The potential for targeted surveillance of live fish movements in Scotland

Darren Michael Green[1*], Marleen Werkman[1]
& Lorna Ann Munro[2]

April 12, 2012

[1] *Institute of Aquaculture, University of Stirling, Stirling, Stirlingshire FK9 4LA, UK.*
[2] *Marine Scotland Marine Laboratory, PO Box 101, 375 Victoria Road, Aberdeen, AB11 9DB, UK.*
[*] *Author for correspondence. email: darren.green@stir.ac.uk; tel: +44 (0)1786 467872; fax: +44 (0)1786 472133*

Abstract

The network structure of the movements of live fish in the Scottish aquaculture industry has recently been demonstrated for 2003. In this paper, we enlarge this analysis to a longer three-year period from 2002 to 2004, the new data allowing complete coverage of at least one production cycle. The resulting network contains slightly more sites than that for a single year, and is denser with more arcs (directed site-to-site connections) present, but otherwise features recognisable in the one-year network are still recognisable in the three-year network. Arc removal algorithms (a proxy for targeted surveillance) were identified that could successfully reduce the portion of the network reachable from a node (a proxy for potential epidemic size) by approximately one third by removing as few as four arcs. This results from the high centrality of particular nodes and arcs. A strong community structure was identified in the network, corresponding with species farmed but only weakly geographical, with a high proportion of arcs occurring between management areas and catchments.

Keywords: aquaculture; network; graph; transmission

Introduction

Three species of salmonid fish dominate aquaculture production in Scotland: brown trout *Salmo trutta*, rainbow trout *Oncorhynchus mykiss*, and Atlantic salmon *Salmo salar*. Of these, Atlantic salmon is by far the largest sector with *c.* $130,000$ tonnes per year of production over the last decade, over a gradually decreasing number of distinct sites (Scottish Government, 2011a). With production aggregated into a relatively small number of sites, in turn clustered in distinct geographic areas, the risk of spread of disease from site to site, and its management, are of importance both to producers and to the government. Under recent EU legislation, EU directive 2006/88/EC (implemented Aug 08), EU countries are required to implement risk-based surveillance. In Scotland, the competent authority for implementing this is Marine Scotland through the Fish Health Inspectorate, to whom records of live fish movements are already required to be submitted by those registered farming enterprises (under *The Registration of Fish Farming and Shellfish Farming Businesses Order* 1985).

In epidemiological systems with complex population structure, network models have been used widely to study patterns of contact through live animal movements, including in aquaculture (e.g. Thrush & Peeler, 2006; Munro & Gregory, 2009; Green et al. 2009). Epidemiological questions that can be asked of networks include *inter alia*, which sites are at risk of spreading infection, should an epidemic arise? And which sites are at risk of being infected? These are not necessarily the same. And where should we concentrate effort to help reduce epidemic spread? Live fish movements are an important potential route for disease transmission, as has been demonstrated in the cases of both infectious salmon anaemia (Murray et al., 2002) and bacterial kidney disease in rainbow trout (Bland, 2007).

Recently, Green et al. (2009) explored the network structure of movements of live fish within the Scottish aquaculture industry for 2003 (see also Munro & Gregory, 2009), with a view to informing targeted surveillance policy for infectious disease prevention and control. They reported how several algorithms could be used to identify contacts between sites that might prove suitable targets for targeted surveillance. This aids efficient application of limited resources towards high-risk farms, locations, and farm types as part of risk-based surveiilance (Stärk et al., 2006). Since publication of this paper, further work has consolidated movement data into an electronic database for the additional years 2002 and 2004, representing the largest dataset for salmonid movements available for Scotland. This brings further sites and connections into the network, and importantly, covers one complete production cycle for marine salmon production. In this paper, we extend the earlier analyses of network structure and targeted surveillance approaches onto this, more complete, dataset. We investigate whether the conclusions of the earlier analysis are robust to being applied on a considerably larger network.

Method

Data

Data were obtained from the Fish Health Inspectorate at Marine Scotland, Aberdeen, and converted into an electronic database. Only 'validated' data were used, i.e. where fish movements were confirmed by paper records from both exporting and receiving sites and could be cross referenced. These data included all life stages from egg to adult, for all three species, for all registered sites within Scotland. Movements to unregistered sites (predominantly freshwater fisheries; please note that this dataset precedes the legislation introduced under EU directive 2006/88/EC which requires the registration of a wider range of aquaculture production businesses) and imports and

exports outwith Scotland (e.g. to England) were not included in the dataset (Munro & Gregory, 2009). The dataset was extended forwards and backwards by a year giving a three year time series, enough to cover complete production cycles of the salmon industry. However, for 2002, only salmon movement data had been converted to electronic format.

Location and river catchment data were available for each site. Inland sites may be categorised according to 'supercatchment' (i.e. sites connected by any route through fresh water: the whole drainage basin) and 11 official Salmon Fishery Statistical Regions (below, 'Statistical Regions', see Fig. 1 inset). Marine sites are divided into 20 Management Areas (Scottish Government, 2011b), which are in turn subdivided (up to four-way). Management Area boundaries are determined by site locations, and are reviewed periodically if sites open and close, but did not straddle the coastline of adjacent Statistical Regions. Here, Management Area boundaries as of 2003 were used. 'Live' management areas are detailed by the Scottish Government (2011b) but the historical data are not reported online.

The contact network representation

Data analysis broadly followed the approach of Green et al.(2009). Each of n sites is represented by a **node**, with potentially infectious contact from a site i to a site j represented by directed **arcs** (i, j). A matrix element $A_{ij} = 1$ indicates that at some point over the period of interest, movement of live fish occurred from site i to j; $A_{ij} = 0$ indicates no such receipt of live fish. Any (erroneous) self loops were removed ($A_{ii} = 0$). The simplest node properties are the numbers of connections – in, out, and total (undirected) – enjoyed by each node, i.e. the node **degree**, k: $k_i^{in} = \sum_j A_{ji}$; $k_i^{out} = \sum_j A_{ij}$; and $k_i^{undir} = k_i^{in} + k_i^{out} - \sum_j A_{ij}A_{ji}$. The means of node statistics averaged across all nodes are denoted using angled brackets, e.g.$< k^{in} >= \frac{1}{n}\sum_{i,j} A_{ij}$.

Some epidemiologically useful measures can be taken from the network topology. The shortest distance in network steps from node i to node j is represented by the matrix of shortest paths L_{ij}, which is zero if i and j are the same node, and undefined where no suitable path exists. The number of such paths passing through arc (i, j) is its **betweenness** B_{ij} (undefined where $A_{ij} = 0$). Additional measures of network structure included the clustering coefficient $\mathcal{C}$ as used by Keeling (1999), amended for a directed network as the proportion of ordered node triples (i, j, k) with arcs (i, j) and (j, k) that also have arcs (i, k). The level of assortativity (preferential mixing between nodes of relatively high or low degree) r_{assort}, was defined as the correlation between the *in* degree of node i and *out* degree of node j across all arcs (i, j) (Newman, 2003).

From these and other values, estimates can be obtained for the implications of network toplogy on the basic reproduction number $\mathcal{R}_0$ of a propogating epidemic, that is the number of secondary cases generated by a typical case over its infectious period, where $\mathcal{R}_0 > 1$ represents the transmission threshold for a large epidemic in a large network (Anderson & May, 1991). Several network measures are demonstrated on the small 'toy' network shown in Box 1.

Targeted surveillance algorithms

Several algorithms, as described by Green et al. (2009), were used to identify arcs of high importance in the network structure, whose removal from the network effectively limits the potential spread of disease. These algorithms are introduced briefly below. One such algorithm identifies arcs which bridge network **communities**. A network community is a group of nodes (sites) with a relative high density of arcs within the group, with relatively few connections to other such groups. As with the earlier study, a 'modularity'-maximising algorithm was used (Newman, 2004; Leicht and Newman,

2008) to identify communities, with each node a member of a single community. The 'greedy' algorithm used is efficient though does not necessarily find an absolute global maximum for modularity Q, which is a hard problem for large networks. Greedy algorithms proceed through a set of steps, always choosing the locally best solution at each step.

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

Calculation of Q requires the total number of arcs $M = \sum_{i,j} A_{ij}$. For two variables x and y, the Iverson bracket $[x = y]$ has the value of one where the condition $x = y$ is true, and zero otherwise. Therefore the summation for Q above is only performed for combinations of nodes i and j which belong to the same community. Modularity Q penalises placing disconnected nodes with high degree in the same community, and rewards placing together those nodes with low degree which are connected.

The resilience of the network to arc removal was expressed in terms of maximum or mean node 'reach' (Green et al. 2009) D, which can be defined as the number of nodes downstream from a focal node, following directed paths, that are potential targets for epidemic spread, $D_i = \sum_{j \neq i} [L_{ij} \neq \infty]$. All the algorithms, listed below (see also Green et al. 2009), attempt to identify an ordering of arcs from the most to the least important for maintaining network structure. A successful algorithm will result in the fastest disassembling of network structure by removing the least number of arcs, either in terms of maximum reach $\max(D)$ (an estimate of worst-case epidemic size) or mean reach $\langle D \rangle$ (an estimate of typical epidemic size).

Arbitrary The null model, with a non-intelligent selection of arcs, was to choose arcs purely at random, representing non-purposeful (but potentially limited) surveillance.

Highly connected (degree) A node with many *in* connections is at high risk of infection; a node with many *out* connections poses a high risk of further connection. The second algorithm therefore ranked arcs (i, j) according to how well they fulfilled both criteria, using the metric $k_i^{in} \times k_j^{out}$. Networks with many arcs with a high value for this expression are assortative (see above), and assortativity leads to a higher value of $\mathcal{R}_0$.

High betweenness Arcs were ranked according to their betweenness, B_{ij}. After each sequential removal of an arc, betweenness must be recomputed for the whole network as shortest paths are frequently rerouted.

Community-bridging Arcs bridging communities as identified above were prioritised for removal. Within-community and between-community arcs were chosen arbitrarily, aside from this criterion.

Greedy max & greedy mean In these algorithms, arcs are removed one at a time, always choosing that arc which causes the greatest reduction to either maximum or mean reach. Though this sounds 'optimal', as with many algorithms—including that for assigning communities above—this locally optimal choice by no means ensures finding a globally optimal solution.

Eigenvector-based Network eigen analysis provides an eigenvector V which is a measure of node centrality, and an eigenvalue λ which in some conditions can be used to obtain an estimate of $\mathcal{R}_0$. Two algorithms as used by Green et al. (2009) chose *out* arcs arbitrarily within nodes, ranking nodes according to highest eigenvector centrality V_i. The adjacency matrix was modified (Bonacich & Lloyd, 2001) in two ways before eigenanalysis, assuming either additional contact between all nodes of strength β/n (**eigen spread**

algorithm), or constant total weight of *out*ward contact from all nodes (**eigen walk** algorithm).

For the most successful algorithms, the ten arcs were identified that appeared most frequently amongst the first ten arcs removed (due to the stochastic nature of the algorithms, the chosen arcs may vary between runs). The properties of these arcs were then compared with the average arc properties across the whole network.

Results

Small- and large-scale network structure

With the new larger movement dataset presented in this paper (compared with Green et al., 2009), a larger network is obtained with $n = 561$ nodes and a higher density of links (Fig. 2). Mean degrees (and their coefficients of dispersion, i.e. the variance-to-mean ratio) were $< k^{in} >= 2.39$ (1.23), $< k^{out} >= 2.39$ (6.36) and $< k^{undir} >= 4.53$ (4.24) for an undirected network (see Fig. A1 in electronic supplementary material for a histogram of node degree). Other simple measures of network shape remained relatively unchanged from the earlier analysis with correlation between *in*- and *out* degree of nodes of $r = 0.238$, a clustering coefficient $\mathcal{C} = 0.069$ and a coefficient of assortativity of $r_{\text{assort}} = 0.164$ (assortative). With a higher k, the eigenvalue approach to estimating $\mathcal{R}_0$ also gives a higher value, with $\mathcal{R}_0 \approx 4.59$ for $\beta = 0.5$. Again, this is higher than the degree-based estimate of $\mathcal{R}_0 \approx \langle k^{in}k^{out}\rangle \div \sqrt{< k^{in} >< k^{out} >} = 3.05$.

The higher arc density has a large impact on the overall connectivity and community structure of the entire network. Considering large-scale measures of network structure, as opposed to the node- and arc-level measures reported above, mean shortest path length (where defined and non-zero) was 5.92, with such paths accounting for 0.105% of the $n(n-1)$ possible. As in Green et al. (2009), a rewiring algorithm was employed to provide null-model networks for comparison, equivalent to a null hypothesis of no large-scale structure to the network and random connection (notwithstanding that nodes have different degree). For rewired networks, mean shortest path length was similar, 5.13, with finite non-zero paths accounting for 0.598% of those possible (see Fig. A1 in electronic supplementary material for histograms of path lengths in the original and rewired networks).

The community-joining algorithm to find an optimal partition of the network to maximise modularity found partitions with maximum modularity of $Q = 0.75$, similar to that found for the one-year network data. The community assignments and joining dendrogram associated with this maximum modularity are shown in Figs. (3) and (A2, in electronic supplementary material). This community distribution consists of several large communities well defined in the dendrogram, with a small number of nodes belonging to smaller communities or disconnected. This community algorithm produces a partition with a higher modularity index than that using shared membership of a supercatchment or management area as criteria for membership of the same 'community' ($Q = 0.41$). This reflects a relatively large proportion of arcs occurring between such communities ($43\,\%$), as can be seen in Fig. (1).

Reducing network reach

The effects on network reach of removing upto 100 arcs is shown in Fig. 4 for both maximum reach and mean reach, and for the eight algorithms explored in Green et al. (2009) and described above. With only a small number of extra nodes, but considerably more connections, network density is higher and this is reflected in different efficiencies amongst the algorithms. The two greedy algorithms perform well for both measures, with betweenness being comparatively successful. The other

algorithms (degree-, community-, and eigenvalue-based) perform relatively poorly. Degree- and community-based measures are somewhat effective but only if a large number of arcs are removed.

Arcs having been ranked in order of ability of importance to network structure, the properties of high importance arcs and the average arc were compared for the greedy (both) and betweenness algorithms. For the greedy max algorithm, four out of ten selected arcs were shared with the greedy mean algorithm, and two with the betweenness algorithm, which in turn shared four with the greedy mean algorithm. Overall, 22 arcs were represented involving 33 nodes with some nodes being both source and destination nodes for these arcs. Compared with the whole network, these arcs left nodes with a higher *in* degree (3.8) than average (3.1) and led to nodes with a considerably higher *out* degree (10.8 versus 3.1). A fraction 0.52 of arcs joined nodes in different communities, compared with 0.12 for the whole network. They are also more likely to join nodes in different regions (0.52 versus 0.43).

Discussion

Compared with analysis of the 2003 data alone (Green et al. 2009), the 2002-4 network shows a modest increase in the number of nodes, but a large increase in the number of arcs. This is demonstrated by the more densely connected network picture, with few poorly connected or disconnected node pairs. Community structure however remained strong. Compared with the 2003 network, the full network has a distribution of path lengths with higher mode, but with a shorter tail, as expected with a more compact network. Maximum degree is also higher. A caveat concerning the movement data raised in Green et al. 2009, remaining important here, is that only movements between registered sites are recorded (Munro & Gregory, 2009). Unregistered sites consist mainly of trout fisheries, which although they may pose a risk of disease spread by other means than fish movements, are likely to be primarily 'sinks' for live fish movements with no onward network connections. These data will be available in future since sites are required to be registered as of August 2008 under EU directive 2006/88/EC. Electronic recording of movement data in a database format will also aid in future analyis: the scope of the study reported here was necessarily limited in breadth by the large amount of effort involved in processing the paper movement records, particularly as these are in duplicate (off and on pairs). The complexities of analysing this data source have been recently discussed by Werkman et al. (2011).

Arc-removal methods of fragmenting the network remain effective, with c. 10 arcs removed reducing both mean and maximum network reach by about one half. This is less effective than with the 2003 data alone (Green et al. 2009), but this is to be expected given the more tightly connected network: its denser structure is held together by single arcs in fewer places. Lower betweenness values for arcs are found for similar reasons (data not shown), as a larger number of paths between nodes exist, lowering the centrality of a particular path. Though the successful arc-removal algorithms targeted arcs with a tendency towards high degree, joining communities, algorithms based on degree or community assignment alone behaved relatively poorly, demonstrating the benefit of a more technical, model-based approach to targetting surveillance. This contrasts with the earlier analysis on the smaller network (Green et al. 2009): there, the degree-based measure performed less poorly, and the eigenvalue-based measures performed at least as well as arbitrary removal. Here, the **eigen spread** measure performed worse than random removal.

The ability to identify high-centrality nodes is only beneficial for disease control if the network structure is sufficiently stable: a rapidly changing network means the predictive power of network properties for future disease control will be low. Recently,

algorithms for determining the stability in time of trading networks, using both the aquaculture network fully described here, and the network for large livestock movements in the UK for 2003-4, have been examined (Green et al. 2011). These authors concluded that though temporal autocorrelation existed in both networks, it is difficult to make a judgement on the utility of this without an objective baseline for comparison. One way of defining this baseline may be by through simulating real disease problems and their control measures on (dynamic) networks.

However, when considering the spread of disease within the aquaculture industry, live fish movements are far from the only means of disease transmission between sites. Other transmission routes include fomite, well-boat associated (Murray et al. 2002), or direct spread through the water column in sea or freshwater, for example for infectious salmon anaemia virus (ISAV; Jarp & Karlsen, 1997, Gustafson et al., 2007), with differences in risk between diseases. Where Management Areas are relatively well sealed due to infrequent movements between them and separation in water distance, synchronisation in fallowing may provide a benefit in clearing disease from specific areas. The effectiveness of such fallowing strategies has been explored by Werkman et al. (2011), which contrasts the effects of partial or fully synchronised fallowing in simulated epidemics of the Scottish salmon industry (marine sites) based on the distribution of sites within Management Areas.

In conclusion, the network of live fish movements in Scotland shows itself to be strongly organised into communities, with potential for targeted surveillance to focus on high centrality arcs, to aid in the development of risk-based surveillance programmes where resources are necessarily limited (Stärk et al., 2006). Nevertheless, utility of such approaches is limited by the lack of real-time data in a continuously changing industry.

Acknowledgements

With thanks to the Fish Health Inspectorate of Marine Scotland for providing access to the movement records. DMG and MW's contributions to this work was funded through Marine Scotland.

References

Anderson, R.M. & May, R.M. (1991) Infectious diseases of humans: dynamics and control. Oxford University Press.

Bland, M. (2007) Epizootic investigation into the presence of bacterial kidney disease (BKD) in rainbow trout farms in Scotland 2005. Fisheries Research Services Internal Report No 14/07. `http://www.scotland.gov.uk/Uploads/Documents/1407.pdf` Accessed 21 January 2011.

Bonacich, P., Lloyd, P. (2001) Eigenvector-like measures of centrality for asymmetric relations. Social Networks 23, 191 – 201.

Green, D.M., Gregory, A., Munro, L.A. (2009) Small- and large-scale network structure of live fish movements in Scotland. Preventive Veterinary Medicine 91, 261 – 269.

Green, D.M., Werkman, M., Munro, L.A., Kao, R.R., Kiss, I.Z. & Danon, L. (2011) Tools to study trends in community structure: application to fish and livestock trading networks. Preventive Veterinary Medicine, in press.

Gustafson, L.L., Ellis, S.K., Beattie, M.J., Chang, B.D., Dickey, D.A., Robinson, T.L., Marenghi, F.P., Moffett, P.J. & Page, F.H. (2007) Hydrographics and the timing of infectious salmon anemia outbreaks among Atlantic salmon (*Salmo salar* L.) farms in the Quoddy region of Maine, USA and New Brunswick, Canada. Preventive Veterinary Medicine 78, 35 – 56.

Jarp J, Karlsen E. (1997) Infectious salmon anaemia (ISA) risk factors in sea-cultured Atlantic salmon Salmo salar. Diseases of Aquatic Organisms 28, 79 – 86.

Keeling, M.J. (1999) The effects of local spatial structure on epidemiological invasions. Proceedings of the Royal Society of London Series B 266: 859 – 867.

Leicht, E.A., Newman, M.E.J. (2008) Community structure in directed networks. Physical Review Letters 100, 118703.

Munro, L., Gregory, A. (2009) Application of network analysis to fish movement data. Journal of Fish Diseases 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. (2003) Mixing patterns in networks. Physical Review E 67, 026126.

Newman, M.E.J. (2004) Fast algorithm for detecting community structure in networks. Physical Review E 69, 066133.

The Scottish Government (2011a) Production surveys `http://www.scotland.gov.uk/Topics/marine/Fish-Shellfish/FHI/surveys` Accessed 21 January 2011.

The Scottish Government (2011b) Management Areas. `http://www.scotland.gov.uk/topics/marine/fish-shellfish/FHI/managementagreement`. Accessed on 21 January 2011.

Stärk, K.D.C., Regula, G., Hernandez, J., Knopf, L., Fuchs, K., Morris, R.S., Davies, P. (2006) Concepts for risk-based surveillance in the field of veterinary medicine and veterinary public health: Review of current approaches. BMC Health Services Research 6, 20.

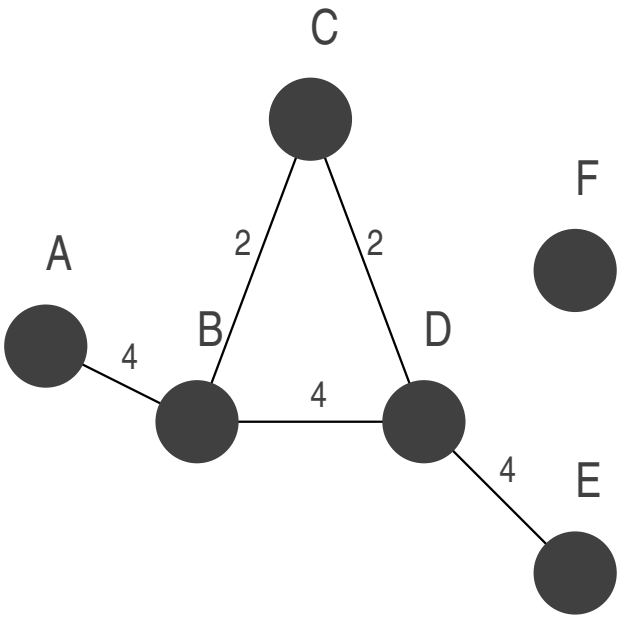
Thrush, M., Peeler, E. (2006) Stochastic simulation of live salmonid movement in England and Wales to predict potential spread of exotic pathogens. Diseases of Aquatic Organisms 72, 115 – 123.

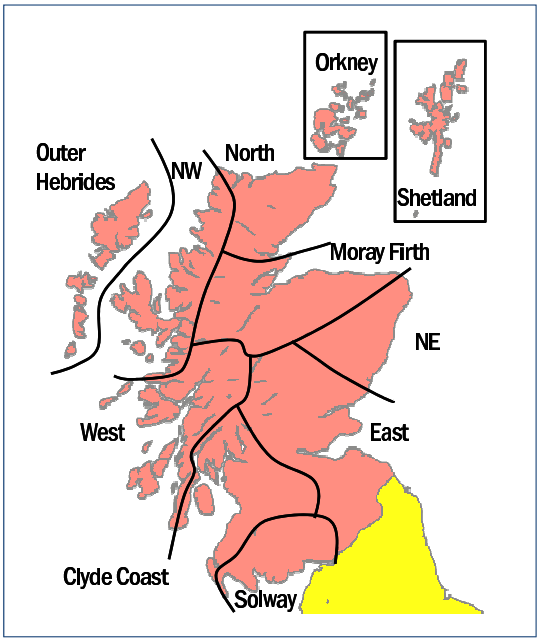
Werkman, M., Green, D.M., Murray, A.G. & Turnbull, J.F. (2011) The effectiveness of fallowing strategies in disease control in salmon aquaculture. Preventive Veterinary Medicine 98, 64 – 73.

Werkman, M., Green, D.M., Munro, L.A., Murray, A.G. & Turnbull, J.F. (2011) Seasonality and heterogeneity of live fish movements in Scottish fish farms. Diseases of Aquatic Organisms, in press.

Display items

Box 1. Demonstration of network statistics.





For simplicity, the small undirected network shown is analysed in the table showing key network statistics referred to in the paper. On the network diagram, nodes are indicated and edges are marked by their betweenness B_{ij}.

	A	B	C	D	E	F	mean
k	1	3	2	3	1	0	$10/6$
$L_{\mathrm{A},i}$	0	1	2	2	3	∞	
D_i	4	4	4	4	4	0	

Summary statistics

$\max(D) = 4$ $\quad\quad$ $\langle D \rangle = 20/6$

$< k >= 1.67$ $\quad\quad$ $< k^2 >= 4$

$\mathcal{R}_0 \approx < k^2 > / < k >= 2.4$

triples: $ABC\ ABD\ BCD\ CBD\ CDB\ BDE\ CDE$

triangles: $BCD\ CBD\ CDB$

$\mathcal{C} = \frac{3}{7}$

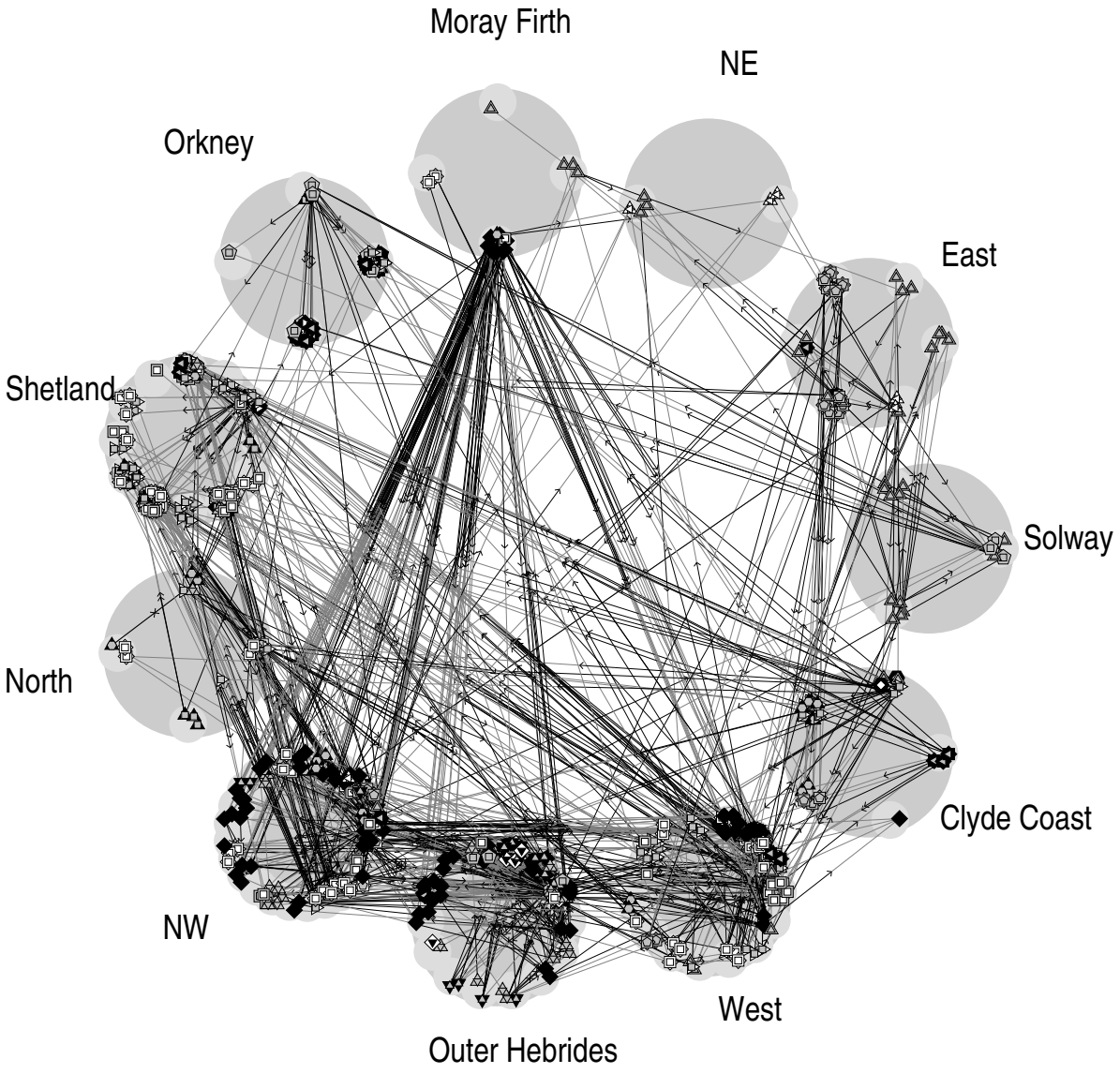


Figure 1: Geographical representation of network community structure. Large circles of nodes represent regions; sites belonging to the same management area (and sub-area) or catchment are drawn in the same small circle. Inset: schematic showing official Salmon Fishery Statstical Regions.

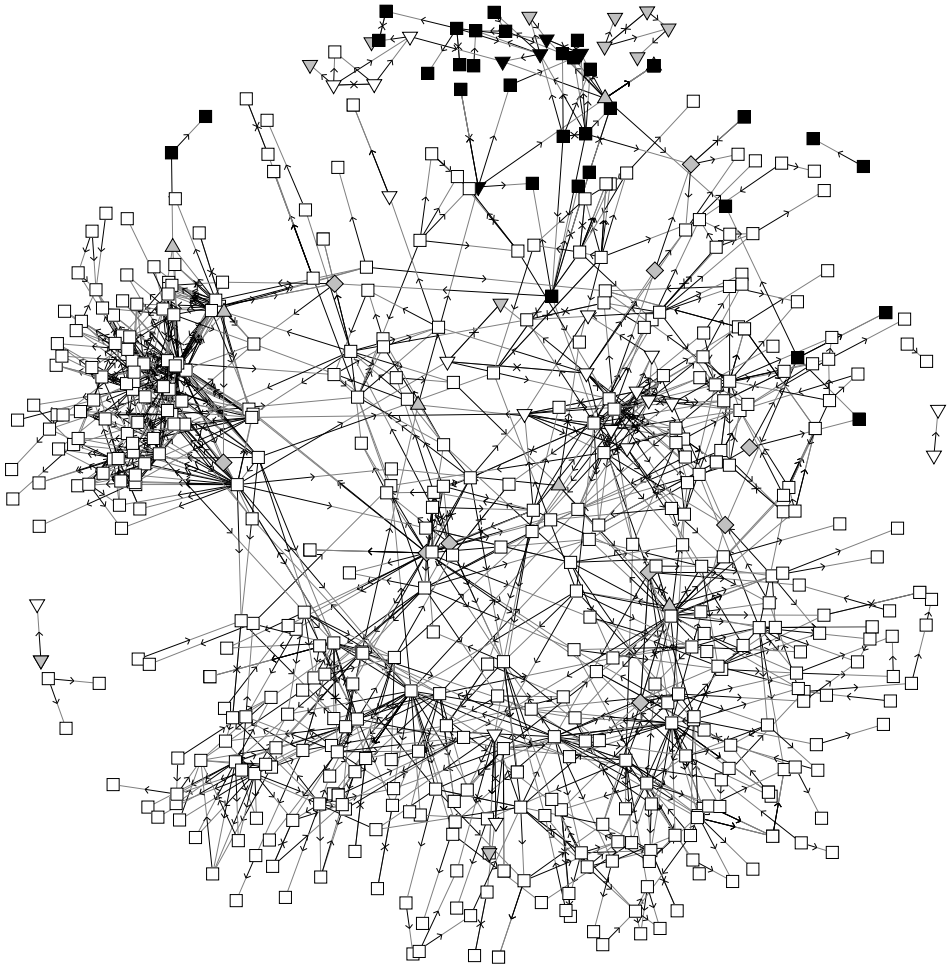


Figure 2: The Scottish live fish movement network (2002-4), according to species moved to/from sites. □ salmon only (S); ■ rainbow trout only (R); ♦ S+R; ▼ brown trout only (T); ▼ T+R; ▽ T+S; ▲ T+S+R. Arrows indicate movement direction (sometimes bidirectional).

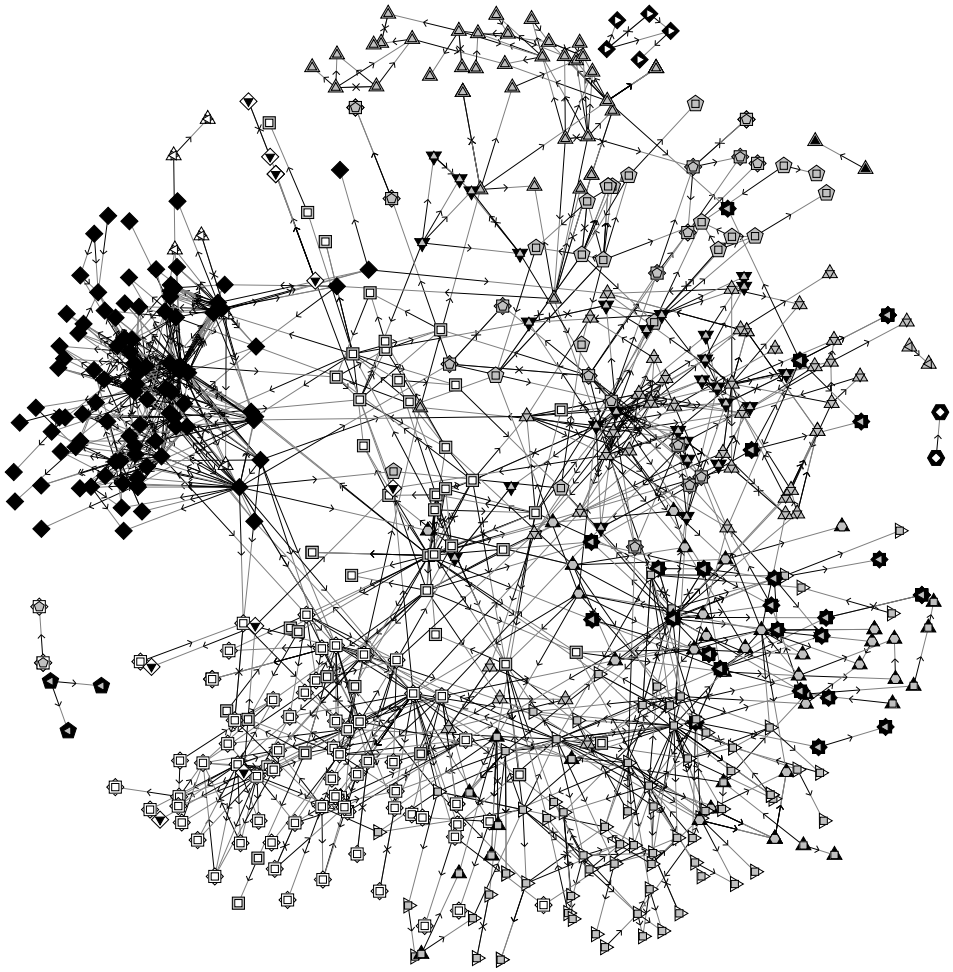


Figure 3: Community assignment for the live fish movement network for Scotland for 2002-4. Community membership is indicated by different symbols.

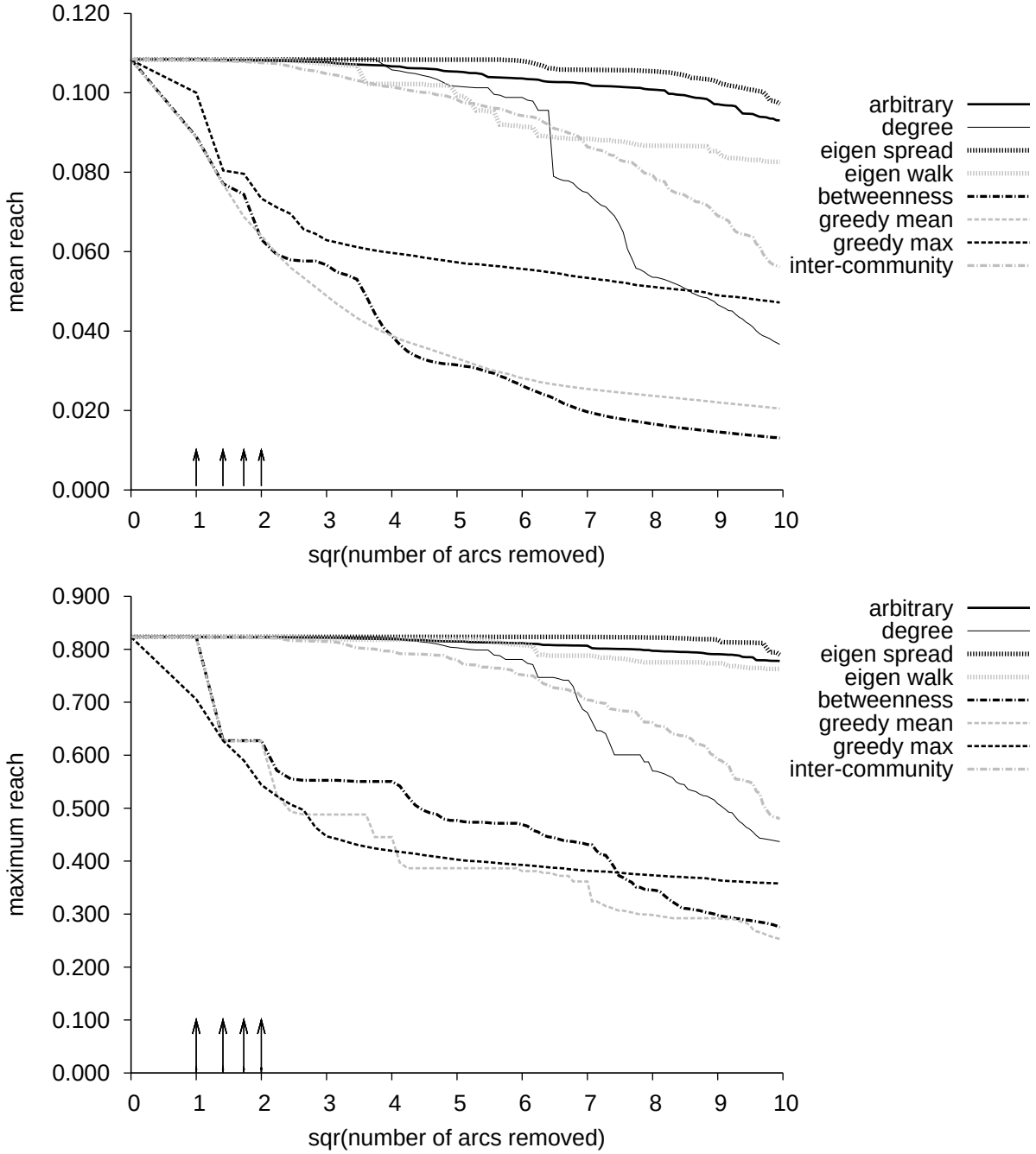


Figure 4: Mean and maximum reach from all nodes, versus proportion of network arcs removed (plotted on a square-root scale), for eight different algorithms for determining precedence of arc removal. Arrows indicate x-axis values corresponding to the removal of 1, 2, 3, and 4 arcs.