Examining dynamic network structures in relation to the spread of infectious diseases

Thomas Prescott

September 29, 2011

Abstract

Dengue fever is a tropical, mosquito-borne disease, currently with no vaccine. We require greater understanding of how its spatial distribution evolves over time, in order to inform policies aiming to prevent epidemics. This report uses data on weekly dengue fever cases recorded in 79 provinces of Peru to form a time-dependent network representing time-dependent correlations between time series. We will investigate the community structure in this network, and attempt to relate the changing community structure to epidemic events by investigating how time-dependent spatial distributions and geographical characteristics of communities correspond to disease outbreaks.

Word count:

5047, according to `http://app.uio.no/ifi/texcount/online.php`.

Contents

1 Introduction

1.1 Dengue fever

Dengue fever, principally carried by the *Aedes aegypti* mosquito, is a viral infection that affects humans. It is prevalent in almost every tropical country, and there is currently no vaccine. Approximately 2.5 billion people are at risk of infection, and 50 million infections per year are estimated to occur worldwide. The virus is most dangerous when a patient is infected a second time: in young children especially, a secondary infection can cause dengue shock syndrome or dengue haemorrhagic fever, both of which are characterised by severe haemorrhaging and further resulting complications [1].

To control the spread of dengue fever, we require greater understanding of how epidemics occur in the places and at the times that they do. Given infection patterns up to a certain point in time, there would ideally be a method for predicting how likely an outbreak is to occur in a particular place in the near future, and its potential extent. One could then envision efficient preventative policies targeting the control of *A. aegypti* in strategically important provinces.

1.2 Epidemics in Peru

The present paper is based on data supplied by Dr Gerardo Chowell at Arizona State University and originally collected by the Peruvian Ministry of Health. The data is a set of 79 time series, each recording the weekly number of dengue fever cases in 79 (of 195) provinces of Peru from 1994 to 2008. The other 116 provinces recorded no cases.

The US Agency for International Development (USAID) define a *dengue fever epidemic* as a disease count that is two standard deviations above the baseline [2]. The dotted red curves in Figure 1 show where this threshold lies; we use the entire time series to define the baseline (i.e. mean) and standard deviation. This estimate is crude: national health agencies use more sophisticated definitions [2]. Nevertheless, from Figure 1(a), we can identify at least two clear Peru-wide epidemics (in 1996 and 2000–2001).

Dengue fever has become more prevalent since the 2000–2001 epidemic, suggesting a key

question: what was different between the two epidemics, such that the dengue fever infection rate did not return to its original level after the second?

There is also a spatial element to the definition of an epidemic. This is illustrated in Figure 1, where we also plot the time series of two different provinces. Figure 1(b) shows that only Utcubamba contributed a significant amount to the 1996 spike. Conversely, Figure 1(c) shows that Alto Amazonas did not experience either epidemic identified in Figure 1(a). However, there were four epidemic events local to that province, but none of them are identifiable from the global time series.

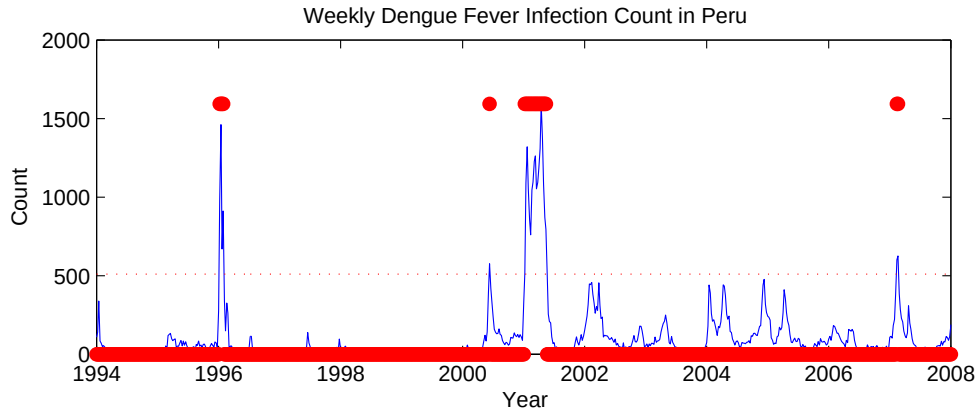
Figure 1 shows that one should take care in defining the context of epidemics. Examining global properties of data is important, but if one wants to understand the reasons epidemics happen in the places that they happen then one must also take local properties into account. Chowell et al. [3] comment that lower spatial resolution of epidemiological data will help to generate new hypotheses about the underlying mechanisms of dengue distribution: for example, they find that dengue fever tends to move from areas of high to low population.

1.3 Networks and communities

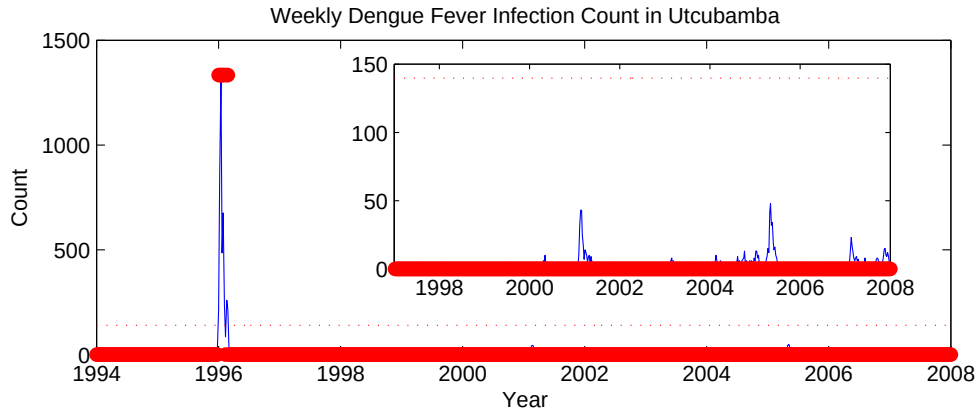
In this study, we convert the Chowell data (79 time series, each over 780 weeks) into a time-dependent network and use techniques of network science to uncover underlying structures in the data. This approach has been applied to the analysis of diverse time series data, including fMRI [4, 5], foreign exchange rates [6], and social networks [7, 8, 9]. Many techniques have to be adapted to their specific application, but Newman [10] summarizes a rapidly expanding toolkit of generally applicable methods.

To create a time-dependent network out of our data, for any given time t, let each node represent the portion of a province's time series of some fixed length up to time t. The edge weights between each pair of nodes are determined by some measure of the correlation between those portions. The time-dependence arises by varying t, creating a network where edge weights vary with t.

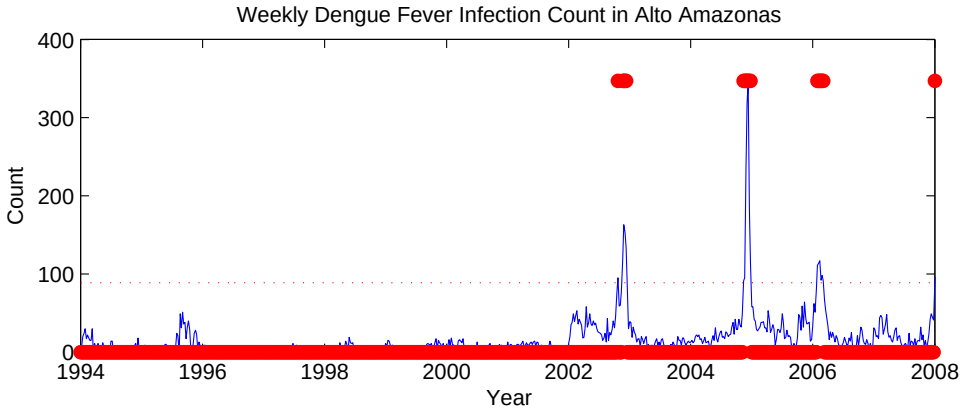
One technique for examining a network is the detection of its community structure [11]. Given a network, a *community* is a group of nodes with relatively denser edge weights between one another than to the other nodes. A community structure is a partition of a



(a) Weekly count of dengue fever infections aggregated across all provinces.



(b) Weekly count of dengue fever infections in Utcubamba. The inset gives a magnified view.



(c) Weekly count of dengue fever infections in Alto Amazonas.

Figure 1: Aggregated time series, and example time series for specific regions.

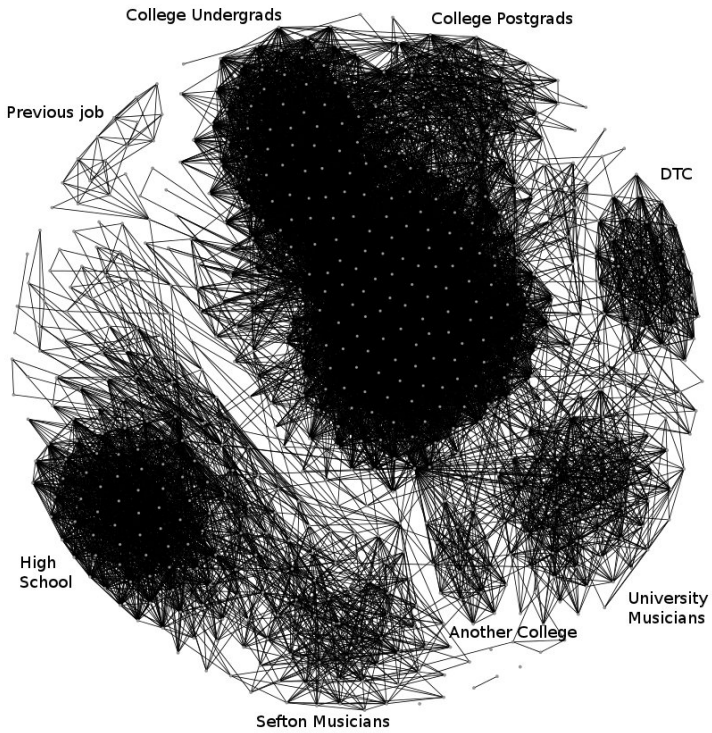


Figure 2: A network in which the author's Facebook friends correspond to the nodes, and an edge exists if there is a corresponding friendship link on Facebook. The network has been laid out using the Fruchterman–Reingold algorithm [12]. The data was downloaded from Facebook using the NameGenWeb application [13] written by Bernie Hogan of the Oxford Internet Institute (`http://apps.facebook.com/namegenweb/`) and visualized in Gephi (`http://gephi.org/`).

network into (usually disjoint) communities, provided such a partition is reasonable.

An illustration of a (social) network with community structure is given in Figure 2. By inspection, we can see densely connected communities (some more distinct than others) with fewer connections to the rest of the network. These communities correspond to the author's assorted affiliations.

Although the notion of communities makes intuitive sense, a precise mathematical definition of communities is difficult to pin down [11, 14, 15]. The framework for most community detection methods is to measure the quality of a partition of the nodes of the network in terms of some quality function (see Section 2.2).

A community in our network corresponds to a distinct group of regions whose infection

patterns are closely correlated. Our aim is to observe how changes in community structure relate to the onset of epidemics of dengue fever. Are certain provinces important for the transfer of infection to the rest of their communities, and why do certain provinces experience epidemics when others do not?

2 Methodology

2.1 Network construction

To form a network from a set of time series $\{X_1, X_2, \ldots, X_N\}$, each of length T, we define a set of N nodes $\{1, 2, \ldots, N\}$, where node i corresponds to X_i. The edge weights A_{ij} between each pair of nodes i and j represent a measure of the similarity between X_i and X_j. There are many choices for how to calculate the *adjacency matrix* A [4].

A simple map from a time series to a network is correlation; we will use a slight modification of the correlation used in [16]. Let $X_i(t)$ denote the value of X_i in week t. If Δ is a positive-integer parameter defining the time window for comparison of time series, then let

$$\beta_{ij}(t) = \sum_{\tau=t-\Delta+1}^{t} X_i(\tau)X_j(\tau) \tag{1}$$

define a set of scalar-valued functions on $t = \Delta, \Delta+1, \ldots, T$. We use $\{\beta_{ij}(t)\}$ to define the entries of a time-dependent adjacency matrix A by putting

$$A_{ij}(t) = \frac{\beta_{ij}(t)}{\sqrt{\beta_{ii}(t)\beta_{jj}(t)}} - \delta(i,j), \tag{2}$$

where we have ensured, using $\delta(i,j)$, that $A_{ii}(t) = 0$ for all i and t. Note that, if $\beta_{ii}(t) = 0$ for some i and t, there were no infections over the time window, so we cannot define A_{ij} or A_{ji} for any j using (2). In these instances, we set $A_{ij}(t) = A_{ji}(t) = 0$ so that node i has no adjacent edges at time t.

The correlation (1) is a simple quantity to use as edge weights, but many other methods also exist. These range from pairwise approaches, such as correlation, to global network inference methods [4].

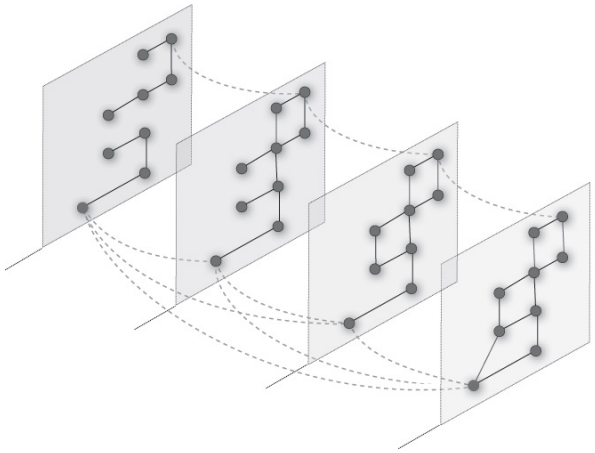


Figure 3: A schematic (taken, with permission, from [8]) of inter-slice coupling: the four squares represent networks in sequence as individual slices, and the dotted curves are the inter-slice edges to make a single multislice network.

2.1.1 Coupling networks in sequence

An extension of this approach is to view each adjacency matrix $A(t)$ for $t \in \{\Delta, \ldots, T\}$ as a slice of a single "multislice" network [7, 8]. For a sequence of time-indexed networks, each slice is coupled to its nearest-neighbouring slices. The inter-slice coupling is introduced by linking together nodes which appear in neighbouring slices (see Figure 3). Inter-slice couplings for a node j appearing in both time t and time s are denoted $C_{ts}(j)$. All applications up to now [e.g. 5, 7, 8] consider $C_{ts}(j) \in \{0, \omega\}$ and vary $\omega \geq 0$.

There is a trade-off between having many slices for good temporal resolution and ensuring that each slice holds enough data that one can be confident of the statistical significance of the adjacency matrix entries [5]. Thus we require a good choice of Δ: for instance, choosing $\Delta = 78$ gives 10 non-overlapping slices $A(78), A(156), \ldots, A(780)$, which does not give sufficient resolution, given that the large outbreaks in Figure 1 are shorter than 78 weeks. However, choosing a Δ that is too small might result in unreliable intra-slice adjacencies.

2.2 Defining modularity

Suppose that we partition the nodes of a network into disjoint communities. To measure the quality of the partition, we use the standard *modularity*

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - P_{ij}\right) \delta\left(c_i, c_j\right), \tag{3}$$

where $2m = \sum_{ij} A_{ij}$, c_i denotes the community containing node i, and δ denotes the Kronecker delta. The quantity P_{ij} is known as the *null model matrix*, which gives a relative value of how densely connected we expect nodes in a community to be under a given *null model* (see below). It can be derived by combinatoric arguments [10] or by examining the statistical properties of random walks on the network [17].

2.2.1 Newman–Girvan null model

The most popular null model used in community detection is $P_{ij} = k_i k_j/(2m)$ [18], where $k_i = \sum_j A_{ij}$. This corresponds to the edge weight between nodes i and j expected to arise by chance, given the degrees of each node. This gives the Newman–Girvan modularity

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m}\right) \delta\left(c_i, c_j\right). \tag{4}$$

By using the framework of statistical properties of random walks on such networks, this null model can be modified to deal with directed networks, or signed edge weights, and so on [17].

One can also bias modularity to favour larger or smaller communities. Reichardt and Bornholdt [15] incorporated a *resolution parameter* γ into (4) to give

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right) \delta\left(c_i, c_j\right), \tag{5}$$

and Lambiotte et al. [17] identified this parameter as the inverse of the timescale of a random walk on the network. Small values of γ mean communities tend to be larger, and vice versa.

2.2.2 Modularity for sequences of coupled networks

The definitions above are made in the context of a single network slice. When considering the multislice setting described in Section 2.1.1, Mucha et al. [8] used an altered Newman–Girvan null model. Communities are formed across the entire multislice network. The modularity is

$$Q = \frac{1}{2\mu} \sum_{ijsr} \left[\left(A_{ij}(s) - \gamma_s \frac{k_{is} k_{js}}{2m_s} \right) \delta_{sr} + \delta_{ij} C_{sr}(j) \right] \delta\left(c_{is}, c_{jr}\right), \tag{6}$$

where c_{is} refers to the community containing node i in slice s. The values γ_s, k_{is}, and m_s refer to γ, k_i, and m as defined in the sections above, but they are now defined separately for each slice s, and $2\mu = \sum_{ijs} A_{ij}(s) + \sum_{srj} C_{sr}(j)$ is the sum of all of the edge weights in the multislice network.

Recall that we assumed that $C_{sr}(j) \in \{0, \omega\}$, where $\omega \geq 0$ is a constant. For our network, if node j is in the large connected component at time t and $t \pm 1$, then $C_{t,t\pm1}(j) = \omega$. All other values of $C_{sr}(j)$ are set to 0. Modularity, given by (6), is then biased by the value of ω such that larger values tend to result in communities containing nodes across more contiguous slices, whereas smaller values allow more nodes in contiguous slices to be placed into different communities.

2.3 Community detection algorithm

Given a function describing the quality of any partition, one can attempt to determine an optimal partition of nodes into communities. This problem is NP-hard [19], so many heuristic methods [10, 11] have been developed to maximize modularity. We will use the popular, locally greedy algorithm known as the Louvain method [20], which is fast and returns competitively high values of modularity [14].

The Matlab implementation of this algorithm was downloaded from Netwiki[1].

[1] See `http://netwiki.amath.unc.edu`.

2.4 Community structure and epidemics

Our goal is to explore the community structure of the Chowell data, and in particular to investigate how the community structure changed through 2000–2001 such that dengue fever became more prevalent after the epidemic in that period [Figure 1(a)]. A secondary question is to examine how the first outbreak, in 1996, was localized to just one node in our network [Figure 1(b)] and did not propagate to others.

2.4.1 Geographical and temporal characteristics of communities

One approach to the investigation of algorithmically-determined communities is to examine their geographical properties, to gain insight into how the geography of Peru determines the different infection patterns they represent. In the multislice setting, communities also exist over time, so we can study whether there are any critical time-points that mark a shift in the distribution of infection patterns.

The inspiration for this approach is the work of Traud et al. [9], who carried out a similar investigation for a set of five (static) online social networks. Algorithmically-detected communities were compared to partitions of the data by demographic information. Our network consists of nodes representing provinces, rather than people, so we will use similar methods but replace demographic classifications with geographical ones.

The basis of the work in [9] is the standardization of pair-counting methods for comparing two partitions. Given two partitions, define w_{11} as the number of node pairs classified in the same community in both partitions, and w_{00} as those classified differently in both partitions. If M is the total number of node pairs in the network, then the Rand coefficient

$$S_R = \frac{w_{00} + w_{11}}{M}$$

measures the proportion of pairs that both partitions classify as either in the same community or in different ones. While easily interpreted, the measure is skewed towards higher values when there are more communities. Hence, the value of this measure for one pair of partitions is not, in general, directly comparable with the value returned for another pair.

Traud et al. [9] also identified similar problems with other (more complicated) pair-counting methods. To overcome this, they compared the value calculated for the measure against

what would be expected at random. This yields a z-score, known as the z-Rand score. Large z-Rand scores point towards a statistically significant value for w, and hence significant agreement between the two partitions.

The formula for the calculation of the z-Rand score is given as follows. Let $w := w_{11}$, and let M_1 and M_2 denote respectively the number of pairs classified the same way in the first and second partition. The z-Rand score is

$$z_R = \frac{1}{\sigma_w}\left(w - \frac{M_1 M_2}{M}\right) \tag{7}$$

for standard deviation σ_w expressed in terms of M_1, M_2, M and total nodes N [9, Equations (2.2) and (2.3)].

This allows us to quantify the agreement between any two partitions of a set of nodes. In Section 3.4 we will quantify how the detected communities reflect node classifications based on geographical characteristics and year.

3 Results

3.1 Time-dependent networks and modularity

In this section, we choose a time window of Δ weeks to create a sequence of network adjacencies $A(t)$, given by (2), for $t = \Delta, \Delta + 1, \ldots, 780$. For each value of t, we use the Louvain algorithm to detect communities.

Figure 4 shows how the maximized modularity given by (5) varies with t for three values of resolution parameter γ. This figure is an extension of Figure 4.1 from [16]. However, in this report, the time t is defined as the end of the time window rather than as the beginning. We made this choice because it only makes sense to look at past data to be predictive. As a result, the large decrease in modularity found in [16] to occur at the start of 2000 is actually found at the peak of the epidemic around the start of 2001. Nevertheless, we can see that this decrease in modularity is not unique to $\gamma = 1$. In fact, a decrease in modularity of over 0.1 occurs for all $\gamma \in [0.8, 2.3]$, so we can be confident of this feature's robustness. Furthermore, this association of t with the end of each time window with does not change the conclusion of [16] that the modularity decrease indicates

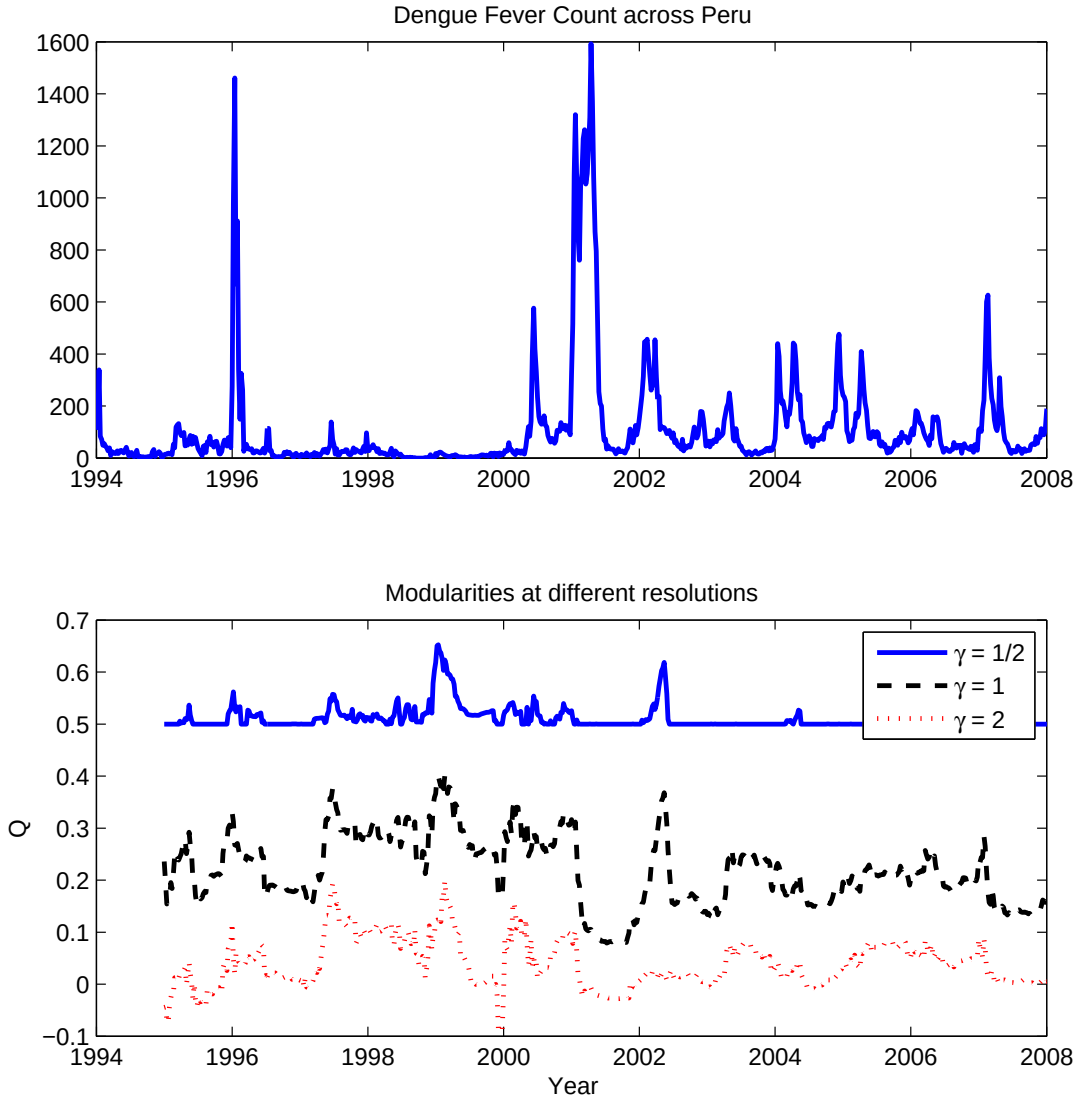


Figure 4: Modularity versus year for various resolution parameter values, compared to total dengue fever cases. We use a window of $\Delta = 52$ (matching [16]).

a possible correlation between community structure and the onset of epidemics. However, the correlation between modularity, as a time series, and the global Peru-wide time series ranges only from -0.1936 to -0.1084 for $\gamma \in [0.8, 2.3]$.

3.2 Spatial properties of communities

Because our networks are embedded in space, we examined the spatial properties of the detected communities. Figure 5 shows the positions of the nodes, collated from the Geographic Names Database[2], and coloured by assigned community, at four times spaced 200 weeks apart. Some nodes are naturally more spaced apart, as they correspond to the central points of larger provinces. Although there is no obvious pattern to track the spatial properties of the communities, we can use the positions of the nodes to quantify the spatial spread of communities.

For a given time t, we have a partition of all 79 nodes into r_t communities. Each community consists of nodes j with associated centroids (x_j, y_j). We measure the spread of a community by taking the mean Euclidean distance of each of its nodes from their mean position. We then average this over all non-singleton communities to give an average spread.

If a community structure is reliant on proximity, one would expect that as the number of non-singleton communities increases, their corresponding mean geographical spread will decrease. This is because each community's size (i.e. number of nodes) will have to decrease, but we would expect spatially adjacent nodes to remain in the same community. This indeed happens in Figure 6, where there is a clear negative correlation of -0.6 between community number and community spread after 2000. In 1999, there is a significant decrease in geographical spread without the expected increase in community number.

We tentatively interpret this result to suggest that the community structure becomes extremely localized in the year leading up to the large, widespread epidemic starting in 2000 [again see Figure 1(a)], and from that point proximity remains important: up to that point, communities were much more dispersed across the country. It is difficult to be certain, as the small numbers of non-singleton communities also add some level of uncertainty to the

[2]Toponymic information is based on the Geographic Names Database, maintained by the USA National Geospatial-Intelligence Agency. More information is available at `www.nga.mil`.

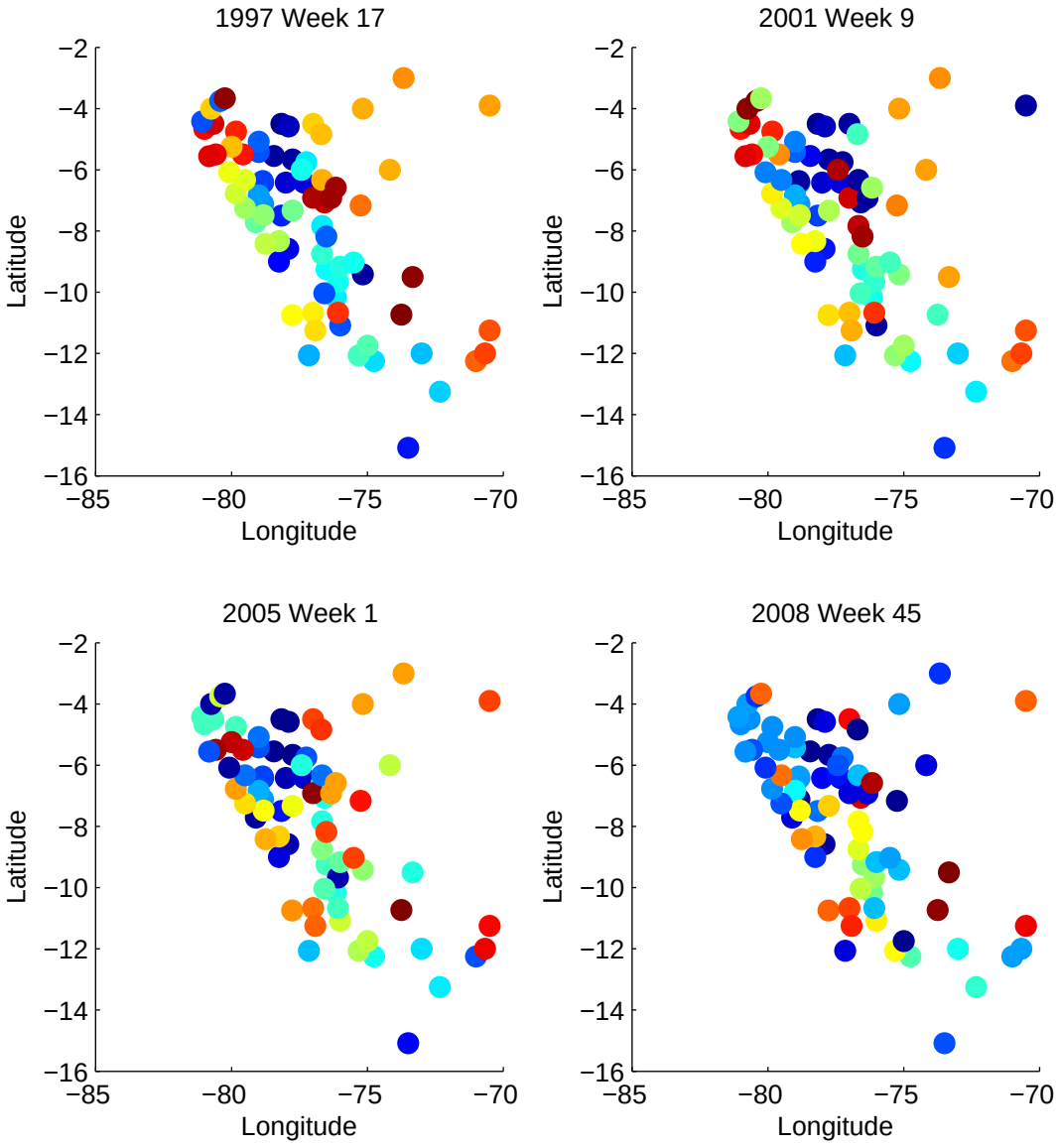


Figure 5: Four snapshots of the detected community structure at different points in the time series, where each node is arranged by the latitude and longitude of the centroid of the province. In each snapshot, each colour corresponds to a different community. There is no correspondence between colour across snapshots.

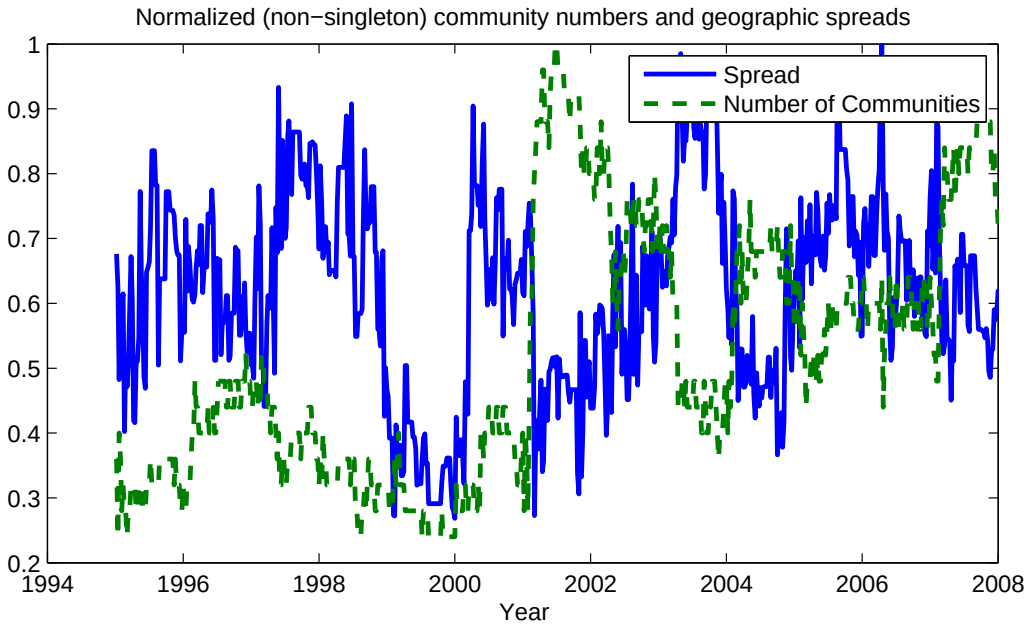


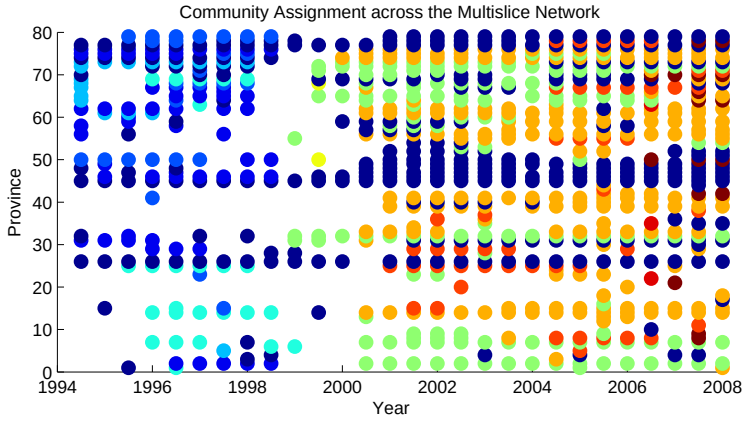
Figure 6: Number of (non-singleton) detected communities, and average geographical spread of communities versus time. We normalize both measures to lie between 0 and 1.

significance of this change. Nevertheless, Figure 6 is strikingly suggestive of such an interpretation, but these results would be more convincing if we could show that the decrease in spread is significant after incorporating the fact that some nodes are closer together than others (see Figure 5).

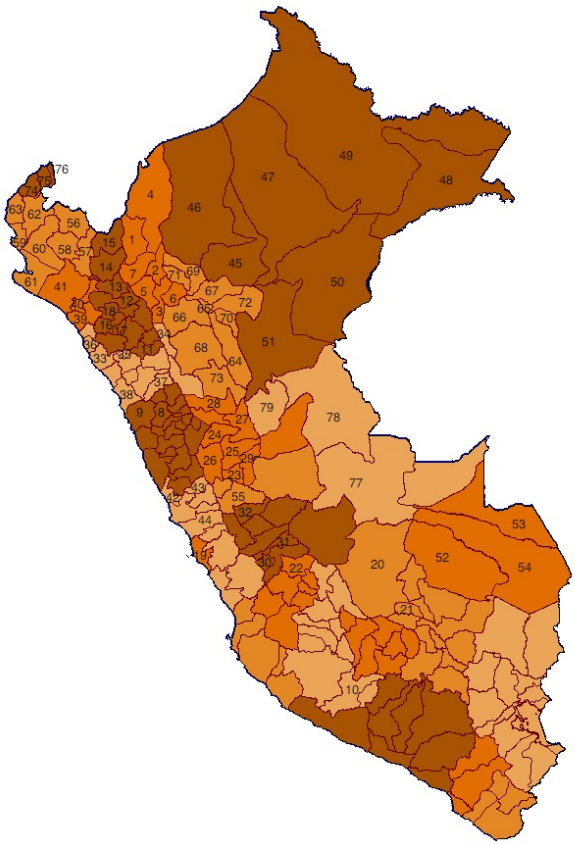
3.3 Multislice network sequence

For further investigation of the global properties of the network, we use the multislice framework of Sections 2.1.1 and 2.2.2. The adjacency matrix of each slice is formed by the definition (1)–(2), where we now choose $\Delta = 26$. We take the slices $A(26), A(52), \ldots, A(780)$, so that the time windows do not overlap, and we consider only the large connected component of the network at each stage. For each pair of consecutive slices, we link nodes that appear in both with an interconnection strength of $\omega \geq 0$.

Figure 7(a) shows the result of the Louvain algorithm on the multislice network described above for resolution and inter-slice parameters $\gamma = \omega = 2$. It appears that the network undergoes a clear transition in community structure just after the year 2000 begins (of



(a) Community assignment in the multislice network. Each colour represents a community (detected with parameter values $\gamma = \omega = 2$).



(b) Numbered nodes arranged on a map of Peru (taken with permission from [16]).

Figure 7: Community assignment of numbered provinces.

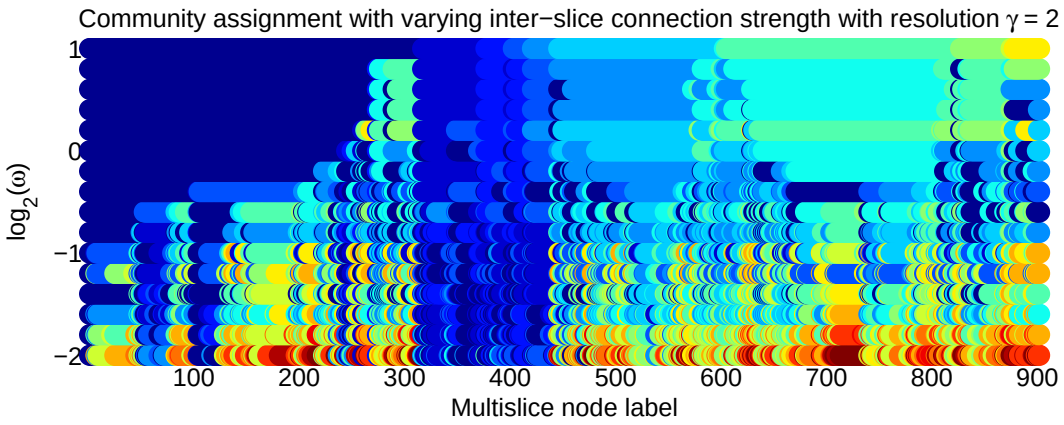
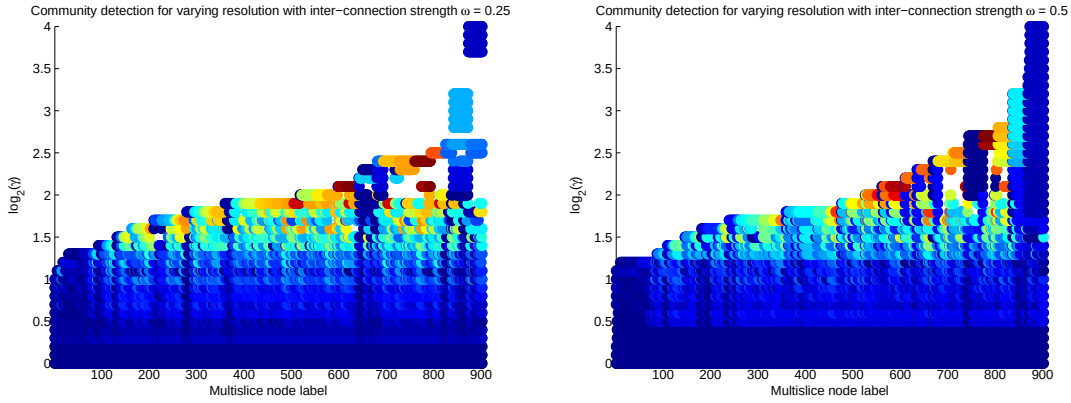


Figure 8: Nodes of the multislice network (i.e. province-times) are arranged along the x-axis, and $\log_2(\omega)$ varies along the y-axis. Each colour represents a different community, detected at varying values of ω, for $\gamma = 2$. Nodes are ordered arbitrarily.

course, this observation is for only one choice of parameter values). It is also apparent that the network itself becomes much larger at this point, as infections appear in more provinces. However, to get meaningful results, we require a systematic approach to ensure that interesting features are robust to different values of the parameters γ and ω. This report will use the "paintdrip" plots of Lewis et al. [21] for this.

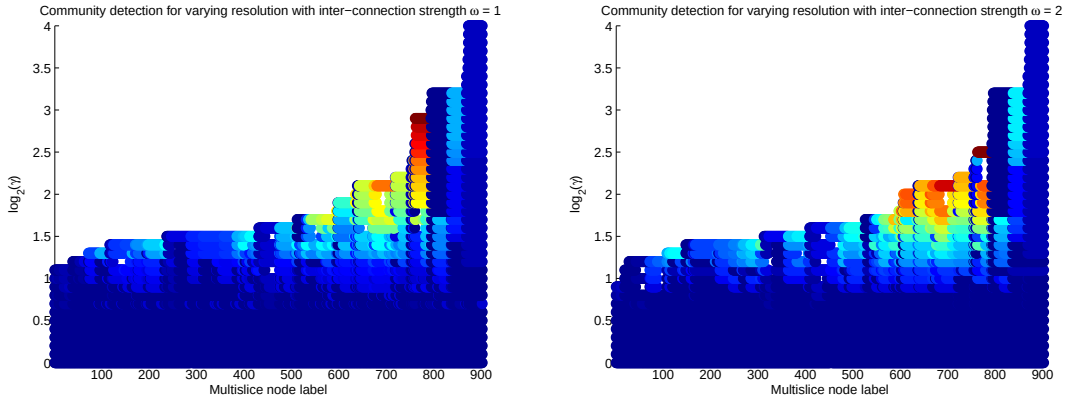
Figure 8 shows such a plot: the top line of points in this figure, where $\log_2(\omega) = 1$, is a reordering of all 902 nodes found in Figure 7(a) into one row (sorted by community assignment). Each row of the figure corresponds to one value of ω and represents the same 902 nodes in the same order (again coloured by community structure for that value of ω). In this figure, $\gamma = 2$ is constant. Figure 9 follows the same procedure (additionally removing any communities of size less than 30 nodes) for varying γ, where each panel corresponds to a fixed ω. The ordering of nodes in the five plots of Figures 8 and 9 is not the same.

These plots bring out the nodes that remain more cohesively grouped together than others for multiple parameter values. For example, the nodes labelled from around 320 to 440 in Figure 8 remain distinct from the progressively more jumbled communities of the other nodes as ω decreases. Figure 10 shows these nodes as node-times. We can clearly see a robust community assignment consisting of most of the pre-2000 nodes. Note that Figure 8 does not put all of nodes 320–440 in the same community, but their various communities



(a) Inter-slice strength $\omega = 0.25$.

(b) Inter-slice strength $\omega = 0.5$.



(c) Inter-slice strength $\omega = 1$.

(d) Inter-slice strength $\omega = 2$.

Figure 9: Nodes of the multislice network (i.e. province-times) are arranged along the x-axis, and $\log_2(\gamma)$ varies along the y-axis. Each colour represents a different community (detected at varying values of γ) for the respective values of ω. Nodes are ordered arbitrarily, and the order is different for each ω. Communities of size less than 30 nodes are not plotted.

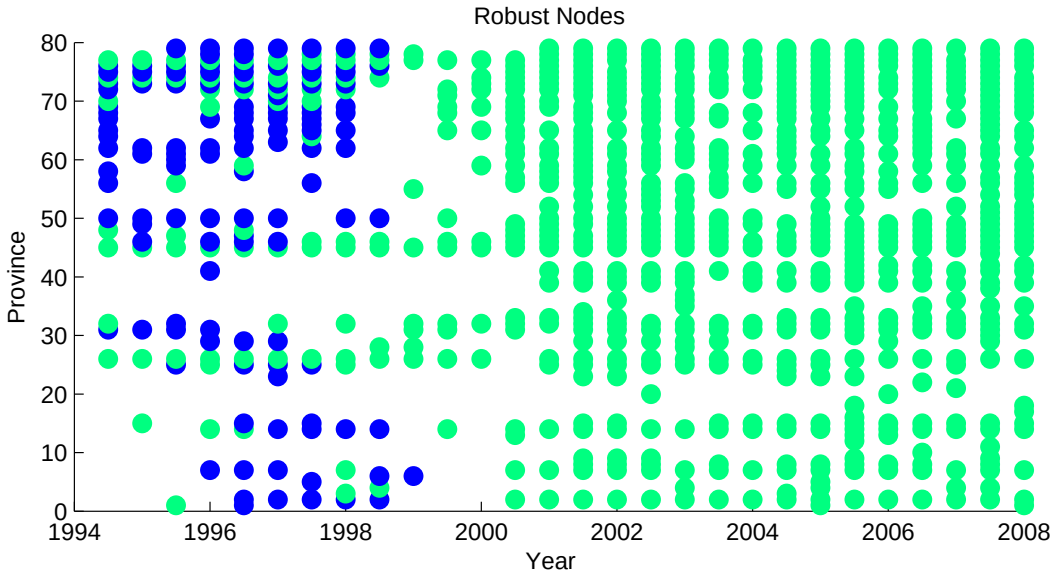


Figure 10: Nodes labelled 320–440 in Figure 8 (a grouping robust to $\omega \in [0.25, 2]$) plotted in blue compared to the other nodes in green.

remain distinct from the rest of the network.

We now fix ω and vary γ. Figure 9 shows several "paintdrips". Again, each of the four panels show a cohesive group that persists for all values of γ considered, although Figure 9(a) has communities that are split into groups too small to be plotted as γ increases, before joining together again. Again, these groups all correspond to pre-2000 nodes (i.e. before the main epidemic). After the epidemic, communities appear to be less robust to parameter variation.

3.4 Comparing multislice communities to manual partitions

3.4.1 A critical time point

To verify that community structure changes significantly after 2000, we now quantify how well the detected communities partition the multi-slice network into two distinct time periods. We use the methods described in Section 2.4.1 to compare the algorithmically-determined partition of the multi-slice network with partitions into two time intervals $[t \leq t_c,\ t > t_c]$, defined by a critical point t_c. We let t_c vary across all values of t (for

$\Delta = 26$, this gives $t_c = 1, \ldots, 29$) and observe for each how well the detected community partition agrees with the $[t \leq t_c,\ t > t_c]$ partition.

Figure 11 shows how the z-Rand score varies by partition time t_c and by inter-slice strength ω for $\gamma = 2$. In particular, Figure 11(a) shows that the peak in z_R at $t_c = 2003.5$ is robust to all values of $\omega \in [0, 3]$. Figure 11(b) zooms into the smaller values of ω; it is apparent that for any $\omega > 0$ (however small) there is an increase in z_R at this value of t_c. As noted in Section 2.1.1, $\omega = 0$ corresponds to no inter-linking of slices. Hence, as soon as slices are linked, we see a signal showing that the detected communities partition the time-dependent network into $[t < 2004,\ t \geq 2004]$.

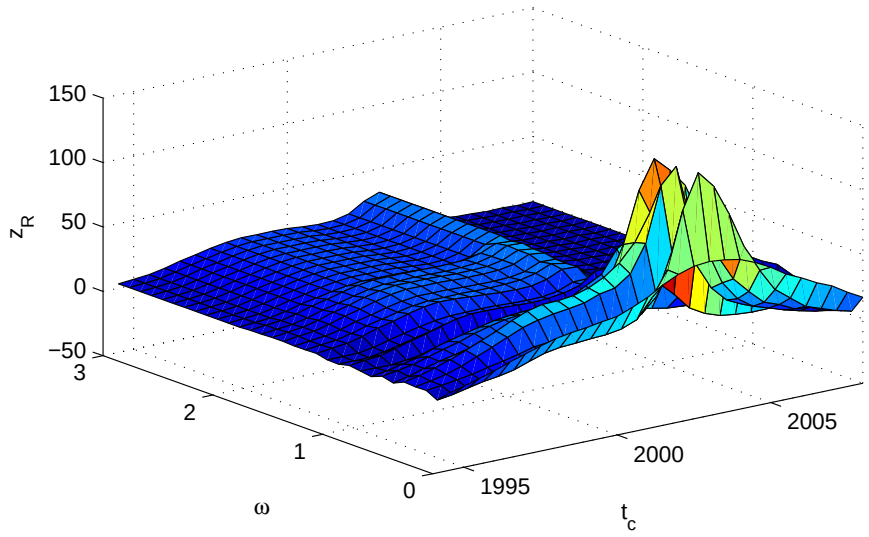
Observe in Figure 1(a) that 2004 is some time after the large epidemic. In fact, by inspection of Figure 7(a) and from the results of Section 3.3, we would expect 2000, not 2004, to form the dividing year. We suspect that, because half of the node-times are before 2004, this is skewing the results such that 2004 will naturally form a better partition. A topic of further investigation is to determine if this is indeed a bias in methodology (and to account for it, if it is).

3.4.2 Geographical partitioning

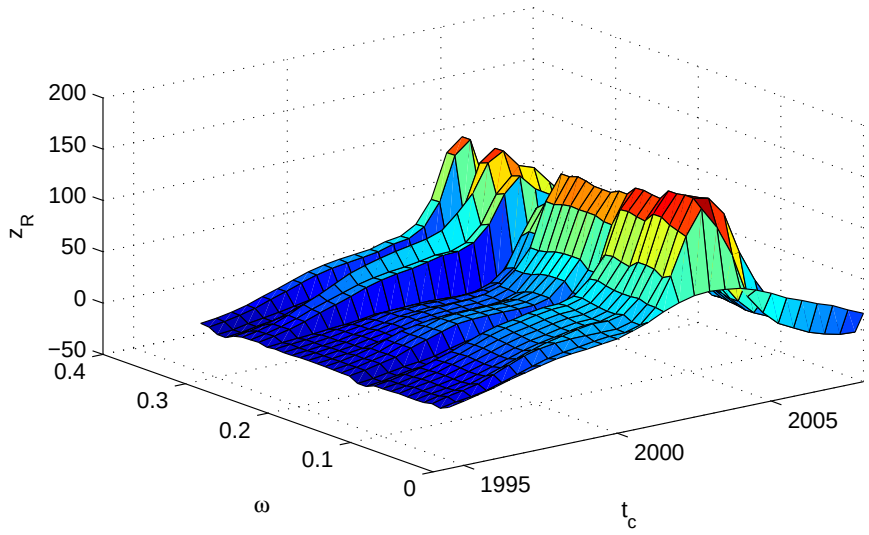
One can also manually partition the node-times using the geographical properties of the nodes. Supplied with the data was a topographical classification of each province into "Mountain", "Coast", and "Jungle", with a further classification of the first two into "north", "central", and "south". Thus we partitioned the nodes into seven communities. Another manual partition we considered came from the geopolitical divisions of Peru: the provinces group together into regions[3].

Figure 12 shows how z_R (see Section 2.4.1) varies for different values of ω. We use the partition into pre- and post-2004 nodes (Section 3.4.1) and the geopolitical and topographical partitions discussed above. For $\omega \geq 1$, the z-Rand scores show that all three manual partitions are significantly correlated with the community structure. The most important is region, which is followed by pre/post-2004, and then topography.

[3]See `http://en.wikipedia.org/wiki/Provinces_of_Peru` for the details of the grouping of provinces into regions.



(a) Variation of z-Rand score z_r with partition time t_c for inter-slice strength parameters $0 \leq \omega \leq 3$.



(b) Variation of z-Rand score z_r with partition time t_c for inter-slice strength parameters $0 \leq \omega \leq 0.3$.

Figure 11: Agreement between algorithmically-determined community partition and partition by a critical time point. We measure this using the z-Rand score z_r (7).

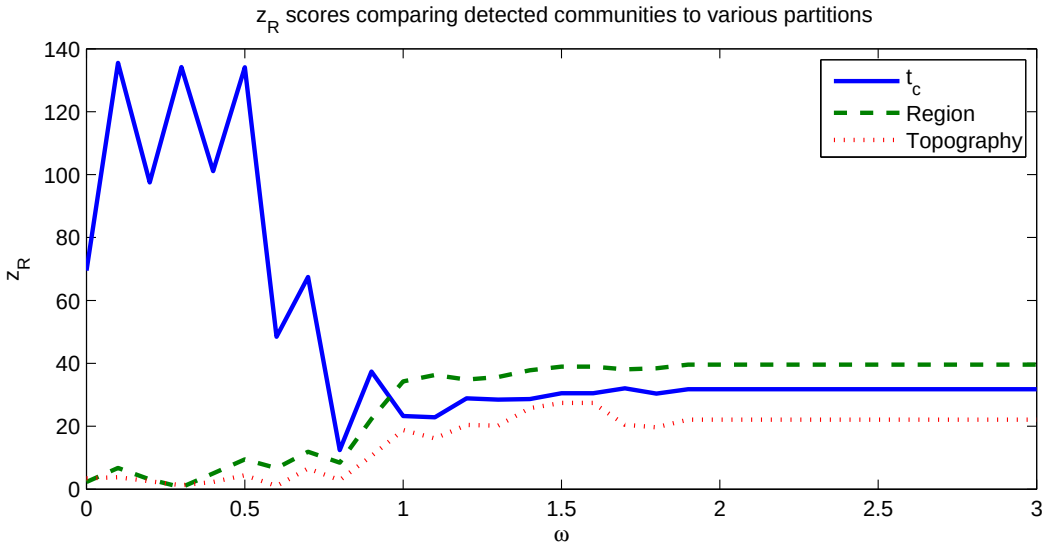


Figure 12: The z-Rand score z_R comparing the algorithmically-detected community partition of the multislice network (with $\gamma = 2$) to the manual partitions into geopolitical regions, topographic classification, and $[t < 2004, t \geq 2004]$

That region is more important than topography is expected: regions are, typically, topographically homogenous (i.e. all jungle, all coast, etc.) but also composed of neighbouring provinces. Therefore, because they share topography *and* proximity, and we have already discovered that proximity is important to community structure, regions are in much greater agreement with the detected communities than topography.

4 Discussion and Conclusions

4.1 Discussion

The results in Section 3 are indicative of connections between the community structure of the network and the time series to which the network corresponds.

Treating the network as time-dependent, we have shown that optimal modularity (calculated by the Louvain algorithm) drops significantly at the peak of the 2000–2001 epidemic. Recall that a decrease in modularity corresponds to a less clear partition into communities. Figure 13 shows a simultaneous rise in total edge weight, corresponding to a rise in aggre-

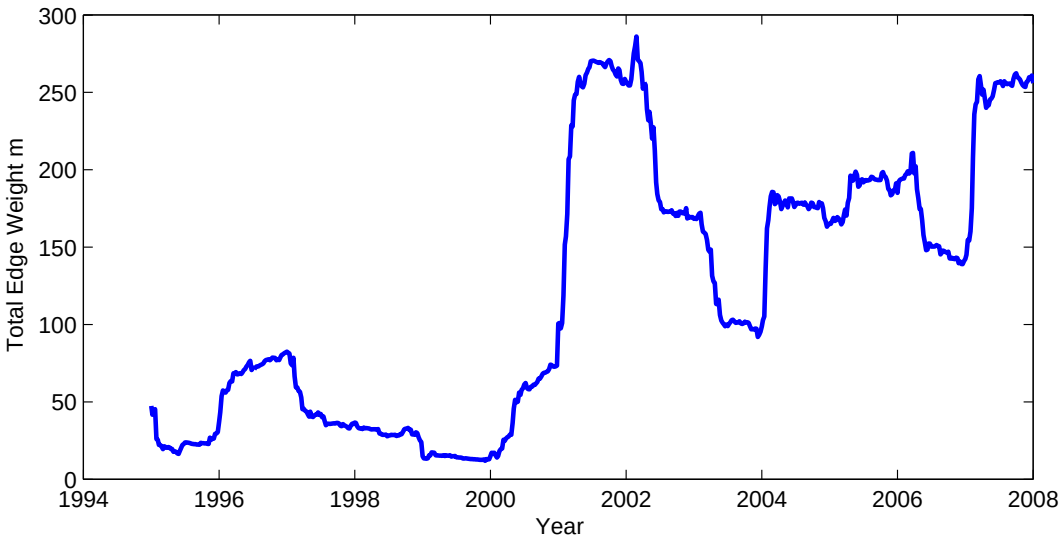


Figure 13: Total edge weights for $A(t)$ with $\Delta = 52$.

gate correlation. Thus, the drop in modularity seems to suggest a greater inter-connection between communities, where those which were once distinct become less so (rather than the other possibility, in which communities become less densely intra-connected).

We also found a significant drop in the mean community spread during 1999. We tentatively interpreted this result to suggest that the community structure becomes extremely localized geographically in the year leading up to the large epidemic starting in 2000. If further work confirms this result as a network feature robust to parameter changes, other formulations of community spread, and other network constructions, then it supplies a potentially retrodictive signal for the onset of an epidemic.

This calculation is also an example of an approach combining spatial with temporal features. The results of Section 3.4 have the weakness of only considering either spatial or temporal features: partitioning the node-times arranged in Figure 7(a) into either horizontal or vertical groups. Ideally, one would manually cluster in both dimensions to attempt to confirm our hypothesis of a major change in the importance of node proximity after 1999.

4.2 Further work

The results discussed above require much more testing if they are to be conclusive. As well as checking robustness with respect to changes in parameter values, we require robustness to alternative definitions of adjacency weights, modularity, community detection algorithms, and so on. Below we describe some ways of of trying to check the robustness of our results.

4.2.1 Introducing time lag

The adjacency given by (2) is symmetric, and it compares the correlations of time series pairs over the same time window. However, a disease takes time to propagate through space. If we offset one time series by some lag parameter λ, then we might be able to capture causality between the infection patterns in each pair of provinces. This change would make the resulting network directed.

For each i, define the lagged time series X_i^λ by $X_i^\lambda(t) = X_i(t-\lambda)$ with time lag λ. Rather than substituting β_{ij} from (1) into equation (2), we use an altered $\beta_{ij}(t)$ defined as

$$\beta_{ij}(t) = \sum_{\tau=t-\Delta+1}^{t} X_i^\lambda(\tau) X_j(\tau) \tag{8}$$

for $t \in \{\Delta + \lambda, \Delta + \lambda + 1, \ldots, T\}$.

This yields a directed network in which a large value of A_{ij} implies that the time series X_i somehow causes the behaviour of X_j. Of course, choosing to follow this method requires us to decide the value of the time lag λ. Equation (8) assumes a constant global parameter λ for simplicity only. There is no reason why the time lag from i to j should be the same as from i to k, or even from j to i. An appropriate choice of λ should be informed by the characteristic speeds of dengue fever spread but could also perhaps be inferred from the data using cross-correlation methods [22].

4.2.2 Incorporating spatial effects in the null model

Modifications of the standard null model have recently been developed in order to deal with networks embedded in space [23, 24]. Closely-located, larger nodes are more likely to

be connected, so the null model should take this into account. One thereby hopes to take into account what features of detected communities are simply artefacts of their nodes' proximity and to detect structures that are not due simply to co-location. We have found that proximity is extremely important in our data, but this might mask other structural factors that we have not identified.

Expert et al. [23] recently proposed a null model that is given in the form

$$P_{ij} = S_i S_j f(d_{ij})$$

for pairs of nodes i, j distance d_{ij} apart. Here S_i is some measure of the size or importance of node i and the function f describes the changing likelihood of edge weight by distance.

4.2.3 Rubella data

Also available is similar data on rubella infections from 1997 to 2009 across more provinces (by province, week, and age). Any future work will be far more convincing if the methods can also be applied successfully to both data sets. Some parameter choices will need changing: for instance, the time lag described in 4.2.1 above will be different if the characteristic timescales of the diseases are different. This is likely, as rubella is transferred through airborne droplets rather than by mosquito [25].

4.3 Conclusions

This report is an exploratory step in investigating community structure in networks of correlated time series of dengue fever infections. We have uncovered several promising directions for further work that have generated some testable hypotheses. We have confirmed and tested the robustness of the main result of Ng [16], that modularity drops during the widespread epidemic as communities become more inter-connected. We have also shown that the mean spatial spread of the detected communities drops significantly in the year leading up to the main epidemic, suggesting a potentially retrodictive signal of the dengue epidemic.

We also used a multislice framework for linking network slices corresponding to consecutive, non-overlapping time windows. We have used methods of measuring the agreement between

two partitions to show how individual geographical and temporal properties of the network correlate with the detected communities. These communities tend to group provinces belonging to the same region and into groups pre- or post-2004. This latter observation is at odds with our findings about the major shift in community structure at the onset of 2000, and it requires further investigation.

Acknowledgements

Many thanks to my supervisors: Mason Porter, Elizabeth Leicht and David Gavaghan; also to Yulian Ng whose work [16] formed the starting point of this report. Code, thanks to Peter Mucha, was downloaded from NetWiki, and Jukka-Pekka Onnela kindly allowed the replication of Figure 3. Thanks also to Sue Bird of the RSL for the geographical data, and Gerardo Chowell for the main data set.

References

[1] M.G. Guzmán, S.B. Halstead, H. Artsob, P. Buchy, J. Farrar, D.J. Gubler, E.A. Hunsperger, A. Kroeger, H.S. Margolis, E. Martínez, et al. Dengue: a continuing global threat. *Nature Reviews Microbiology*, 8:7, 2010.

[2] L.S. Lloyd. Best practices for dengue prevention and control in the Americas. Strategic report, Environmental Health Project Office of Health Infectious Diseases and Nutrition. US Agency for International Development, Washington DC, 2003. URL `http://www.ehproject.org/PDF/Strategic_papers/SR7-BestPractice.pdf`.

[3] G. Chowell, C.A. Torre, C. Munayco-Escate, L. Suárez-Ognio, R. López-Cruz, J.M. Hyman, and C. Castillo-Chávez. Spatial and temporal dynamics of dengue fever in Peru: 1994–2006. *Epidemiology and Infection*, 136(12):1667, 2008.

[4] S.M. Smith, K.L. Miller, M. Webster, G. Salimi-Khorshidi, C.F. Beckmann, T.E. Nichols, J.D. Ramsey, and M.W. Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54 (2):875, 2011.

[5] D.S. Bassett, N.F. Wymbs, M.A. Porter, P.J. Mucha, J.M. Carlson, and S.T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641, 2011.

[6] D.J. Fenn, M.A. Porter, M. McDonald, S. Williams, N.F. Johnson, and N.S. Jones. Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis. *Chaos*, 19(3):033119, 2009.

[7] P.J. Mucha and M.A. Porter. Communities in multislice voting networks. *Chaos*, 20(4): 041108, 2010.

[8] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876, 2010.

[9] A.L. Traud, L.D. Kelsic, P.J. Mucha, and M.A. Porter. Comparing community structure to characteristics in online social networks. *SIAM Review*, 53(3):526, 2011.

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

[11] M.A. Porter, J-P. Onnela, and P.J. Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082, 1164, 2009.

[12] T.M.J. Fruchterman and E.M. Reingold. Graph drawing by force-directed placement. *Software: Practice & Experience*, 21(11):1129, 1991.

[13] B. Hogan. Visualizing and interpreting Facebook networks. In B. Shneiderman and M.A. Smith, editors, *Analyzing Social Media Networks With NodeXL: Insights From a Connected World*, page 165. Elsevier Science, 2010.

[14] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75, 2010.

[15] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[16] Y. Ng. Community detection in relation to the spread of epidemics. Master's thesis, University of Oxford MMath, 2011. URL `http://people.maths.ox.ac.uk/porterm/research/yuli_Dissertation_final.pdf`.

[17] R. Lambiotte, J-C. Delvenne, and M. Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv:0812.1770*, 2009.

[18] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[19] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172, 2007.

[20] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008(10):10008, 2008.

[21] A.C.F. Lewis, N.S. Jones, M.A. Porter, and C.M. Deane. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100, 2010.

[22] P. Denbigh. *System Analysis & Signal Processing*. Addison–Wesley Longman, 1998.

[23] P. Expert, T.S. Evans, V.D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19): 7663, 2011.

[24] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S.H. Strogatz. Redrawing the map of Great Britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.

[25] P.A. Tookey and J.M. Best. Rubella. In D.A. Warrell, T.M. Cox, and J.D. Firth, editors, *Oxford Textbook of Medicine*, page 561. Oxford University Press, 5th edition, 2010.