Classification

Major category: Biological Sciences; minor category: Agricultural Sciences

Title: Sheep movement network and the transmission of infectious diseases

V.V. Volkova, R. Howey, N.J. Savill and M.E.J. Woolhouse

Author affiliation: Centre for Infectious Diseases, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, Kings Buildings, West Mains Road, Edinburgh, EH9 3JT, UK

Corresponding author:

Dr. Victoriya Volkova,

Centre for Infectious Diseases, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, Kings Buildings, West Mains Road, Edinburgh, EH9 3JT, UK

Phone: +44-(0)131-650-5446

Fax: +44-(0)131-650-6564

e-mail: Victoriya.Volkova@ed.ac.uk

Manuscript information: abstract page and further 15 text pages (including references, and figure and table legends), two composite figures, and three tables.

Data deposition: None; records of sheep movements in Scotland 2003 to 2007 used in these analyses can be requested from the Scottish Government.

ABSTRACT

Livestock farms can be considered as a network connected by movement of animals, each movement between a pair of farms representing a 'contact'. The potential for transmission of an infectious disease through this network can be expressed in terms of the basic reproduction number, R_0. The value of R_0 is related to the mean contact rate, the variances in contact rates, the co-variance between contact rates to and from farms, and so-called 'higher order' effects which can be evaluated by calculating the dominant eigenvalue of the contact matrix, ε. Here, we calculate these quantities for the population of ~15000 sheep farms in Scotland. ε has not previously been calculated for such large contact matrices. We find that, depending on precisely how contact rates are defined, higher order effects can have a very substantial impact, increasing R_0 by up to a factor of 6. Previous studies have suggested a rule that 20% of the population typically contribute at least 80% of R_0. We consider different algorithms for identifying the 20% making the greatest contribution (noting that computing ε for all possible 20% subsets is not feasible). The best performing algorithm confirms that the 20-80 rule applies to this population. We conclude that it is not possible to quantify R_0 for this system (and presumably other contact networks too) without calculating ε, requiring knowledge of the complete network. Our methodology provides a means of identifying farms where interventions are best targeted, even for large populations where this has not previously been possible.

\body

INTRODUCTION

Understanding the structure of contact networks is important for predicting and controlling the spread of infectious diseases (1-3). Examples of contact networks are provided by records of movements of livestock between farms. Livestock movements play an important role in the spread of many diseases (4). In Britain, comprehensive, computerized record keeping systems for movements have been in place for cattle since 1998 and for sheep since 2002. These data have been widely used in studies of the epidemiology of a variety of livestock diseases: foot-and-mouth disease in cattle and sheep (5), bovine tuberculosis in cattle (6, 7); scrapie in sheep (8, 9). In parallel with these disease-specific studies there have also been a number of studies of the generic properties of livestock movement networks relating to the spread of infectious disease. These have taken two approaches to characterising movement networks.

One approach is to calculate the size of the giant connected components of the network. These measure the size of the largest linked subsets of the population. For directed networks (where the link may be in one direction or the other or both) the giant strongly connected component (GSCC) is the largest subset of the population linked by bi-directional contacts, and the giant weakly connected component (GWCC) is the largest subset of the population linked by any contact (10). Therefore the GSCC and GWCC provide lower and upper bounds respectively to maximum epidemic size. The giant out-component (GOC) is the subset of the population approachable from the GSCC by a direct path (10); it therefore includes the GSCC itself and all the farms which can be reached directly from the GSCC. An increase in the size of the GSCC of the network of British cattle farms was reported after new regulations governing the movement of cattle in the UK were introduced between 2001 and 2003 (11). This result implies that the potential scale of infectious disease epidemics in British cattle may have subsequently increased rather than decreased.

A second approach is to calculate the basic reproduction number, R_0. In this context R_0 is a measure of the expected average number of secondary cases generated from a single primary case introduced into a naïve population (12). Using this measure, Woolhouse, Shaw *et al.* (13) concluded that cattle movement networks in Scotland were consistent with the '20-80' rule, which states that 20% of the population

contribute at least 80% of R_0. Interventions targeted at these farms could therefore be particularly effective in reducing the size of epidemics or the level of endemic infection. The relationship between R_0 and the giant connected components is discussed by Kao, Danon *et al.* (14). An important distinction is that R_0 is a function of the *rates* of contact between members of the population whereas the giant components are static measures of the connectedness of a network (15).

In general, R_0 is related to the dominant eigenvalue of the contact matrix for the population (12). Calculation of this eigenvalue requires knowledge of the complete network (i.e. all contacts by all members of the population). Eigenvalue calculations for very large, sparse contact matrices are challenging (see below) and even though the necessary information is available for British livestock movement networks such calculations have not previously been attempted. Earlier work on livestock movement and other (e.g. human sexual contact) networks has focused on the contribution of the variance in contact rates and, for networks with bi-directional links, the covariance between contact rates in either direction (3, 13, 14, 16). More generally, for any contact network, contributions to R_0 can be partitioned into a first order moments (relating to the mean contact rate), second order moments (relating to the variances and co-variances in contact rates) and higher order moments (17).

Here, we analyse a contact network based on the movements of sheep between farms in Scotland. Given knowledge of the complete network we calculate the sizes of the weakly and strongly connected giant components, the giant out-component, and the basic reproduction number; and quantify the contributions to the latter of the first, second and higher order moments. These calculations allow us to identify which features of the network structure and which individual farms contribute the most to the potential for infectious disease spread through the network. We do not focus here on specific infections; however, because we choose a long time span (one year) and do not attempt to capture the early dynamics of disease outbreaks (which requires knowledge of the ordering of contacts) our results are most directly relevant to endemic chronic infections.

RESULTS

Descriptive network statistics. Descriptive population and network statistics for Scottish sheep network for the four years studied are given in full in Table 1. In summary, each year the number of farms in the network, N, was greater than 15,000, with approximately 70,000 uni-directional connections between the farms. Over 100,000 sheep batches were moved within the network per year, totalling more than 2,000,000 sheep. Approximately half of the farms that recorded moving sheep within Scotland each year were part of the GSCC, two-thirds were part of the GOC, and over 98% were part of the GWCC.

The quantities making up the calculations of R_0 (Models 1-3) are given in full in Table 2. The mean numbers of contacts per year were within the range 4.3 to 4.7 for the four years of interest (Table 2*A*). The distributions of the numbers of in-contacts, β_{in}, and out-contacts, β_{out}, made by individual farms in one year were highly over-dispersed (Fig. 1*A*), with a small fraction of the farms making large numbers of contacts. The variances in the numbers of in-contacts were much greater than that of out-contacts (Table 2*A*). The linear correlations between the numbers of annual in-contacts and out-contacts of the farms, $r_{\beta in \beta out}$, were positive but weak over the four years studied (Pearson correlation coefficient +0.07 to +0.11, all $p<0.001$) (Fig. 2*A*, and Table 2*A*).

The mean numbers of batches of sheep received by (or sent from) a farm was in the range 6.7 to 7.8 over the four years (Table 2*B*). The variances in the numbers of batches received by a farm were much greater than that in the numbers of batches moved off (Fig. 1*B,* and Table 2*B*). The linear correlations between the numbers of batches moved on and off the farms in a year were slightly lower (Pearson correlation coefficient +0.04 to +0.07, all $p<0.001$) (Fig. 2*B*, and Table 2*B*) than the correlations between the numbers of annual in-contacts and out-contacts.

The mean numbers of sheep received by (or sent from) a farm was in the range 137 to 144 per year (Table 2*C*). The variances in the numbers of sheep received by a farm annually were much greater than that in the numbers of sheep moved off (Fig. 1*C*, and Table 2*C*). The linear correlations between the numbers of sheep moved on and off the farms in a year were higher (Pearson correlation coefficient +0.18 to +0.36, all $p<0.001$) (Fig. 2*C*, and Table 2*C*) than the correlations between the annual numbers

of batches moved on and off or between the numbers of annual in-contacts and out-contacts.

Impact of network properties on basic reproduction number. Using unweighted a_{ij} values (Model 1) the net contribution of the second order moments of the contact network was to increase (from that contributed by the first order moments alone) the magnitude of R_0 by up to a factor of 2 (Table 2*A*, *Column 6 versus Column 2*). However, the higher order moments acted to decrease the magnitude of R_0 by as much as 13% (Table 2*A*, *Column 7 versus Column 6*). Using a_{ij} values weighted by numbers of batches (Model 2) the net contribution of the second order moments of the contact network was to increase the value of R_0 by up to a factor of 3 (Table 2*B*, *Column 6 versus Column 2*). The higher order moments further increased the magnitude of R_0 by up to 71% (Table 2*B*, *Column 7 versus Column 6*). Using a_{ij} values weighted by numbers of sheep (Model 3) the net contribution of the second order moments was to increase the value of R_0 by up to a factor of 6 (Table 2*C*, *Column 6 versus Column 2*). The higher order moments of the network acted to further increase the magnitude of R_0 by a factor of between 2 and 6 (Table 2*C*, *Column 7 versus Column 6*).

Method to identify farms contributing the most to R_0. Of the six methods considered for identifying a 0.2*N* subset of farms contributing the most to R_0 based on current year contact information, Method 6 consistently performed the best in terms of the reduction achieved in R_0 when the contacts made by identified farms were removed. Targeting the 20% subset of farms identified by Method 6 from the current year's network resulted in 86.8% to 88.0% reduction in the magnitude of R_0 for unweighted contacts (no other method achieved 80%), 92.6% to 95.1% for contacts weighted by numbers of batches (no other method achieved 91%), and over 99% for contacts weighted by numbers of sheep moved (no other method achieved 99%) (Table 3, *Column 2*).

Using preceding year's information to identify farms contributing the most to R_0. When sheep movement data from the preceding year were used to identify farms to include in the 20% subset, the resulting reductions in the value of R_0 were consistently smaller, and also more variable, compared with using data from the current year (Table 3, *Column 3 versus Column 2*).

DISCUSSION

Although there are numerous studies of contact networks as they relate to transmission of infectious diseases reported in the scientific literature very few of these investigate complete networks, and those that do have generally dealt with small populations (17). Livestock movement databases allow analyses of complete networks (here, covering the entire population of Scottish sheep farms). Another feature of the majority of studies of contact networks is that they consider bi-directional, often symmetrical, contacts. Again, livestock movement databases are unique in recording uni-directional contacts where movement of livestock from farm j to farm i is associated with risk of disease transmission only in that direction (13). This paper therefore provides information on contact network structure and its relationship to the potential spread of infectious diseases not available from previous studies.

Calculation of the giant weakly connected component of the network, given a relatively small number of movements from farms outside Scotland, confirms that Scottish sheep farms can be regarded as a single population connected by sheep movements for the purposes of these analyses. The size of this component relative to the size of the network confirms that the Scottish sheep industry is inter-connected: in contrast, for example, to the commercial pig industry where movements are constrained within sub-networks. Notably, this large connected component emerges even though the contract matrix itself is very sparse (with approximately 0.03% non-zero entries in a year) reflecting that, on average, in a given year each farm moves sheep to or receives sheep from less than five other Scottish farms. The size of the giant out-component confirms that a long-lasting infection introduced into this farm population within a year can be transmitted directly to nearly 70% of the farms through the movements of sheep alone.

We can then use calculation of the basic reproduction number, R_0, as a method to characterise the properties of the network of contacts between Scottish farms through sheep movements and how these properties relate to the spread of infectious diseases within that population of farms. However, for a number of reasons these calculations do not represent formal estimates of R_0 for any specific infectious disease. First, as indicated in Expressions [1]-[4], we generate relative, not absolute, measures of the

magnitude of R_0. Nor are the different contact formulations (unweighted, batch-weighted and individual animal-weighted contact) directly comparable amongst themselves (each being most relevant to certain epidemiological scenarios, as discussed above). Secondly, we aggregate all movements over a one-year interval to provide a measure of relative contact rates for the farms. This does not account for temporal heterogeneities within the year, in particular marked seasonality in Scottish sheep movements; these can significantly affect R_0 (22) and could influence the results reported here if temporal variations in contact rates were poorly correlated across the farms. Finally, although movement of livestock is an important risk factor for the spread of many livestock infections, it is not typically the only factor; other routes of transmission between farms may be relevant for specific applications.

The size of Scottish sheep farm network and the sizes of its weakly and strongly connected giant components were broadly consistent across the four years studied (Table 1). However, there were year-to-year fluctuations in the relative magnitude of R_0, particularly when contacts were weighted by the numbers of sheep moved: more than doubling over the period of study (Table 2*C*). These fluctuations were not clearly related to changes in the contribution of the first or second order moments of the network (Table 2*C*). We conclude that during the four years studied there were changes in the higher order structure of the network associated with the numbers of sheep moved by individual farms which more than doubled the potential for transmission of infection through this population of farms. Similar changes in the magnitude of R_0 across the four years were not apparent for unweighted contacts or contacts weighted by the numbers of batches moved (Table 2*A* and 2*B*).

Previous studies of contact networks have reported increases in the value of R_0 associated with heterogeneities in contact rates between individuals (17). Here we find that the size of such effects vary greatly according to how contacts are weighted. The effect (indicated by ratio of Expression [4] to Expression [1]) is modest (up to a 2-fold increase) if contacts between the farms are unweighted, larger (up to a 4-fold increase) if contacts are weighted by numbers of batches of sheep, and very substantial (up to a 30-fold increase) if contacts are weighted by total numbers of sheep moved. In other words, if the rate of transmission of infection between farms is related to the numbers of batches or, especially, to the numbers of individual sheep

moved between the farms then calculations based on knowledge of the mean contact rate alone would massively underestimate the true value of R_0.

Furthermore, we are able to partition the impact of heterogeneities in contact rates into the effects of second order moments (relating to the variances and covariances of movements to and movements from individual farms) and higher order moments (relating to further features of network structure). The second order moments significantly contribute to the overall impact of heterogeneities in contact patterns on R_0. However, this is still far less than might be anticipated from the very high variances in farm contact rates (3). The explanation is that there is only a weak correlation between the movements on and movements off individual farms (Table 2). Nonetheless, because these correlations are positive (if negative, the effect would be to reduce R_0, see Expression [2]) and the variances of contact rates are so high, the net effect is still substantial.

Importantly, we find that higher order properties of the Scottish sheep network greatly influence the overall value of R_0 in this farm population (Table 2). For contacts weighted by the numbers of sheep moved between the farms, these effects can increase R_0 by more than 5-fold. Hence, even calculations based on knowledge of the means, variances and covariances of contact rates (as previously considered, for example, for cattle movements (13)) would massively underestimate the true magnitude of R_0. Smaller, but still substantial, increases (up to 71%) due to the effects of higher order properties were found for contacts weighted by the numbers of batches moved. Interestingly, if contacts between farms are unweighted (just present or absent in a given direction) then the higher order effects show no impact or act to slightly decrease (by as much as -13%) the value of R_0. This last result implies that some features of the higher order structure of the unweighted contacts acts to decrease the potential for transmission of infection through this farm population.

Given the importance of heterogeneities in farm contact rates in determining R_0, it is apparent that targeting interventions at farms contributing the most to R_0 is likely to be highly efficient (17). In practice, targeted control can include such measures as livestock movement restrictions or pre-movement testing. The theoretically ideal method for ranking the farms in their contribution to R_0 (in terms of effects on the

dominant eigenvalue of the farm contact matrix) is not feasible for the reasons discussed above. Of the six alternative methods considered here, the best performing was to iteratively identify the farm with the highest value of the corresponding component in the dominant eigenvector of the contact matrix. For the Scottish sheep farm network, this method performed reasonably well across the four years of study for all three contact formulations (at least 86.8% reduction in the magnitude of R_0 when the top 20% of farms were targeted) (Table 3). The reduction was greatest (over 99%) for the contacts weighted by total numbers of sheep moved between the farms (Table 3).

However, information on contacts of farms in the preceding year was consistently less valuable for identifying the 20% of farms to target in the current year (Table 3). This result presumably reflected year-to-year variation in individual farms' contact rates, even though mean contact rates for a farm were relatively constant across the four years studied (Table 2). As to the processes underlying such variation in the Scottish sheep network, characterising the farms repeatedly or intermittently appearing in the 20% contributing the most to the potential for transmission of infections each year may provide further insights.

The key conclusions arising from this work are as follows. First, the higher order properties of a contact matrix (i.e. those not quantifiable from knowledge of the means, variances and covariances of contact rates) can have a substantial impact on the magnitude of R_0. Quantification of such effects requires knowledge of the complete network, which is rarely available for large populations. Second, the way in which contacts are weighted makes a very substantial difference to quantification of R_0 and its components. When and how contacts should be weighted is relatively straightforward for livestock movements, perhaps less so for other kinds of 'contact' between individuals in a population. Third, contact matrices may vary through time not only in terms of contact rates of individual members of the population but also in terms of other, higher order, properties, as has been reported previously for the UK cattle movement network (11) and observed here for Scottish sheep movement network. The wider applicability of these conclusions depends on how representative the livestock farm networks are of contact networks in general, but we conjecture that similar issues will arise in many other contexts.

METHODS

Sheep Movement Data. Records of sheep movements among Scottish holdings were obtained from the Scottish Animal Movement System (SAMS), operated by the Scottish Government. All SAMS entries for sheep from 2003 to 2007 were processed in the Python programming environment and then in SAS® 9.1.3 software for Windows (SAS Institute Inc., Cary, NC, USA). Up-to-date lists of sheep markets, show-grounds, abattoirs and other industry units registered in Scotland were collated with help from Livestock Traceability Policy of Animal Health and Welfare Division of the Scottish Government Rural Directorate and Animal Health agency in Scotland. The data were processed, including definitions of types of holdings and movements, as previously described (18). In short and pertinent to these analyses: the vast majority (99.6%) of the SAMS entries for sheep 2003 to 2007 were logical movement records, and the number of sheep movements not reported to SAMS during this period is believed to be low. A farm was included in the analyses if it either sent or received sheep from another Scottish farm directly or via a Scottish market during the time interval studied (movements to and from designated show-grounds and to slaughter were excluded). The movement data were divided into four one-year intervals: Year 1, 01/07/2003 to 30/06/2004; Year 2, 01/07/2004 to 30/06/2005; Year 3, 01/07/2005 to 30/06/2006; and Year 4, 01/07/2006 to 30/06/2007. The June/July dividing date precedes the major annual movement of sheep in the autumn. Seasonality in movement patterns is not considered further in these analyses.

During the period of study, the sheep identification and traceability regulations in Scotland did not require specification of individual animals in the movement documents (the Sheep and Goats Movement Interim Measures Scotland Order 2002 and Amendments; the Sheep and Goat Identification and Traceability Scotland Regulations 2006 and Amendments). Therefore the length of stay of an individual sheep on a given farm could not be determined. The required legally standstill period was 13 days, i.e. no sheep should have been moved off the farm earlier than 13 days after a sheep on-movement unless to slaughter, although certain categories of movements were exempt from the standstill. Sheep housed on mixed livestock farms were also subject to standstill after an on-movement of cattle (13 days), pigs (20 days) or goats (13 days).

The focus of these analyses was the network of Scottish sheep farms. For the purpose of these analyses the network was treated as closed and movements outside Scotland were ignored. In practice, cross-border movements onto Scottish farms, primarily from England and Wales, did occur, but at low rates (less than 2% of movements onto Scottish farms during the study period). Movements off Scottish farms to locations outside Scotland were much more frequent, but are not relevant here.

Within Scotland, the majority of sheep movements between the farms (>80% in each of the four years analysed) occurred via Scottish livestock markets. Since we are considering a relatively long time period (full year) we assume the potential for disease transmission during brief stays at markets to be negligible (noting that this assumption would not hold for acute infections which are transmitted over short time scales). Therefore, we treat any indirect movement from farm j to farm i via a market as equivalent to a direct movement from farm j to farm i.

Weighting of contacts. Let a_{ij} be the directed contact rate from farm j to farm i in a particular year. We assign values to a_{ij} in one of three ways. 1) contact scored as 0 (no movement of sheep from farm j to farm i) or 1 (any movement of sheep from farm j to farm i); 2) as (1) but weighted by the number of batches of sheep moved from farm j to farm i (noting that this is equivalent to the frequency of contact from j to i); and 3) as (1) but weighted by the number of sheep moved from farm j to farm i. Model 1 is most appropriate for a highly transmissible infection with high on-farm prevalence (i.e. likely to be transmitted by any contact). Model 3 is most appropriate for a rare infection with low on-farm prevalence (so the probability of transmission is low and linearly dependent on the number of sheep moved). Model 2 is intermediate between 1 and 3.

Giant network components. Connectedness of the farm network in each of the four years was evaluated by calculating the giant strongly connected component (GSCC), the giant weakly connected component (GWCC) and the giant out-component (GOC) of the network. The GSCC and GWCC were calculated with Tarjan's algorithm (19) implemented in C++. The GOC was calculated by choosing a farm from the GSCC and performing a depth-first search excluding cycles to identify every farm reachable

from the chosen farm by a direct path; this was implemented in C++. For a given year, the GSCC encompassed all farms linked by bi-directional contacts; the GOC encompassed the GSCC plus all farms reachable from the farms in GSCC by a direct path ('sinks'); and the GWCC encompassed the GSCC plus all farms connected to the farms in the GSCC by any uni-directional contact (both 'sources' and 'sinks').

Basic reproduction number. For all three models discussed above, the total in-contact rate for farm i is $\beta^i_{\text{in}}=\sum_j a_{ij}$, and the out-contact rate is $\beta^i_{\text{out}}=\sum_j a_{ji}$. R_0 is related to the mean contact rate. In a closed network $\sum_i \beta^i_{\text{in}}=\sum_i \beta^i_{\text{out}}$, therefore:

$$R_0 \propto \sqrt{\overline{\beta_{in}}\,\overline{\beta_{out}}} = \overline{\beta_{in}} = \overline{\beta_{out}} \qquad [1]$$

R_0 is further influenced by the second order moments of the contact matrix. We denote the standard deviation of in-contact rates as $\sigma(\beta_{\text{in}})$, the standard deviation of out-contact rates as $\sigma(\beta_{\text{out}})$, and the Pearson product-moment correlation coefficient between in-contact rates and out-contact rates as $r_{\beta in \beta out}$. As previously shown (13), R_0 (ignoring higher order moments – see below) is a function of these terms as follows:

$$R_0 \propto \sqrt{\overline{\beta_{in}}\,\overline{\beta_{out}} + \sigma(\beta_{in})\sigma(\beta_{out})r_{\beta_{in}\beta_{out}}} \qquad [2]$$

Therefore, non-zero variances of β_{in} and β_{out} can increase R_0 if β_{in} and β_{out} are positively correlated. Expression [2] can be written in terms of the cross-product of β_{in} and β_{out}:

$$R_0 \propto \frac{\overline{\beta_{in}\beta_{out}}}{\overline{\beta_{in}}} = \frac{\overline{\beta_{in}\beta_{out}}}{\overline{\beta_{out}}} \qquad [3]$$

The contribution of second order moments to R_0 was evaluated as the ratio of the quantity calculated in Expression [3] to the quantity calculated in Expression [1].

However, R_0 is further influenced by higher order moments, the evaluation of which requires calculation of the dominant eigenvalue of the full contact matrix. Let **A** be

the contact matrix with elements a_{ij} and the magnitude of the dominant eigenvalue be ε. According to Barbour (20) and Diekmann and Heesterbeek *et al.* (12)

$$R_0 \propto \varepsilon \qquad [4]$$

For each year's contact matrix, the contribution of higher order moments to R_0 was evaluated as the ratio of the dominant eigenvalue to the quantity calculated in Expression [3].

Quantities [1] to [4] were calculated for the contact matrices where contacts between farms were weighted according to each of the three models described above.

Farms contributing the most to transmission. The contribution of a set of farms to R_0 was evaluated by calculating ε of a year's complete contact matrix and ε' of the matrix with the contacts made by the designated set of farms removed. Noting the '20-80' rule (3, 13) we focus here on the subsets of farms of size $0.2N$ in each year's network. Ideally we would compare all possible subsets of size $0.2N = M$ from the total population N, to identify the subset targeting which achieves the greatest reduction in R_0 (the smallest ε' of the resultant contact matrix). However, this would require $N!/M!(N-M)!$ calculations of ε, which is not feasible for N as large as the number of farms in annual Scottish sheep network.

We considered the following six methods for identifying $0.2N$ farms that contribute the most to R_0. We compared the reductions in R_0 achieved by targeting the set of M farms identified using each method for all three models and for each of the four years studied.

Method 1. The i=1 to M farms with the largest values of β^i_{in}.

Method 2. The i=1 to M farms with the largest values of β^i_{out}.

Method 3. The i=1 to M farms with the largest cross-products ($\beta^i_{\text{in}}\beta^i_{\text{out}}$).

Method 4. The i=1 to M farms with the highest individual contributions to ε, evaluated as the reduction in ε when all the farm's contacts are removed from the contact matrix.

Method 5. Identify the farm making the largest contribution to ε using Method 3, repeating the procedure for the remaining N-1 farms, and continuing until N-M farms remain.

Method 6. Identify the farm with the largest component in the eigenvector corresponding to ε, repeating the procedure for the remaining N-1 farms, and continuing until N-M farms remain.

Using the best performing method for identifying 0.2N of farms that contribute the most to R_0, we compared the reductions in R_0 when this set of farms was identified using their contact information for the year of interest (current year) versus such information from the preceding year - which is likely to be more readily available in practice.

The dominant eigenvalues of all contact matrices were calculated using the ARPACK FORTRAN77 code libraries written to solve large-scale eigenvalue problems, in particular for structured or sparse matrices (21). Each of the complete annual matrices had a dimension over 15,000, with around 70,000 (0.03%) non-zero entries. The approximate calculation time to identify the 0.2N set of farms that contribute the most to R_0 and quantify the effects of removal of the contacts made by these farms for one-year network on a given contact formulation, using a UNIX platform with GNU-complied C++ code, was less thank 10 seconds for Methods 1 to 3, 25 minutes for Method 5, and 16 hours for Methods 4 or 6.

ACKNOWLEDGEMENTS. This work was undertaken through the Centre of Excellence in Epidemiology, Population Health and Infectious Disease Control funded by the Scottish Government. We thank Gary Ferguson for continued assistance, and Kevin Duffy and other Scottish Government personnel for work on the SAMS extracts. We are thankful to Derek Wilson and Paul Honeyman for the information on livestock markets, abattoirs, show-grounds and dealerships in Scotland. We appreciate the support of Nick Ambrose and Mike Lamont. We thank Martin Miller for comments on the manuscript. RH and MEJW acknowledge additional support from the BBSRC and the Wellcome Trust.

References

1. Galvani AP,May RM (2005) Epidemiology: dimensions of superspreading. *Nature* 438:293-295.
2. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438:355-359.
3. Woolhouse ME*, et al.* (1997) Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc Natl Acad Sci U S A* 94:338-342.
4. Fèvre EM, Bronsvoort BM, Hamilton KA, Cleaveland S (2006) Animal movements and the spread of infectious diseases. *Trends Microbiol* 14:125-131.
5. Ortiz-Pelaez A, Pfeiffer DU, Soares-Magalhaes RJ, Guitian FJ (2006) Use of social network analysis to characterize the pattern of animal movements in the initial phases of the 2001 foot and mouth disease (FMD) epidemic in the UK. *Prev Vet Med* 76:40-55.
6. Gilbert M*, et al.* (2005) Cattle movements and bovine tuberculosis in Great Britain. *Nature* 435:491-496.
7. Green DM, Kiss IZ, Mitchell AP, Kao RR (2008) Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. *Proceedings of the Royal Society B-Biological Sciences* 275:1001-1005.
8. Green DM*, et al.* (2007) Demographic risk factors for classical and atypical scrapie in Great Britain. *Journal of General Virology* 88:3486-3492.
9. Kao RR, Green DM, Johnson J, Kiss IZ (2007) Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK. *Journal of the Royal Society Interface* 4:907-916.
10. Dorogovtsev SN, Mendes, J.F.F. (2003) Evolution of Networks. From Biological Nets to the Internet and WWW (Oxford University Press, New York).
11. Robinson SE, Everett MG, Christley RM (2007) Recent network evolution increases the potential for large epidemics in the British cattle population. *Journal of the Royal Society Interface* 4:669-674.
12. Diekmann O, Heesterbeek JAP, Metz JAJ (1990) On the Definition and the Computation of the Basic Reproduction Ratio R_0 in Models for Infectious-Diseases in Heterogeneous Populations. *Journal of Mathematical Biology* 28:365-382.
13. Woolhouse MEJ*, et al.* (2005) Epidemiological implications of the contact network structure for cattle farms and the 20-80 rule. *Biology Letters* 1:350-352.
14. Kao RR, Danon L, Green DM, Kiss IZ (2006) Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings of the Royal Society B-Biological Sciences* 273:1999-2007.
15. Vernon MC,Keeling MJ (2009) Representing the UK's cattle herd as static and dynamic networks. *Proceedings* 276:469-476.
16. Kiss IZ, Green DM, Kao RR (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.

17. Woolhouse MEJ, Watts CH, Chandiwana SK (1991) Heterogeneities in Transmission Rates and the Epidemiology of Schistosome Infection. *Proceedings of the Royal Society of London Series B-Biological Sciences* 245:109-114.
18. Volkova V, Savill NJ, Bessell PR, Woolhouse MEJ (2008) Report on seasonality of movements and spatial distribution of sheep, cattle and pigs in Scotland. Report to Animal Health and Welfare Division of the Scottish Government's Rural and Environment Research and Analysis Directorate.:ISBN 978 970 7559 1718 7551.
19. Tarjan R (1972) Depth-first search and linear graph algorithms. *SIAM J Comput* 1:142-160.
20. Barbour AD (1978) Macdonald's model and the transmission of bilharzia. *Trans Roy Soc Trop Med Hyg* 72:6-15.
21. Lehoucq RB, Sorensen DC, Yang C (1997) ARPACK Users' Guide Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, Technical Report, Computational and Applied Mathematics, Rice University, USA.
22. Lord CC, Woolhouse MEJ, Heesterbeek JAP, Mellor PS (1996) Vector-borne diseases and the basic reproduction number: A case study of African horse sickness. *Medical and Veterinary Entomology* 10:19-28.

Figure legends

Figure 1a. Distribution of numbers of Scottish farms that a Scottish farm either received sheep from (in-contacts) or sent sheep to (out-contacts) from 01/07/2006 to 30/06/2007.

Figure 1b. Distributions of numbers of sheep batches moved on and off a Scottish farm from 01/07/2006 to 30/06/2007.

Figure 1c. Distributions of numbers of sheep moved on and off a Scottish farm from 01/07/2006 to 30/06/2007.

Figure 2. Co-distribution of a) out-contacts and in-contacts, b) numbers of batches moved off and on, and c) numbers of sheep moved off and on (all transformed as $\log_{10}[x+1]$) for Scottish sheep farms from 01/07/2006 to 30/06/2007.

Table legends

Table 1.Summary statistics for the network of Scottish farms connected by sheep movements.

Table 2. Properties of contact matrices for Scottish farms based on sheep movements.

Table 3. Contribution to R_0 of top 20% of farms identified from current year versus preceding year contact information.

Table 1. Summary statistics for the network of Scottish farms connected by sheep movements.

Year	Number of farms	Number of uni-directional contacts between farms	Number of sheep batches moved between farms	Number of sheep moved between farms	Size of giant strongly connected component (fraction of farms)	Size of giant out-component (fraction of farms)	Size of giant weakly connected component (fraction of farms)
Year 1	15,788	72,067	116,973	2,217,940	0.516	0.670	0.989
Year 2	15,314	71,999	118,957	2,118,099	0.505	0.666	0.989
Year 3	15,762	68,952	108,978	2,162,764	0.486	0.651	0.986
Year 4	15,750	68,347	105,500	2,266,971	0.491	0.669	0.986

Table 2. Properties of contact matrices for Scottish farms based on sheep movements.

Year	Mean contacts	Variance in-contacts	Variance out-contacts	Correlation in-contacts and out-contacts	Ratio of mean cross-product in-contacts*out-contacts to mean contacts	Dominant eigenvalue of contact matrix
A) Unweighted contacts between farms						
Year 1	4.6	1,302.5	48.4	+0.081	9.0	8.3
Year 2	4.7	1,289.5	50.1	+0.077	8.9	6.2
Year 3	4.4	826.6	44.7	+0.074	7.6	7.6
Year 4	4.3	663.2	46.4	+0.109	8.8	7.6

B) Contacts weighted by numbers of batches moved between farms						
Year 1	7.4	29,893.8	148.3	+0.054	22.7	25.5
Year 2	7.8	31,203.7	172.2	+0.035	18.3	19.2
Year 3	6.9	7,750.8	120.8	+0.059	15.2	25.9
Year 4	6.7	5,830.6	124.3	+0.072	15.8	23.8
C) Contacts weighted by numbers of sheep moved between farms						
Year 1	140.5	2,600,728.0	96,718.7	+0.200	855.3	1,981.4

Year 2	138.3	2,203,887.0	82,795.6	+0.183	704.0	2,596.9
Year 3	137.2	483,045.2	85,900.0	+0.360	670.9	3,800.6
Year 4	143.9	555,775.1	168,741.1	+0.303	789.5	4,364.8

Table 3. Contribution to R_0 of top 20% of farms identified from current year versus preceding year contact information.

Year	% reduction in R_0 based on removal of top 20% of farms of current year	% reduction in R_0 based on removal of top 20% of farms of preceding year
A) Unweighted contacts between farms		
Year 1	88.0	-
Year 2	87.4	64.2
Year 3	86.8	66.0
Year 4	86.9	61.3
B) Contacts weighted by		

numbers of batches moved between farms		
Year 1	94.5	-
Year 2	92.6	68.0
Year 3	95.1	32.8
Year 4	94.1	43.6
C) Contacts weighted by numbers of sheep moved between farms		
Year 1	99.3	-

Year 2	99.5	59.0
Year 3	99.7	67.8
Year 4	99.7	82.9

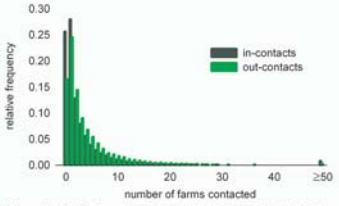

Figure 1a. Distribution of numbers of Scottish farms that a Scottish farm either received sheep from (in-contacts) or sent sheep to (out-contacts) from 01/07/2006 to 30/06/2007.

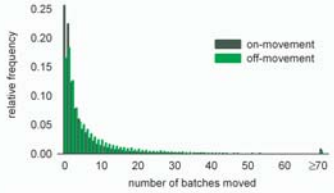

Figure 1b. Distributions of numbers of sheep batches moved on and off a Scottish farm from 01/07/2006 to 30/06/2007.

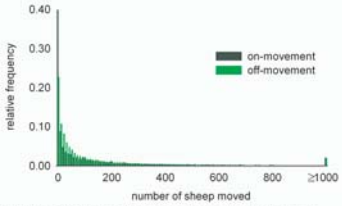

Figure 1c. Distributions of numbers of sheep moved on and off a Scottish farm from 01/07/2006 to 30/06/2007[a].

[a] Numbers of sheep were devided by 10 to reduce the number of frequency classes; the corresponding exact numbers are displayed on the plot.

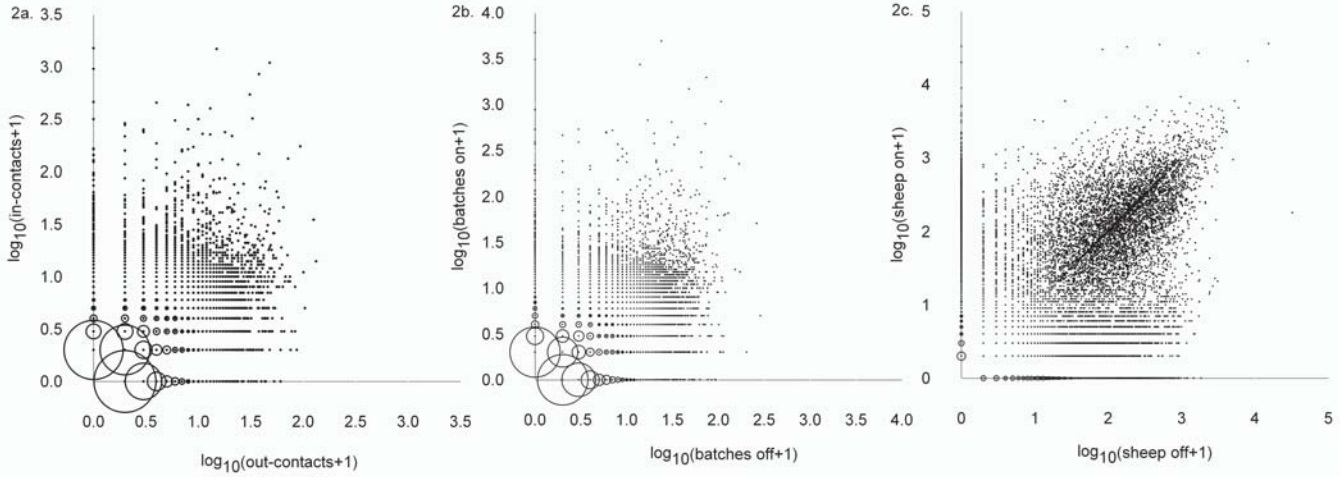

Figure 2. Co-distribution of a) out-contacts and in-contacts, b) numbers of batches moved off and on, and c) numbers of sheep moved off and on (all transformed as $\log_{10}[x+1]$) for Scottish sheep farms from 01/07/2006 to 30/06/2007. Diameter of the mark is proportional to the number of observations.