Livestock market data for modeling disease spread among US cattle

Ian T. Carroll*[1,3], Jason E. Lombard[2], and Shweta Bansal[1,3]

[1] *Department of Biology, Georgetown University, Washington, DC, USA*
[2] *United States Department of Agriculture, Animal and Plant Health Inspection Services, Veterinary Services, Center of Epidemiology and Animal Health, Fort Collins, CO, USA*
[3] *Fogarty International Center, National Institutes of Health, Bethesda, MD, USA*

August 3, 2015

Abstract

Transportation of livestock carries the risk of spreading foreign animal diseases throughout a susceptible population, leading to costly public and private sector expenditures on disease containment and eradication. Individual animal tracing systems that exist in countries other than the US have allowed epidemiologists and veterinarians in those countries to model the risks engendered by livestock movement and prepare responses designed to protect the livestock industry. Within the US, data on livestock movement is not sufficient for direct parameterization of disease models, but network models that assimilate limited data provide a path forward in model development to inform preparedness for disease outbreaks in the US. Here, we report on a novel data stream, the information publicly reported by US livestock markets on the origin of cattle consigned at live-auctions, and demonstrate such potential. By aggregating weekly auction reports from markets in several states, some spanning multiple years, we obtain an ego-centric sample of edges from the dynamic cattle transportation network in the US. We first demonstrate how the sample might be used to infer shipments to unobserved livestock markets in the US, although we find the assumptions of edge prediction by generalized linear models too restrictive. The sample itself, however, can still be used to parameterize simplified disease models; which we use to demonstrate that the temporal resolution of the data is sufficient to reveal seasonal trends in the risk of disease outbreaks. We conclude that future work on statistical models for dependence between edges will improve the inference of a complete cattle movement network model from market data, one able to addresses the capacity of markets to spread or control livestock disease.

*Corresponding author: *itc2@georgetown.edu*

Author Summary

We have "crowd-sourced" the collection of previously unavailable cattle movement data, benefiting from buyers interest in the origins of cattle sold at live-auction markets, to implement a minimum level of movement surveillance. Using our novel dataset, we demonstrate potential to infer a complete dynamic transportation network and model national-scale livestock epidemics.

Introduction

Livestock operations within the United States (US) must be vigilant against transboundary animal diseases, including the critical threat to cattle producers posed by a re-introduction of foot-and-mouth disease [FMD; 1]. The 2001 FMD outbreak in the United Kingdom (UK) cost their agricultural sector £3 billion, and 5% of the nation's 11 million cattle were culled to control the disease [2]. A study on FMD risk to California's 5 million beef and dairy cattle predicts economic losses in the tens of billions of dollars, even for an outbreak artificially terminated at California's border [3]. The potential impact of a full-blown epidemic in the US, putting at risk a 90 million strong cattle herd [4], compels us to study the likely patterns of disease spread from an initially infected cattle operation [5]. Mechanistic models that incorporate livestock transportation are needed to help guide prevention and control of FMD-like diseases, which are known to cause massive burdens on livestock industries, require costly public interventions, raise public health concerns and impact food security [6].

Studies of past livestock epidemics and disease simulations reveal that network models provide a useful abstraction of data on animal shipments between livestock operations [7]. Network models typically emphasize heterogeneity in the number of disease transmitting contacts attributed to infectious nodes, a pattern that emerged strongly during the initial spread of FMD in the UK's 2001 epidemic [8] and one usually absent from simulations where transmission depends on distance alone. Network representations of cattle trade exist for several European livestock industries [9–14], where data for model development is generated from animal tracking systems mandated by the European Parliament [15]. Availability of these data have allowed for several advances in surveillance and control strategies: for example, (1) identification of "sentinel" livestock premises projected to become infected early during an outbreak in Italy [16], (2) validation of risk reduction from the standstill rules implemented in the UK after 2001 [17], and (3) evaluation of targeted movement bans that selectively eliminate network edges based on node centrality in transportation networks [14]. Network models for the UK cattle production system have additionally provided a foundation for livestock transportation strategies that promise efficient control of *endemic* diseases [18].

The US has opted against individual animal tracking in favor of animal disease traceability, which only requires that a paper trail on individual movements can be unearthed subsequent to disease detection. With respect to the development of models for disease prevention and preparation, the traceability principle promotes inadequate and belated data collection, and is significantly limited compared to the point-to-point data on individual animal movements that drive models of European systems. In addition, the findings gleaned from epidemiological models based on European livestock transportation patterns are not transferable to the US due to the nature and scale of the US industry: the US industry has developed an unparalleled feedlot system that relies on widely distributed calving and back-grounding farms to supply the 25 million head sold annually by feeding operations [4], and the US Department of Agriculture (USDA) census shows that 22% of the US herd are transported across state lines in a year [19].

The best data currently available is maintained by individual state agricultural agencies, which collect shipment origin and destination locations on interstate certificates of veterinary inspection (ICVIs) for cattle entering the state. One cattle transportation network model has been estimated from a 10% sample of year 2009 ICVIs acquired from 48 state agencies [20, 21]. The epidemiological network model constructed from these data [22] represents the state of the art for analysis of a nation-wide epidemic in the US, but limitations in the underlying data are substantial. Given that on average only 19.2 (SEM 3.1) percent of shipments onto US beef operations travel over 100 miles [23], shipping within states likely occurs at much higher rates than shipments documented on ICVIs. Interannual variability in the transportation network cannot be observed without repeating a major collection effort, nor can the type of origin or destination facility be determined from ICVIs alone. Finally, and of most relevance to the present study, ICVIs are not required for shipments to exempt facilities, including certain federally approved livestock markets [24].

Cattle in the US are commonly sold at live auction markets between stages of beef or dairy production (Box 1). In year 2007-08 surveys by the USDA, over half of beef producers sold non-breeding stock at auction markets in the US (60.7% steers, 58.3% cows), with internet auctions and private treaties being the main alternatives [25]. Dairy contributes fewer US cattle shipments, but cows removed from dairy operations are also predominantly sent to auction markets or stockyards [21, 26]. Local epidemiological studies have also found direct contacts with livestock markets prevalent in Colorado and Kansas [27] and California [28]. The economic reality now, and for the foreseeable future, is that cattle owners regularly buy and sell cattle at particular stages of production and rely on live-auction markets to obtain the best price [29, 30].

In the UK, and quite possibly in other countries with recent outbreaks of non-endemic FMD [31], livestock markets played a central role in early, rapid expansion of the 2001 FMD epidemic [8]. The epidemiological importance of livestock markets arises from their potentially high degree in both contact and epidemic networks—the same epidemic phenomena airports create as hubs for transmission and spread of human influenza [32]. When a livestock operation ships infected animals to market, two processes spread the disease: splitting of the original group of infected animals among multiple buyers, and transmission to susceptible animals passing through the same market [33]. Both processes act to give livestock markets high out-degree in an epidemic network for FMD [8,34], while a less contagious disease would primarily be affected by splitting up infected animals arriving from one premises. Livestock markets can also have high in-degree within the contact network, or a large number of operations from which cattle are sourced [35]. High in-degree markets are a natural point of surveillance for disease detection; indeed, markets seeking USDA approval are required to provide veterinary inspection of cattle at auction [24].

On the premise that transportation of cattle to and from markets is an important class of potential disease-transmitting contacts between farms and other longer-term animal holdings, we studied the potential for data-driven modeling of a market-based contact network for infectious livestock disease in the US. To this end, we collected data on the locations of origin for individual cattle sold at livestock markets that publicly share this information as a form of advertising. Because the data represent an opportunistic sample of livestock transportation, we first report basic trends in the data alongside some potential sources of bias in the sample. We then demonstrate two methods for inference of a network model of contacts between livestock operations from these data: (1) using the sample to estimate degree distributions of an otherwise random contact network, and (2) fitting coefficients to covariates that may predict the presence and weight of unobserved network edges. The results firmly establish that market-bound cattle shipments are dominated by intra-state movements, and are consistent with the possibility that transportation to markets also drives interstate flows. The daily resolution of this data source allows detection of sub-annual variation in trade volume and network degree distributions, and as a continuously updated data stream creates potential for both inter-annual trend and recent-event detection. We demonstrate high potential for inferring the properties of unobserved edges connected to non-reporting livestock markets, and conclude with a discussion of the critical gaps for building a complete epidemic network model that includes markets acting as hubs for the spread of livestock disease.

Methods

Data Collection

As an integral part of livestock production, stockyards are distributed across all parts of the US with beef or dairy operations, i.e. throughout the US [Box 1; 36]. Prices obtained at live auction are rapidly publicized to help consignors and buyers decide when and where to trade cattle. In some cases, professional market reporters attend sales and distribute volume and price information through the USDA Market News Service. Where market reporters are unavailable, or to provide additional information, sale reports might be generated by the market itself and publicized on its own website. A subset of markets list specific lots of cattle sold in their sale reports, including a location of origin, number of cattle, and other attributes. These data, sometimes labeled “representative sales” as we refer to them here, indicate cattle were transported from the origin to market on, or very near, the sale date.

We aggregated representative sales from several livestock markets and georeferenced each location of origin to a US county or county-equivalent (hereafter “county”). Overlapping and incomplete directories of US livestock markets are maintained by multiple regulatory agencies or business associations: we compiled four such directories to identify target markets [36]. The directory released by the Livestock Marketing Association [37] uniquely provides websites of livestock markets, when available. We manually searched the 322 listed websites for representative sales, and wrote software to parse data from sites that regularly (usually weekly) publish market reports and which permit crawling by the Robots Exclusion Protocol. For each lot provided as a representative sale, the software attempts to parse the consignor's location along with cattle type (e.g. steer or heifer), number, average weight and price (either per head or per hundred weight). A single animal per lot was assumed whenever the number of cattle could not be parsed. We tuned the parser for each website until two researchers found no data extraction errors in independent spot checks of representative sales. Websites were subsequently checked twice weekly for new reports, and parsers returning data of the wrong type (i.e. string or numeric) were promptly corrected. This study addresses sale reports obtained between June 2014 and June 28, 2015, including some archived reports on sales dating from the first week of 2012.

Acceptable locations of origin are given as the name of a city or other populated place, with or without a state, and can be ambiguous. We matched each location, substituting common abbreviations for full words as needed to obtain a match, to names of populated places in North America using the GeoNames web-service [38] in order to identify the encompassing county. The county closest to the reporting market, as determined by the great-circle distance (GCD) between county centroids [39], is recorded as the true location of origin.

Comparison with Interstate Shipments

Interstate cattle shipments are present among representative sales, allowing a comparison to cattle shipping data obtained from state ICVI records. For each state with at least one market in our study, we correlated the number of cattle in representative sales originating in every other state with the analogous interstate flows reported by Shields and Mathews [5]. The ICVI data sampled shipments occurring in the 2001 calendar year, which pre-dates all sale reports collected for this study. To match the time scale of the certificate-derived data, we aggregated representative sales over the year preceding June 28, 2015 before calculating correlations. However, the comparison necessarily reflects over a decade of change in the livestock system on top of any differences between market shipments and shipments accompanied by ICVIs.

Analysis of Sampling Rate

We analyzed variation in the sample size for each sale report by fitting a GLMM to the number of head in representative sales, given the total head of cattle sold ("receipts"), using covariates from the agricultural census [4]. The analysis intends to address two issues: estimation of sampling rate for reports where the total receipts is unknown, and detection of potential bias among representative sales. Representative sales are not randomly sampled with equal weight from all cattle shipped to reporting markets, but are effectively stratified by sale report. Because markets may vary the proportion of sales listed as representative, each sale report should have an associated sample weight for shipments found in that report. Estimation that involves aggregation across sale reports should take this weight into account, but receipts are unknown for roughly one third of sale reports. By taking covariates into account while estimating unknown receipts with a fitted GLMM, resulting estimates will address certain biases that show up as statistical associations between known sampling rates and covariates for a given sale.

Receipts for a given sale are taken from USDA Market News Service reports [40] or, when available, from the market's own report. Covariates included as fixed effects include the inventory and sales of cattle, as well as the number of cattle operations, in the county where the market is located [4], the number of representative sales, the sale year and the sale week. Numeric covariates were first log transformed and standardized. Each market is additionally allowed a random intercept, representing unexplained variation attributed to average behavior of individual markets. To avoid possible parameter bias, we added an observation level random intercept to eliminate overdispersion [41]. We fit a binomial family GLMM with logit link function using the 'lme4' package [42], and performed Wald χ^2 tests for significance of fixed effects using the 'car' package [43], in R version 3.1.3 [44].

Network Inference: Edge Prediction

The definition of nodes and the meaning of edges in contact networks for disease models are flexible, facilitating data-driven modeling approaches. Models for disease spread among livestock incorporate transportation data as edges in the network of contacts between susceptible and infected individuals [e.g. 10]. Representative sale data is compatible with a model having two types of nodes: one representing all farms, ranches and other long-term animal holdings located within a given county, and one representing a single market. In order to study the contribution of markets to the livestock transportation network, and its consequence for disease spread, we ignore edges between counties that arise from fence-line contact, private sales, transportation for grazing, and other mechanisms that might transmit disease directly between counties. Observing no indication of market-to-market transportation in the representative sales, we assumed their absence as well. As a result, the only edges in a network derived from representative sales are between nodes of different types, yielding a bi-partite contact network.

Edge prediction is any process for inferring the properties of unobserved edges, which must be made explicit for disease models that use contact networks to drive infectious interactions between nodes [e.g. 22]. A primary goal of edge prediction is to build a model that reflects clustering within the transportation network, or the propensity of livestock operations within different counties to trade cattle at the same two markets, without directly observing these second-degree interactions. A model that represents higher-order structural attributes of the network, including clustering, may yield different predictions for the spread of disease, but direct estimation of these attributes requires particular sampling methods [45]. For example, a random sample of nodes is not appropriate for estimating clustering coefficients; a first-wave link tracing approach [*sensu* 46] is needed to avoid underestimating the number of triangles touching each focal node.

We applied a regression approach to the problem of edge prediction, using observed edges to estimate whether county and market covariates predict their connectivity. The response variable is the number of cattle shipped from a given county to a market, which we assume to arise from a zero-inflated negative binomial (ZINB) distribution. This GLM includes the sales, inventory and number of farms for cattle (including calves) for each county of origin, GCD distance between the centroids of each county and the county of each market, the square of this distance, the number of livestock markets giving an address in each origin county, a boolean factor indicating whether the market is in the county, and a boolean factor indicating whether both are in the same state. All numeric predictors are normalized to unit variance with zero mean. Finally, the model includes a fixed effect of market: in the simpler case of a Poisson GLM, this fixed effect would have no effect on the resulting multinomial probability of cattle originating from each county, given the total number of cattle by market. The more complicated ZINB, necessary to obtain a good fit to the observed edges, comes at the cost of incorporating a meaningful market effect which will interfere with edge prediction for non-reporting markets. We fit the ZINB GLM using the 'pscl' package in R [47].

Disease Consequence of Degree Distribution

A key insight from network epidemiology is that the degree distribution for contacts among individuals, or nodes, is of primary importance for disease spread [48, 49]. Edge prediction is not needed to infer the degree distribution among livestock markets; we may assume the number of counties appearing in the representative sales data for each market are independent samples from this distribution, and then specify the remaining network properties parametrically. We define k_i as the sampled degree for market i of m markets. Empirical estimates for the market degree distribution generating function, $G_{M,0}(x)$, and the generating function for market "excess degree" [e.g. 50 eq. 12], $G_{M,1}(x)$, are

$$G_{M,0}(x) \approx \frac{1}{m}\sum_{i=1}^{m} x^{k_i} \tag{1}$$

$$G_{M,1}(x) \approx \frac{\sum_{i=1}^{m} k_i x^{k_i-1}}{\sum_{i=1}^{m} k_i}. \tag{2}$$

We specify properties of the full network by choice of the algorithm for connectivity and the probability distribution for county degree: we assume a bi-partite configuration model for edges and a Poisson distribution on county degree. In a static network with these properties, the expected size of a disease outbreak in terms of the number, or proportion, of counties affected can be calculated exactly [51]. With τ_C and τ_M the disease transmission probabilities from counties and markets, respectively, and the mean county degree equal to λ_C, the epidemic threshold is a fixed value of $\phi = \tau_C \tau_M \lambda_C$ determined by the market degree distribution. The threshold occurs at

$$\phi^{-1} = G'_{M,1}(1). \tag{3}$$

For parameterizations below this threshold, the expected number of counties affected by an outbreak is

$$1 + \frac{\phi G'_{M,1}(1)}{1 - \phi G'_{M,1}(1)}. \tag{4}$$

For parameterizations above this threshold, the proportion of counties in the epidemic is

$$1 - e^{\lambda_C(u-1)}, \tag{5}$$

where u is the smallest root of

$$u = (1 - \tau_C) + \tau_C G_{M,1}\left((1 - \tau_M) + \tau_M e^{\lambda_C(u-1)}\right). \tag{6}$$

See S1 Text for an explanation of these equations.

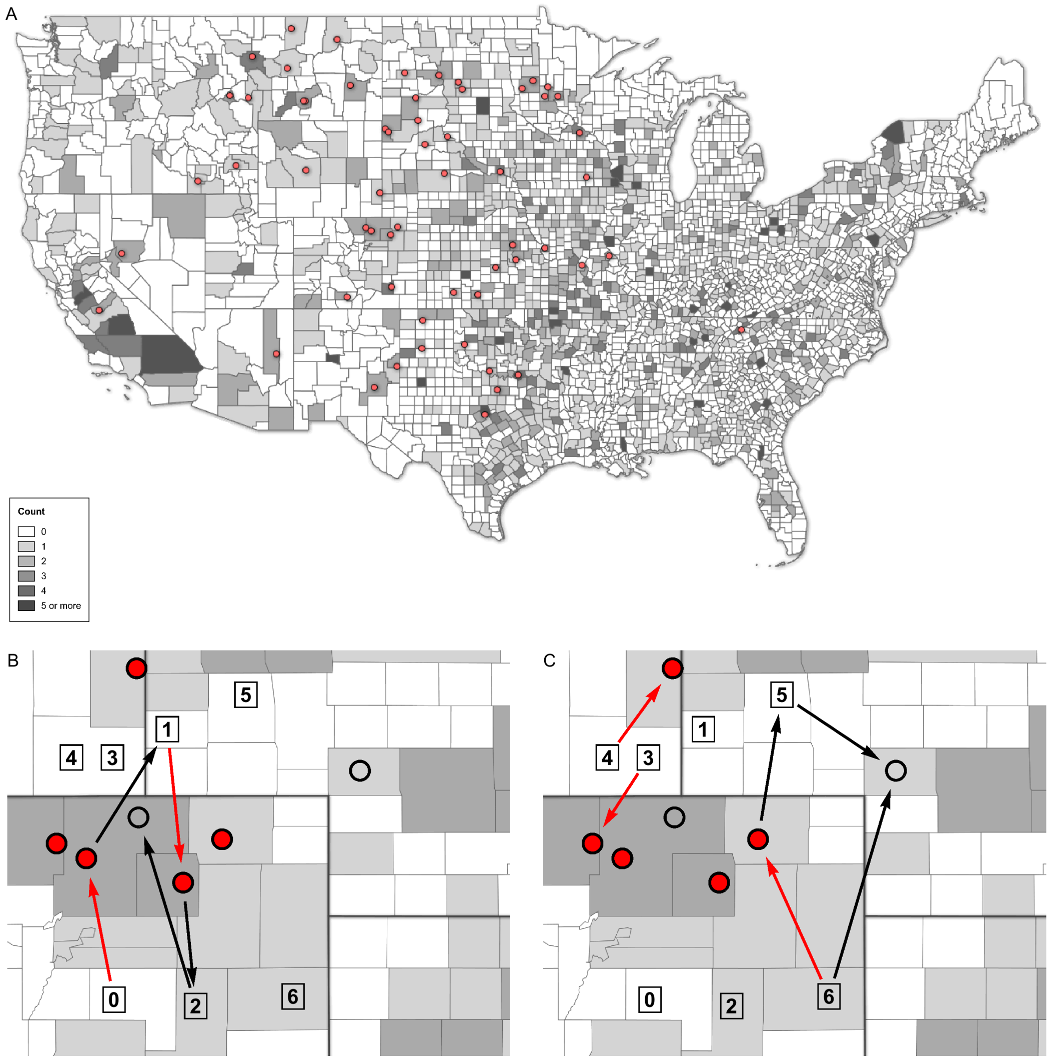
Two features of representative sales data conflict with applying this 'random graph' approach to network modeling of livestock disease spread. First, the degree of each market potentially varies from sale to sale, admitting possible temporal variation in the observed degree distribution. We examine the extent of variation in degree by visualizing temporal variation and calculating seasonal estimates of epidemic size. Second, the representative sales may not include all the counties of origin, potentially biasing the observed distribution toward smaller degrees. Using the iNEXT R-package developed for analysis of species accumulation curves [52,53], we calculate complete degree estimates for each sale using extrapolation of the county accumulation curve as a function of the number of cattle in representative sales.

Results

Cattle transported to 55 markets located in 53 counties in 16 states are represented in this analysis (Box 1). The first section below describes seasonal trends observed in the representative sales, the fair to strong correlation with published records of inter-state shipments, and quantifies the unexplained variation in the proportion of sales that different markets report as representative of a live auction. In the next section, we relate edge presence and weight to their distance and county covariates from the agricultural census, as well as unexplained differences between markets. The last section provides a demonstration of network epidemiological inferences that incorporates degree distributions from representative sales data. Data on each movement, summarized in tables suitable for reproducing our analyses, are freely available online (S2 Text).

Characteristics of Representative Sales

The average number of cattle movements reported in representative sales for a live auction increases to a peak of 1000-1500 head in late fall and decreases to a few hundred during summer months (Fig 1A). This seasonal trend persists between the period for which representative sales come from a handful of markets with accessible archives and the period since mid-2014, when we began capturing representative sales posted weekly (Fig 1B). At a given sale, livestock markets report receiving cattle from 11.2 (SD 1.5) counties on average, with weekly average market degree showing seasonal variation peaking in late autumn (Fig 1C). While the timing of peak degree closely corresponds to the time of year when the number of representative sales is also greatest, the troughs in degree are flatter, broader and less pronounced than the summertime lows in representative sales. The seasonal fluctuation in market degree is weakest during the most recent year, for which the sample size is larger as well as geographically more expansive. Aggregating across all markets, the proportion of sales that originate in-state shows no trend in deviations from an average of 0.84 (SD 0.07) (Fig 1C). Markets selling the majority of out-of-state cattle appear to be clustered in Oklahoma and South Dakota, where it is not uncommon for less than half of representative sale cattle to originate in-state (Fig 1E).



Box 1: Cattle trade at live auction markets is an integral part of the US livestock system. Markets are distributed widely throughout the US because different operations specialize on different stages of production and trade cattle at live auction to obtain the best price. By collecting representative sales from markets with online reports, we obtain a sample of the cattle transportation network. A) The number of livestock markets per county (grayscale) in a compilation of public and private market directories [36], as well as the location (red points) of the livestock markets reporting representative sales collected for this study. B & C) In northeastern Colorado and adjacent states, four markets (red points) post representative sales online, revealing movement of cattle from unspecified farms (numbered squares) to specified markets (red arrows). Black arrows represent un-sampled edges representing movement of cattle leaving a market or sold at non-reporting markets (black circles). B) Individual animals may pass through multiple markets and farms; for example, a cow-calf operation (0) will sell weaned steer calves for purchase by a stocker/back-grounder (1), who subsequently sells the animal to a buyer for a feedlot (2), which sells the fed steer at auction for slaughter. C) Our method of data collection aggregates farms within counties, so farms 3 and 4 would be combined in a single node connected to two markets. Another limitation of the data is that closed cycles in the transportation network, the loop involving farms 5 and 6 for example, cannot be observed directly.

bioRxiv preprint doi: https://doi.org/10.1101/021980; this version posted August 3, 2015.

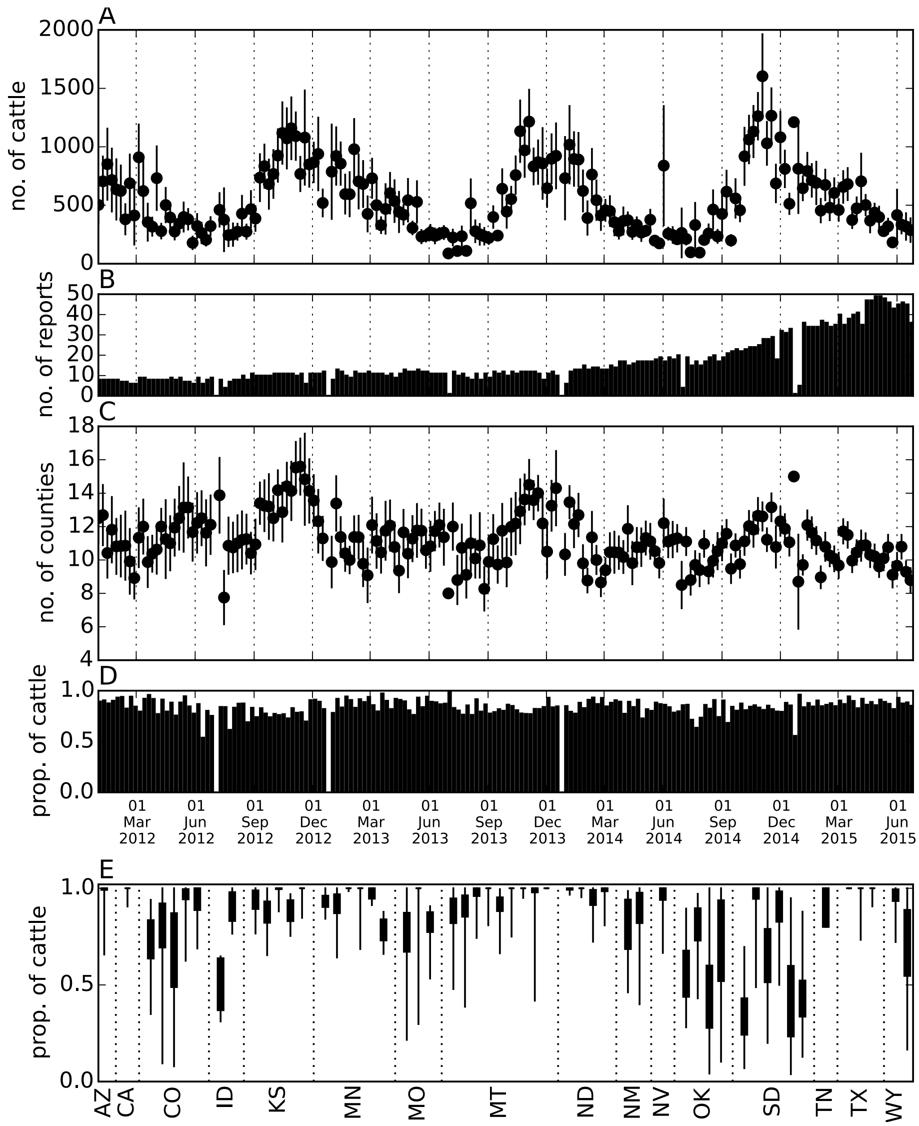


Figure 1: Representative sales captured for a 3.5 year time span that include the county of origin for each shipment to a given market. A) The average (+/- SEM) head of cattle reported with locations of origin each week by a single livestock market. The outlying observation in the first week of June, 2014 includes a massive sale conducted by the World Livestock Auctioneer Champion in Ft. Pierre, SD. B) The total number of reports collected for cattle sales occurring within a given week. C) The average (+/- SEM) number of counties from which cattle arrived for sale at a market on a given week. D) The proportion of representative sales from all reporting markets that originate within state for a given week. E) The proportion of representative sales that are transported from a county within the same state to each reporting market.

The state of origin for interstate shipments show fair to strong correlation with certificate-derived data from 2001 (Table 1). Among the states we could compare to this published summary of interstate transportation, and aside from Idaho, the proportion of cattle shipped within state is above 71%, so the number of cattle shipments used for each correlation is relatively small. For example, a nearly exact correlation results for New Mexico, but the number of head in representative sales available for comparison is only a few thousand head. South Dakota has the largest number of cross-border representative sales that we observed, however, and also shows a strong correlations of 0.79. Montana and Colorado have similarly large sample sizes, the first showing a strong correlation of 0.88 while the latter is among the weakest at 0.46. Shipments into Texas and California are insufficient for a meaningful comparison, while the lowest correlation (0.24 for Idaho) is driven by one strong connection to Nye County in Nevada. Variation in the strength of correlation across sample sizes suggests the presence of real variation in the kind of interstate shipments sampled by different data streams. Differences here could be due to the relative proportions of shipments that are market bound versus non-market bound as well as heterogeneity in state requirements for health certificates.

Table 1: Correlations between inter-state representative sales and inshipment data from Shields & Mathews [5]. For each destination state with market data, the table shows the Pearson correlation (r) for the number of cattle from n origin states, along with the total head of cattle in the representative sales used and the proportion (p) of representative sales used.

Dest.	p	Head	n	r
CA	1.00	6	43	0.30
CO	0.85	10557	23	0.46
ID	0.49	1155	43	0.24
KS	0.93	3386	38	0.58
MN	0.96	2124	37	0.49
MO	0.81	6367	17	0.41
MT	0.93	11996	16	0.88
ND	0.97	4714	3	0.79
NM	0.84	3722	17	0.99
NV	0.95	232	28	0.89
OK	0.71	3927	18	0.73
SD	0.74	95570	39	0.79
TX	1.00	35	33	0.82
WY	0.86	7506	28	0.73

Uncertainty about the total number of animals shipped to market for a given sale is amplified by unknown sources at the market level and to a lesser extent for each individual report. For roughly one third of market reports, the total receipts is not available for use in weighting the sale's affect on estimates aggregating across sales or markets. In other reports, the proportion of sales given as representative is associated with the number of representative sales ($\chi^2_1 = 2.40 \times 10^3$), with a positive regression coefficient (approx. 95% CI 0.88 to 0.95). Both year ($\chi^2_3 = 17.7$) and week of year ($\chi^2_{50} = 4.64 \times 10^2$) are associated with variation in the sampling probability, but none of the covariates taken from agricultural census data are significant. Overall, the fixed effects contribute the majority of variation in the fitted GLMM, leaving the random effect of market (SD = 0.82) and the observation level random effect (SD = 0.20) to explain the remainder of the variance between sampling probabilities. Average differences between markets account for 20% of the variance; however, the observed proportions remain overdispersed with respect to the model fitted without observation level random intercepts. In other words, variation in the binomial sampling probability predicted by the fitted model for each market underestimates actual variation observed in the representative proportion (Fig 2). Observation level random intercepts are included to account for the remaining 12% of variance in the proportion of cattle reported in the representative sales, but the random intercept for each market is the greater source of uncertainty.

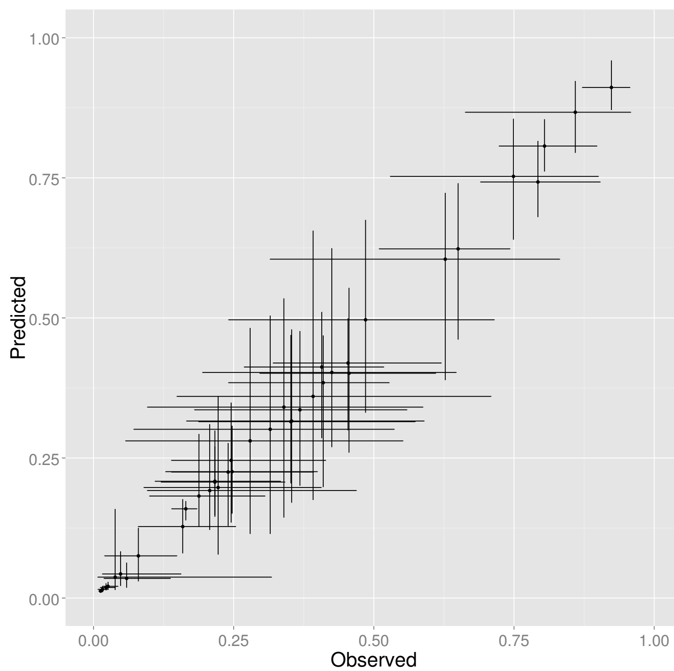


Figure 2: Observed versus predicted proportions of cattle listed as representative sales, grouped by market. Horizontal bars show within-market variability in the observed ratios of representative sales to receipts. Vertical bars show corresponding variability in the binomial sampling probability predicted by the GLMM described in the main text, but with the observation level random effect dropped. Bars intersect at the mean and extend from the 5^{th} to 95^{th} percentiles.

Edge Prediction

The regression model returns a negative coefficient for the impact of distance (approx. 95% CI -2.07 to -1.42) on the average number of cattle shipped from a county to market, as well as a positive coefficient (approx. 95% CI 6.88 to 7.61) that associates greater distances with zero-inflation, or the absence of an edge between county and market. This confirms intuition that cattle are preferably shipped to nearby markets, and quantifies the effect of distance to use when extrapolating edges for non-reporting markets. In addition, both distance-related factors for edges linking counties to markets within that county ($\chi^2_1 = 65.5$) or within the same state ($\chi^2_1 = 92.6$) are significant predictors. The covariates extracted from the agricultural census have inconsistent results, possibly due to their strong inter-correlations. The effect of sales (approx. 95% CI -0.19 to -0.05) and inventory (approx. 95% CI 0.93 to 1.78) on the average head of cattle shipped are of opposite sign, while the number of farms is insignificant. Because the covariates are standardized, we can interpret the result to mean that the size of cattle operations measured by head is of greatest importance and is consistent with the hypothesis that counties with larger inventory contribute more heavily weighted edges. Among the census covariates, only the number of farms has a non-zero (approx. 95% CI -0.62 -0.40) effect on zero-inflation.

Judged by simulated response variables generated by the fitted ZINB model, the model shows a good fit to the observed transportation network (Fig 3). Market degree distributions obtained with simulated response variables are uniformly similar to the observed distribution for market in-degree aggregated over the full study period (Fig 3A). While a single realization of simulated edge data cannot reveal the model's degree of uncertainty, mapping the edges provides visual confirmation of the role of distance and distance related factors on the weight of network edges (Fig 3C&D). The most striking difference between the observed and simulated response is the weight of long-distance edges, suggesting that observed shipments are either more clustered on a fewer number of edges (including long-distance edges) or are even more commonly from nearby counties than simulated shipments.

Inclusion of the fixed effect of market in the ZINB model greatly improves the fit ($\Delta\text{AIC} = -3340$ on 55 degrees of freedom), but eliminates direct application of the model in predicting cattle shipments to non-reporting livestock markets. The fitted coefficients for market effects could instead be modeled as random effects, and extrapolation to a full network carried out under the assumption that reporting markets are an unbiased sample with respect to network attributes. Based on the fitted intercepts for each market, however, the usual assumption of normality for random intercepts may not be justified (Fig 3B).

Epidemic Size on Random Graphs

A model for disease spread that does not include a full contact network is possible under the assumption that epidemics develop as a tree-like graph, and is related to the representative sales data through estimates of the market degree distribution. The majority of seasonal variation in the distribution on market degree exists between a peak season (from the 39th (the last week of September) through years

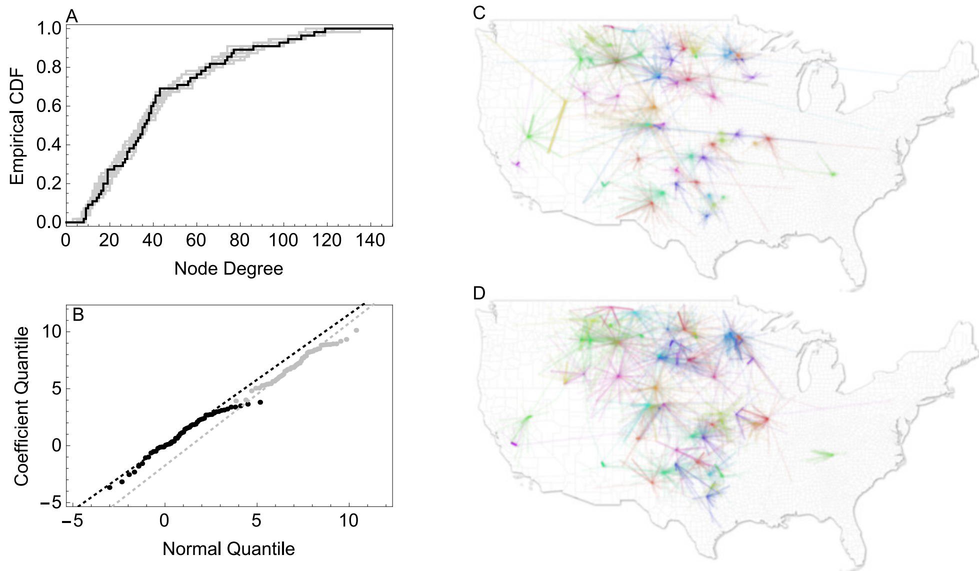


Figure 3: Comparison between the origin of representative sales observed and simulated using the fitted zero-inflated GLM. A) The (unweighted) node degree distribution observed (black) and returned in ten simulations of the fitted model (gray). B) Normal quantile plot of the estimated fixed effect of each market on the mean head count (black) and zero-inflation probability (gray). C) The observed proportion of cattle from each market's representative sales by county of origin, aggregating all available sale reports. D) The simulated proportion of cattle from each of county of origin, again grouped by market, in one simulation from the fitted GLM. Edge coloration corresponds to individual markets, and opacity is linearly scaled between a maximum proportion or probability of 0.8 (transparency at 0%) and a minimum of 0.004 (transparency at 80%).

end) and the remaining off-peak portion of the year, which exhibit distinct empirical cumulative distribution functions (ECDFs; Fig 4A). The former is indistinguishable from a negative binomial distribution by Pearson's goodness-of-fit test ($\chi^2_{15} = 14.0$) while the latter, although similar in shape, is not ($\chi^2_{17} = 60.9$).

For a disease spreading on a bi-partite random graph with these market degree distributions, seasonal variation effects the location of the epidemic threshold with respect to the unknown parameters. Roughly 20% lower values of ϕ, the product of market and county transmisabilities and mean county degree, prompt an epidemic for the peak time of year relative to off-peak (Fig 4C). Above the epidemic threshold, the difference between seasons becomes negligible as transmisability increases; it is overwhelmed by the overall high degree of livestock markets. Even with the average excess degree of counties equal to one, nearly two-thirds of counties are affected in the extreme case that every contact between susceptible and infective cattle leads to successful disease transmission (Fig 4D).

The number of counties included in representative sales is, on average, 8.6% of the estimated number of counties extrapolated from county accumulation curves. In over half of sale reports, the estimate is less than 1% greater than the observed

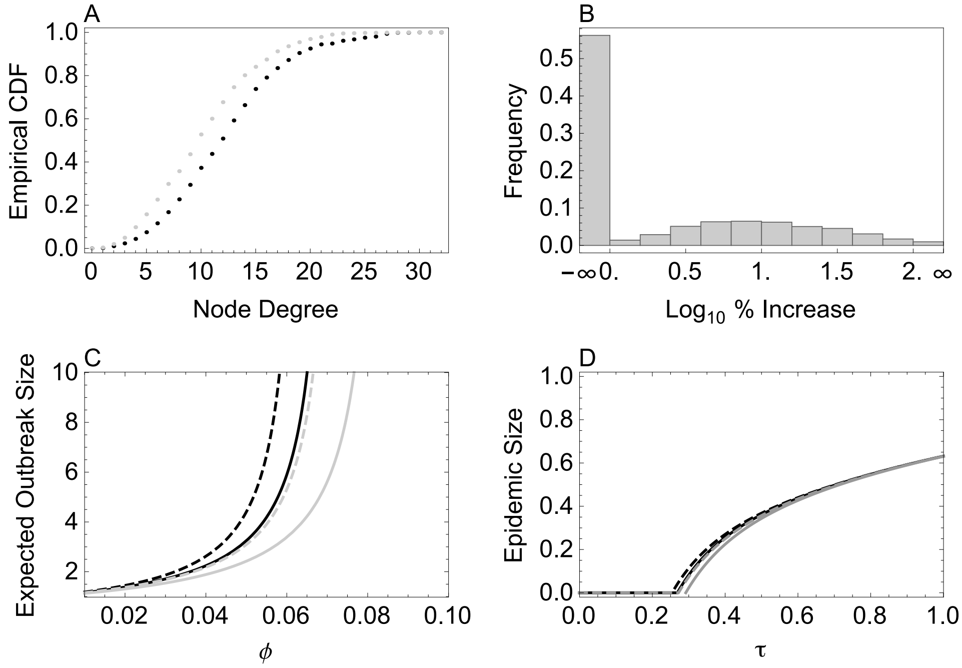


Figure 4: Empirical market degree distributions inferred from representative sales and implications for disease spread on the corresponding random graph. A) Raw degree distribution for sales occurring during peak (black) and off-peak (gray) seasons. B) Percent increase in extrapolated degree relative to sampled degree, with extrapolation up to the total receipts. C) Expected number of counties affected by a disease outbreak during the "peak" season (black) or "off-peak" time of year (gray), using either observed (solid) or extrapolated (dashed) degrees. ϕ is a product of τ_M, τ_C, and λ_C. D) Proportion of counties affected in the four cases shown in panel C but above the epidemic threshold, assuming $\tau_M = \tau_C = \tau$ and $\lambda_C = 1$.

number of counties, and among the rest the most common increase in degree is just 10% (Fig 4B). The rarity of singleton counties (i.e. counties with only one individual in representative sales) and the sufficiently high sampling rates (Fig 2) are responsible for the completeness of the sample for counties of origin. Using the extrapolated values in calculating disease spread on a random graph has the same qualitative effect as the shift from off-peak to peak season market degree distributions. Quantitatively however, the difference between observed and extrapolated market degree distributions has less impact than season on the estimates of disease spread (Fig 4C&D).

Discussion

The epidemiological contact network is a fundamental component of models for the spread of diseases, and sale reports publicized by livestock auction markets contribute urgently needed data to support inference of such networks within the

US livestock system. For this study, we initiated an ongoing process to archive representative sales as an opportunistic sample of cattle transported from counties with beef or dairy operations to livestock markets distributed across the US. The study complements previous efforts to summarize transportation of cattle within the US using data derived from certificates of veterinary inspection [5, 21], but extends our ability to model within-state shipping patterns. We demonstrate how inference of a bi-partite contact network, between nodes representing either cattle holding operations aggregated within a US county or a particular auction market, can allow for new models for the spread of economically disruptive livestock diseases.

Representative sales extracted from livestock market reports provide a reliable sample of cattle shipments and the corresponding potential for disease transmitting contacts. Seasonal variation in the volume of representative sales is consistent with beef cattle production systems, where calves are produced in spring and weaned cattle or yearlings sold to pasturing or feedlot operations in the fall and subsequent spring [54]. The proportion of receipts at a given sale whose origin can be identified is anywhere from a negligible fraction to around three quarters, and understanding this variability is important for scaling up assessments of transportation networks. The dominant source of uncertainty is variation between markets, but this can be quantified for future modeling efforts despite having no identified deterministic source. Covariates taken from the agricultural census on the county where markets are located do not influence the proportion of sales reported, which reduces concern about biasing population estimates from the representative sales.

Interstate shipments among representative sales correlate fairly well with ICVI data, while the remaining majority of representative sales provide unmatched data on cattle shipments that remain within states. Intrastate shipment data were previously unavailable and dominate market directed shipments at typically over 80% on any week. Although transportation of infectious cattle within a state would not immediately spark a regional epidemic, cattle movements at this scale could spread disease beyond the 10km control radius to be established around infected premises in response to FMD detection within the US [1]. Sale at livestock markets is not the only impetus for cattle transportation, but the correlation between representative sale origins and ICVI origins for transportation between states demonstrates its importance. Indeed, if the certificate-derived data do sample all movements without bias, then the magnitude of correlation with representative sales supports the hypothesis that most cattle (excepting slaughter animals) shipped between states are bound for a livestock market. Shipments leaving livestock markets would have to primarily remain in-state, and therefore be absent from certificate-derived data, for this hypothesis to hold: for example, it implies the testable conclusion that feedlots and back-grounders obtain most cattle born out-of-state from in-state livestock markets.

The market-derived data allows estimates for contact networks ranging in complexity from random graphs, which have many analytically tractable properties, to networks with non-trivial clustering, modularity, assortativity and other non-random features. A collection of edge data, resulting from sampling random nodes without tracing its edges to sample additional nodes, is an ego-centric sample that allows straightforward estimation of node, but not edge, attributes [45]. From this sample, we find market degree distributions that fit a negative binomial distribution with variance roughly twice the mean, which has more dispersion than a Poisson distribution but less than a power law. We also find seasonal shifts in the degree distribution that lower the epidemic threshold of a random graph during the peak cattle trading season. However, the overlapping marketsheds apparent in this sample suggest high potential for network clustering, which tends to dampen the spread of disease but is not easily assessed from an ego-centric sample [55]. Estimating this kind of structural attribute requires inference about unobserved edges, and a first analysis shows potential for incorporating linear effects of county and market attributes in exponential family likelihoods for edge weight. Extensions to this likelihood that include multiple response variables, particularly market degree and squares in the bi-partite network, may achieve a reliable fit to the representative sales that provides a data-driven, non-random graph for livestock disease simulations [46].

Despite the increased availability of data on livestock transportation in the US that our study provides, disease models here lag behind the relatively data rich European livestock systems. Research in these systems on the optimal spatial and temporal resolution at which to model contact networks is critical for efficient use of limited information available in the US and targeted development of new data streams. In a spatially embedded contact network, each node represents a geographically constrained subpopulation within an interacting metapopulation [56]. The constraint should reflect where mixing of susceptible and infected individuals occurs in proportion to their frequencies, but no theory exists for transferring constraints developed in one region (e.g. the UK) to any other (e.g. Pennsylvania). The abstraction of temporally discrete livestock shipments into static edges, representing the potential for disease transmission over time, is better understood [7,57]. An additional challenge, for a contact network distinguishing livestock markets from longer term animal holdings, is the synchronicity of shipments of cattle between two counties arriving at the same market. Extreme cases of complete segregation of cattle from different origins versus within market mixing should bracket the range of disease outcomes [33].

The greater purpose of collecting data on livestock transportation is to improve surveillance for disease outbreaks and to guide prevention or control of epidemics. The sources of nation-wide data on US livestock movements contributing to these goals have previously included health certificate records accompanying interstate movements [5, 21] and owner/operator surveys on animal health and management practices for representative animal holdings [58]. Future research should aim to combine these sources with representative sales data to jointly infer contact networks, because each data source addresses network attributes absent from the others. The primary deficiency of representative sales data is the absence of out-going shipment information, or the destination of cattle purchased at auction. Surveys of livestock operations presently include information on the in-shipment degree, source type and distance, which may provide evidence about the missing outgoing edges for livestock markets. Representative sales also only include market-directed shipments, while health certificate data provides information on network edges that may not have a livestock market at either end. Especially in combination, which we recognize to be a difficult task both conceptually and statistically, inference from multiple data sources will dramatically improve awareness of the network of potentially disease spreading contacts between livestock.

Supporting Information

S1 Text

Disease Percolation on a Directed Bi-Partite Random Graph.

S2 Text

Description of Data Release.

Acknowledgments

The authors want to acknowledge several undergraduate or graduate assistants who contributed to software development: Daniel Anderson, Adam Graves, Ching-Hao Hu, and Xinyang Jiang. We thank Nancy Robinson at the Livestock Marketing Association for sharing a member directory, and Centennial Livestock Auction of Fort Collins, CO for answering questions on cattle market practices. Funding for contributions by ITC and SB was provided through DHS Contract #HSHQDC-12-C-0014, with additional support from the RAPIDD Program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. The analyses, views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the regulatory opinions, official policies, either expressed or implied, of the USDA-APHIS-Veterinary Services or the U.S. Department of Homeland Security.

References

[1] US Department of Agriculture. Foot-and-Mouth Disease Response Plan: The Red Book. USDA:APHIS:VS, FAD PReP; 2014. Available from: `http://www.aphis.usda.gov/animal_health/emergency_management/downloads/fmd_responseplan.pdf`.

[2] Thompson D, Muriel P, Russell D, Osborne P, Bromley A, Rowland M, et al. Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. Revue scientifique et technique (International Office of Epizootics). 2002;21(3):675–687.

[3] Carpenter TE, O'Brien JM, Hagerman AD, McCarl BA. Epidemic and economic impacts of delayed detection of foot-and-mouth disease: a case study of a simulated outbreak in California. Journal of Veterinary Diagnostic Investigation. 2011;23(1):26–33.

[4] US Department of Agriculture. 2012 Census of Agriculture. USDA:NASS; 2012. Available from: `ftp://ftp.nass.usda.gov/quickstats/qs.census2012.txt.gz`.

[5] Shields DA, Mathews KH. Interstate Livestock Movements. US Department of Agriculture, Economic Research Service; 2003. Available from: http://www.ers.usda.gov/media/312234/ldpm10801_1_.pdf.

[6] Anderson I. Foot and Mouth Disease 2001: Lessons to be Learned Inquiry Report; 2002. Available from: http://webarchive.nationalarchives.gov.uk/20100807034701/http://archive.cabinetoffice.gov.uk/fmd/fmd_report/report/index.htm.

[7] Vernon MC, Keeling MJ. Representing the UK's cattle herd as static and dynamic networks. Proceedings of the Royal Society B: Biological Sciences. 2009;276(October 2008):469–476.

[8] Shirley MDF, Rushton SP. Where diseases and networks collide: lessons to be learnt from a study of the 2001 foot-and-mouth disease epidemic. Epidemiology and infection. 2005 Dec;133(6):1023–32. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2870335&tool=pmcentrez&rendertype=abstract.

[9] Bigras-Poulin M, Thompson RA, Chriel M, Mortensen S, Greiner M. Network analysis of Danish cattle industry trade patterns as an evaluation of risk potential for disease spread. Preventive Veterinary Medicine. 2006 Sep;76(1-2):11–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16780975.

[10] Kao RR, Danon L, Green DM, Kiss IZ. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. Proceedings of the Royal Society B: Biological Sciences. 2006 Aug;273(1597):1999–2007. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1635475&tool=pmcentrez&rendertype=abstract.

[11] Nöremark M, Hå kansson N, Lewerin SS, Lindberg A, Jonsson A. Network analysis of cattle and pig movements in Sweden: measures relevant for disease control and risk based surveillance. Preventive veterinary medicine. 2011 May;99(2-4):78–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21288583.

[12] Dutta BL, Ezanno P, Vergu E. Characteristics of the spatio-temporal network of cattle movements in France over a 5-year period. Preventive Veterinary Medicine. 2014;117(1):79–94. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167587714002943.

[13] Lentz HHK, Konschake M, Teske K, Kasper M, Rother B, Carmanns R, et al. Trade communities and their spatial patterns in the German pork production network. Preventive Veterinary Medicine. 2011;98(2-3):176–181. Available from: http://dx.doi.org/10.1016/j.prevetmed.2010.10.011.

[14] Natale F, Giovannini A, Savini L, Palma D, Possenti L, Fiore G, et al. Network analysis of Italian cattle trade patterns and evaluation of risks for potential disease spread. Preventive Veterinary Medicine. 2009;92(4):341–350.

[15] Regulation (EC) No 1760/2000 of the European Parliament and of the Council of 17 July 2000 establishing a system for the identification and registration of bovine animals and regarding the labelling of beef and beef products and repealing Council Regulati; 2000. Available from: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000R1760.

[16] Bajardi P, Barrat A, Savini L, Colizza V. Optimizing surveillance for livestock disease spreading through animal movements. Journal of the Royal Society, Interface / the Royal Society. 2012 Nov;9(76):2814–25. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3479905&tool=pmcentrez&rendertype=abstract.

[17] Green DM, Kiss IZ, Kao RR. Modelling the initial spread of foot-and-mouth disease through animal movements. Proceedings of the Royal Society B: Biological Sciences. 2006 Nov;273(1602):2729–35. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1635508&tool=pmcentrez&rendertype=abstract.

[18] Gates MC, Woolhouse MEJ. Controlling infectious disease through the targeted manipulation of contact network structure. Epidemics. 2015;Available from: http://linkinghub.elsevier.com/retrieve/pii/S1755436515000304.

[19] US Department of Agriculture, National Agricultural Statistics Service. Meat Animals Production, Disposition, and Income: 2014 Summary; 2015. Available from: http://www.usda.gov/nass/PUBS/TODAYRPT/meat0415.pdf.

[20] Lindström T, Grear DA, Buhnerkempe MG, Webb CT, Miller RS, Portacci K, et al. A bayesian approach for modeling cattle movements in the United States: scaling up a partially observed network. PloS One. 2013 Jan;8(1):e53432. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3537632&tool=pmcentrez&rendertype=abstract.

[21] Buhnerkempe MG, Grear DA, Portacci K, Miller RS, Lombard JE, Webb CT. A national-scale picture of U.S. cattle movements obtained from Interstate Certificate of Veterinary Inspection data. Preventive Veterinary Medicine. 2013 Nov;112(3-4):318–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24035137.

[22] Buhnerkempe MG, Tildesley MJ, Lindström T, Grear DA, Portacci K, Miller RS, et al. The impact of movements and animal density on continental scale cattle disease outbreaks in the United States. PloS One. 2014 Jan;9(3):e91724. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3966763&tool=pmcentrez&rendertype=abstract.

[23] US Department of Agriculture. Beef 2007-08, Part II: Reference of Beef Cow-calf Management Practices in the United States, 2007-08. Fort Collins, CO: USDA:APHIS:VS, CEAH; 2009. Available from: http://www.aphis.usda.gov/animal_health/nahms/beefcowcalf/downloads/beef0708/Beef0708_dr_PartII.pdf.

[24] US Department of Agriculture. Traceability for Livestock Moving Interstate. USDA:APHIS; 2013. 6. Available from: http://www.gpo.gov/fdsys/pkg/FR-2013-01-09/pdf/2012-31114.pdf.

[25] US Department of Agriculture. Beef 200708, Part IV: Reference of Beef Cow-calf Management Practices in the United States, 200708. Fort Collins, CO: USDA:APHIS:VS, CEAH; 2010.

[26] US Department of Agriculture. Dairy 2007, Part II: Changes in the U.S. Dairy Cattle Industry, 19912007. Fort Collins, CO: USDA:APHIS:VS, CEAH; 2008.

[27] McReynolds SW, Sanderson MW, Reeves A, Sinclair M, Hill AE, Salman MD. Direct and indirect contact rates among livestock operations in Colorado and Kansas. Journal of the American Veterinary Medical Association. 2014;244(9):1066–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24739117.

[28] Bates TW, Thurmond MC, Carpenter TE. Direct and indirect contact rates among beef, dairy, goat, sheep, and swine herds in three California counties, with reference to control of po- tential foot-and-mouth disease transmission. American Journal of Veterinary Research. 2001;62:1121–1129.

[29] RTI International. GIPSA Livestock and Meat Marketing Study. Research Triangle Park, ND; 2007. Available from: http://www.gipsa.usda.gov/Publications/psp/live_meat_market.html.

[30] Bailey D, Brorsen BW, Thomsen MR. Identifying Buyer Market Areas and the Impact of Buyer Concentration in Feeder Cattle Markets Using Mapping and Spatial Statistics. American Journal of Agricultural Economics. 1995;77(2):309–318. Available from: http://ajae.oxfordjournals.org/content/77/2/309.abstract.

[31] McLaws M, Ribble C. Description of recent foot and mouth disease outbreaks in nonendemic areas: exploring the relationship between early detection and epidemic size. Canadian Veterinary Journal. 2007;48(10):1051–1062.

[32] Colizza V, Barrat A, Barthélemy M, Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(7):2015–2020.

[33] Dawson PM, Werkman M, Brooks-Pollock E, Tildesley MJ. Epidemic predictions in an imperfect world: modelling disease spread with partial data. Proceedings of the Royal Society B: Biological Sciences. 2015;282(1808):20150205–20150205. Available from: http://rspb.royalsocietypublishing.org/content/282/1808/20150205?etoc.

[34] Gibbens JC, Sharpe CE, Wilesmith JW, Mansley LM, Michalopoulou E, Ryan JBM, et al. Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. The Veterinary Record. 2001;149(24):729–743.

[35] Robinson SE, Christley RM. Exploring the role of auction markets in cattle movements within Great Britain. Preventive Veterinary Medicine. 2007 Sep;81(1-3):21–37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17482296.

[36] Carroll IT, Bansal S. Cattle Auction Markets in the USA: a De-duplicated List Compiled from Trade Association and Federal Agency Directories. V1. Harvard Dataverse. 2014;.

[37] Livestock Marketing Association. Find an Auction/Dealer; 2014. Available from: http://www.lmaweb.com/find-an-auctiondealer-2/.

[38] GeoNames. Web Service Documentation; 2014. Available from: http://www.geonames.org/export/web-services.html.

[39] Oak Ridge National Laboratory. County-to-County Distance Matrix; 2011. Available from: http://cta.ornl.gov/transnet/SkimTree.htm.

[40] US Department of Agriculture. Livestock, Poultry, and Grain Market News. USDA:AMS; 2015. Available from: http://www.ams.usda.gov/AMSv1.0/LPSMarketNewsPage.

[41] Harrison XA. Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ. 2014 Jan;2:e616. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4194460&tool=pmcentrez&rendertype=abstract.

[42] Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7; 2014. Available from: http://cran.r-project.org/package=lme4.

[43] Fox J, Weisberg S. An {R} Companion to Applied Regression. 2nd ed. Thousand Oaks {CA}: Sage; 2011. Available from: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

[44] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: http://www.r-project.org.

[45] Morris M. Network epidemiology: A handbook for survey design and data collection. Oxford University Press; 2004.

[46] Handcock MS, Gile KJ. Modeling social networks from sampled data. The Annals of Applied Statistics. 2010;4(1):5–25. Available from: http://arxiv.org/abs/1010.0891.

[47] Zeileis A, Kleiber C, Jackman S. Regression Models for Count Data in R. Journal of Statistical Software. 2008;27(8). Available from: http://www.jstatsoft.org/v27/i08/.

[48] Bansal S, Grenfell BT, Meyers LA. When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society, Interface / the Royal Society. 2007;4(16):879–891.

[49] Ames GM, George DB, Hampson CP, Kanarek AR, McBee CD, Lockwood DR, et al. Using network properties to predict disease dynamics on human contact networks. Proceedings of the Royal Society B: Biological Sciences. 2011;278(1724):3544–3550.

[50] Newman M. Spread of epidemic disease on networks. Physical Review E. 2002 Jul;66(1):016128. Available from: http://link.aps.org/doi/10.1103/PhysRevE.66.016128.

[51] Meyers LA, Newman MEJ, Martin M, Schrag S. Applying network theory to epidemics: Control measures for Mycoplasma pneumoniae outbreaks. Emerging Infectious Diseases. 2003;9(2):204–210.

[52] Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, et al. Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. Ecological Monographs. 2014;84(1):45–67.

[53] Hsieh TC, Ma KH, Chao A. iNEXT: iNterpolation and EXTrapolation for species diversity. R package version 2.0.; 2014. Available from: http://chao.stat.nthu.edu.tw/blog/software-download.

[54] RTI. GIPSA Livestock and Meat Marketing Study. Report Volume 3. 2007 Feb;3. Available from: http://www.gipsa.usda.gov/Publications/psp/live_meat_market.html.

[55] Smieszek T, Fiebig L, Scholz RW. Models of epidemics: when contact repetition and clustering should be included. Theoretical Biology & Medical Modelling. 2009;6:11.

[56] Schumm P, Scoglio C, Zhang Q, Balcan D. Global epidemic invasion thresholds in directed cattle subpopulation networks having source, sink, and transit nodes. Journal of Theoretical Biology. 2014;367C:203–221. Available from: http://dx.doi.org/10.1016/j.jtbi.2014.12.007http://www.sciencedirect.com/science/article/pii/S0022519314006961.

[57] Valdano E, Ferreri L, Poletto C, Colizza V. Analytical computation of the epidemic threshold on temporal networks. arXiv preprint. 2014;p. 19. Available from: http://arxiv.org/abs/1406.4815.

[58] US Department of Agriculture. Analysis of Bovine Tuberculosis Surveillance in Accredited Free States. USDA:APHIS:VS; 2009. Available from: https://www.aphis.usda.gov/vs/nahss/cattle/tb_2009_evaluation_of_tb_in_accredited_free_states_jan_09.pdf.

S1 Text: Disease Percolation on Directed, Bi-Partite Random Graphs

Equations 4 through 6 in the main text summarize results for the spread of disease transmitted with constant probability through a bi-partite random graph. The theory leading to these results is summarized by Newman [1] and Meyers et al. [2]. For convenience we re-present an extension of the basic theory to a directed, bi-partite random graph. But in doing so, we also clarify how the number of infected nodes of just one type, "counties" in our case, may be followed through the derivation.

In a directed, bi-partite graph, where nodes have type M or C for market or county, respectively, edges are either $M \to C$ or $C \to M$. Let $G_{C,0}(x)$ be the generating function for the probability distribution on the number of $C \to M$ edges leaving a C node, marginalizing its in-degree. The county "excess degree" distribution is the probability distribution on the number of $C \to M$ edges departing from the county at the end of a randomly chosen $M \to C$ edge. Its generating function is

$$G_{C,1} = \frac{G'_{C,0}(x)}{G'_{C,0}(1)}, \tag{1}$$

as usual. The market "excess degree" distribution is the same, with M instead of C.

The key random variable of interest is $S_{C\to M}$, the number of infected counties in the cluster of nodes reached by tracing the out-going edge of a particular infected county. Let's denote the generating function for $S_{C\to M}$ by $H_{C,1}(x)$, and highlight that we're neither counting the number of markets mixed up in this cluster nor the original county. The function $H_{M,1}(x)$ will generate the distinct distribution on the number of infected *counties* reached by tracing a $M \to C$ edge. We determine these functions by deriving self-consistency equations from the following two observations. First, using superscipt (i) to indicate i of N independent samples of the random variable, N as the given out-degree of a market, and T as the given boolean variable for successful disease transmission:

$$S_{C\to M}|T, N = \begin{cases} 0 & T = \text{False} \\ S^{(1)}_{M\to C} + S^{(2)}_{M\to C} + \dots S^{(N)}_{M\to C} & T = \text{True}. \end{cases} \tag{2}$$

This bookeeping equation results from following the instructions, "choose among all $C \to M$ edges going to markets with out-degree N and add up the number of counties infected, assuming the market either is or is not infected." The second observation is

$$S_{M\to C}|T, N = \begin{cases} 0 & T = \text{False} \\ 1 + S^{(1)}_{C\to M} + S^{(2)}_{C\to M} + \dots S^{(N)}_{C\to M} & T = \text{True}. \end{cases} \tag{3}$$

In the next step, we achieve the desired generating functions on the right hand side:

$$\langle x^{S_{C\to M}}|N\rangle = (1-\tau_C) + \tau_C H_{M,1}(x)^N \tag{4}$$

and

$$\langle x^{S_{M\to C}}|N\rangle = (1-\tau_M) + \tau_M x H_{C,1}(x)^N, \tag{5}$$

where τ_C and τ_M are the probabilities that $T = \text{True}$ in Eq. 2 and Eq. 3, respectively. Averaging each equation over the appropriate "excess degree" distribution for N gives the coupled system:

$$H_{C,1}(x) = (1-\tau_C) + \tau_C G_{M,1}(H_{M,1}(x)) \tag{6a}$$
$$H_{M,1}(x) = (1-\tau_M) + \tau_M x G_{C,1}(H_{C,1}(x)). \tag{6b}$$

These must be solved in order to obtain the distribution on the outbreak size starting in a randomly infected county, which is generated by $H_{C,0}(x)$ and derived starting from the observation that

$$\langle x^{1+S^{(1)}_{C\to M}+S^{(1)}_{C\to M}+\cdots+S^{(N)}_{C\to M}}|N\rangle = xH_{C,1}(x)^N. \tag{7}$$

The derivation is completed by averaging over the county out-degree distribution (*not* the county "excess degree" distribution) to obtain

$$H_{C,0}(x) = xG_{C,0}(H_{C,1}(x)). \tag{8}$$

The outbreak size and epidemic proportion calculations follow in the usual way. Let $u = H_{C,1}(1)$ and $v = H_{M,1}(1)$. Using the $u = v = 1$ solution to Eq. 6, the expected outbreak size reduces to

$$H'_{C,0}(1) = 1 + \frac{\tau_C\tau_M G'_{C,0}(1)G'_{M,1}(1)}{1+\tau_C\tau_M G'_{C,0}(1)G'_{M,1}(1)}. \tag{9}$$

Numerically finding other solutions for u and v, with $0 < u < 1$ and $0 < v < 1$, leads to the epidemic size proportion as

$$1 - H_{C,0}(1) = 1 - G_{C,0}(u). \tag{10}$$

To obtain Eqs. 4 through 6 in the main text, we assume that county out-degree is Poisson distributed with mean λ_C and that market in- and out-degree are perfectly correlated. Imperfect correlations could be included with an additional parameter.

References

[1] Newman M. Spread of epidemic disease on networks. Physical Review E. 2002;66(1):016128. doi: 10.1103/PhysRevE.66.016128.

[2] Meyers LA, Newman MEJ, Martin M, Schrag S. Applying network theory to epidemics: Control measures for Mycoplasma pneumoniae outbreaks. Emerging Infectious Diseases. 2003;9(2):204–210. doi: 10.3201/eid0902.020188.

S2 Text: Description of Data Release

Note: The data release will coincide with publication of the report – the DOI given below will remain inactive until release.

Representative sales data collected from several livestock market websites, as aggregated for the analyses in this report, are available for download from the Bansal Lab Dataverse [1]. Accompanything the data are scripts for R [2] that reproduce the results presented in the main text.

volume.csv

date (char as YYYY-MM-DD) date of cattle auction given on each sale report

market (int) unique market identifier

orig_location (char) FIPS code for county at market street address

dest_location (char) FIPS code for nearest county containing the origin city[, state]

head (int) number of cattle in all lots (each defaults to 1 for missing data)

A disaggregated version of the representative sales data, sufficient to re-create the panels of Fig. 1 in the main text. Note that the first two characters of a FIPS code correspond to the state, allowing for in-state proportion calculations.

proportion.csv

year (char) year of cattle auction

week (char) ISO week of year

market (int) unique market identifier

receipts (int) total head sold at auction from sale report or [3] (if unreported)

head (int) head given as representative sales (lot size defaults to 1 for missing data)

sales (int) head in county-wide sales*

inventory (int) head in county-wide inventory*

farms (int) farms in county-wide inventory*

The script **proportion.R** reads this file and fits a binomial family GLMM, associating the sampling probability for representative sales in each report with the covariates provided.

certificate.csv

orig (char) State abbreviation for cattle origin

dest (char) State abbreviation for cattle destination

rep_sales (int) Head from representative sales, (see Methods for time-span)

flows (int) Head from certificate-derived data†

Note that the data from [5] are available in electronic form at *http://webarchives.cdlib.org/sw12j6951w/http://www.ers.usda.gov/Data/InterstateLivestockMovements/View.asp*. No script is provided to calculate the correlations between interstate rep_sales and flows for each dest.

*Reproduced for convenience from [4], without endorsement by the USDA.

†Reproduced for convenience from [5]

edge.csv

market (char) year of cattle auction

dest_location (char) FIPS code for nearest county containing the origin city[, state]

sales (int) head in county-wide sales

inventory (int) head in county-wide inventory

farms (int) farms in county-wide inventory

distance (real) great-circle distance between orig_location and dest_location county centroids‡

head (int) head given as representative sales (lot size defaults to 1 for missing data)

instate (bool) zero if and only if dest_location and market are in different states

The script **edge.R** reads this file and fits a zero-inflated negative-binomail family GLM, associating a zero-inflation probability and mean head of cattle for each county-market pair with the covariates provided. The script additionally simulates counts for each pair, with the same random seed used in this report, and writes the counts to a new file.

References

[1] Carroll, IT, Bansal S. Replication Data for: Livestock market data for modeling disease spread among US cattle. Harvard Dataverse 2015. Available from: `http://dx.doi.org/10.7910/DVN/MAXZAY`.

[2] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: `http://www.r-project.org`.

[3] USDA\AMS. Livestock, Poultry, and Grain Market News;. Available from: `http://www.ams.usda.gov/AMSv1.0/LPSMarketNewsPage`.

[4] USDA\NASS. 2012 Census of Agriculture; 2015. Available from: `ftp://ftp.nass.usda.gov/quickstats/qs.census2012.txt.gz`.

[5] Shields DA, Mathews KH. Interstate Livestock Movements. USDA:ERS; 2003. Available from: `http://www.ers.usda.gov/media/312234/ldpm10801_1_.pdf`.

[6] County-to-County Distance Matrix; 2011. Available from: `http://cta.ornl.gov/transnet/SkimTree.htm`.

‡Reproduced for convenience from [6]