observe the evolution of the communities over time by applying a spectral optimisation algorithm at each time slice. Having plotted the modularity function over the time period, there appeared to be a significant dip in modularity at the time of the major 2000 - 2001 epidemic. Finally I identified nodes that play a significant role within their communities during this time with the aim of assessing their role within their community.

5.2 Possible Further Explorations

The techniques that I have looked at use many new ideas, and they have been applied to very few time-dependent data sets. Even though we have found several interesting correlations in this study, there is still plenty of work to be done. If we can find more structural patterns and if we can find causal explanations for the results, then we might be able to find preventative measures for a disease for which there is currently no known vaccine. The further possible explorations that I will give merely skim the surface of work that can be done.

5.2.1 Other Applications

Having found what appears to be a correlation between community structure and dengue outbreak, applying the same techniques to other similar high frequency time series data could be insightful. One possible study involves analysis of the spread of rubella in Peru.[16] This data set also includes the number of reported cases collected on a weekly basis by province, and was collected over a similar time period: from 1997 to 2009. Although we would need to account for rubella-specific characteristics, such as different time scales relating to the disease, having information on the population and spatial location of the provinces could make it easier to make comparisons between the two data sets.

5.2.2 Refining the Null Model

Thus far, I have defined the adjacency matrix by assuming a weighted but undirected network. However, given that dengue fever is a vector-borne disease, a more accurate representation would incorporate the fact that it is a directed network. To take this into account, one could use a version of modularity by Leicht and Newman[13] that has been developed for directed networks.

If the *in-degree* of node i is the number of edges entering node i, and the *out-degree* of node j is the number of edges leaving node j, define[14]

$$k_i^{in} = \sum_{j=1}^{n} A_{ij},$$

and

$$k_j^{out} = \sum_{i=1}^{n} A_{ij}.$$

The null model in (3.1) is then

$$P_{ij} = \frac{k_i^{in} k_j^{out}}{m}$$

so that

$$Q = \frac{1}{2} \sum_{ij} \left(A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right) \delta(c_i, c_j)$$

where k_i^{in} is the weighted in-degree of node i and k_j^{out} is the weighted out-degree of node j.

Recent developments in the role of modularity in spatial networks by Expert *et al.*[6] also suggests that it might be beneficial to require some dependence of the null model on spatial factors, perhaps using

$$P_{ij} = N_i N_j f(d_{ij})$$

where N_i is some measure of importance of the node i in terms of its location within the network and f is a function such as

$$f(d) = \frac{\sum_{ij|d_{ij}=d} A_{ij}}{\sum_{ij|d_{ij}=d} N_i N_j},$$

where d_{ij} is the distance between nodes i and j. This allows one to find underlying patterns in the network that do not solely depend on the spatial characteristics of the data.

Appendix A

List of Provinces

Node	Province	Region	Description
1	Bagua	Amazonas	Mountains, North
2	Bongara	Amazonas	Jungle
3	Chachapoyas	Amazonas	Mountains, North
4	Condorcanqui	Amazonas	Jungle
5	Luya	Amazonas	Mountains, North
6	Rodriguez de Mendoza	Amazonas	Mountains, North
7	Utcubamba	Amazonas	Jungle
8	Corongo	Ancash	Coast, Central
9	Santa	Ancash	Coast, Central
10	Parinacochas	Ayacucho	Mountains, Central
11	Cajabamba	Cajamarca	Mountains, North
12	Chota	Cajamarca	Mountains, North
13	Cutervo	Cajamarca	Mountains, North
14	Jaen	Cajamarca	Jungle
15	San Ignacio	Cajamarca	Jungle
16	San Miguel	Cajamarca	Mountains, North
17	San Pablo	Cajamarca	Mountains, North
18	Santa Cruz	Cajamarca	Mountains, North
19	Callao	Callao	Coast, Central
20	La Convencion	Cusco	Mountains, South
21	Urubamba	Cusco	Mountains, South
22	Tayacaja	Huancavelica	Mountains, Central
23	Ambo	Huanuco	Mountains, Central
24	Huamalies	Huanuco	Mountains, Central
25	Huanuco	Huanuco	Mountains, Central
26	Lauricocha	Huanuco	Mountains, Central
27	Leoncio Prado	Huanuco	Mountains, Central
28	Marañon	Huanuco	Mountains, Central
29	Pachitea	Huanuco	Mountains, Central

Node	Province	Region	Description
30	Chupaca	Junin	Mountains, Central
31	Concepcion	Junin	Mountains, Central
32	Junin	Junin	Mountains, Central
33	Ascope	La Libertad	Coast, North
34	Bolivar	La Libertad	Coast, North
35	Gran Chimu	La Libertad	Coast, North
36	Pacasmayo	La Libertad	Coast, North
37	Santiago de Chuco	La Libertad	Coast, North
38	Viru	La Libertad	Coast, North
39	Chiclayo	Lambayeque	Coast, North
40	Ferreñafe	Lambayeque	Coast, North
41	Lambayeque	Lambayeque	Coast, North
42	Barranca	Lima	Coast, Central
43	Cajatambo	Lima	Coast, Central
44	Huaral	Lima	Coast, Central
45	Alto Amazonas	Loreto	Jungle
46	Datem Del Marañon	Loreto	Jungle
47	Loreto	Loreto	Jungle
48	Mariscal Ramon Castilla	Loreto	Jungle
49	Maynas	Loreto	Jungle
50	Requena	Loreto	Jungle
51	Ucayali	Loreto	Jungle
52	Manu	Madre de Dios	Jungle
53	Tahuamanu	Madre de Dios	Jungle
54	Tambopota	Madre de Dios	Jungle
55	Pasco	Pasco	Mountains, Central
56	Ayabaca de Mendoza	Piura	Coast, North
57	Huancabamba	Piura	Coast, North
58	Morropon	Piura	Coast, North
59	Paita	Piura	Coast, North
60	Piura	Piura	Coast, North
61	Sechura	Piura	Coast, North
62	Sullana	Piura	Coast, North
63	Talara	Piura	Coast, North
64	Bellavista	San Martin	Jungle
65	El Dorado	San Martin	Jungle
66	Huallaga	San Martin	Jungle
67	Lamas	San Martin	Jungle
68	Mariscal Caceres	San Martin	Jungle
69	Moyobamba	San Martin	Jungle
70	Picota	San Martin	Jungle
71	Rioja	San Martin	Jungle
72	San Martin	San Martin	Jungle
73	Tocache	San Martin	Jungle

Node	Province	Region	Description
74	Contralmirante Villar	Tumbes	Coast, North
75	Tumbes	Tumbes	Coast, North
76	Zarumilla	Tumbes	Coast, North
77	Atalaya	Ucayali	Jungle
78	Coronel Portillo	Ucayali	Jungle
79	Padre Abad	Ucayali	Jungle

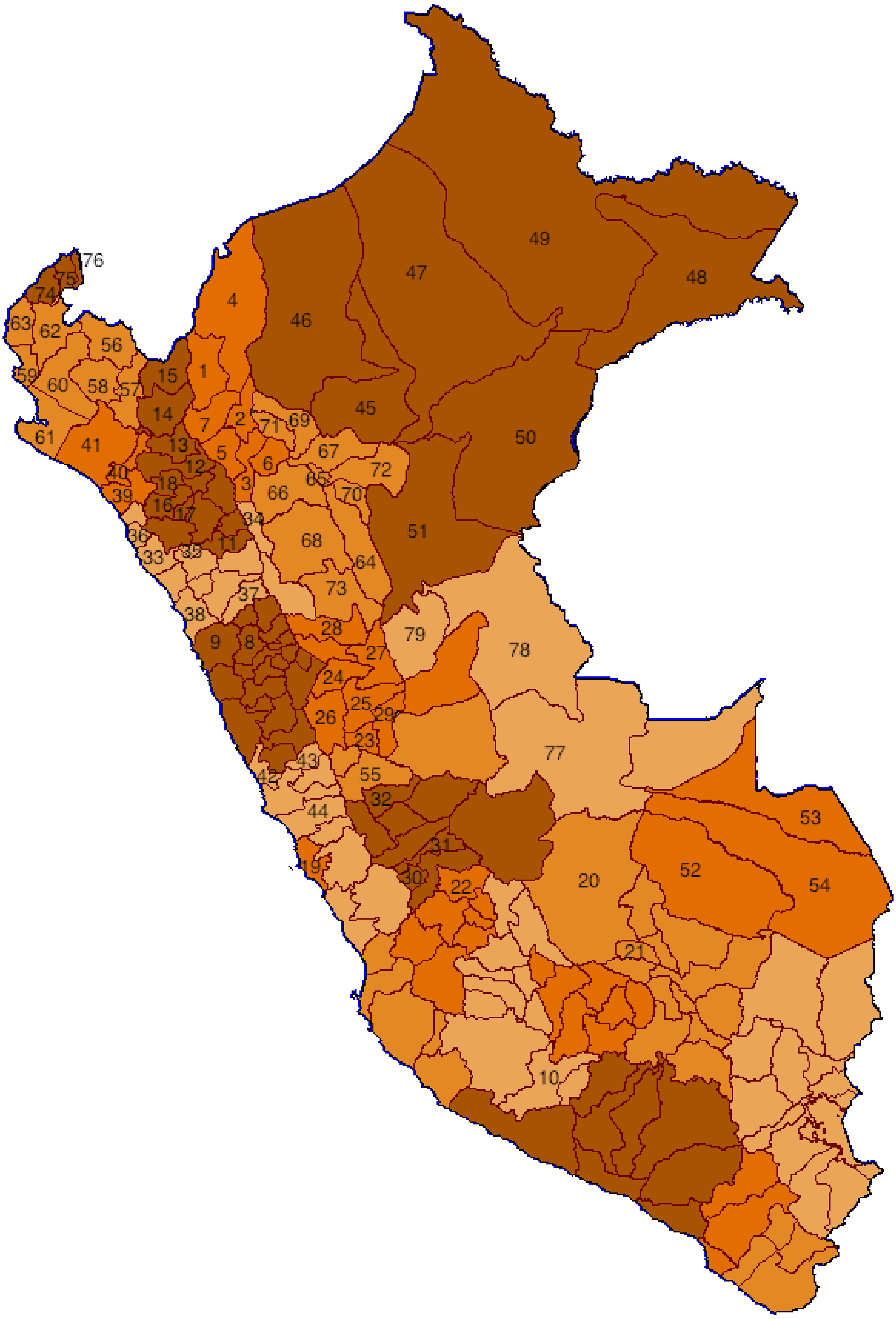


Figure A.1: Map illustrating location of provinces.

Bibliography

[1] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10):10008, 2008.

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, **424**(4-5):175–308, 2006.

[3] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, **20**(2):172–188, 2008.

[4] Jacqueline L Deen, Eva Harris, Bridget Wills, Angel Balmaseda, Samantha Nadia Hammond, Crisanta Rocha, Nguyen Minh Dung, Nguyen Thanh Hung, Tran Tinh Hien, and Jeremy J Farra. The who dengue classification and case definitions: time for a reassessment. *The Lancet*, **368**(9530):170–173, 2006.

[5] Floyd W. Denny, Wallace A. Clyde Jr., and W. Paul Glezen. Mycoplasma pneumoniae disease: Clinical spectrum, pathophysiology, epidemiology, and control. *The Journal of Infectious Diseases*, **123**(1):74–92, 1971.

[6] Paul Expert, Tim Evans, Vincent D. Blondel, and Renaud Lambiotte. Beyond Space For Spatial Networks. *arXiv:1012.3409*, 2010.

[7] Daniel J. Fenn, Mason A. Porter, Peter J. Mucha, Mark McDonald, Stacy Williams, Neil F. Johnson, and Nick S. Jones. Dynamical Clustering of Exchange Rates. *arXiv:0905.4912*, 2010.

[8] Santo Fortunato. Community detection in graphs. *Physics Reports*, **486**(3-5):75–174, 2010.

[9] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, **104**(1):36–41, 2007.

[10] Arthur Getis, Amy C. Morrison, Kenneth Gray, and Thomas W. Scott. Characteristics of the spatial pattern of the dengue vector, *Aedes ae-*

gypti, in iquitos, peru. *The American Journal of Tropical Medicine and Hygiene*, **69**:494–505, 2003.

[11] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, **81**(4):046106, 2010.

[12] Jason A. L. Jeffery, Nguyen Thi Yen, Vu Sinh Nam, Le Trung Nghia, Ary A. Hoffmann, Brian H. Kay, and Peter A. Ryan. Characterizing the *Aedes aegypti* population in a vietnamese village in preparation for a *Wolbachia*-based mosquito control strategy to eliminate dengue. *PLoS PLoS Neglected Tropical Diseases*, **3**(11):552, 2009.

[13] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Physical Review Letters*, **100**(11):118703, 2008.

[14] M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[15] Naomi Mapstone. Peru on red alert after dengue outbreak. *The Financial Times*, 2011.

[16] C J E Metcalf, C V Munayco, G Chowell, B T Grenfell, and O N Bjørnstad. Rubella metapopulation dynamics and importance of spatial coupling to the risk of congenital rubella syndrome in peru. *Journal of the Royal Society Interface*, **8**(56):369–76, 2011.

[17] Lauren Ancel Meyers, M. E. J. Newman, Michael Martin, and Stephanie Schrag. Applying network theory to epidemics: Control measures for outbreaks of mycoplasma pneumoniae. *Emerging Infectious Diseases*, **9**(2):204–210, 2003.

[18] Lauren Ancel Meyers, M E J Newman, and Babak Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, **240**(3):400–418, 2006.

[19] J. Murray. *Mathematical biology: I. An introduction.* Springer, 2005.

[20] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, **74**(3):036104, 2006.

[21] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, **103**:8577, 2006.

[22] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, **69**:026113, 2004.

[23] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *World Wide Web Internet And Web Information Systems*, **56**(9):1082–1097, 2009.

[24] Steven T. Stoddard, Amy C. Morrison, Gonzalo M. Vazquez-Prokopec, Valerie Paz Soldan, Tadeusz J. Kochel, Uriel Kitron,

John P. Elder, and Thomas W. Scott. The role of human movement in the transmission of vector-borne pathogens. *PLoS Neglected Tropical Diseases*, **3**(7):e481, 2009.

[25] Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter. Community structure in online collegiate social networks. *arXiv:0809.0690v3*, 2008.

[26] Wayne Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**(4):452–473, 1977.