Supplementary information for:

Analysing livestock network data for infectious diseases control: an argument for routine data collection in emerging economies

G.L. Chaters[1,a], P.C.D. Johnson[1,a], S. Cleaveland[1], J. Crispell[2], W.A. de Glanville[1], T. Doherty[3], L. Matthews[1], S. Mohr[1], O.M. Nyasebwa[4], G. Rossi[3], L.C.M. Salvador[3, 4, 5], E. Swai[6], R.R.Kao[3,b]

[1] Boyd Orr Centre for Population and Ecosystem Health, Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK
[2] School of Veterinary Medicine, University College Dublin, Ireland
[3] Royal (Dick) School of Veterinary Studies and Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, Scotland, UK
[4] Department of Infectious Diseases, University of Georgia, Athens, Georgia, USA
[5] Institute of Bioinformatics, University of Georgia, Athens, USA
[6] Department of Veterinary Services, Ministry of Livestock and Fisheries, Tanzania;

[a] These authors contributed equally to this work
[b] Corresponding author

Network dynamics

Eigenvector centrality and dynamic infection processes.

If the contact network between nodes in a network is represented as an unweighted adjacency or contact matrix M (where a link between two individuals i and j is represented by a 1 at position m_{ij} or 0 otherwise), then for some initial disease vector v_0, and some probability of infection p, the product $p.M.v_0$ estimates the distribution of infection probability across all nodes in the first generation so long as the clustering coefficient (i.e. the proportion of triplet nodes where A is connected to B is connected to C and A and C are also connected) and/or p are sufficiently small. For an *SIS* infection process where susceptible individuals when exposed become immediately infectious and upon recovery, become completely susceptible again and the probability of recovery before re-infection is high (e.g. if the density of infected locations is always low, or the recovery time is shorter than the intergeneration time), then $p^n.M^n.v_0$ is an estimate of the distribution of infection probability in the n^{th} generation. We shall assume here that the matrix is square and irreducible (this is only the case if all nodes are part of the same strong component), and note that it is also positive definite (i.e. real and non-negative); in this case, the Perron-Frobenius theorem applies and the lead eigenvalue of the matrix is guaranteed to be real and positive. By definition, for an eigenvector e_i of M (i = 1 to n_p, where n_p is the dimension of the matrix),

$$\begin{aligned}\lim_{n\to\infty} \sqrt[n]{(p.M)^n}\,.\,e &= p.\lim_{n\to\infty} \sqrt[n]{M^n}\,.\,e \\ &= \lambda.\,e\end{aligned}$$

where λ is an eigenvalue of M, and $e_i = \sum_{j=1}^{n_p} \alpha_j u_j$ where α_j are coefficients associated with the unit vectors u_j. As any vector v can be written as a sum of eigenvectors, provided v includes the lead eigenvector, then not only is α_j the eigenvector centrality value associated with node j, it also represents the relative proportion of time that the node is infected over the long-term evolution of the epidemic. This is analogous to the next generation matrix (NGM) definition of the basic reproduction number R_0 [136], except that the contact matrix considers individuals, while the NGM definition considers populations and thus issues of node clustering and the requirement that infectious nodes quickly become susceptible again are not issues in the NGM definition.

3.3. Mulitplexes, multi-layer networks and multi-host pathogen systems

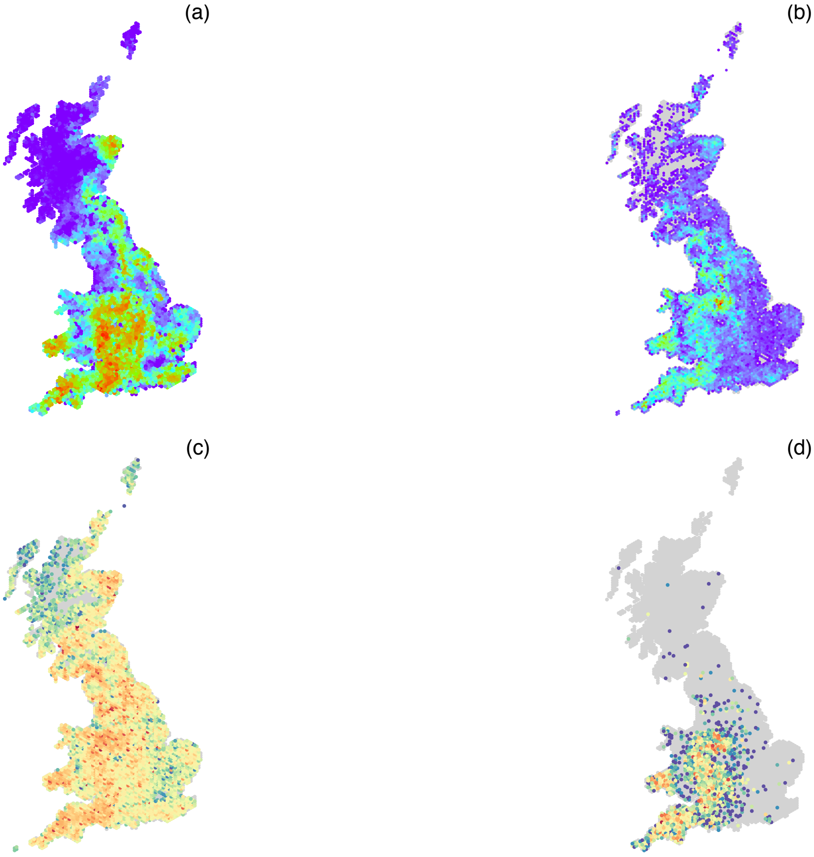

Figure S1. Maps of Great Britain, showing density (per 100 km^2 hexagonal tile) of activity associated with three layers contributing to incidence of cattle with positive bovine Tb test results (reactors). Colour scale reflects proportional reduction compared to the largest calculated value (red highest, blue lowest). (a) Estimated density of badger main setts across GB, based on [1] (b) Cattle numbers as of 01/01/2013. (c) Geometric mean of total number of inward cattle movements multiplied by total number of outward cattle movements. (d) Number of test positive cattle.

Livestock movement permit data analysis methods

Data source

Access was granted to all archived government movement permit receipt books from the study regions (Arusha, Manyara and Kilimanjaro) at the Northern Zonal Veterinary Office, Arusha. Movement permit receipt books were selected for analysis from 2009, 2011, 2013 and 2015. A permit receipt book consists of 50 consecutively numbered permit receipts (referred to here as permits).

Data processing

Data from all of the available receipt books from 2009, 2011, 2013 and 2015 were entered directly into a spreadsheet (n = 5,045) or photographed and stored as JPEG files (n = 56,849) between September 2016 and March 2017. The data recorded were date (year, month and day), origin, destination, number of animals of each species (cattle, sheep and goat) moved, and permit number. Owing to the considerable effort required to enter data from thousands of permits, data was entered from only 50% of the permit JPEGs, as selected follows. Prior to data entry, permit JPEGs were ordered by district, then by year within each district. Consecutive permit JPEGs were allocated to each of twenty batches in turn so that each batch contained a representative subsample. Data from ten of the twenty batches of JPEGs were entered into spreadsheets, resulting in a database from 28,421 (50%) photographed permits. The 5,045 directly entered permits were allocated to batches and subsampled down to 2525 (50%) records in the same way as the JPEGs, to avoid these permits being over-represented in the final raw data set, which contained 30,946 records.

Data cleaning

Data were cleaned using an R program to detect anomalous data, correcting it where possible using the stored JPEGs, and deleting records that could not be corrected. First, removal of records where the data entry technicians had indicated that the permit was blank or unreadable reduced the database to 26,855 permits. The origin and destinations recorded on the permits were matched, using a fuzzy text-matching program written in R, against a database of Tanzanian geographic names compiled from the Geographic Names Database (http://geonames.nga.mil/gns/html/namefiles.html; file dated 10 April 2017) and the National Bureau of Statistics 2012 Population and Housing Census of Tanzania. Central point coordinates were assigned to locations that matched to origins and destinations. Fuzzy matching was used as a guide only; all origins and destinations were checked visually against the JPEG, where available, and ambiguous matches were adjudicated with guidance from Tanzanian colleagues with local knowledge. Where possible, missing year was imputed from the preceding and subsequent permits, ordered by permit number, where the two permit numbers differed by less than 100 and bore the same year. Month was imputed similarly, except that where the subsequent permit bore a later month than the preceding permit, the mean month was imputed. Following imputation of missing dates, 22,538 records with all of the following data were retained: year, month, origin, destination, and number of each species moved. Duplicate records were identified either by having the same permit number or highly similar data, verified by comparing JPEGs, then removed, leaving 21,316 records. Finally, permits from years outside the four target years were removed, leaving 19,438 complete permit records, recording 112,531 cattle movements (mean 8.1 cattle per permit, range 1-337), 11,900 sheep movements (mean 6.0 sheep per permit, range 1-85) and 47,201 goat movements (mean 10.4 goats per permit, range 1-180).

For the present analysis, only cattle movements were analysed. Data were aggregated temporally within 48 (12 months × 4 years) calendar months and spatially within the 398 wards (administrative units of around 12,000 people across the study regions), resulting in a database recording the number of cattle moved in each of the 48 months between each pair of wards. Local (within-ward) movements were not analysed because of suspected non-compliance with the permit system for short movements, and movements to outside the three study regions were also omitted.

Detection of missing data

A major obstacle to inferring the movement network from the permit data was the large number of non-randomly missing permits. Frequently permits were missing from locations and time periods that were known to be active from local knowledge and trade volume data (Livestock Information Network Knowledge System; http://www.lmistz.net). To distinguish true from artefactual absence of movement (months where an origin ward sent out no cattle) a zero-inflated negative binomial (ZINB) generalised linear model (GLM) was fitted to each origin ward. Significant zero-inflation ($P < 0.05$) was detected at 16 of the 112 origin wards (wards with at least one outward movement of any species), and in each case a clear majority of zeroes were predicted to be false (range 89-100%), therefore all zero months for these wards were assumed to be due to missing data, and were removed to allow them to be imputed in the subsequent statistical modelling. The final data set recorded the movement of 86,195 cattle from 98 origin wards to 239 destination wards over the 4 sampled years.

Statistical model of inter-ward cattle movement

Inter-ward livestock movement was modelled using a hurdle model, which represented movement between each pair of wards in a given month as two processes: the binary event of any cattle being moved, modelled as a binomial generalised linear mixed-effects model (GLMM); and the number of animals moved, given that at least one animal was moved, modelled as a zero-truncated negative binomial (ZTNB) GLMM. Both parts of the hurdle model allowed movement to depend on the distance between origin and destination wards and their masses (human and cattle population sizes), in addition to other characteristics (Table S1), so the combined model components can be viewed as a gravity model of the livestock movement network. Unexplained spatial and temporal variation was modelled by fitting normal random effects for origin and destination ward and for time period (48 months). The same fixed and random effects were fitted in both parts of the model. Characteristics of origin and destination wards that were included as fixed effects were: $\log_{10}$ human population size, $\log_{10}$ cattle population size, $\log_{10}$ area in km^2, all of which were continuous, and presence of a primary (N=81) or secondary market (N=3), and production system classification (agropastoral N=159; pastoral N=55; smallholder N=150; urban N=34), which were categorical. In addition, log distance in km between ward centroids (continuous), calendar month (continuous), and year (categorical) were fitted as fixed effects. To allow their relationships with movement to deviate from linearity, all continuous variables were fitted as natural cubic splines with three degrees of freedom. Models were fitted using the glmmTMB [2] package for R version 3.5.0 [3]. Distributional assumptions were checked by inspecting plots of residuals against fitted values. The predictive performance of each stage of the hurdle model was assessed by calculating a modified R^2 statistic which we term R^2_{LATENT} because it focusses on the variance components on the latent scale (i.e. the transformed scale where the model is linear). R^2_{LATENT} is the fixed effects variance as a proportion of the total linear predictor variance (composed of the fixed effects plus the three random effects), and can be interpreted as gauging the predictive power of the fixed effects to explain spatial and temporal variation in cattle movements.

Predictor variables for the hurdle model

- **Ward area** shapefile with ward boundaries from 2012 Tanzania national census data [4].
- **Human population size** from 2012 Tanzania national census data [4].
- **Cattle population size** Estimates of cattle population numbers in 313 of the 398 wards were provided by local District Veterinary Officers via the Directorate of Veterinary Services (DVS) of Tanzania. Cattle numbers for the remaining 85 wards were imputed by linear regression (R^2 = 42%) of the log-transformed 313 DVS estimates onto log-transformed estimates from Gridded Livestock of the World [5] (map: "Predicted global cattle density (2005), corrected for unsuitability, adjusted to match observed totals)", downloaded from *http://www.fao.org/ag/againfo/resources/en/glw/ GLW_dens.html* on 2017-08-11).
- **Ward classification** created using recently updated Tanzania northern zone village classification data developed by W.A.d.G. The village classification model assigns a classification based on the highest probability of it being ‘agripastoral’, ‘pastoral’, or ‘smallholder’ generated by the model. Ward classification was determined by the most common classification among constituent villages.
- **Primary or Secondary market presence**; A list of all active or recently active markets in the study regions was taken from The Zonal Veterinary Centre in Arusha and two binary variables were created for each ward; ‘presence of primary market’ and ‘presence of secondary market’.

Simulation of livestock movement

Three quantities (fitted probability of any movement between each pair of wards; fitted rate of cattle movement given any movement; and the estimated dispersion parameter of the ZTNB distribution) were used to simulate livestock movements out of markets among the 398 wards in the study area for each of the 12 months of the year 2015, conditioning on both the fixed and random effects. Before simulating the movements, the probabilities and rates of movement were amplified twofold to account for the 50% subsampling of the permit data. This was achieved by first multiplying the probability of movement using the expression $1 - (1 - p)^2$, which approximately doubles low probabilities ($p < 0.1$) but has less effect on larger probabilities, then scaling the rate of movement so that the overall increase in expected number of movements was exactly twofold.

We simulated five kinds of market-related movement: from herd to primary market; from primary market to secondary market; from primary market to ward; from secondary market to secondary market; and from secondary market to ward. The three wards containing secondary markets (Bwawani, Machame Kusini and Meserani) were treated as if the ward was a secondary market, so that these wards had to be empty of cattle after each round of movements. Where the model predicted imbalances in cattle flows through secondary markets (for example, typically inflow to a secondary market was less than outflow, probably because permits stored at large secondary markets are less likely to be lost than those stored at primary markets), this was dealt with by boosting the deficient cattle flow rate so that inflow and outflow were balanced. Because permits are only generated by markets, wards with no recorded outflow of cattle over the four years studied were assumed not to contain markets. Such wards were assumed to export cattle only locally, supplying the nearest market. The assumption that most cattle sold at primary markets come from local herds is based on the experiences of E.S. and O.M.N., who have worked in the livestock industry and for the government veterinary services for several years, and G.C., who has conducted an unpublished survey investigating the origins of livestock at 24 primary livestock markets. These wards with no markets ($n = 287$) were therefore linked as feeder wards to the nearest active ward containing a primary market (n = 108), and the outflow from this active primary market ward was balanced by creating inflow divided evenly among the nearest feeder wards. Any remaining imbalances were evened out by creating births and deaths. By following this scheme, the cattle population remained stable at 3.7 million, with approximately 30,000 movements, 1,200 births and 1,200 deaths occurring each month.

Network measures for targeting interventions

The simulated livestock movement data were used to calculate three measures of network centrality (betweenness centrality, degree centrality and eigenvector centrality) for each ward with the aim of targeting the two types of intervention (market movement ban and vaccination at 70% coverage) to influential wards that are potentially important for disease transmission. The package *igraph* [6] for R [3] to derive a year-aggregated, static, directed, weighted movement network for cattle. A spatial contact layer was added to the market movements network as a simplified means of accounting for contacts that happen at waterholes and grazing points and via the transfer of animals between households as gifts or financial support. For this example, each ward was connected to all spatially adjacent wards via a single link with a probability one. Betweenness centrality, degree centrality and eigenvector centrality were calculated for each ward from the resulting multiplex network.

Simulating disease outbreaks

To conduct an example simulation of pathogen transmission on the network and investigate the effects of targeted interventions of epidemic spread, some simplifying assumptions were made. Each ward was assumed to have a homogeneously mixing population of cattle, with the population size estimated as described above. To avoid underflow in the simulations, wards with fewer than 1000 cattle (n = 38) were assumed to have 1000 cattle, except for the three secondary market wards which were assumed to hold zero cattle. One cattle per month was moved between adjacent wards (along the spatial network) to capture short-distance non-market movements across wards boundaries. The choice of one animal moved per month between adjacent wards was taken from a separate unpublished model of Rift Valley fever (RVF) spread among cattle across the same regions and using the same movement data. In the RVF model, we calibrated the number of cattle moved across ward boundaries so that when combined with longer-range market movements we observed a realistic rate of spread across the study area when compared with published data [7].

Random introduction and transmission of both fast and slow transmitting pathogens was simulated on the cattle movement multiplex network using the SimInf [8] package for R. A stochastic SIR model with frequency-dependent transmission was simulated within each ward. R_0 was 3 for the fast disease and 1.5 for the slow disease, and the mean infectious period was 7 days, corresponding to transmission rates (β) of 0.429 (fast) and 0.214 (slow) and a recovery rate (γ) of 0.143. In each simulated epidemic, 10 infected cattle were introduced into each of five "seed" wards, and the epidemic was simulated over 12 months. To allow the average intervention effects to be estimated, each epidemic scenario was run 79 times, each time starting from a different set of five seed wards, after which all the disease had been seeded into all 395 wards with cattle populations (i.e. not including the three secondary markets), allowing the effect of seed wards to be balanced between scenarios. To further reduce the effect of sampling variation when comparing interventions, each set of 79 simulations was repeated 3 times, so that each scenario was run a total of 237 times. For each type of intervention, market movement ban and vaccination at 70% coverage, seven intervention scenarios were simulated, including worst- and best-case scenarios, three interventions targeted using network measures, and two non-network-targeted controls:

- No intervention (worst case scenario)
- Intervention applied to all wards (best case scenario)
- Targeting of the intervention in 5% all 398 wards (n = 20), selected for:
    - Highest betweenness centrality
    - Highest degree centrality (geometric mean of indegree and outdegree)
    - Highest eigenvector centrality
    - Highest number of cattle ("common sense" network-free intervention scenario)
    - Random selection (non-targeted negative control to gauge the effect of reducing effort from all wards to 5% of wards)

The effect of each intervention was estimated as the percentage reduction in population cumulative incidence (PCI) at 1 year relative to the no-intervention scenario. This was calculated as one minus the geometric mean across the 237 simulations of the intervention PCI divided by the no-intervention PCI, expressed as a percentage. The standard error of the geometric mean was calculated according to [9].

Results

Table S1. Predictors of monthly inter-ward cattle movement fitted in the two stages of the hurdle model of cattle movement. Continuous variables were fitted as natural cubic splines with three degrees of freedom. Effect size: reduction in model mean sum of squares when predictor is removed. R^2_{LATENT}: proportion of spatial and temporal variation explained by the fixed effects. LRT: likelihood ratio test.

	Model stage predicting probability of movement (R^2_{LATENT} = 40%)		Model stage predicting no of animals moved (R^2_{LATENT} = 24%)	
Predictor	**Effect size**	**LRT P-value**	**Effect size**	**LRT P-value**
$\log_{10}$(distance/km) [spline]	58%	<0.001	0%	<0.001
$\log_{10}$(origin human pop. size) [spline]	4%	0.053	16%	0.125
$\log_{10}$(destination human pop. size) [spline]	14%	<0.001	0%	0.197
$\log_{10}$(origin cattle pop. size) [spline]	0%	0.952	-1%	0.349
$\log_{10}$(destination cattle pop. size) [spline]	0%	0.968	4%	0.304
$\log_{10}$(origin area/km^2) [spline]	1%	0.274	6%	0.910
$\log_{10}$(destination area/km^2) [spline]	2%	0.028	0%	0.280
Calendar month [spline]	0%	0.054	5%	<0.001
Year [categorical]	3%	<0.001	2%	<0.001
1ary/2ary market in origin/destination [categorical]	27%	<0.001	27%	<0.001
Origin production system [categorical]	6%	0.036	29%	0.039
Destination production system [categorical]	6%	<0.001	2%	0.759

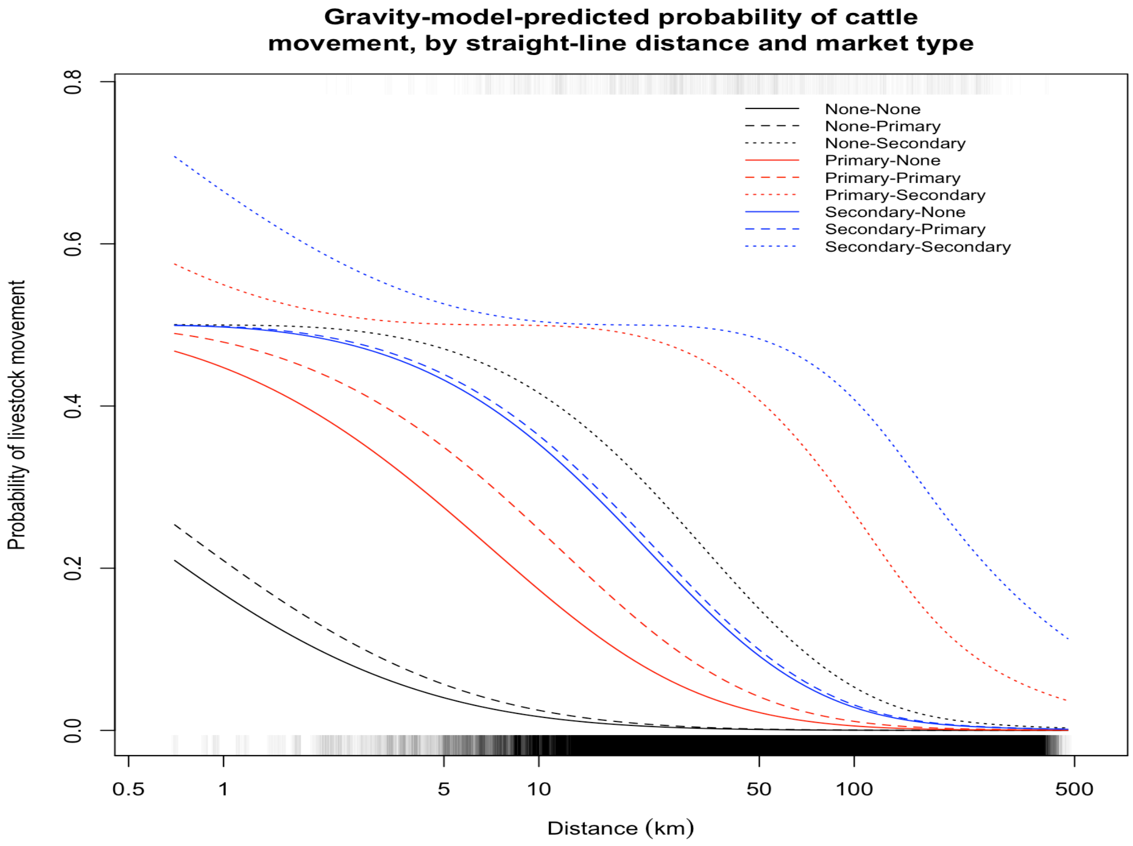

Figure S2. Probability of any cattle movement between wards in a given month, by straight line distance and market type in the origin and destination wards. Predictions are conditional on all continuous variables except month being set at their geometric means. Month is set to January 2015, while production system in both origin and destination wards is agropastoral.

Network measures

Table S2. Network measures calculate from the cattle market movement network, the spatial contact network, and the multiplex network

Network measure	Cattle	Spatial	Multiplex
Giant weakly connected component	344	398	398
Giant strongly connected component	143	398	398
Diameter	8	18	12
Transitivity	0.17	0.39	0.21
Reciprocity	0.19	1.00	0.66
Number of edges	1760	2222	3792
Edge weight	89,229	2,222	91,451
Edge density	0.011	0.014	0.024

Table S3. Mean percentage population cumulative incidence (PCI) of the fast-transmitting pathogen ($R_0 = 3$) after one year, with no intervention and under six strategies for targeting interventions to wards. The two targeted interventions were a ban on cattle movements through markets, and vaccination of 70% of the cattle in a ward. Mean (SE) absolute reduction in cumulative incidence relative to the no-intervention scenario is also given. The simulated scenarios are: no intervention; application of the intervention to all wards; and targeting of the intervention to 20 wards (5% of the total 398 wards) selected either for network centrality (betweenness, degree and eigenvector), the total number of cattle, or randomly. The total cattle population size in each simulation was 3,707,830.

Targeting method	**Movement ban**		**Vaccination**	
	Mean PCI (%)	**Mean % reduction in PCI (SE)**	**Mean PCI (%)**	**Mean % reduction in PCI (SE)**
No intervention	23.9	-	23.5	-
All wards	3.8	82.7 (1.3)	0.0	99.9 (0.0)
Betweenness centrality	6.3	75.3 (2.0)	12.5	50.7 (4.4)
Degree centrality	5.4	77.4 (1.6)	11.0	57.9 (3.6)
Eigenvector centrality	7.8	70.1 (2.2)	16.4	38.8 (5.3)
Number of cattle	19.7	17.4 (6.6)	14.4	47.3 (4.8)
Random	17.9	31.0 (5.7)	19.5	20.8 (6.5)

Table S4. Mean percentage population cumulative incidence (PCI) of the slow-transmitting pathogen ($R_0 = 1.5$) after one year, with no intervention and under six strategies for targeting interventions to wards. The two targeted interventions were a ban on cattle movements through markets, and vaccination of 70% of the cattle in a ward. Mean (SE) absolute reduction in cumulative incidence relative to the no-intervention scenario is also given. The simulated scenarios are: no intervention; application of the intervention to all wards; and targeting of the intervention to 20 wards (5% of the total 398 wards) selected either for network centrality (betweenness, degree and eigenvector), the total number of cattle, or randomly. The total cattle population size in each simulation was 3,707,830.

Targeting method	**Movement ban**		**Vaccination**	
	Mean PCI (%)	**Mean % reduction in PCI (SE)**	**Mean PCI (%)**	**Mean % reduction in PCI (SE)**
No intervention	1.7	-	1.7	-
All wards	0.9	37.0 (2.9)	0.0	99.8 (0.0)
Betweenness centrality	1.1	27.0 (3.4)	1.1	27.1 (3.4)
Degree centrality	1.1	28.6 (3.1)	1.0	31.0 (3.3)
Eigenvector centrality	1.2	20.5 (3.4)	1.3	16.9 (3.4)
Number of cattle	1.5	10.3 (3.6)	1.1	28.1 (3.7)
Random	1.5	12.4 (4.1)	1.5	12.4 (3.9)

References

1. Croft S, Chauvenet ALM, Smith GC. 2017 A systematic approach to estimate the distribution and total abundance of British mammals. *PLoS One* **12**, e0176339. (doi:10.1371/journal.pone.0176339)
2. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Mächler M, Bolker BM. 2017 *glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling*. R Foundation for Statistical Computing. See https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.
3. R Core Team. 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
4. World Bank, FAO, ILRI, AU-IBAR. 2011 Numbers for Livelihood Enhancement The Tanzania National Sample Census of Agriculture 2007/2008: A Livestock Perspective. *Livest. Data Innov. Africa Br.*
5. Wint GRW, Robinson TP. 2007 Gridded livestock of the world 2007. *FAO Rome*, 131.
6. Csardi G, Nepusz T. 2006 The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695**.
7. Sindato C, Karimuribo ED, Pfeiffer DU, Mboera LEG, Kivaria F, Dautu G, Bernard B, Paweska JT. 2014 Spatial and Temporal Pattern of Rift Valley Fever Outbreaks in Tanzania; 1930 to 2007. *PLoS One* **9**, e88897. (doi:10.1371/journal.pone.0088897)
8. Widgren S, Bauer P, Eriksson R, Engblom S. 2016 SimInf: An R package for Data-driven Stochastic Disease Spread Simulations.
9. Norris N. 1940 The Standard Errors of the Geometric and Harmonic Means and Their Application to Index Numbers. *Ann. Math. Stat.* **11**, 445–448. (doi:10.1214/aoms/1177731830)