Short communication:

Tools to study trends in community structure: application to fish and livestock trading networks

Darren Michael Green[a*], Marleen Werkman[a],
Lorna Ann Munro[b], Rowland Raymond Kao[c],
István Zoltán Kiss[d] & Leon Danon[e,f]

January 11, 2011

[a] *Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling, Stirlingshire FK9 4LA, UK.*
[b] *Marine Scotland Marine Laboratory, PO Box 101, 375 Victoria Road, Aberdeen, AB11 9DB, UK.*
[c] *Institute of Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Glasgow, G61 1QH, UK*
[d] *Department of Mathematics, University of Sussex, Falmer, Brighton BN1 9RF, UK.*
[e] *Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.*
[f] *Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston 02115, USA.*
[*] *Author for correspondence. email: darren.green@stir.ac.uk*

Abstract

Partitioning of contact networks into communities allows groupings of epidemiologically related nodes to be derived, that could inform the design of disease surveillance and control strategies, e.g. contact tracing or design of 'firebreaks' for disease spread. However, these are only of merit if they persist longer than the timescale of interventions. Here, we apply different methods to identify concordance between network partitions across time for two animal trading networks, those of salmon in Scotland (2002-4) and livestock in Great Britain (2003-4). Both trading networks are similar in that they moderately agree over time in terms of their community structures, but this concordance is higher – and therefore community structure is more consistent – when only the 'core' network of nodes involved in trading over the whole time series is considered. In neither case was higher agreement found between partitions close together in time. These measures differ in their absolute values unless appropriate standardisation is applied. Once standardised, the measures gave similar values for both network types.

Keywords aquaculture; community; network; graph; movements

1. Introduction

Movement of farmed animals is an important route for disease spread in what are highly structured industries. For example, sheep, cattle, and pigs were all involved in the UK epidemic of foot-and-mouth disease in 2001 (Shirley & Rushton, 2005), and movements of salmon were involved in the spread of infectious salmon anaemia (Murray et al. 2002). A network representation, where farm sites are represented by 'nodes', and potentially infectious contact by directed 'arcs' or undirected 'edges' is a powerful tool for studying the potential for disease spread and control (for a review of networks in preventive veterinary medicine, see Martínez-López et al., 2009).

Network communities represent partitions of nodes with a high level of within-partition connectivity (for a review, see Fortunato, 2010). In a strongly community-organised network, contact between communities may be relatively weak, and community algorithms can provide us with natural groupings of epidemiologically related nodes, derived from the network itself rather than artificially imposed. Uncommon inter-community links might furthermore be considered as potential targets for proactive targeted surveillance, or reactively in disease control (Kao et al. 2006; Green et al. 2009; Salathé & Jones, 2010). That is to say, removing the disease transmission risk of such contacts could reduce the size of potential epidemics by creating 'firebreaks', particularly where these contacts are long distance. However, these analyses are only of merit if partitions can be used predictively; that is, if community structure changes more slowly than we collect data in order to inform surveillance or disease control strategy.

A key problem here is that objective measures of the rate of change of large-scale network structure are not clearly defined, nor how large a change must be to heavily compromise disease control strategies. In this short paper, we consider the first part of this question, by comparing different methods for determining how network community structures change, or not, over time. We apply these methods to two movement networks of farmed animals, to investigate whether networks closer in time have more similar network structure. The two networks are that of live Atlantic salmon *Salmo salar* movements within Scotland 2002 to 2004, and that of livestock (pigs, sheep, cattle) in Great Britain for 2003 to 2004.

2. Method

2.1. Data

The network of live fish movements in Scotland has been described for salmonid species (brown trout *Salmo trutta*, rainbow trout *Onchorhynchus mykiss*, Atlantic salmon *S. salar*) by Green et al. (2009) and Munro & Gregory (2009). Here, we extend and refine their analysis to a three-year dataset of Atlantic salmon alone for 2002-4. In brief, these data comprise movements of live fish (egg to adult) between registered sites in Scotland, where paper records of both off and on movements were legible and in agreement. Data are held by the Fish Health Inspectorate of Marine Scotland.

For the network of livestock movements, the partitions used here are derived from the data extract used by Kao et al. (2006). Their data set comprised data from January 2003 to December 2004 for cattle (Cattle Tracing System) and sheep and pigs (Animal Movements Licence System, England and Wales; Scottish Agricultural Movements System, Scotland). A full description is given by Kao et al. (2006).

Both data sets provide source and destination premises, species and number moved, and date. Data were segregated into time periods (years for fish, four-week periods for livestock), with each network described by an adjacency matrix A. Here, $A_{ij} = 1$ implies movement of animals from node (site) i to node j (zero for no contact). The number of in and out connections for node i are given by k_i^{out} and k_i^{in}, the total number of nodes by n, and the total number of arcs by M.

2.2. Graph partitioning

Communites were identified for the two datasets using related partitioning algorithms. For the fish network, the measure of community fit used is that defined by

$$Q = \frac{1}{M} \sum_{i,j} \left(A_{ij} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{M} \right) [X_i = X_j]$$

where X_i is the community 'label' of node i. The Iverson bracket $[\cdot = \cdot]$ returns one if the condition inside is true, and zero otherwise. This formulation – as described by Kao *et al.*

(2006) and Leicht & Newman (2008) – accounts for the strong directed nature of the fish network. Higher Q indicates a larger fraction of arcs within communities. 'Lone' nodes in a network, with no movements during the period of interest, gain a unique label. However, it could be argued that without network activity, such nodes are not part of the network at all (further discussed in the Results section). The livestock network was treated similarly, except that the partition data available were based on undirected edges.

For both systems, we employ a 'hill-climbing' algorithm (Newman, 2004; Danon et. al. 2005). This begins by assigning each node a unique community label $X_i = i$. Each possible merger of two communities is considered, with that providing the largest positive change in Q accepted. This step is repeated until a maximum Q is reached, for which the corresponding community assignments are taken as the 'best fit'. Though other algorithms may find improved partitions, this one has the benefit of being practicable on very large networks such as that for livestock movements.

2.3. Entropy measures

Borrowing concepts from information theory, entropy-based measures can be used to compare multiple partitions of the same network (Strehl et al. 2002; Vinh et al. 2009). Beginning with two vectors X and Y containing community labels for two partitions, two vectors U and V are built containing the number of nodes present in each community in X and Y: $U_u = \sum_i [X_i = u]$; $V_u = \sum_i [Y_i = u]$. Also, an $n \times n$ matrix is defined containing the frequency combinations of communities in both X and Y: $W_{uv} = \sum_{i,j} [X_i = u][Y_j = v]$. For two networks with congruent partitions, this matrix contains only a single non-zero element in each row and column. The Shannon entropy (a measure of the information content of a dataset) is calculated for the partitions of each network ($H(U)$ and $H(V)$), and that of the matrix of community combinations, the 'joint entropy' $H(U,V)$.

$$
\begin{aligned}
H(U) &= -\sum_u \frac{U_u}{n} \log \frac{U_u}{n} \\
H(U,V) &= -\sum_{u,v} \frac{W_{uv}}{n} \log \frac{W_{uv}}{n}
\end{aligned}
$$

Choice of logarithm base does not affect the end result below, and by definition, $0 \times \log 0 = 0$. The mutual information $I(U,V) = H(U) + H(V) - H(U,V)$ then measures the amount of information shared between the two partitions – and thus their similarity – with a lower bound of zero, but no upper bound. For comparison between networks, a normalised measure of similarity is required. A simple approach is to scale I by its maximum potential value (it cannot exceed the minimum of $H(U)$ and $H(V)$), giving the normalised mutual information $0 \leq NMI_1 \leq 1$:

$$NMI_1 = \frac{I(U,V)}{\min(H(U), H(V))}.$$

Alternatively, we can scale by the geometric mean of these two quantities (Strehl et al. 2002), $0 \leq NMI_2 \leq NMI_1$:

$$NMI_2 = \frac{I(U,V)}{\sqrt{H(U)H(V)}}.$$

For correlation coefficients such as Pearson's or Spearman's, a value of zero is obtained where there is no relationship, i.e. under the null hypothesis. However here, under a reasonable null hypothesis that communities are assigned randomly, the expectation of $I(U,V)$, $E_0(I(U,V))$ is not generally zero and depends upon the size distribution of communities (Vinh et al. 2009). A further approach is to normalise I against this expectation, providing the adjusted mutual information AMI (Vinh et al. 2009), with a maximum of one, zero under the null hypothesis, and negative where there is less agreement between network communities than would be expected by chance.

$$AMI = \frac{I(U,V) - E_0(I(U,V))}{\min(H(U), H(V)) - E_0(I(U,V))}$$

This definition of AMI is similar in form to that of Cohen's Kappa statistic, and has a lower value than NMI_1 except where $E_0(I(U,V))$ is vanishingly small. Vinh et al. (2009) suggest using $\max(\cdot,\cdot)$ not $\min(\cdot,\cdot)$, however the $\min$ term has more in common with the formula for NMI_2 above. Unlike correlation coefficients, its minimum possible value is not defined to be -1. A permutation test was employed to determine the mean and distribution of $E_0(I)$ allowing for calculation of AMI and its significance. One of the vectors X and Y is repeatedly shuffled, removing association between the node labels in X and Y. On each

permutation, $I(U,V)$ is recalculated. The original $I(U,V)$ can be compared with the distribution of these permuted versions.

2.4. Pair-based measures

Pairs of nodes can be examined with respect to whether or not they are in the same communities. Pairs of nodes that were in the same community in the two partitions were counted: $x = \sum_{i,j\neq i}[X_i = X_j]$ and $y = \sum_{i,j\neq i}[Y_i = Y_j]$, as well as pairs that were in the same community in both partitions: $a = \sum_{i,j\neq i}[X_i = X_j][Y_i = Y_j]$, or in different communities in both: $b = \sum_{i,j\neq i}[X_i \neq X_j][Y_i \neq Y_j]$.

From these values, the probability that a pair of nodes present in the same community in partition X are also in the same community in partition Y was calculated: $P(\text{pair in } Y|\text{pair in } X) = \frac{a}{x}$. However, this metric is not necessarily symmetric with respect to X and Y, unlike the earlier measures. Instead, the geometric mean of both possible probabilites was taken: $\bar{P} = \frac{a}{\sqrt{xy}}$ (Wallace, 1983; quoted in Meilă, 2007). These probabilities benefit from being easily interpretable. A further pair-based measure of clustering similarity, the Rand index $R = \frac{a+b}{n(n-1)}$ (Rand, 1971), was also calculated.

Again, these measures do not equal zero under the null hypothesis that the two partitions are independent. The statistical significance of both was determined through a permutation test and – as with the mutual information – standardised according to $[\phi - E_0(\phi)]/[1 - E_0(\phi)]$, where ϕ is the measure of interest, giving an adjusted Rand index AR and an adjusted probability related to $\bar{P}$, $A\bar{P}$.

3. Results and Discussion

For the salmon movement network ($n = 502$), the unadjusted indices NMI_1, $\bar{P}$ and R gave numbers of different magnitude, despite their apparent normalisation (Table 1). This reflects their different values under their null models. A $\bar{P}$ index of ~ 0.3 is easily interpretable as the proportion of same-community node pairs that persist across both partitions. Once 'adjusted', the range of values was narrower, with the pair-based indices giving almost coincidental values (Table 1). This coincidence was also evident for the

livestock network, thus in Figure 1 only the index R is shown. The null model for the permutation test was amended for the fish network to account for variation in the activity of nodes between years: Those nodes with no links were not considered during the reshuffling process to prevent their single-node communities being spuriously reassigned to other nodes.

For the much larger livestock network ($n = 141607$; see supplementary animation), networks were built from four-week periods of data. As with the fish network, all correlations were statistically significant ($P < 0.05$). These networks show a marked seasonal pattern (Kao et al. 2006) with a higher density of arcs due to an autumn peak in sheep trading. This seasonality was still noticeable despite normalisation as a peak in AMI values for networks 13 four-week periods (i.e. one year) apart (Fig. 1). Though this peak may represent a real similarity in the trading structure at particular times of year, Meilă (2007) raises concerns over the use of adjusted indices for comparison purposes where the baseline and actual values may vary non-linearly.

To explore this further, we accounted for seasonality in trading volume by considering only a 'core' sub-network of nodes that were active in each of the 25 networks examined ($n = 6424$). The AMI values together with the Rand index R are shown for this core network in Figure 1, showing close agreement between the three statistics and much reduced seasonality in community structure. Taking the 'core' network of $n = 208$ nodes for the salmon network, a similar result is found as for the whole fish network, albeit with higher values (Table 1).

Though both sets of networks show moderate agreement between partitions at different time points, in neither case was a higher agreement between networks closer in time apparent. One possible explanation of this is that there are no significant long-term trends in community structure for either network, or that any such trends operate on timescales either longer or shorter than examined in this study. There may also be other trends and patterns within the data that remain observed. For example, the partitions above are not absolutes: different measures and algorithms could produce different groupings. Also, no allowance is made in this approach for the potential for sub- and super-community network structure (Kao et al. 2006; Green et al. 2009).

The unadjusted indices give a wide selection of values for the same network, however once adjusted they are more similar. Those for $\bar{P}$ coincided with R. However, whether this is in general the case or is network dependent remains to be established. The computational efficiency of the measures varies: Despite their apparent complexity, the entropy-based measures are relatively fast to compute, particularly for large networks, since they do not rely on counting edges.

4. Conclusions

In conclusion, for both networks a significant and non-trivial level of concordance between network partitions over time was seen. Dissimilarity in partitions, however, appears to represent random variation rather than decay in partition similarity over time for both networks. Characterising the way networks change over time remains a challenging problem. Our results suggest that despite the fact that many features change, a large part of the intermediate structure is conserved over time, particularly in the core network. Nevertheless, the how stable a contact network must remain over time to be epidemiologically useful for disease surveillance and control remains to be explored, potentially through simulation of dynamic disease control measures on dynamic network epidemic models.

Conflict of interest

None declared.

Acknowledgements

With thanks to the Fish Health Inspectorate for providing access to the movement records, and to Malcolm Hall for comments on the manuscript. DMG and MW are supported by Marine Scotland.

References

Danon, L., Duch, J., Díaz-Guilera, A., Arenas, A., 2005, Comparing community structure identification, J. Stat. Mech. P09008

Fortunato S., 2010, Community detection in graphs. Physics Reports, 486, 75 – 174.

Green, D.M., Gregory, A., Munro, L.A. 2009. Small- and large-scale network structure of live fish movements in Scotland. Prev. Vet. Med. 91, 261 – 269.

Kao, R.R., Danon, L., Green, D.M., Kiss, I.Z. 2006. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. Proc. R. Soc. B 273, 1999 – 2007.

Leicht, E.A., Newman, M.E.J. 2008. Community structure in directed networks. Phys. Rev. Lett. 100, 118703.

Martínez-López, B., Perez, A.M., Sánchez-Vizcaíno, J.M., 2009. Social network analysis. Review of general concepts and use in preventive veterinary medicine. Transboundary and Emerging Diseases 56, 109 – 120.

Meilă, M., 2007. Comparing clusterings – an information based distance. Journal of Multivariate Analysis 98, 873 – 895.

Munro, L., Gregory, A. 2009. Application of network analysis to fish movement data. J. Fish Dis. 32, 641 – 644.

Murray, A.G., Smith, R.J., Stagg, R.R., 2002. Shipping and the spread of infectious salmon anemia in Scottish aquaculture. Emerging Infectious Diseases 8, 1 – 5.

Newman, M.E.J., 2004. Fast algorithm for detecting community structure in networks, Phys. Rev. E. 69, 066133.

Rand, W.M. 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66, 846 – 850.

Salathé M., Jones, J.H. 2010. Dynamics and control of diseases in networks with community structure. PLoS Computational Biology 6, e10000736.

244 Shirley, M.D.F., Rushton, S.P. 2005. Where diseases and networks collide: lessons to be
245 learnt from a study of the 2001 foot-and-mouth disease epidemic. Epidemiol.
246 Infect. 133, 1023 – 1032.

247 Strehl, A., Ghosh, J., Cardie, C. 2002. Cluster ensembles – A knowledge reuse framework
248 for combining multiple partitions. Journal of Machine Learning Research 3, 583 –
249 617.

250 Vinh, N.X., Epps. J., Bailey, J. 2009. Information theoretic measures for clustering
251 comparison: Is a correction for chance necessary? Proceedings of the 26th
252 international conference on machine learning. Montreal, Canada, 2009.

253 Wallace, D.L. 1983. Comment. J. Amer. Statist. Assoc. 78, 569 – 576.

254 **Caption for figure**

Figure 1: Measures of agreement for network communities based on livestock movements of cattle, sheep and pigs in Great Britain (2003-4). Means and standard errors of measures for all possible combinations of four-week periods are shown, stratified by time difference in periods (1 to 24). Shown are the adjusted Rand index R (solid line) and entropy measure AMI (dashed line) for the 'core' network, with AMI for the entire network (dotted line). Probability $\bar{P}$ coincided with R and is not shown.

255 **Caption for animation file [electronic]**

256 Animation of livestock movement network communities in Great Britain, 2003-4; nodes
257 (sites) sharing the same community label are indicated by the same colour.