Small- and large-scale network structure of live fish movements in Scotland

Darren Michael Green[1*], Alison Gregory[2]
& Lorna Ann Munro[2]

July 23, 2009

[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

Abstract

Networks are increasingly being used as an epidemiological tool for studying the potential for disease transmission through animal movements in farming industries. We analysed the network of live fish movements for commercial salmonids in Scotland in 2003. This network was found to have a mixture of features both aiding and hindering disease transmission, hindered by being fragmented, with comparatively low mean number of connections (2.83), and low correlation between inward and outward connections (0.12), with moderate variance in these numbers (coefficients of dispersion of 0.99 and 3.12 for in and out respectively); but aided by low levels of clustering (0.060) and some non-random mixing (coefficient of assortativity of 0.16). Estimated inter-site basic reproduction number R_0 did not exceed 2.4 at high transmission rate. The network was strongly organised into communities, resulting in a high modularity index (0.82). Arc (directed connection) removal indicated that effective surveillance of a small number of connections may facilitate a large reduction in the potential for disease spread within the industry. Useful criteria for identification of these important arcs included degree- and betweenness-based measures that could in future prove useful for prioritising surveillance.

Keywords: aquaculture; network; graph; transmission

1 Introduction

Scotland is the third largest producer of Atlantic salmon *Salmo salar*, with an ex-farm value of £400 million per annum (http://www.scottishsalmon.co.uk/economics/economics.asp) and 143 000 tonnes produced in 2007 (http://www.marlab.ac.uk/FRS.Web/Uploads/Documents/Scottish%20Fish%20Farm%20Production%20Survey%202006.pdf). This dominates the other commercially produced species, which include rainbow trout *Oncorhynchus mykiss* and brown trout *Salmo trutta*, with some 7000 tonnes of rainbow trout produced in 2006 (*ibid.*). As with other farming industries, long-distance spread of disease via movements of live or dead fish are of concern. Well boats have been implicated in facilitating the spread of infectious salmon anaemia (ISA) in Scotland (Murray *et al.* 2002), and road haulage of live rainbow trout for the spread of bacterial kidney disease (BKD) UK wide in 2005 (http://www.marlab.ac.uk/FRS.Web/Uploads/Documents/1407.pdf). Furthermore, there is ongoing concern over the potential for introduction of *Gyrodactylus salaris* into countries which are currently free from it. Over a variety of industries, similar epidemic problems in the UK and elsewhere have led to the recording of live animal movements for a number of species including large mammals and fish, allowing their study (Thrush & Peeler 2006; Kao *et al.* 2007).

The epidemiological risk posed by movements between sites lends itself well to a network representation, as has been used for a number of species including pigs (Bigras-Poulin *et al.* 2007; Ribbens *et al.* 2008), sheep (Webb 2005; Kiss *et al.* 2006), chickens (Truscott *et al.* 2007; Dent *et al.* 2008) and cattle (Christley *et al.* 2005a; Green *et al.* 2008). Networks are a powerful epidemiological tool allowing one to investigate potential for disease spread, the structure of the industry, and its implications for biosecurity, both in terms of the risk of a site being infected, and the risk posed by a site, should it become infected. A network consists of nodes – here, the epidemiological unit, or site – connected by bidirectional edges or unidirectional arcs – here representing potentially infectious contact between sites due to movement. Potentially infectious contact does not imply infection, but it is nevertheless a prerequisite for it. Sites may be sources of infection (posing a risk of onward spread), sinks for infection (at risk from disease spread), or both. Links between sites differ in their contribution to potential epidemics, in a manner that is not easy to predict by examining the behaviour of individual nodes, but only by examining the network as a whole.

Fish farms in Scotland are required by legislation (The Registration of Fish Farming and Shellfish Farming Businesses Order 1985) to be registered, keep records of all live fish movements, and submit these records to the Fish Health Inspectorate. Furthermore, under the new EU directive 2006/88/EC (implemented from August 2008), all EU member states must implement risk-based surveillance for diseases of aquatic animals. Live fish movements are an important focus for such surveillance. Here, we analyse earlier data with a view to informing surveillance strategy applicable to data collected in future. Focused selection of surveillance targets will improve efficiency in terms of cost–benefit ratio, important where resources are limited (Stark *et al.* 2006).

Previous authors have applied different criteria to contact networks in an attempt to identify high-risk edges or nodes, whose removal reduces the potential for the network to support an epidemic (Albert *et al.* 2000; Kiss *et al.* 2006). In terms of a network representation, node or edge removal is equivalent to removal of the risk of potentially infectious contact either from a node (node removal) or between two nodes (edge removal), and does not in any way imply that a node or edge itself has ceased to exist. For example, effective immunisation of a farm could be interpreted as node removal, whereas pre-movement testing for bovine tuberculosis as has been implemented for cattle in GB could be considered edge

removal. Identifying high-risk nodes has frequently been considered as a problem of finding network nodes with high *centrality*, for which various metrics exist (Bell *et al.* 1999; Bonacich & Lloyd 2001; Christley *et al.* 2005b; Zemljič & Hlebec 2005). Broadly, centrality measures can be fitted into three categories: those dependent on the properties of individual nodes; those dependent upon paths to, from, or through a particular edge or node; and those based upon eigendecomposition of a network adjacency matrix, explained further below.

In terms of the aquaculture industry, edge (specifically, directed 'arc') removal rather than node removal can be considered more appropriate: Concentrated surveillance, such that particular links between sites are without epidemiological risk, is equivalent to removing that edge from the network. Node removal is equivalent to removing a whole site from the epidemiological network, and is both less achievable, and less appropriate. This paper investigates methodologies for identifying both high-risk sites, and high-risk interactions between sites – here, movements – for commercial salmonids. A previous study has described the internal movement structure of one of the major salmon companies in addition to the combined trout (rainbow and brown) movement record for 2003 (Munro & Gregory 2009). However, our study here presents the first mathematical study of the whole of the Scottish salmonid industry using the records of live fish movements.

2 Method

2.1 Data

The data source was the official fish movement records for Scotland 2003, held by the Fish Health Inspectorate at Marine Scotland, Aberdeen, which is the agency of the Scottish Government responsible for regulating and enforcing legislation concerning the aquaculture industry in Scotland. More recent data are not yet available. These data comprise validated movements of live fish, from egg to adult, between registered fish farm sites within Scotland, considered validated where the paper records of the off and on movement were legible and could be matched. Movements to fisheries (predominantly freshwater) are not included as they are not registered under current legislation. Neither are imports or exports outwith Scotland (including to or from England and Wales) included in the data. The paper records were transferred to an electronic database recording source and destination sites for each of 3696 movements of Atlantic salmon, rainbow trout and brown trout amongst $n = 422$ sites.

2.2 Network properties

All network and arc removal algorithms were programmed using C++. The network is represented by the adjacency matrix A where $A_{ij} = 1$ indicates at least one (potentially several) directed movement from node (site) i to node j ($1 \leq i, j \leq n$) occurred, and zero, no connection. Rare self -loops (movements from site i back to site i) were removed. Each node is described by its in and out degrees $k_i^{\text{out}} = \sum_j A_{ij}$ and $k_i^{\text{in}} = \sum_j A_{ji}$, and the undirected degree $k_i^{\text{undir}} = k_i^{\text{out}} + k_i^{\text{in}} - \sum_j A_{ij}A_{ji}$, giving a total of $M = \sum_{ij} A_{ij}$ arcs (directed connections) in the network. The network can also be characterised by the matrix of minimum path lengths L_{ij}, the minimum number of steps along arcs required to move from node i to node j (infinite if no possible path exists; $L_{ii} = 0$), and the arc 'betweenness' B_{ij}, defined as the number of shortest paths amongst all pairs of nodes that pass through arc $i \to j$ (or zero where no arc exists). A network with low mean shortest path length, as is found with 'small world'-type networks (Watts & Strogatz 1998), will be subject to rapid epidemic spread compared to networks with longer mean shortest path length, holding all other network properties constant. The distribution of path lengths can be usefully compared with that of

'rewired' networks where higher-scale structure is removed. Rereouting of pairs of arcs of the form $A \to B$ and $C \to D$ to give arcs $A \to D$ and $B \to C$ was performed for 10 replicate networks, following the procedure of Kiss et al. (2006).

The clustering coefficient $\mathcal{C}$ defines the degree to which 'any friend of yours is a friend of mine'. For a directed network, a simple and epidemiologically appropriate definition is that of the proportion of 'triples' – three distinct, ordered nodes U, V, and W with directed arcs $U \to V$ and $U \to W$ – which are also 'triangles', with an additional connection $V \to W$: i.e. $\mathcal{C} = \sum_{uvw} A_{uv}A_{uw}A_{vw} / \sum_{uvw} A_{uv}A_{uw}$. Networks with higher clustering, all other things being equal, are more resilient to epidemic spread (Keeling 1999), and this definition of $\mathcal{C}$ is compatible with various contact models of disease transmission.

The coefficient of assortativity of a network measures the extent to which edges join nodes of similar degree (Newman 2003). For directed networks, a set of potential correlation measures exist of the form

$$r(k_i^{\mathrm{X}}, k_j^{\mathrm{Y}} | i \to j) = \frac{M \sum_{i\to j} k_i^{\mathrm{X}} k_j^{\mathrm{Y}} - \left(\sum_{i\to j} k_i^{\mathrm{X}}\right)\left(\sum_{i\to j} k_j^{\mathrm{Y}}\right)}{\sqrt{\left[M \sum_{i\to j} (k_i^{\mathrm{X}})^2 - \left(\sum_{i\to j} k_i^{\mathrm{X}}\right)^2\right]\left[M \sum_{i\to j} (k_i^{\mathrm{Y}})^2 - \left(\sum_{i\to j} k_i^{\mathrm{Y}}\right)^2\right]}}$$

Where each network node belongs to one community, and there is no restriction on how many nodes may belong to each community, the number of possible arrangements of n nodes within communities is given by the Bell number B_n (Bell 1934), which rises faster than exponential with increasing n, combinations outnumbering the atoms in the universe for quite modest n. An exhaustive search is therefore not possible. The community structure algorithm we used was based on measures of 'modularity', using the 'greedy' algorithm introduced by Newman (2004), amended to account for the strongly directed nature of the network as discussed by Kao *et al.* (2006) and Leicht & Newman (2008). A measure of community fit is given by

$$Q = \frac{1}{M} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i^{\mathrm{out}} k_j^{\mathrm{in}}}{M} \right) [c_i = c_j] \quad 0 \le Q \le 1$$

where c_i is the community label for node i, $[x = y]$ returns unity where $x = y$ and zero otherwise, and γ is a constant, implicitly equal to unity in Leicht and Newman (2008) (see Kumpula *et al.* 2008). Higher Q indicates a larger fraction of arcs within communities. The algorithm proceeded by first assigning each node a unique community label $1 \ldots n$. Then, each possible merger of two communities was considered, with the merger that resulted in the greatest increase in Q (or smallest decrease) accepted and nodes of both communities assigned the same unique label. This process was repeated until only a single community remained, with Q reaching a maximum at some intermediate point.

2.3 Arc removal

To determine the susceptibility of a network to an (unknown) epidemic, the size of the giant strongly connected component (GSCC), or simulation modelling are frequently used (e.g. Kiss *et al.* 2006; Thrush & Peeler 2006). The GSCC represents the largest set of nodes such that any two nodes can be connected by directed paths (a strong

component, SC). With the highly directed structure of the fish network, strong components become small and less useful. Instead, here, the epidemiological risk posed by a node is defined in terms of its 'reach', i.e. the number of other nodes that can be arrived at from node i by following directed paths, in terms of the maximum reach for any node, and the mean reach across all nodes. This is related to the chains of infection discussed by Dubé et al. (2008).

Of interest is the resilience of the network to the removal of small numbers of arcs, corresponding to the concentration of surveillance effort onto particular movements, resulting in the potentially infectious contact associated with these movements becoming negligible. The extent to which networks are thus disrupted depends greatly on the choice of arc or edge to be removed (Kiss *et al.* 2006), implying that different criteria for targeting surveillance will vary considerably in their efficiencies. The effect of choosing different surveillance strategies was explored by sequentially removing one arc at a time according to one of a set of criteria and reevaluating the properties of the pruned network, in particular, reachability. Arc removal proceeded until no arcs remained. These criteria varied between simple and easily implemented, through to more complex methods requiring computer time to solve with reevaluation necessary after each arc removal. These are described below in brief. Where the criterion statistics were tied between arcs, the removed arc was chosen at random. Due to such stochastic components to the algorithms, each algorithm was repeated 80 times and the means of network properties are reported below. The different strategies are not intented to represent different on-site procedures, but different methods of selecting which sites or movements to concentrate these procedures upon.

Arbitrary Arcs were chosen at random - the control.

Inter-community Arcs connecting nodes in different communities were weighted higher than those between, but within these two sets, chosen arbitrarily.

Degree A simple degree-based measure was tested, denoted for an arc $i \to j$ by $d_{ij} = k_i^{\text{in}} k_j^{\text{out}}$. Arcs with high d produce a two-node 'unit' whose combined contribution to potential epidemic spread is large. Networks with more arcs of higher d than expected by chance are assortative according to the measure of assortativity introduced above.

Greedy 'Greedy' algorithms are those which always make the locally optimal choice, which in some cases is sufficient to find the globally optimal solution (Cormen *et al.* 2001). For either maximum reach (**greedy max**) or mean reach (**greedy mean**), at each step, the arc is chosen that would cause the greatest reduction in the selected measure, once removed. All arcs are thus examined at each step.

Betweenness Edge betweenness B_{ij} was calculated for each edge and the arc with the highest value selected. This is closely related to node betweenness, a frequently used measure of centrality.

Eigenvector Two measures of eigenvector-type centrality were implemented (**eigen spread** and **eigen walk**) as described in the following section, with out-arcs chosen at random from the node with the highest such centrality.

The **greedy**, **betweenness** and **eigen** methods all require recomputation of the relevant statistics after each step.

2.4 Network eigen analysis

Given an adjacency matrix A, eigen decomposition can provide us with a dominant eigenvalue λ and corresponding eigenvector V. The eigenvector has one entry per node and gives an estimate of the centrality of each node – a measure of the

relative risk of incidence across nodes for a disease that is relatively rare. The dominant eigenvalue is related to R_0 given a per-link transmissibity τ, with $R_0 \sim \tau\lambda$ (Diekmann & Heesterbeek, 2000; Kiss et al. 2008). Inter-site R_0 is therefore limited by λ, subject to disease-specific parameters encapsulated by τ which we do not attempt to parameterise here. Because the analysis does not consider the infection state of an infected node's neighbours, this estimate does not account for network clustering and is an overestimate where many bidirectional edges are present.

Bonacich and Lloyd (2001) note features of network construction – frequently encountered in the live fish movement network – where standard eigenvector centrality approaches fail. We follow a similar approach to what they recommend by applying eigen analysis to two modified adjacency matrices A, using a simple power iteration method. This modification assumes that a small amount of additional contact $\beta = \frac{1}{2}$ occurs outside of the documented connections.

In the first case (**eigen spread**; A^{spread}) all arcs were weighted equally. Alternatively, we consider a simple random walk (**eigen walk**) by assuming that outward contact is weighted such that the total outward contact from any node is constant (A^{walk}):

$$\begin{aligned} A_{ij}^{\text{spread}} &= \frac{\beta}{n} + A_{ij} \\ A_{ij}^{\text{walk}} &= \frac{\beta/n + A_{ij}}{\beta + k_i^{\text{out}}}. \end{aligned}$$

The eigenvector V is then obtained simply by iterating the expression $V^{s+1} = AV^s$ until convergence, normalising after each step, starting with $V_i^0 = 1/n$. The dominant eigenvalue λ is then obtained by solving the equation $A\lambda = AV$. For A^{spread}, this eigenvalue provides an upper estimate for R_0.

3 Results

3.1 Small- and large-scale network structure

The network of $n = 422$ nodes is shown in figure 1, indicating site type and the direction of movements. Betweenness is indicated by line width, demonstrating site-to-site links that are important connections between parts of the network. Visible are a large number of disconnected pairs and small groups of nodes, and more tightly connected clusters of nodes.

Examination of the degree distribution confirms the mostly directed nature of the network: mean node degrees (and coefficient of dispersion, i.e. the variance to mean ratio) were $\langle k^{\text{in}} \rangle = 1.48\ (0.99)$, $\langle k^{\text{out}} \rangle = 1.48\ (3.12)$, and $\langle k^{\text{undir}} \rangle = 2.83\ (2.16)$ for an undirected network (figure 2). Only $8.7\,\%$ of site–site connections were bidirectional. Correlation between the *in*- and *out*-degree of nodes was weakly positive: $r(k_i^{\text{in}}, k_i^{\text{out}}) = 0.12$. Mean shortest path length (where not zero or infinite) was 4.4 (figure 2), with 2.6 % of potential paths existing. This compares with the rewired networks, with a mean shortest path length of 7.1 and 3.2 % of potential paths existing. This reflects the low mean degree and the fragmented nature of the network seen in figure 1. There was relatively little difference in degree amongst sites moving different species, but a tendency for lower *out* degree for site types further down the production chain, as might be expected (table A1 in electronic supplementary material), with notable net flow from freshwater to seawater sites (figure A1 in electronic supplementary material).

For the movement network, $r(k_i^{\text{in}}, k_j^{\text{out}} | i \to j) = 0.16$, indicating a small degree of assortative mixing. The clustering coefficient was also low, with a ratio of triangles to triples of $\mathcal{C} = 0.060$. The network therefore has a mixture of properties that both encourage epidemic spread (assortativity, overdispersion in degree) and

discourage spread (clustering at the node and community level, low *in–out* degree correlation).

Eigen analysis provided an estimate for R_0 of 2.4τ for $\beta = \frac{1}{2}$. This is sensitive to the value of β chosen, reaching asymptotes of $R_0 \approx \beta\tau$ for large β and $R_0 = 2.2\tau$ as β approaches zero. This is higher than the traditional estimate of R_0 taken from degree statistics, of $\frac{\langle k^{\text{in}} k^{\text{out}} \rangle}{\langle k^{\text{in}} \rangle}\tau = 1.7\tau$. Relatively little difference was found between site types in the associated eigenvectors, unless very small values for β were used (table A1 in electronic supplementary material).

The best-fit community assignment is shown in figure 3, corresponding to a high modularity index of $Q = 0.82$. The dendrogram showing community structure above and below this level of joining is shown in figure 4. Interpretation of the dendrogram requires care: two structural features indicate lack of identifiable sub-structure in the data (identified by symbols in figure 4). First, the algorithm joins the isolated pairs of nodes into a binary tree visible on the dendrogram in an arbitrary way. Second, many communities are constructed by sequentially adding on single nodes forming a 'plume'-like pattern also seen by Newman (2004). However, elsewhere in the tree, communities close together can be seen that are also noticeably close together in the network picture (figure 3).

3.2 Reducing network reach

The mean number of nodes reachable by following directed paths from another node was 12 (2.8 % of the network), however node reach was highly variable with many nodes being sinks only, and had a maximum of 130 nodes (31 %) (figure 2b). This contrasts with the smaller GSCC of 17 (and next-largest SCC of 9), again reflecting the strongly directed nature of the network. The eight algorithms presented above vary in their efficiency in reducing network reach (figure 5). For maximum reach, the corresponding greedy method is initially the most effective, but later performs less well compared with the betweenness method. The two eigenvalue-based methods perform almost as poorly as arbitrary arc removal, and less well than the simple degree-based or inter-community approaches. In reducing mean reach, ordering of effectiveness of the eight methods is somewhat different: the corresponding greedy algorithm is consistently the most effective, but equally effective is betweenness. The R_0-like eigenvalue measure and the simple degree-based measure are moderately effective.

4 Discussion

The data presented above include only movement records between registered sites confirmed by validation of the paper record for both sender and receiver. In addition, *circa* 500 unconfirmed records were excluded where paper records did not match, as well as a further 4100 or so from or to unregistered sites (Munro & Gregory 2009). The majority of these latter records are to unregistered fisheries, mostly untraceable, which are sinks for movements only. Epidemiologically, these can be considered 'dead ends', at least via movements of live fish, though potentially not through their impact on the aquatic environment. Their exclusion from the dataset will not therefore affect the 'core' of the network. Furthermore, though these movements are large in number, they are of smaller numbers of fish; they are also predominantly of trout (both species). Under EU directive 2006/88/EC, fisheries are required from August 2008 both to be registered, and to record live fish movements, making these data available in future.

A large number of sites are linked to only one other site. These sites likely do interact more within the network, but do not do so within the single year (2003) of data examined in this study. In particular, the natural timeframe for salmon production is the two-year cycle from egg to smolt to adult.

Future extension to further years of data will show a more tightly connected network, at least for salmon. In datasets of longer timespan, modelling not only direction of movements, but also their relative timings will become increasingly important. For example, if movements exist from A to B and B to C, then C is only at risk of infection from A should the movement A to B occur earlier. This added complexity is difficult to incorporate into a network approach, but manageable through simulation models, as have been developed for sheep and cattle (Green et al. 2006; Green et al. 2008). Considering networks as static or undirected can lead to overestimation of the numbers of nodes that are at risk of infection, given a disease incursion, as examined for cattle movement networks by Dubé *et al.* (2008) and Vernon & Keeling (2009).

The most salient feature of the network is its strong community structure. This partly reflects low interaction between different species, but also within the salmon industry there are loosely connected portions of the network and disconnected fragments. If this pattern persists over multiple years, it provides benefits to surveillance and disease control, limiting the extent of potential epidemic spread through movements. Future work will investigate the structure of other routes for disease transmission that are not present in these data, such as shared use of well boats or presence in the same river catchment.

Arc removal demonstrates that the removal of a small number of arcs can have a disproportionally large reduction in the size of any potential epidemic, if the correct arcs are chosen. This was also demonstrated for the GB network of sheep movements by Kiss *et al.* (2006). For our network, the betweenness measure was found to work well as a strategy for selecting arcs, whereas eigenvalue measures performed relatively poorly. This is in contrast to the results of Bell *et al.* (1999), who found betweenness measures to perform worse than eigenvector centrality, albeit for identifying vulnerable nodes, rather than edge removal. Both Bell *et al.* (1999) and Christley *et al.* (2005b) found degree-based measures of vulnerability to perform well, and our degree-based measure for arc removal also performed acceptably. Zemljič and Hlebec (2005) note that centrality measures differ in their robustness, and are more reliable for dense networks. That the greedy algorithms are not optimal is unsurprising since such approaches often do not obtain a global optimum. (See figure A2 in electronic supplementary material.)

The two estimates for R_0 above differ, with the eigenvalue-based estimate higher than that based on node degree. Though the low correlation between *in* and *out* degree is accounted for by both estimates, the models differ in which other network features they capture. First, clustering decreases R_0 (Keeling 1999) and is accounted for by neither estimate; but levels of clustering here are small. Second, node degree-based measures do not take into account assortativity, which increases R_0 (Newman 2003; Kiss *et al.* 2008), and the fish network is slightly assortative. Third, the eigenvalue-based measure is less appropriate where the network is deeply cleft into distinct components. *In extremis*, where a network is completely divided into unlinked subnetworks, eigenvalue R_0 would represent only that subnetwork with the highest intrinsic R_0, whereas the degree-based measure would average across the entire network, giving a lower value as is seen here. Given the regional and sector differences in different farming industries, one must be therefore careful when estimating R_0, and avoid applying single-figure values as descriptors of processes which are too complex to allow them.

In summary, though the live fish movement network is comparatively small compared with other industries, it is nevertheless demonstrable that application of network-based statistical methods is more informative than simply examining the behaviour of nodes as individuals. Inter-site links identified as important through the arc-removal procedures above might be considered as a particular focus for targetted

surveillance, giving a more efficient use of limited surveillance resources. However for this, up-to-date, accurate data must be available, and specialised software is required. Additionally, one must consider disease-specific factors such as the timescale of disease, the relative size of different nodes in terms of the number of animals stocked, and other transmission routes: Water-borne, fomite, or airborne transmission will require the inclusion of other inter-site contact structures, crossing the boundaries between network modelling and metapopulation-based approaches.

Acknowledgements

With thanks to the Fish Health Inspectorate for providing access to the movement records. DMG's contribution to this work was funded through Marine Scotland. With thanks to Istvan Kiss for helpful comments on the manuscript.

5 References

Albert, R., Jeong, H., Barabási, A.-L. 2000. Error and attack tolerance of complex networks. Nature 406, 378 – 382.

Bell, C.B., Atkinson, J.S., Carlson, J.W. 1999. Centrality measures for disease transmission networks. Soc. Net. 21, 1 – 21.

Bell, E.T. 1934. Exponential Numbers. Amer. Math. Monthly 41, 411 – 419.

Bigras-Poulin, M., Barfod, K., Mortensen, S., Greiner, M. 2007. Relationship of trade patterns of the Danish swine industry animal movements network to potential disease spread. Prev. Vet. Med. 80, 143 – 165.

Bonacich, P., Lloyd, P. 2001. Eigenvector-like measures of centrality for asymmetric relations. Soc. Net. 23, 191 – 201.

Christley, R.M., Robinson, S.E., Lysons, R., French, N.P. 2005a. Network analysis of cattle movement in Great Britain. In: Mellor, D.J., Russell, A.M., Wood, J.L.N. (Eds.) Proc. Soc. for Veterinary Epidemiology and Preventive Medicine, Nairn, Scotland, 30 March–1 April, pp. 234 – 244.

Christley, R.M., Pinchbeck, G.L., Bowers, R.G., Clancy, D., French, N.P., Bennett, R., Turner, J. 2005b. Infection in social networks, Using network analysis to identify high-risk individuals. Am. J. Epidemiol. 162, 1024 – 1031.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C. 2001. Introduction to Algorithms, 2nd edition. MIT Press. Chapter 16.

Dent, J.E., Kao, R.R., Kiss, I.Z., Hyder, K., Arnold, M. 2008. Contact structures in the poultry industry in Great Britain: exploring transmission routes for a potential avian influenza virus epidemic. BMC Vet Res. 4, 27.

Diekmann, O., Heesterbeek, J.A.P. 2000. Mathematical epidemiology of infectious diseases, model building, analysis and interpretation. Wiley, Chichester, UK.

Dubé, C., Ribble, C., Kelton, D., McNab, B. 2008. Comparing Network Analysis Measures to Determine Potential Epidemic Size of Highly Contagious Exotic Diseases in Fragmented Monthly Networks of Dairy Cattle Movements in Ontario, Canada. Transboundary and Emerging Diseases 55, 382 – 392.

Green, D.M., Kiss, I.Z.,, Kao, R.R. 2006. Modelling the initial spread of foot and mouth disease through animal movements. Proc. R. Soc. B 273, 2729 – 2735.

Green, D.M., Kiss, I.Z., Mitchell, A.P., Kao, R.R. 2008. Estimates for local and movement-based transmission of

bovine tuberculosis in British cattle. Proc. R. Soc. B 275, 1001 – 1005.

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

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.

Kao, R.R., Green, D.M., Johnson, J., Kiss, I.Z. 2007. Disease dynamics over very different time-scales, Foot-and-mouth disease and scrapie on the network of livestock movements in the UK. J. R. Soc. Interface 4, 907 – 916.

Kiss, I.Z., Green, D.M., Kao, R.R. 2006. The network of sheep movements within Great Britain, Network properties and their implications for infectious disease spread. J. R. Soc. Interface 3, 669 – 677.

Kiss, I.Z. Green, D.M., Kao, R.R. 2008. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. J. R. Soc. Interface 5, 791 – 799.

Keeling, M.J. 1999. The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. B 266, 859 – 867.

Kumpula, J.M., Saramaki, J., Kaski, K., Kertesz, J. 2008. Limited resolution and multiresolution methods in complex network community detection. Fluct. Noise. Lett. 7, 209 – 214.

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. 2003. Mixing patterns in networks. Phys. Rev. E. 67, 026126.

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

Ribbens, S., Dewulf, J., Koenen, F., Mintiens, K., de Kruif, A., Maes, D. 2008. Type and frequency of contacts between Belgian pig herds. In: Peeler, E.J., Alban, L., Russell, A. (Eds.) Proc. Soc. for Veterinary Epidemiology and Preventive Medicine, Liverpool, UK, 26 – 28 March, pp. 155 – 171.

Stark, 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, 1 – 8.

Thrush, M., Peeler, E. 2006. Stochastic simulation of live salmonid movement in England and Wales to predict potential spread of exotic pathogens. Dis. Aqua. Org. 72, 115 – 123.

Truscott, J., Garske, T., Chis-Ster, I., Guitain, J. Pfeiffer, D., Snow, L., Wilesmith, J., Ferguson, N.M., Ghani, A.C. 2007. Control of a highly pathogenic H5N1 avian influenza outbreak in the GB poultry flock. Proc. Roy. Soc. B 274, 2287 – 2295.

Vernon, M.C., Keeling, M.J. 2009. Representing the UK's cattle herd as static and dynamic networks. Proc. R. Soc. B 276, 469 – 476.

Watts, D.J., Strogatz, S.H. 1998. Collective dynamics of 'small-world' networks. Nature 393, 440 – 442.

Webb, C. R. 2005. Farm animal networks, Unraveling the contact structure of the British sheep population. Prev. Vet. Med. 68, 3 – 17.

Zemljič, B., Hlebec, V. 2005. Reliability of measures of centrality and prominence. Soc. Net. 27, 73 – 88.

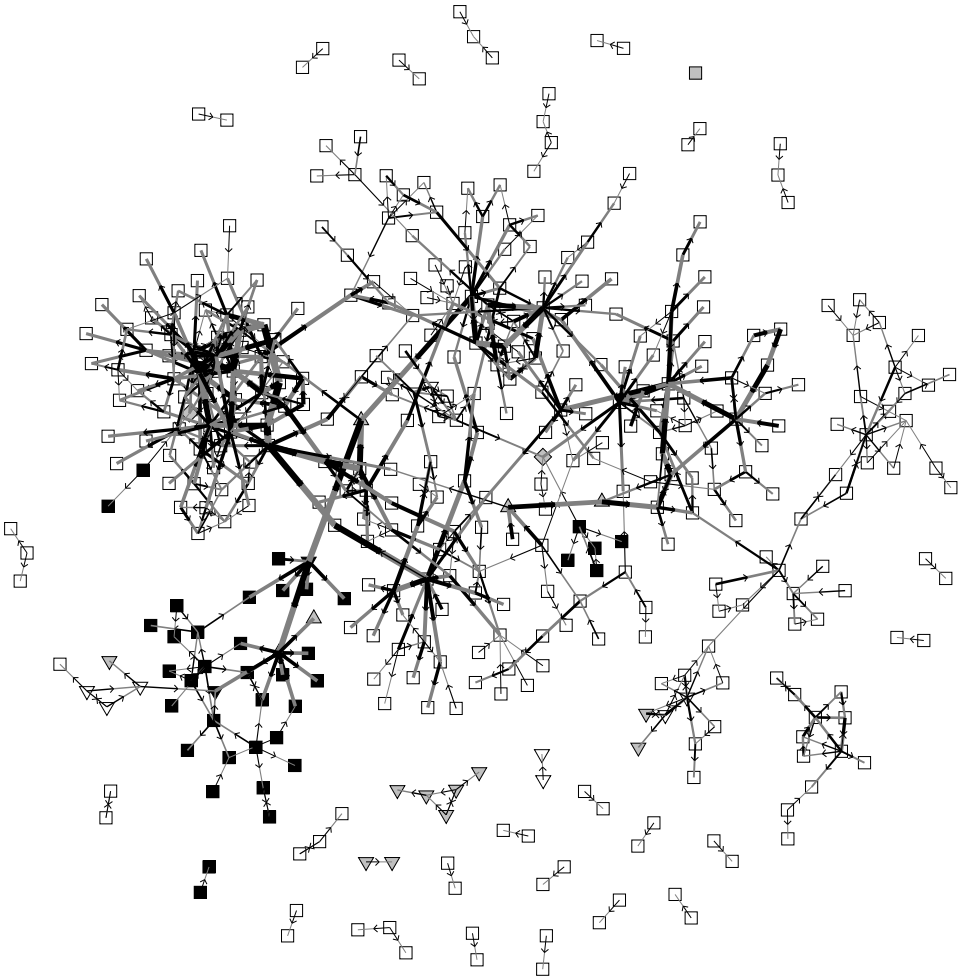
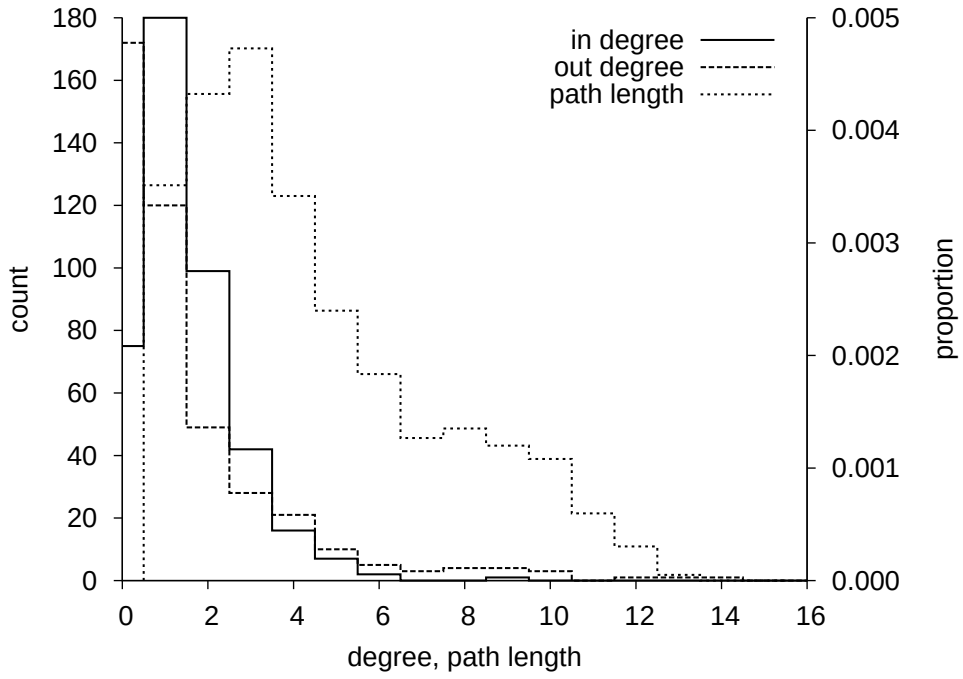


Figure 1: The 2003 Scottish live fish movement network, with sites coded according to species moved between sites. □ salmon (S); ■ rainbow trout (R); ♦ S+R; ▼ brown trout (T); ▼ T+R; ▽ T+S; ▲ T+S+R; ■ self-loops only. The direction of arcs is indicated by arrows and shading (darker half-arc for source), and their relative betweenness ($\log_e$-scale) indicated is by line width.

a)



b)

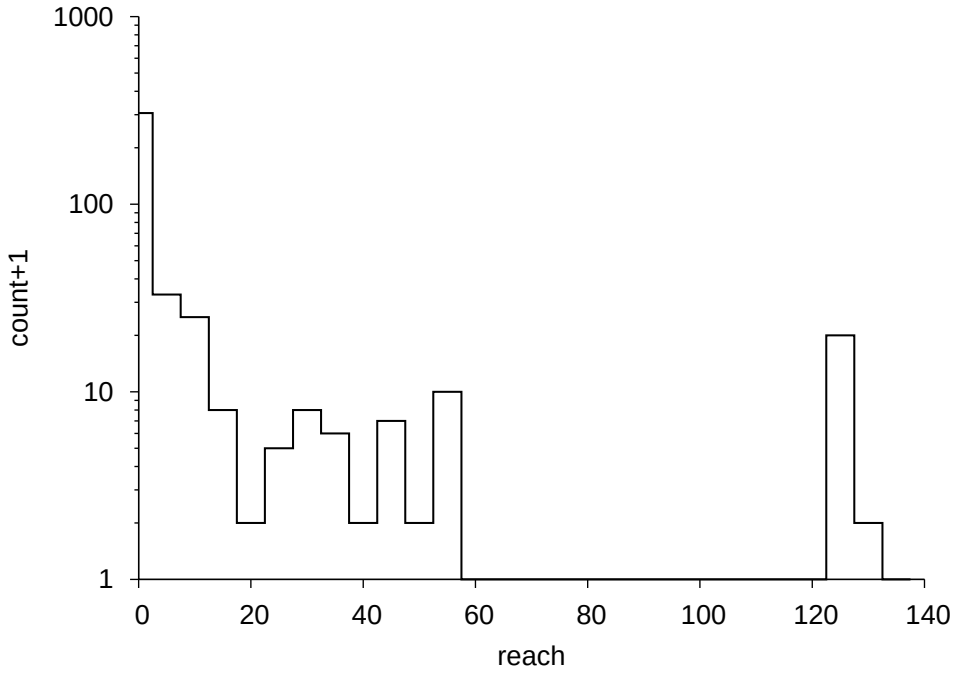


Figure 2: Descriptive statistics for the movement network. a) Histograms of *in* and *out* degree distribution (left axis) and of path length distribution (right axis). Number of paths of length l are expressed as a proportion of the potential total $n(n-1)$. b) Histogram of node reach.

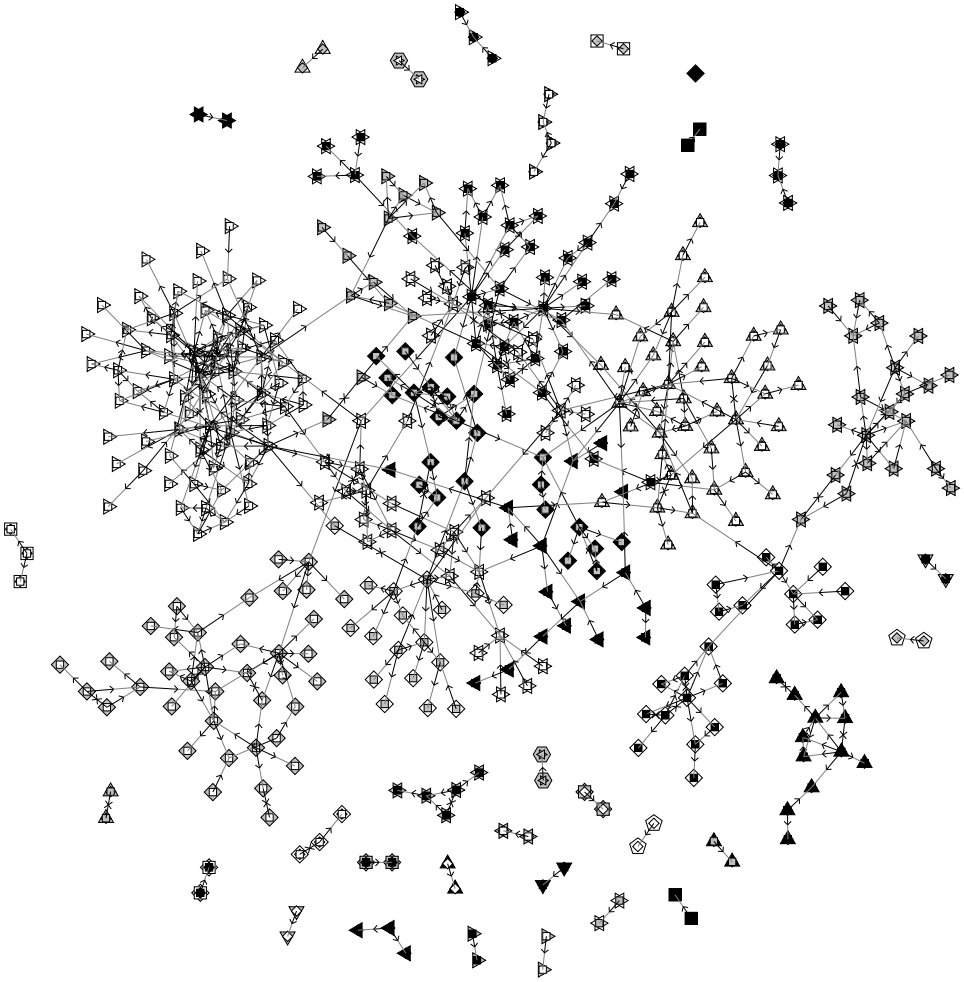


Figure 3: Community assignment for the live fish movement network for Scotland in 2003. Community membership is indicated by different symbols.

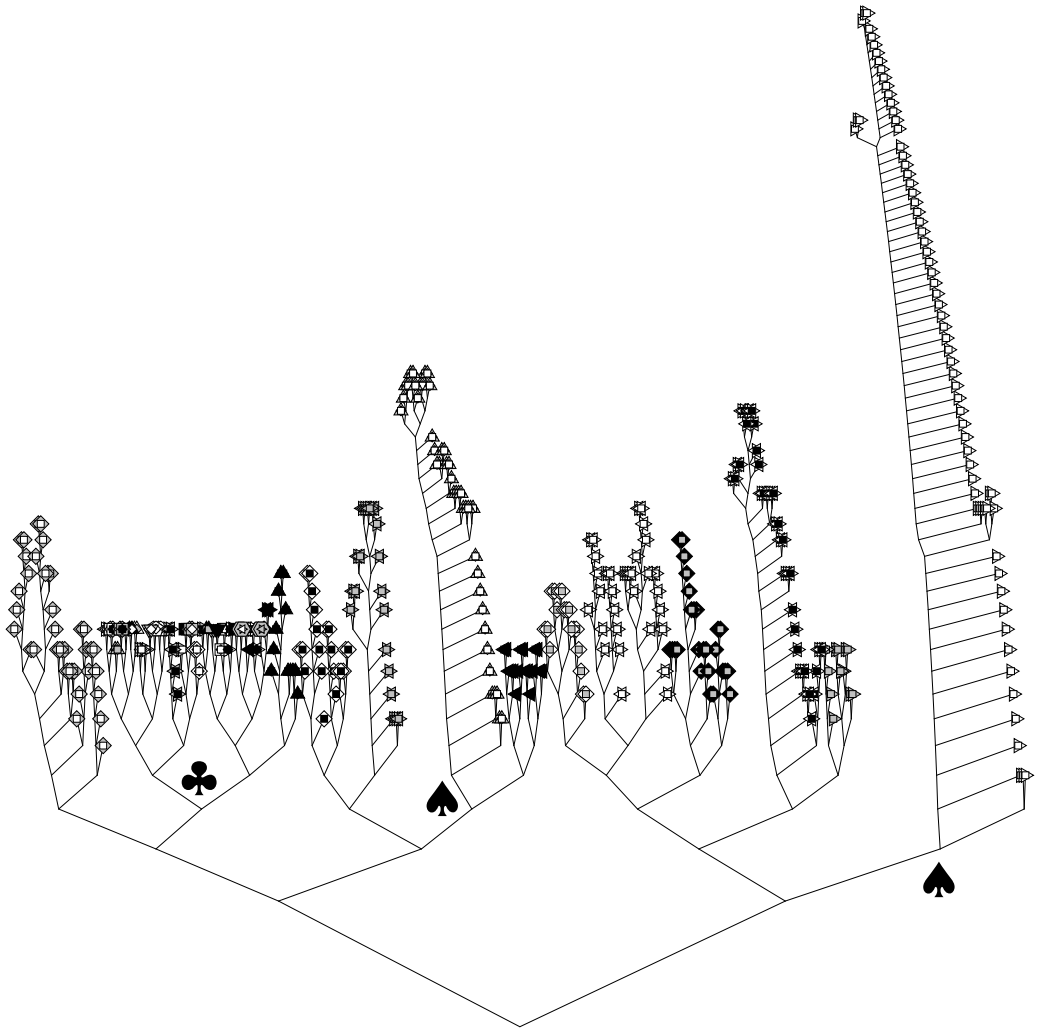


Figure 4: Dendrogram for the community algorithm. Each branch represents a group of nodes that are merged by the algorithm into the same community before they are merged into another such group, reading top-down. Best-fit communities are as shown on Figure 3. Symbols indicate 'plume' ♠ and 'binary branching' ♣ structures described in the text.

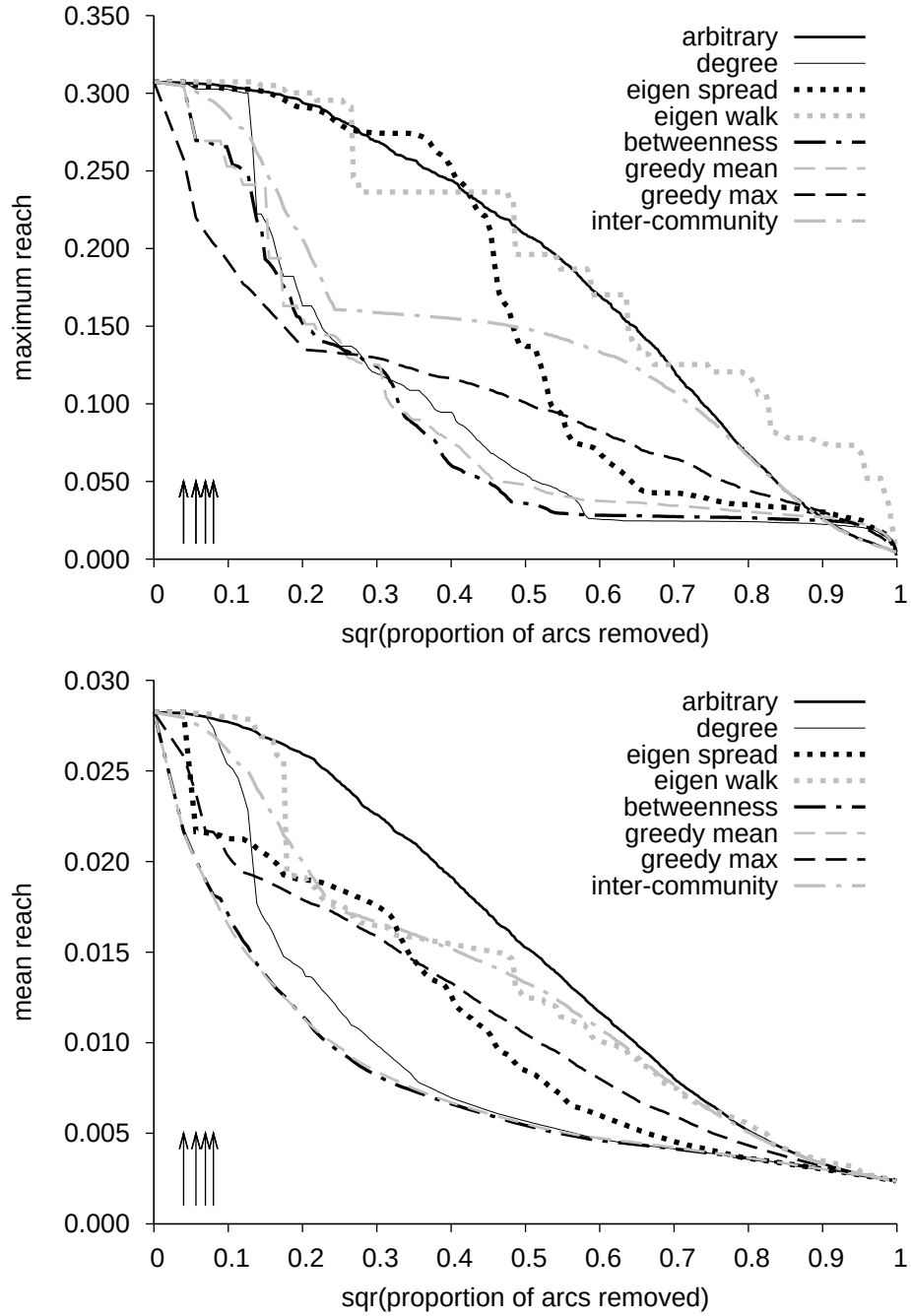


Figure 5: Mean and maximum reach from all nodes, versus proportion of network arcs removed (plotted on a square-root scale), for eight different algorithms for determining precedence of arc removal. Arrows indicate x-axis values corresponding to the removal of 1, 2, 3, and 4 arcs.