-Electronic Supplementary Material- Optimizing surveillance for livestock disease spreading through animal movements

Paolo Bajardi[1,2], Alain Barrat[2,3], Lara Savini[4], Vittoria Colizza[5,6,7]

1 Network aggregation and simulation procedure

The data describing bovines' displacements are recorded on a daily basis, making it possible to construct different network representations according to different assumptions or aggregating on different time window lengths [1] [2]. In particular, we simulated the spread of an infectious disease on networks aggregated on a daily, weekly, monthly and yearly scale. The choice of the aggregating time window length Δt affects the underlying mobility structure, leading to denser displacement networks for longer time windows, while the time step used in the numerical simulations of the spreading dynamics is kept fixed to 1 day. In this perspective, when the spreading process takes place on the daily dynamical networks, at every time step of the spreading the snapshot of the static network is different, while for a longer aggregating window length the network topology remains unchanged for exactly Δt time steps. In Figure 2 of the main paper, the unfolding of the spreading for different aggregating time window lengths is followed by plotting the temporal evolution of the number of infected premises for every spreading time step (=1 day). In the following, if not otherwise specified, we present results corresponding to spreading phenomena simulated on networks aggregated on time windows of length $\Delta t = 1$ day, starting at t_0 =Jan 1, and with an infectious period $\mu^{-1} = 7$ days.

2 Overlap values and clusters

The initial conditions similarity network (ICSN) is a fully connected network where all the information resides in the links weight corresponding to the overlap value of the invasion

[1]Computational Epidemiology Laboratory, Institute for Scientific Interchange (ISI), Turin, Italy
[2]Centre de Physique Théorique, Aix-Marseille Univ, CNRS UMR 6207, Univ Sud Toulon Var, 13288 Marseille cedex 9, France
[3]Data Science Laboratory, Institute for Scientific Interchange (ISI), Turin, Italy
[4]Istituto Zooprofilattico Sperimentale Abruzzo-Molise G. Caporale, Teramo, Italy
[5]INSERM, U707, Paris F-75012, France
[6]UPMC Université Paris 06, Faculté de Médecine Pierre et Marie Curie, UMR S 707, Paris F75012, France
[7]Institute for Scientific Interchange (ISI), Turin, Italy

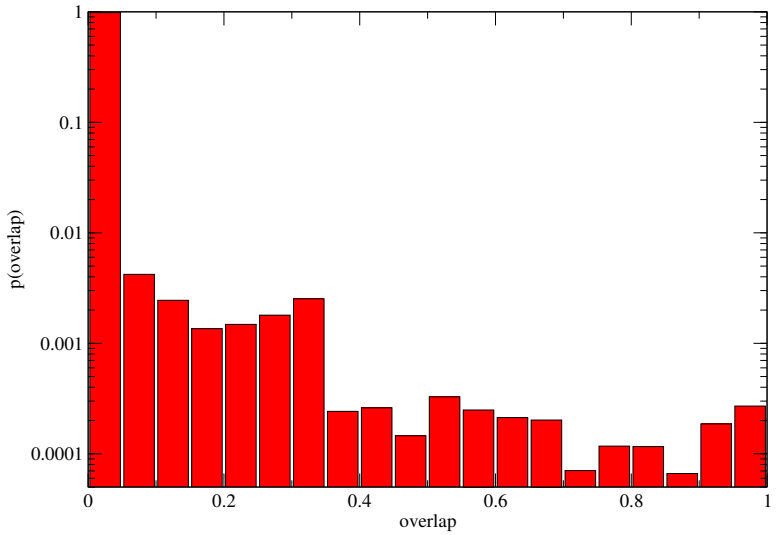

Figure 1: Distribution of the overlaps between invasion paths for the initial conditions similarity network resulting from the simulations of a deterministic spreading phenomenon starting at t_0 =Jan 1, with an infectious period $\mu^{-1} = 7$ days, and performed on a dynamic network corresponding to an aggregating time window length of $\Delta t = 1$ day.

paths generated by the different seeding nodes. Figure 1 shows the corresponding overlap distribution. Most nodes, when taken as seed of a spreading phenomenon, yield very short infection paths and very few infected nodes, so that the resulting overlap with most other paths is zero. Moreover, some particular sub-structures lead to an abundance of overlap values equal to 0.33: some nodes, such as the slaughterhouses, have a large in-degree and no out-links; two seeding nodes that are only linked (during their infectious period) to a slaughterhouse yield therefore two infection paths having each two nodes, among which one is common (the slaughterhouse), leading to an overlap of 1/3. In order to avoid taking into account such structures, we focus in our study on larger overlap values, and often restrict the computations to spreading paths containing at least 10 nodes. As described in the main text, considering in the ICSN only the links corresponding to an overlap larger than a given threshold Θ_{th} separates the ICSN into several connected components, each of whom defines a cluster of nodes. The schematic illustration of the clustering procedure is shown in Figure 3 of the main text. In each cluster, all nodes are however not a priori connected with each other: the fact that two nodes i and j belong to the same cluster simply means that there exists a set of other nodes $i_1, ...i_p$ such that the overlaps $\Theta_{ii_1}, \Theta_{i_1i_2}, \Theta_{i_pj}$ are all larger than the threshold, but it does not imply that the overlap between i and j is larger than Θ_{th}. It is therefore important to measure the distribution of overlaps of all pairs of nodes inside a cluster. Figure 2 shows these distributions for several clusters constructed for a deterministic spreading starting at t_0 =Jan 1, with infectious period $\mu^{-1} = 7$days, and a threshold value $\Theta_{th} = 0.8$: even if the distributions extend to values lower than the threshold, no small overlap values are observed. The clusters therefore group nodes yielding pairwise similar invasion paths.

The clustered structures emerge by pruning links with weights smaller than a certain

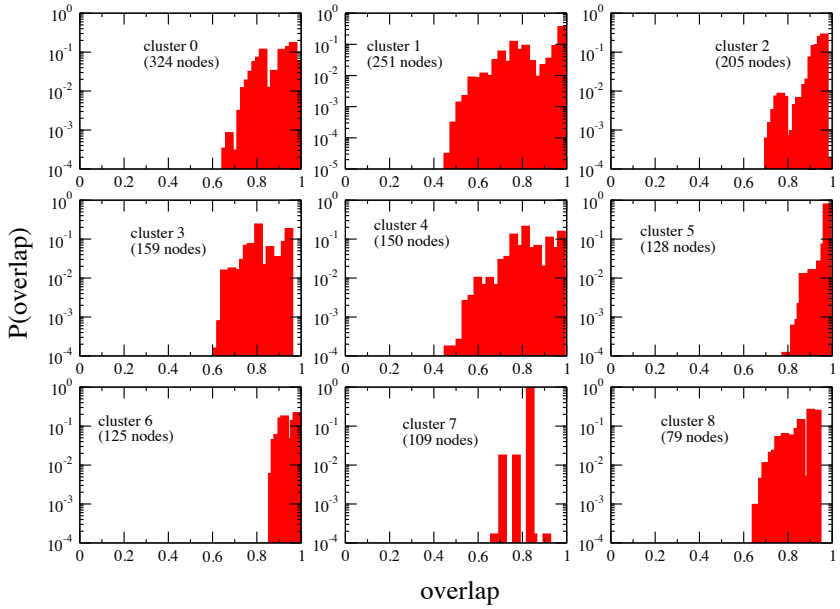

Figure 2: Distribution of the overlaps between invasion paths rooted in nodes belonging to a given cluster. The clusters correspond to the invasion paths of a deterministic spreading phenomenon starting at t_0 =Jan 1, with an infectious period $\mu^{-1} = 7$ days, performed on a dynamic network corresponding to an aggregating time window length of $\Delta t = 1$ day, and obtained with a threshold value of 0.8. The large values of the overlap between all pairs of nodes belonging to the same cluster indicates that all the nodes of a cluster lead to a similar spreading behavior, even if some pairs of nodes have an overlap smaller than the threshold.

threshold in the initial condition similarity network. As the choice of the threshold is arbitrary, it is important to check the robustness of the obtained cluster structure with respect to changes in the threshold value. We investigate this point in Figure 3 by measuring the intersection of clusters obtained with different threshold values. For large enough threshold values, the structure of resulting clusters is stable with respect to small variations in the thresholding criterion. The results described in the main text are obtained by fixing the threshold value to 0.8. Since, in principle, there could be as many clusters as there are seeding nodes, in Figure 4 the cluster size distribution is shown. Most of the clusters are isolated nodes, i.e. seeding nodes leading to invasion paths with overlap smaller than 0.8 with any other invasion path, but some non-trivial clusters with large sizes clearly emerge.

3 Initial Conditions Similarity Network from directed paths

As described in the Methods section of the main text, the measure we adopted to weight the links in the Initial Conditions Similarity Network (ICSN) described above is based on the Jaccard index evaluated as the number of common infected nodes normalized by the total number of nodes reached by the infection paths of the two seeds.

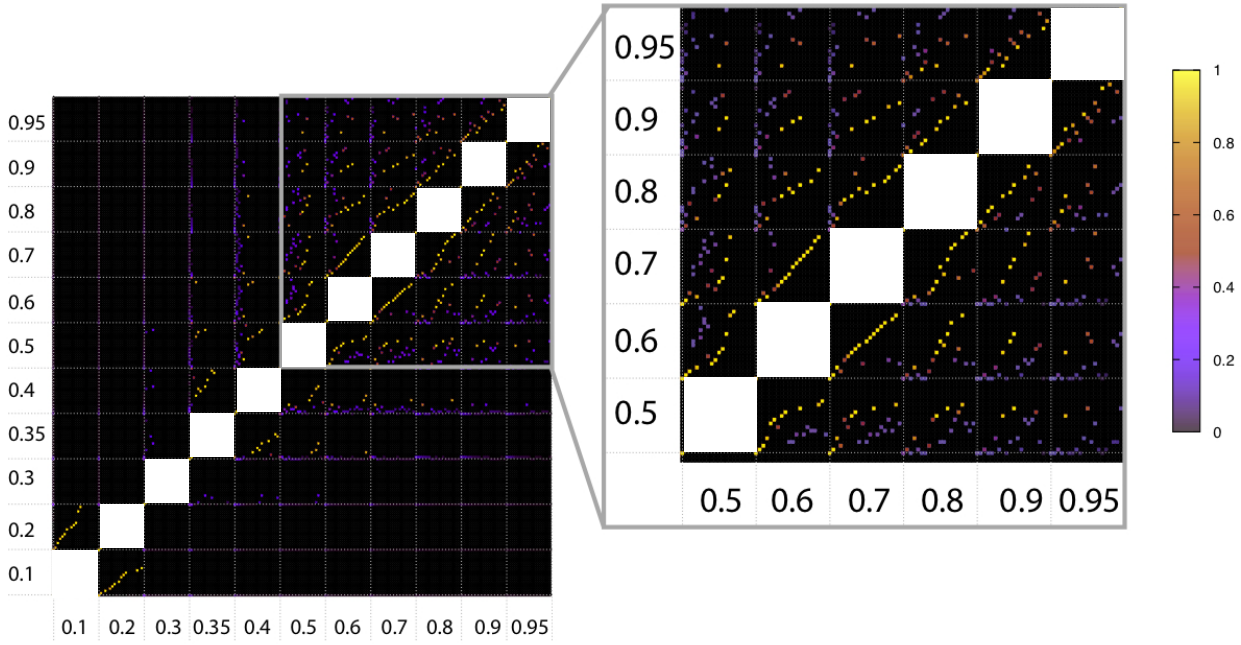

Figure 3: Jaccard indices between clusters constructed using different threshold values. For each couple of threshold values a and b, we show at row a and column b a color-coded matrix of the Jaccard indexes between the two sets of 20 largest clusters of the ICSN obtained for the thresholds values a and b. The Jaccard indices of two clusters is computed as the number of common nodes divided by the number of nodes in the union of the clusters. The cases $a = b$ are not shown as they trivially have a diagonal equal to 1 and zero off-diagonal elements. The violet-to-yellow color scale indicate how much the clusters obtained with different thresholds have in common. For threshold values larger than 0.6, the cluster structure is rather stable with respect to small changes in the thresholding criterion. Note that for each threshold value, the 20 largest clusters are ranked by size; as this ranking may change from one cluster value to the next, the yellow dots are not all on the diagonal of the corresponding matrix.

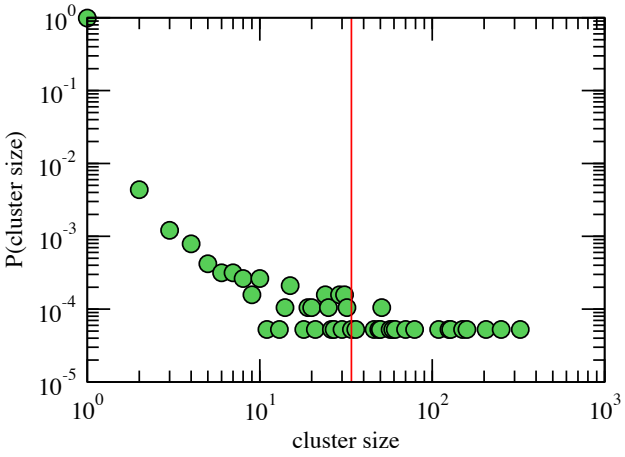

Figure 4: Cluster size distribution. The plot refers to the partition obtained imposing a threshold value of the Jaccard index of the invasion paths equal to 0.8. 98% of the clusters are represented by isolated nodes, but the probability distribution is fat tailed showing that a non-negligible number of clusters includes a large number of seeding nodes. The 20 largest clusters correspnd to the sizes on the right of the red line.

It is also possible to consider an ICSN that takes into account the directedness and thus the causality of the paths of infection. To this aim, we define the modified overlap Θ^l_{12} between two paths Γ_1 and Γ_2, composed by the sets of directed links $\vec{l}_1$ and $\vec{l}_2$, as the Jaccard index $\frac{|\vec{l}_1 \bigcap \vec{l}_2|}{|\vec{l}_1 \bigcup \vec{l}_2|}$, measuring the number of common directed links with respect to the total number of links in the two paths. In Figure 5, we compare the clusters obtained by weighting the links of the ICSN with the overlap of the nodes and with the overlap Θ^l of the directed links of the spreading paths. The partitions of the largest 20 clusters are very similar, but it is worth to notice that the construction of the ICSN using an overlap measure between paths based only on the intersection and union of the set of nodes composing the paths (as explained in the main text) relies on much less information and is therefore easier to achieve in real settings.

We therefore use, both in the main text and in the rest of the ESM, the clusters obtained through the definition of the weight of an ICSN link as the overlap between the sets of nodes of the spreading paths.

4 Clusters features

As explained in the main text, the construction of the clusters is based on a criterion of similarity of invasion paths. In the following we investigate several other characteristics of the clusters. Figures 6, 7 and 8 show some properties of the spreading starting from nodes belonging to the various clusters: the distribution of sizes are typically peaked around a well-defined characteristic value for each cluster. The maximal geographic distance distributions are also narrow or present only a small number of peaks with respect to the number of nodes in the cluster. Figure 8 moreover highlights the very strong similarity between the prevalence

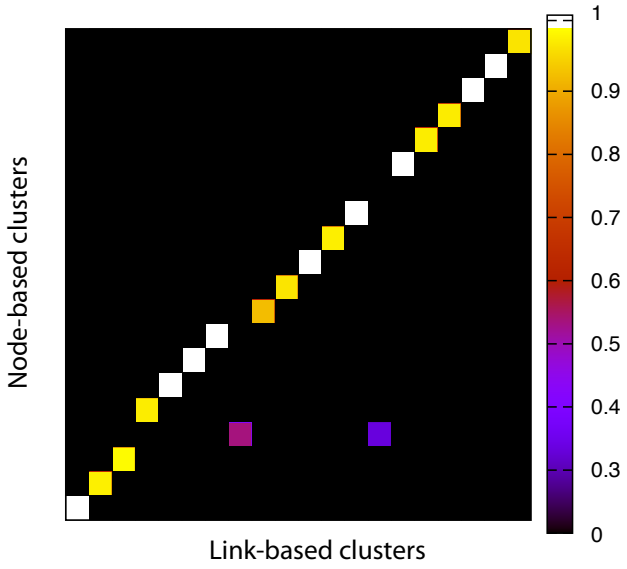

Figure 5: Jaccard indices between clusters constructed using two different definitions of the overlap of two spreading paths Γ_1 and Γ_2, either based only on the nodes composing these paths (Θ_{12}, defined in the main text), or taking into account the directed links (Θ^l_{12}, defined in the ESM text). The Jaccard index of two clusters is computed as the number of common nodes divided by the number of nodes in the union of the clusters, and is shown through a violet-to-yellow color scale.

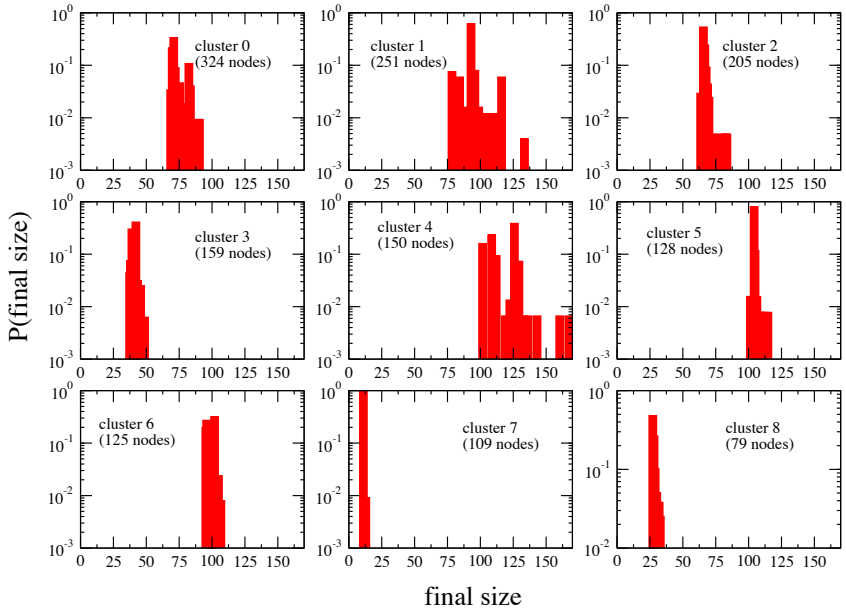

Figure 6: Final size distributions of epidemics rooted in nodes of various clusters.

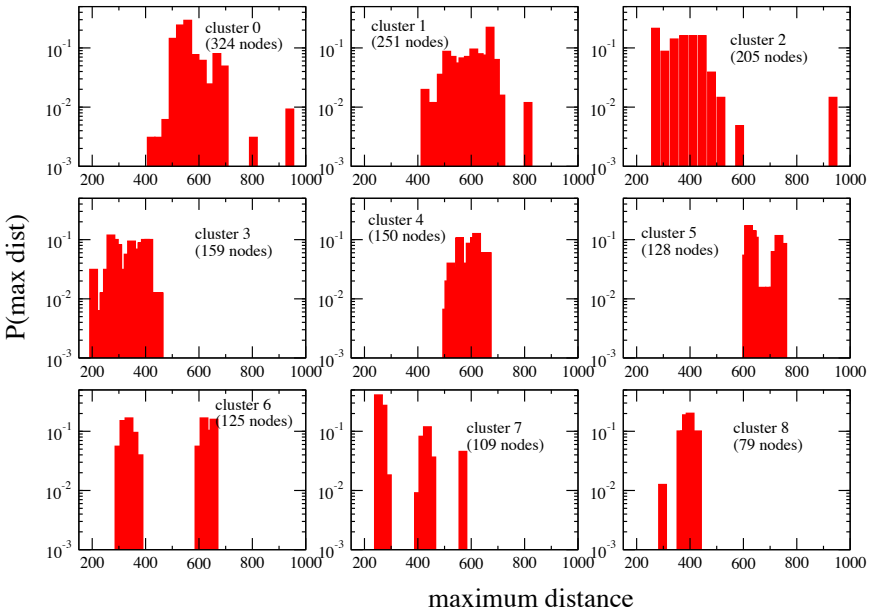

Figure 7: Distribution of the maximum geographical distance (in km) covered by the invasion paths rooted in nodes of various clusters. The long-range disease infections are a major concern in facing an emerging infectious disease. The existence of trade connections spanning several hundreds of kilometers makes possible a rapid and wide spread of the disease.

curves (shape, peak time, duration) of spreading phenomena starting from different nodes of a given cluster.

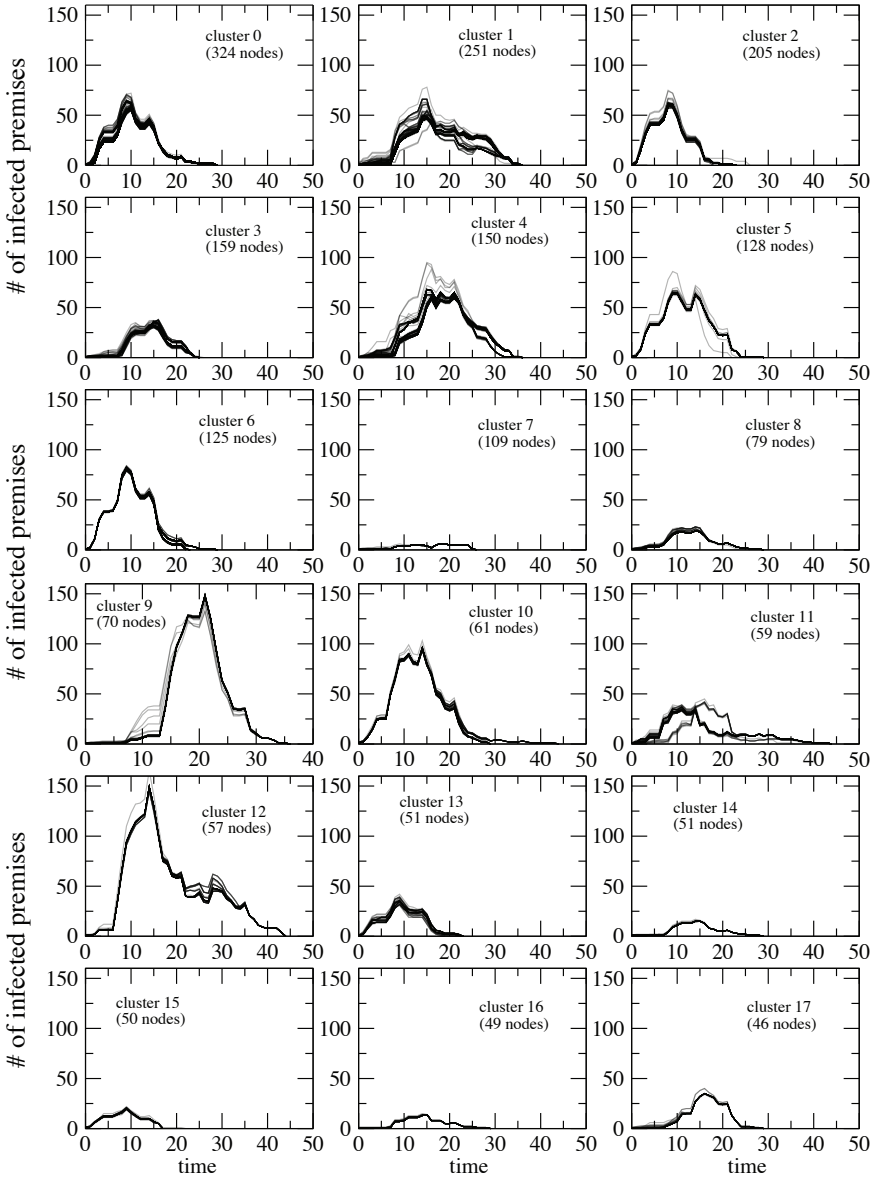

Figure 8: Prevalence curves of spreading phenomena rooted in nodes belonging to various clusters. This figure is similar to Figure 6(a) of the main text: here, the prevalence curves for nodes belonging to the 18 largest clusters are shown.

5 Longitudinal stability of the seeds' clusters

In the main text, the longitudinal stability of the clusters obtained by considering different starting times t_0 of the spreading has been evaluated using an entropy-like function.

In the following, we first present additional results on the clusters' stability measured through this entropy-like function, and then consider further methods to assess the temporal stability of clusters, in order to better characterize their evolution and to verify that such stability is not measure-dependent.

5.1 Entropy-like stability

To supplement the examples shown in Figure 6b of the main text, in Figure 9 we show the results of the cluster's stability for the other clusters, grouped according to their temporal behavior.

5.2 Sensitivity analysis

The entropy function introduced in Equation 1 of the main text evaluates the level of fragmentation of the clusters when different starting dates for the epidemic spreading simulations are chosen. Since most of the clusters have small sizes (see Figure 4), we restricted our analysis to the largest $C = 20$ clusters, and we normalize therefore the entropy function accordingly. In Figure 10, we report a sensitivity analysis with respect to considering different values of C. The results are shown not to be sensitive to the precise value of C. When all the clusters are considered however, the entropy function is shifted towards lower values because of the large number of clusters composed by isolated nodes: the large value of C increases therefore strongly the normalizing factor $\log C$. Using this normalization would therefore artificially enhance the *apparent* stability of the clusters.

5.3 Rand index

In order to compare two partitions of the seed nodes into clusters corresponding to two different seeding dates, we use the Rand index [3], which quantifies to which extent nodes that were clustered together in the first partition are still together in the second one, and, analogously, nodes that belonged to different clusters in the first partition are still classified separately in the second.

More precisely, given two partitions $\mathcal{P}(t_0) : \{C_1^0, C_2^0, ..., C_r^0\}$ and $\mathcal{P}(t_1) : \{C_1^1, C_2^1, ..., C_q^1\}$, the Rand index of $\mathcal{P}(t_0)$ and $\mathcal{P}(t_1)$ is defined by $R = \frac{t+s}{\binom{n}{2}}$ where t is the number of pairs of nodes belonging to the same cluster in both $\mathcal{P}(t_0)$ and $\mathcal{P}(t_1)$, s is the number of pairs of nodes classified in different clusters both in $\mathcal{P}(t_0)$ and $\mathcal{P}(t_1)$ and n is the number of nodes present in both partitions. In Figure 11 the Rand index of the partition $\mathcal{P}(t_0)$ with the partitions obtained for different starting times t_1 are shown.

If we consider all the clusters, we obtain extremely high values for this index. The reason lies in the very large number of clusters with size ≤ 2, shown in Figure 4: most isolated nodes in one partition remain isolated in the successive ones, leading to very high values of

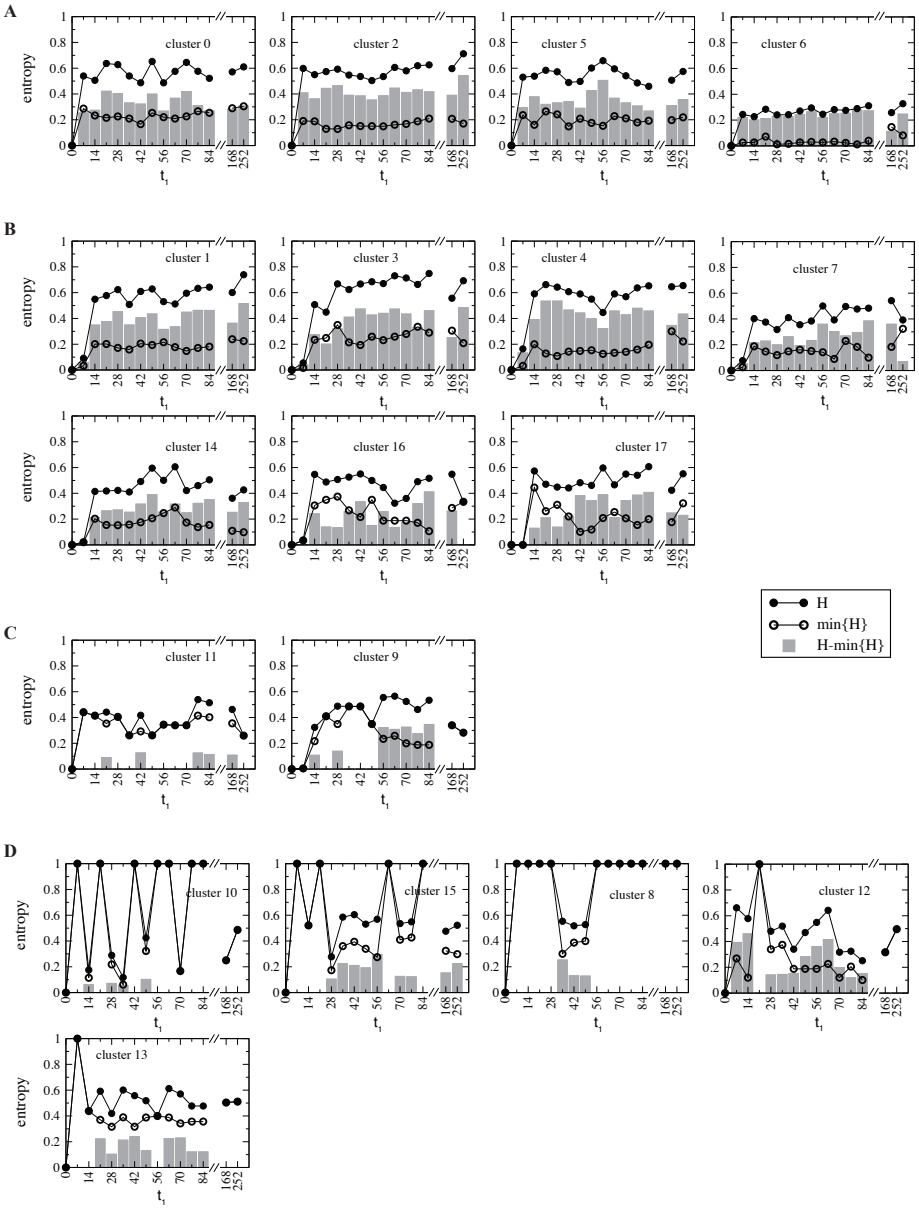

Figure 9: Entropy H of the partition into clusters as a function of time, for the same clusters as in Figure 8. H measures the fragmentation of the largest $C = 20$ clusters obtained for the starting time $t_0 =$ January 1^{st} in the partitions obtained in the following weeks. The difference $H - \min_H$ (grey bars) represents the robustness of the cluster (the smaller the difference and the more robust is the cluster), given that only part of it may be present in the partition obtained for a later starting condition (as measured by $\min_H$). Beyond the four typical behaviors described in the text, a hybrid evolution may also emerge: for instance, cluster 6 (the last in the first row of the figure) remains quite stable over the whole year.

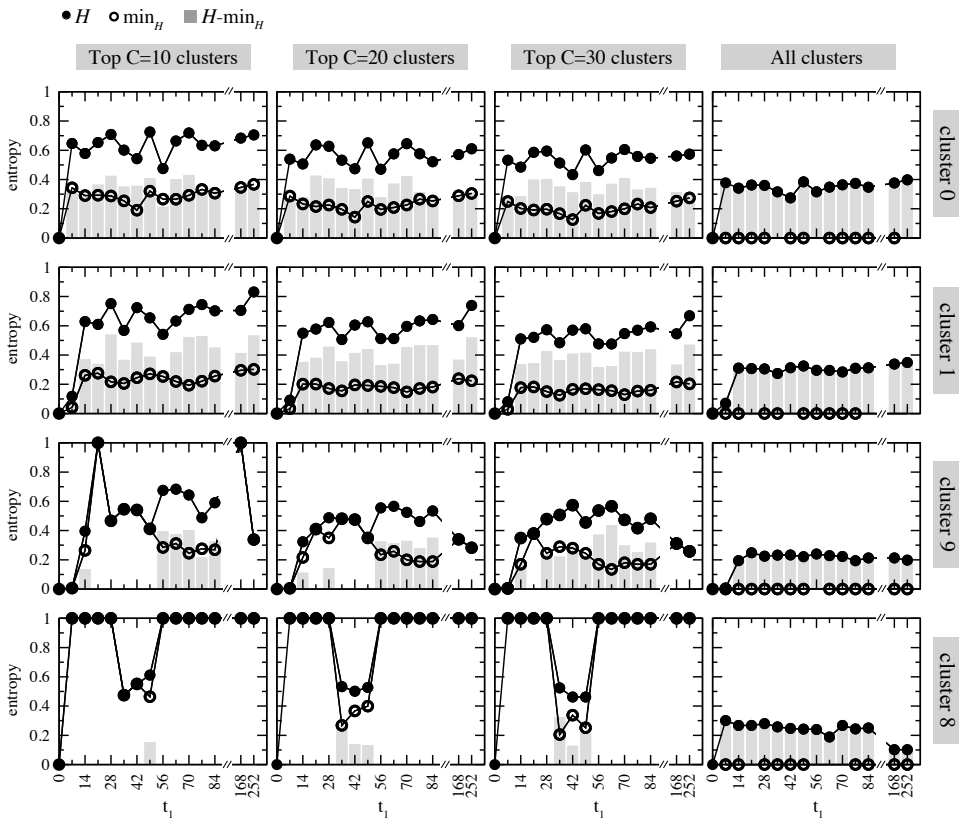

Figure 10: Longitudinal stability sensitivity analysis. The longitudinal stability of the clusters presented in Figure 6b of the main text is assessed by considering different values of C, representing the number of largest clusters considered for the evaluation of the normalized entropy. The results are robust and do not depend on the choice of the parameter C. When all the clusters are considered, the entropy is biased towards lower values because of the large number of small clusters. As in Figure 9 the entropy H, the minimum entropy $\min_H$ and the difference $H - \min_H$ are shown.

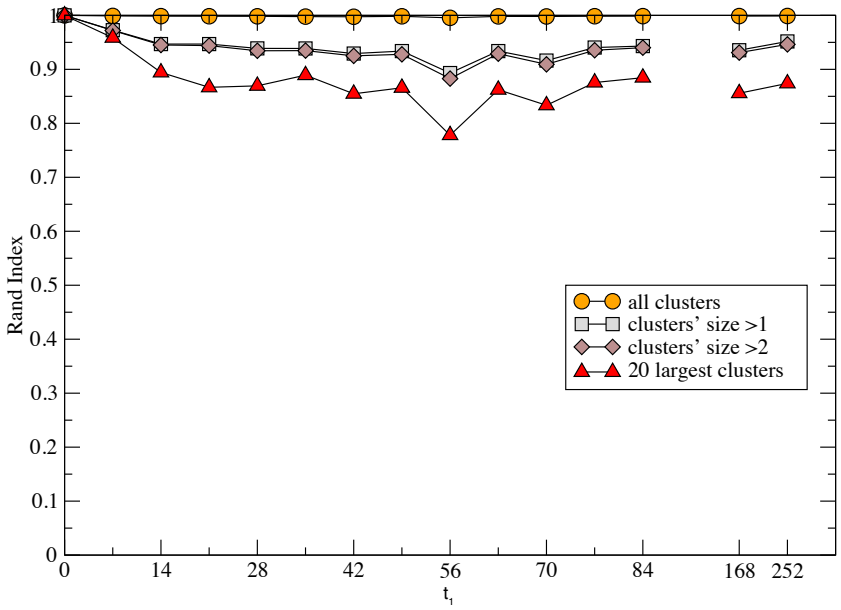

Figure 11: Rand index of the partition obtained for the starting time t_0 =Jan 1 with the partitions obtained for different starting times t_1 of the spreading simulations.

the Rand index. We therefore consider several modified versions of the Rand index, which take into account only a subset of clusters: respectively all clusters of size at least 2, all clusters of size at least 3, and the 20 largest clusters. In all cases, the large values of the Rand index indicate that the partitions obtained for different starting times remain fairly stable.

5.4 Best match

The Rand index is a *global* measure of the stability of the partition of the nodes into clusters, and does not give access to the individual stability of the different clusters. In order to go beyond this evaluation of the overall stability of the partitions and to obtain an alternative measure with respect to the entropy-like function studied above and in the main text, we can also investigate the longitudinal stability of the single clusters. To this aim, we compute the Jaccard index of each cluster of the partition obtained for a spreading starting time t_0, with all the other clusters corresponding to the partition obtained for starting time t_1.

In Figure 12 the Jaccard indices of the largest $C = 20$ clusters obtained with spreading simulations starting at $t_0 = 0$ with the largest 20 clusters obtained at subsequent starting times t_1 are shown. For each value of t_1 and for each cluster of the partition corresponding to t_0, it is then possible to find the best matching cluster in $\mathcal{P}(t_1)$ by focusing on the largest Jaccard index. Figure 13 shows for each of the 18 largest clusters of $\mathcal{P}(t_0)$ the evolution of this largest Jaccard index with t_1.

We also note that the Jaccard index can either be evaluated as the crude ratio of the intersection over the union of the nodes in the two clusters, or can be computed by considering only the nodes present in both partitions (i.e. the ratio of the intersection and the union

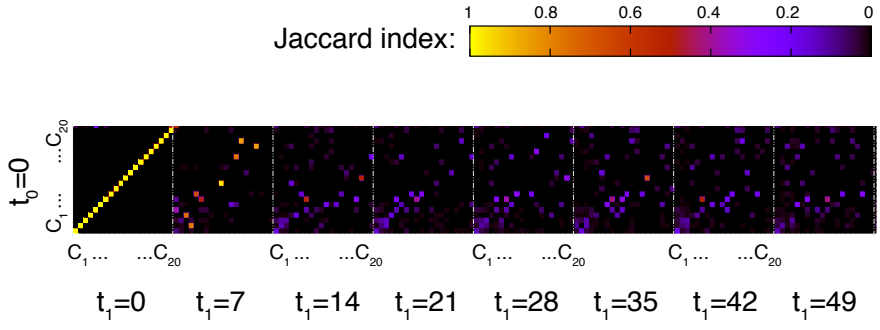

Figure 12: Jaccard indices between clusters obtained for different starting times of the spreading. The Jaccard index of two clusters is computed as the number of common nodes divided by the number of nodes in the union of the clusters, and is shown with a violet-to-yellow color scale.

of the nodes in the two clusters, given that the nodes are present in both $\mathcal{P}(t_0)$ and $\mathcal{P}(t_1)$). Both cases lead to similar results and are shown in Figure 13.

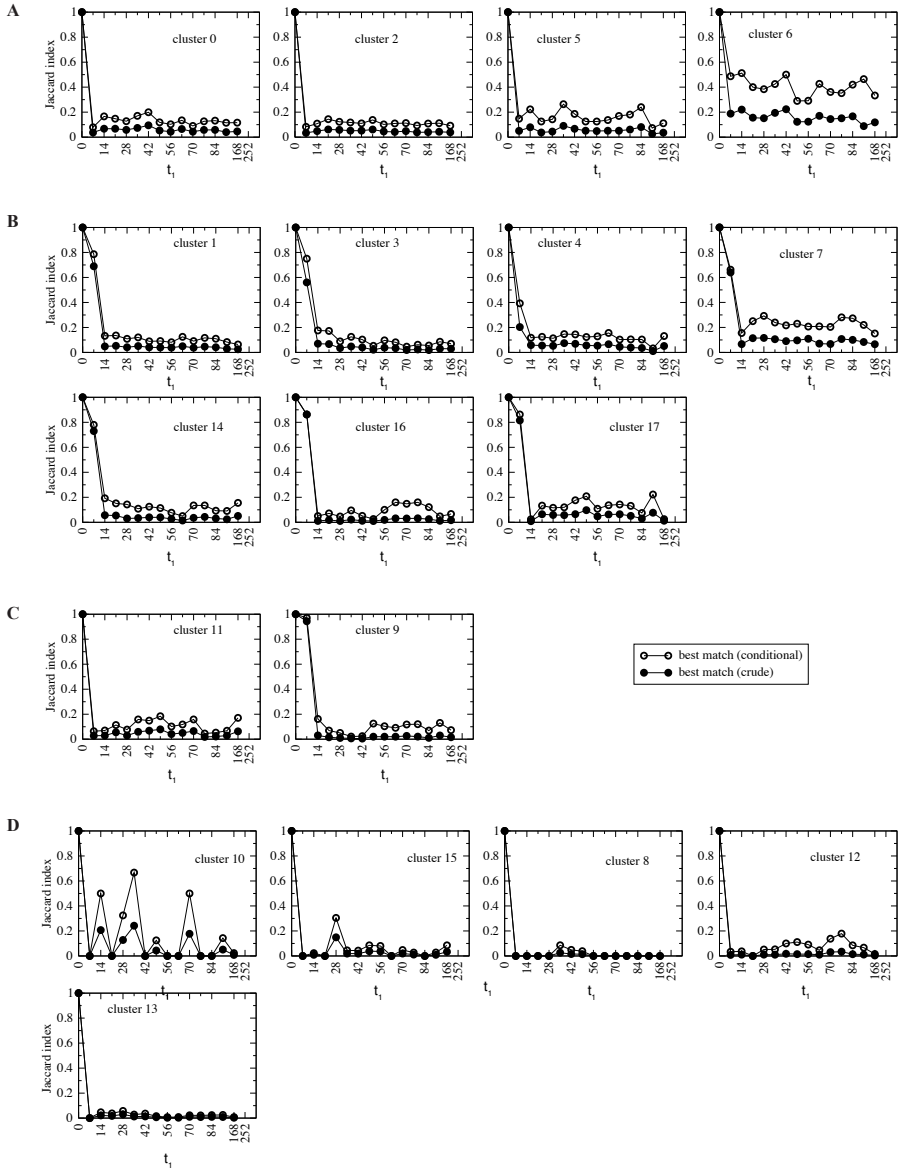

Figure 13: Best matching cluster. The highest value of the Jaccard indices evaluated between the clusters obtained at t_0 and those obtained for subsequent starting times are shown. The crude Jaccard index considers the ratio of the intersection and the union of the nodes of the two clusters, while in the conditional Jaccard index only the nodes that are present at both t_0 and t_1 are considered.

		n_s							
		30	40	50	60	70	80	90	100
ξ_s	0.50	42	40	37	34	31	29	25	25
	0.45	32	31	28	25	22	20	18	18
	0.40	15	14	11	10	9	7	5	5
	0.35	9	8	6	6	5	4	4	4

Table 1: Deterministic simulations. Number of sentinels for different thresholds in the $n - \xi$ space.

6 Sentinels

As described in the main text, the definition of sentinel nodes is flexible. In particular, depending on the surveillance system and on the resources availability it is possible to be more or less conservative in the choice of ξ_s and n_s. It is also possible to tune these values in order to restrict to nodes that provide a better characterization of the epidemic origin (low ξ_s) and that have a higher probability of being reached by an outbreak (high n_s). In table 1 the number of sentinel nodes that should be monitored for different choices of ξ_s and n_s are shown.

6.1 Normalizing factor

In the main text we defined the seeder uncertainty as an entropy-like function normalized with n_k. The presence of clusters of small sizes (isolated nodes or pairs of nodes) raises the total number of clusters, so that the condition $n_k < C$ is always satisfied in our simulations. In this context the most homogeneous infection patterns for the node k is represented by the probability $1/n_k$ of being infected by n_k different clusters. In the following we show the results by normalizing the seeder uncertainty with $log(C)$. The new normalization leads to very small values of the seeder uncertainty as induced by the large number of clusters C, but results are not overall affected by such rescaling. In Figure 14b we consider the 15 sentinels identified in the main text with the normalization $1/n_k$, and plot their trajectories in the plane uncertainty vs. n, where the uncertainty is now calculated with the new normalization $log(C)$ of the entropy function. The sentinels trajectories span a rather compact space demonstrating that changes in the normalizing factor do not alter our sentinel identification method, and just correspond to a rescaling of the threshold imposed for the new definition of the uncertainty.

7 Stochastic case

The high dimensionality of the phase space of the initial conditions has led us to prefer a deterministic computational approach for the presentation of the proposed methodology and main results. We show however in the following how our approach can be extended to stochastic simulations that take into account the intrinsic stochasticity of epidemic propagation phenomena. Results similar to the deterministic case are recovered. In the stochastic

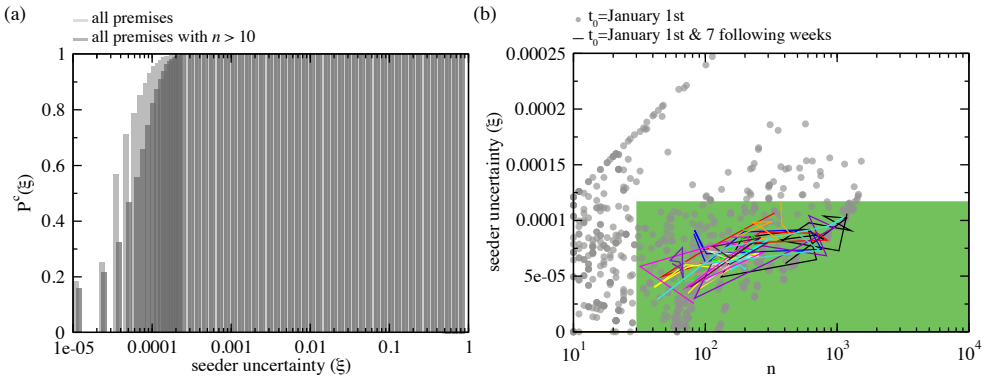

Figure 14: Sentinel premises identification by normalizing the seeder uncertainty with log(C). (a) Cumulative probability distribution of the uncertainty of a given premises in the identification of the seeding cluster. Slaughterhouses are discarded from the analysis, as they cannot spread the disease further to other farms. Note that the x-axis is in the log-scale and more than the 99% of the premises have a seeder uncertainty below 0.001, as induced by the new normalization factor. (b) For a set of initial conditions (x0,t0),with t0=January 1st, each infected holding is represented by a dot in the n-uncertainty phase space, with n being the number of times the farm is reached by an infection. Note that the y-axis is compressed within a small range of uncertainty values. The trajectories of the same 15 sentinels identified in Figure 7 of the main text are shown.

simulations, each spreading event is a probabilistic process which occurs with a probability of infection βdt per time interval dt. We keep for simplicity a deterministic recovery process. Moreover, while the weights of the links are irrelevant in the case of a deterministic spreading, they need to be taken into account in a stochastic modeling. In this context, two definitions of weights can be used. We denote by w_{ij}^H the number of herds of cattle and by w_{ij}^B the number of bovines displaced from i to j in the time window Δt. At most one herd is displaced each day, so that w_{ij}^H is at most equal to Δt. Each weight definition leads to a different definition of the probability of infection of a susceptible node by a neighboring infectious node, with different underlying assumptions. A first possibility is to define the rate of infection as $P(S_i + I_j \rightarrow I_i + I_j) = \beta * (w_{ij}^H)/\Delta t$: for $\Delta t = 1$, this means that an infectious node infects a neighboring node to whom it sends a herd with probability β. On the other hand, one can assume that the spreading power is proportional to the number of displaced bovines during each time window. Since the weights w_{ij}^B are broadly distributed [2], we choose to model the probability of infection by a function that saturates to 1 at large values of the weight, namely $P'(S_i + I_j \rightarrow I_i + I_j) = 1 - exp(-\beta' w_{ij}^B/\Delta t)$. We use the finest time scale $\Delta t = 1$ day for the stochastic simulations. In this case, the number of herds displaced between two nodes at each time step is either 0 or 1. In order to compare the two possible assumptions underlying the stochastic simulations, we use values of β and β' such that the probabilities P and P' are equal on average. This condition is satisfied if $\beta' = ln(1-\beta)/ < w_{ij}^B >$. We explore two values of β: high transmission rate ($\beta = 0.9$, corresponding to $\beta' = 0.63$) and intermediate transmission rate ($\beta = 0.5$, corresponding to

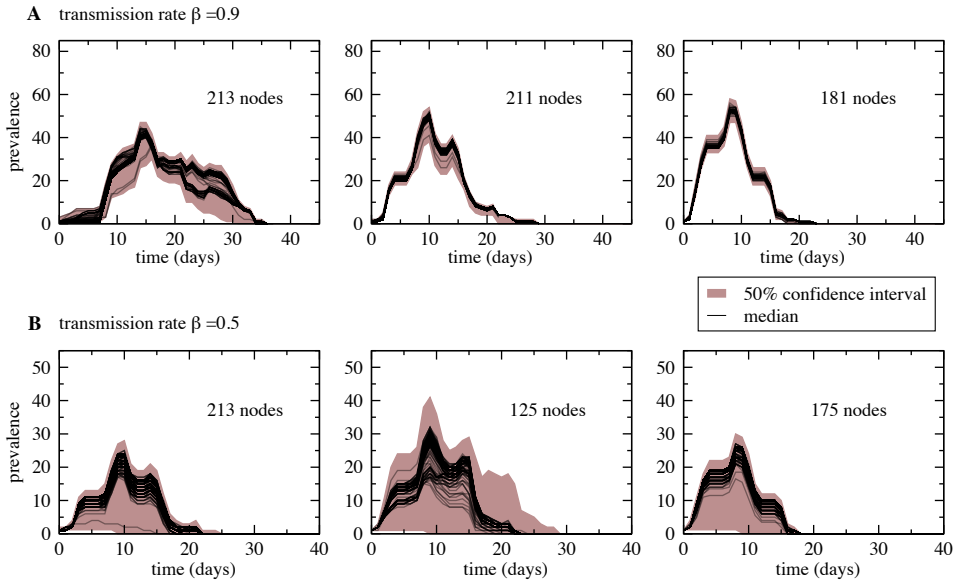

Figure 15: Prevalence curves of stochastic spreading starting from nodes belonging to various clusters. For each seed 500 stochastic simulations have been performed and the medians (black curves) and 50% confidence intervals are shown. In order to give an estimate of the fluctuations we evaluated the 50% confidence interval for each curve and we plot the maximum of the upper bounds and the minimum of the lower bounds (brown shaded area). Higher transmission rate (A) leads to lower fluctuations, while for the intermediate transmission rate (B), the 50% confidence intervals are quite broad. It is worth to stress that hundreds of stochastic curves with different seeding nodes are compared, so that fluctuations are expected; nevertheless the clustering procedure is still able to capture the similarity in the overall behavior (shape, peak time, duration) of the prevalence curves.

$\beta' = 0.19$). Once the transmission probabilities are defined and the transmission rate fixed, each stochastic simulation produces an invasion path. The union of all the paths yields a risk probability associated to each node, given by the fraction of runs in which the node has been infected. We generalize the construction of the ICSN and of the clusters defined in the main text for deterministic spreading as follows. For each seed x, we define a vector r whose element r_i is given by the fraction of runs starting in x for which i was infected, and the set ν of nodes i such that $r_i > 0$ (i.e., the set of nodes which were reached at least once by a spreading issued from x). For each pair of seeds x_1 and x_2, we build in this way the two vectors r_1 and r_2 and the sets of nodes ν_1 and ν_2, and we consider the similarity $\Omega_{12} = \sum_i (1 - |r_{1,i} - r_{2,i}|^2)/|\nu_1 \bigcup \nu_2|$. In the case of a deterministic spreading, we recover the definition of the overlap Θ as the elements of the vectors r_i are 0 or 1.

The cluster detection method described in the main text can now be applied using this measure by retaining in the ICSN only the links with similarity Ω larger than a certain threshold.

We show in Figure 15 the prevalence curves for seeds belonging to various clusters. As in the deterministic case, spreading phenomena originated in nodes of the same cluster show

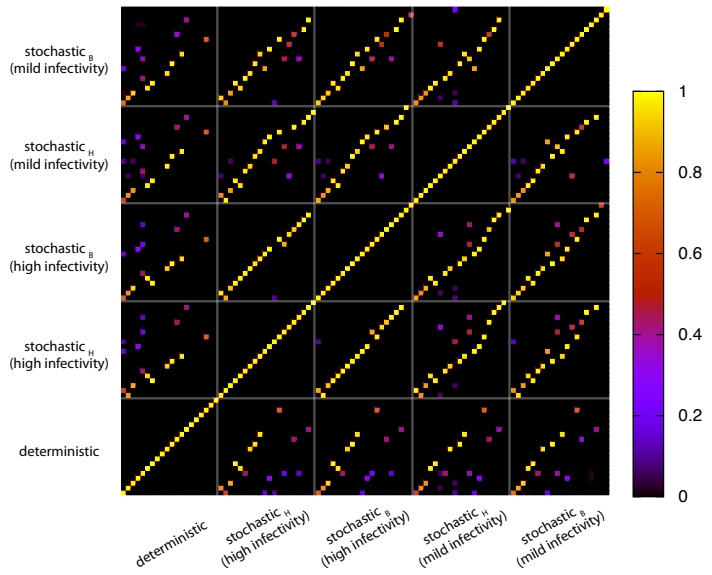

Figure 16: Jaccard indices of the 20 largest clusters obtained with different simulation procedures. We refer with stochastic$_H$ and stochastic$_B$ to simulations with transmission probabilities respectively determined by w_{ij}^H and w_{ij}^B. Each point represents the intersection of clusters obtained with different transmission probability definitions and transmission rates and is color-coded, on a violet-to-yellow color scale, according to the value of the Jaccard index computed as the number of common nodes divided by the number of nodes in the union of the clusters.

		n_s							
		30	40	50	60	70	80	90	100
ξ_s	0.45	39	34	32	27	24	21	20	16
	0.40	27	24	22	18	16	14	13	10
	0.35	15	13	12	10	8	6	5	3
	0.30	4	4	4	2	2	1	1	0

Table 2: Stochastic simulations. Number of sentinels for different thresholds in the $n - \xi$ space.

very similar temporal patterns. The cluster detection procedure leads thus to an efficient grouping of the potential seeds of an epidemics not only for deterministic spreading but also for more realistic stochastic simulations.

We also evaluate and show in Figure 16 the overlap between the clusters obtained with the various types of stochastic simulations and the deterministic ones. Strikingly, many clusters are very stable when the type of simulation and the infectivity parameter are changed.

Once the nodes are grouped in different clusters, the sentinel identification is a rather straightforward procedure. Using the same definition of seeder uncertainty described in the main paper, it is thus possible to identify sentinel nodes starting from stochastic disease spreading simulations. It is worth to notice that for each seeding node we simulate 100 stochastic runs, so that the number of times n_k that a node k has been reached by the disease is potentially much larger than in the deterministic case. In order to make comparable the stochastic and the deterministic scenarios we rescale n in Figure 17 and table 2 of a factor 100. In Figure 17 we show the results obtained from stochastic simulations, similarly to Figure 7-8 of the main text. Surprisingly, the stochastic simulations lead to a lower uncertainty than the deterministic case. This unintuitive behavior can be linked to the fact that including some heterogeneities due to the links' weight the less probable invasion paths contribute very little in the seeder uncertainty evaluation, naturally reducing the noise of the measure. Since the choice of ξ_s and n_s are arbitrary, we report in table 2 the number of sentinels identified with different values of these parameters.

References

[1] Vernon M C, Keeling M J (2009) Representing the UKs cattle herd as static and dynamic networks. Proc R Soc B 276: 469-476

[2] Bajardi P, Barrat A, Natale F, Savini L, Colizza V (2011) Dynamical patterns of cattle trade movements. PLoS ONE 6(5): e19869.

[3] Rand W M (1971) Objective criteria for the evaluation of clustering methods. JASA 66 (336): 846850.

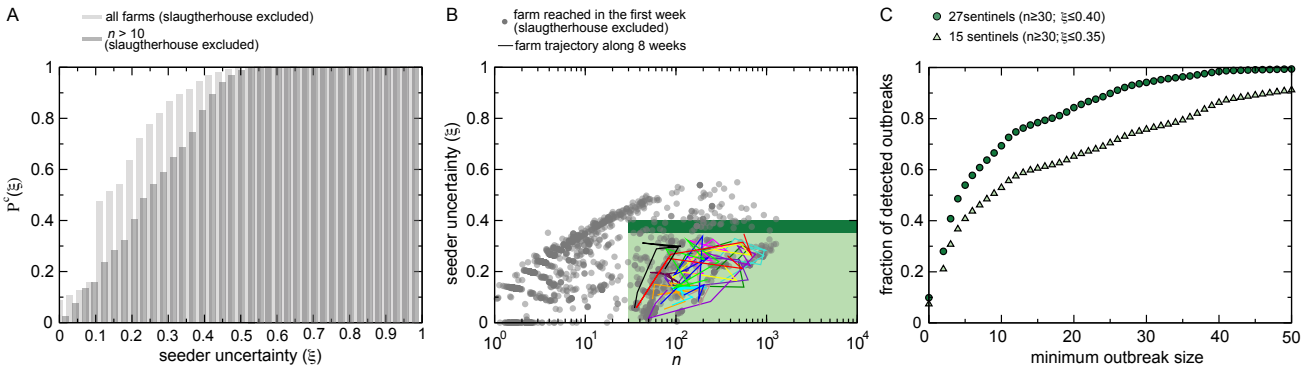

Figure 17: Identification of the initial conditions, and corresponding uncertainty, for the stochastic simulations. (A) Cumulative probability distribution of the uncertainty ξ in the identification of the seeding cluster, once a given node of the network has been detected as infected. Slaughterhouses are discarded from the analysis, as they cannot spread the disease further to other farms (they are the end points of the livestock movements and gather bovines from different sources). (B) For a set of initial conditions, each infected farm is represented by a dot in the $n - \xi$ phase space, with n being the number of times the farm is reached by an infection, and ξ the uncertainty in the identification of the corresponding seeding cluster. Eight consecutive weeks starting from January 1^{st} are considered as temporal initial conditions. Sentinel nodes are defined as the farms that are often reached by epidemics (i.e., $n > n_s$) and that have a low degree of uncertainty in the identification of the seeding cluster that led to the outbreak (i.e., $\xi < \xi_s$). The plot shows the trajectories in the $n - \xi$ phase space of the 15 sentinels obtained by imposing $n_s = 30$ and $\xi_s = 0.4$. (C) Fraction of detected outbreaks as a function of the minimum outbreak size of the epidemic, where an outbreak is considered detected if at least one of the sentinels has been reached by the infection. Two sets of sentinel farms are considered, of 15 and 27 sentinels, corresponding respectively to ($n_s = 30$, $\xi_s = 0.35$) and ($n_s = 30$, $\xi_s = 0.40$).