Effects of modifying weights on epidemiologically relevant network properties

## Aims and objectives of this report

This report provides livestock movement data owners and data managers with information about how enhancing the privacy of these data affects epidemiologically relevant network properties. The aim is to facilitate finding an appropriate balance between privacy for the livestock industry, and utility for veterinary public health practitioners and modellers.

To generate this report, a movenet-format movement data tibble was provided by the user, movement weights were modified using various functions and parameters, and the effects of these modifications were analysed on a selection of weighted network measures. These analyses were carried out for each 28 days within the data (as set via the time\_unit argument). The report presents the results of these analyses in the form of figures with some interpretive guidance.

## Static network analyses

### Basic network summaries

*Add figures for average batch sizes? (post project end)*

### Global network properties

This section describes the effects of movement weight modifications on two global network properties, average shortest path length and strength assortativity, calculated for each 28 days within the data.

#### Average shortest path length

The shortest path length or geodesic distance between a pair of nodes (holdings) in the network is the smallest number of edges (movements) needed to reach one holding from the other. The most common weighted version of this algorithm interprets edge weight as the distance (or cost) associated with an edge, and compares path lengths based on their total weighted distance (Dijkstra 1959). Considering that movement batch size is a measure of the strength of a connection, rather than a distance or cost, here edges were weighted by the reciprocal of the movement batch size () (Newman 2001). The smaller the mean weighted shortest path length is, the faster you can expect a disease to transmit between holdings, due to shorter paths and/or greater movement batch sizes.

**Alternative:** Weighting of edges according to the reciprocal of movement batch size, normalised by the average batch size in the network (). *Use of this metric is suggested by Opsahl et al., as the resulting unit of distance can be easily interpreted (as “one step with the average weight in the network”) and makes distances comparable across networks with different ranges of batch sizes (Opsahl, Agneessens, and Skvoretz 2010; Opsahl 2011).*

*What from here should I move to methods and what to keep? I know these bits are mostly methods but I need some intro here too…*

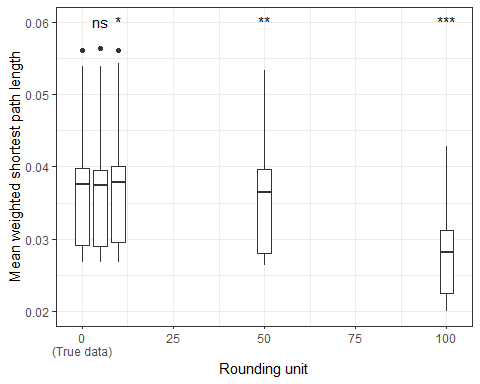
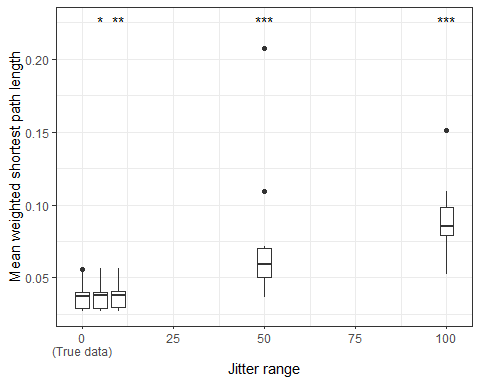


Figure 1: Effects of modifying movement batch sizes on average weighted shortest path lengths.  
Each individual box represents the distribution of average shortest path lengths in the network for the 13 time periods, calculated for either the true data (left-most box in both figures) or data with batch sizes modified according to a certain treatment. In the figure on the left, batch sizes have been jittered, with the jitter range increasing from left to right along the x-axis. In the figure on the right, batch sizes have been rounded to the nearest multiple of a specified unit, with this rounding unit increasing from left to right. In the true and rounded datasets, each data point represents the average shortest path length for a single periodic sub-network; in the jittered datasets, each data point is an average across 3 sub-networks for the same period, as obtained by 3 simulations of jitter\_weight(). Statistics performed by two-tailed paired Wilcoxon signed rank sum tests, comparing each modified dataset to the true data, \*\*\* p≤0.001, \*\* p≤0.01, \* p≤0.05, ns p>0.05.

**Option 1 (generic interpretive guidance):** Where batch size modifications result in average shortest path lengths that are higher than in the true data, this indicates that using these particular modifications would lead to overestimating the real average shortest path length and underestimating the transmission potential of the network. Conversely, where average shortest path lengths are lower than in the true data, this indicates that using these particular modifications would lead to underestimating the real average shortest path length and overestimating the transmission potential of the network. If the variability in average shortest path lengths is higher in the privacy-enhanced datasets than in the true data, this indicates that … *Add something about variability of the data?*

**Option 2 (guidance based on linear regression trend):** The analysed data in Figure 1 show a positive trend for increasing amounts of jitter, suggesting that, in general, using larger jitter ranges would result in an overestimation of the real average shortest path length and an underestimation of the transmission potential of the network. Additionally, the data show a negative trend for increasing rounding units, suggesting that, in general, using greater rounding units would result in an underestimation of the real average shortest path length and an overestimation of the transmission potential of the network. *Add something about variability of the data?*

#### Strength assortativity

The strength of a node (holding) in a network is the sum of the weights of its edges. Here, this is taken to mean the sum of all movement batch sizes. Strength assortativity is a measure of the tendency of holdings with similar strength to be connected to each other. Here, assortativity is considered in a directed manner: the tendency of holdings with a particular total number of outgoing animals (out-strength) to be connected to holdings with a similar number of incoming animals (in-strength).

*Jess/Sibylle/Gianluigi: any particular note to add here on relevance to disease spread?*

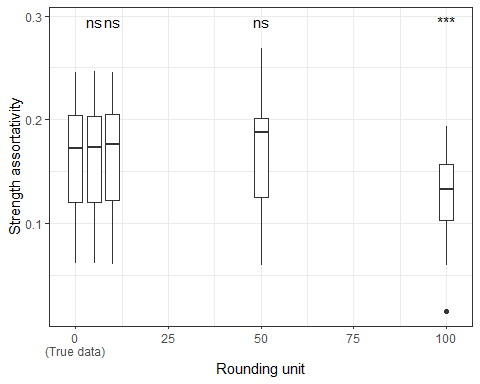
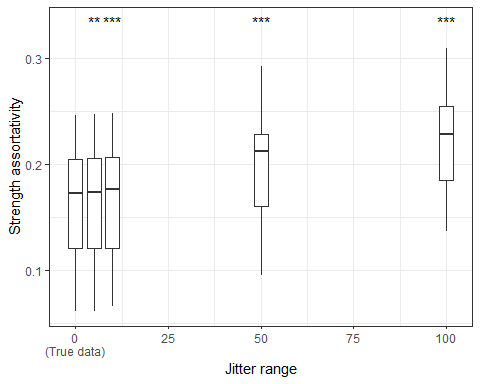


Figure 2: Effects of modifying movement batch sizes on strength assortativity.  
Each individual box represents the distribution of strength assortativities in the network for the 13 time periods, calculated for either the true data (left-most box in both figures) or data with batch sizes modified according to a certain treatment. In the figure on the left, batch sizes have been jittered, with the jitter range increasing from left to right along the x-axis. In the figure on the right, batch sizes have been rounded to the nearest multiple of a specified unit, with this rounding unit increasing from left to right. In the true and rounded datasets, each data point represents the strength assortativity for a single periodic sub-network; in the jittered datasets, each data point is an average across 3 sub-networks for the same period, as obtained by 3 simulations of jitter\_weight(). Statistics performed by two-tailed paired Wilcoxon signed rank sum tests, comparing each modified dataset to the true data, \*\*\* p≤0.001, \*\* p≤0.01, \* p≤0.05, ns p>0.05.

**Option 1 (generic interpretive guidance):** Where batch size modifications result in strength assortativities that are higher than in the true data, this indicates that using these particular modifications would lead to overestimating the real strength assortativities in the network. Conversely, where strength assortativities are lower than in the true data, this indicates that using these particular modifications would lead to underestimating the real strength assortativities in the network. If the variability in strength assortativities is higher in the privacy-enhanced datasets than in the true data, this indicates that … *Add something about relevance to disease spread, or variability of the data?*

**Option 2 (guidance based on linear regression trend):** The analysed data in Figure 3 show a positive trend for increasing amounts of jitter, suggesting that, in general, using larger jitter ranges would result in an overestimation of the real strength assortativity. Additionally, the data show no significant trend for increasing rounding units, suggesting that, in general, using greater rounding units would result in an approximately accurate estimation of the real strength assortativity. *Add something about relevance to disease spread, or variability of the data?*

### Ranking of holdings according to local network properties

This section describes the effects of movement weight modifications on the ranking of holdings according to three local centrality measures, providing information about the importance of holdings in the network in terms of their connections with other holdings. Rankings are calculated and compared for each 28 days within the data.

#### Strength centrality

The strength of a node (holding) in a network is the sum of the weights of its edges. Here, this is taken to mean the sum of all movement batch sizes. Holdings were ranked according to the geometric mean of their total number of incoming animals (in-strength) and outgoing animals (out-strength), in decreasing order. The ranking in each privacy-enhanced dataset was then compared to the ranking in the true dataset.

*Jess/Sibylle/Gianluigi: any particular note to add here on relevance to disease spread?*

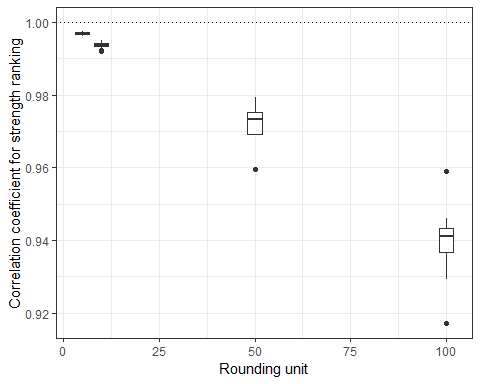
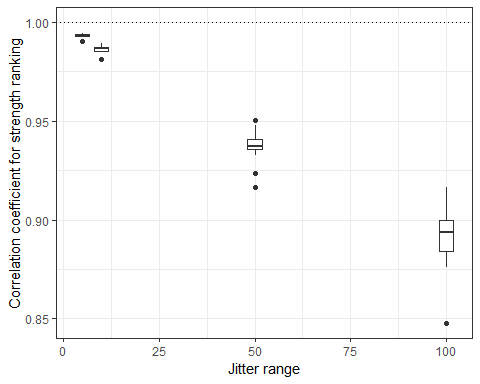


Figure 3: Effects of modifying movement batch sizes on the ranking of holdings according to strength.  
Each individual box represents the distribution of Kendall’s tau correlation coefficients, comparing the ranking according to strength of holdings in a dataset with modified batch sizes to the ranking in the true data, for the 13 periodic sub-networks created from a single modified dataset. In the figure on the left, batch sizes have been jittered, with the jitter range increasing from left to right along the x-axis. In the figure on the right, batch sizes have been rounded to the nearest multiple of a specified unit, with this rounding unit increasing from left to right. In the rounded datasets, each data point represents the correlation coefficient for a single periodic sub-network; in the jittered datasets, each data point is an average across correlation coefficients for 3 sub-networks for the same period, as obtained by 3 simulations of jitter\_weight(). The dotted line at y = 1 represents perfect agreement between the ranking in the true data and the ranking in the modified data.

*Theo: would these rank-comparison figures be improved by significance testing? if so, what kind?*

*Theo: you previously suggested using (mean) rank differences, but the mean rank difference for each periodic subnetwork is always 0 so there’s no point comparing across modifications by doing this. Any suggestions as to how this could still be useful, if the aim is to compare how rankings are affected by different amounts of jitter/rounding?*

Where batch size modifications result in correlation coefficients that are much lower than 1, this indicates that the ranking of holdings according to strength is very different than in the true data. Using these particular modifications could lead to inaccurate identification of the most (and least) important holdings in the network based on their strength.

#### Betweenness centrality

The betweenness of a node (holding) in a network is the number of shortest paths between all pairs of holdings that pass through that holding. Here, the shortest paths were determined by weighting edges according to the reciprocal of movement batch size (). Holdings were ranked according to their betweenness, in decreasing order. The ranking in each privacy-enhanced dataset was then compared to the ranking in the true dataset.

*Jess/Sibylle/Gianluigi: any particular note to add here on relevance to disease spread?*

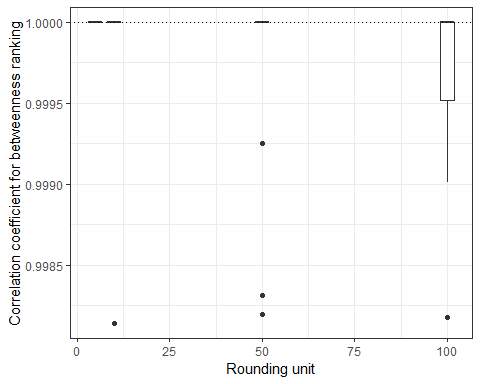
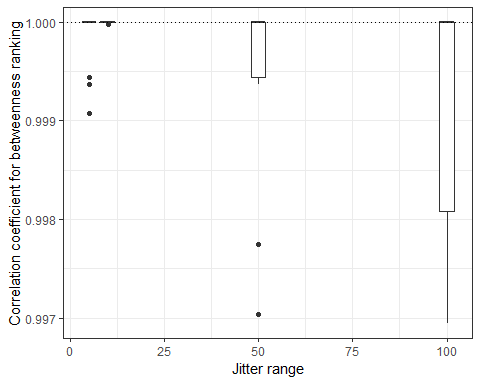


Figure 4: Effects of modifying movement batch sizes on the ranking of holdings according to betweenness.  
Each individual box represents the distribution of Kendall’s tau correlation coefficients, comparing the ranking according to betweenness of holdings in a dataset with modified batch sizes to the ranking in the true data, for the 13 periodic sub-networks created from a single modified dataset. In the figure on the left, batch sizes have been jittered, with the jitter range increasing from left to right along the x-axis. In the figure on the right, batch sizes have been rounded to the nearest multiple of a specified unit, with this rounding unit increasing from left to right. In the rounded datasets, each data point represents the correlation coefficient for a single periodic sub-network; in the jittered datasets, each data point is an average across correlation coefficients for 3 sub-networks for the same period, as obtained by 3 simulations of jitter\_weight(). The dotted line at y = 1 represents perfect agreement between the ranking in the true data and the ranking in the modified data.

Where batch size modifications result in correlation coefficients that are much lower than 1, this indicates that the ranking of holdings according to betweenness is very different than in the true data. Using these particular modifications could lead to inaccurate identification of the most (and least) important holdings in the network based on their betweenness.

#### PageRank

The PageRank of a node (holding) in a network is a measure of its importance based on the importance of the nodes that link to it (Brin and Page 1998). To determine this, edges were weighted according to their batch sizes. Holdings were ranked according to their PageRank, in decreasing order. The ranking in each privacy-enhanced dataset was then compared to the ranking in the true dataset using Kendall’s tau rank correlation coefficient.

*Jess/Sibylle/Gianluigi: any particular note to add here on relevance to disease spread?*

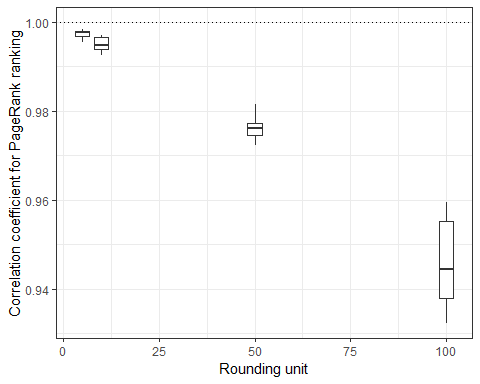
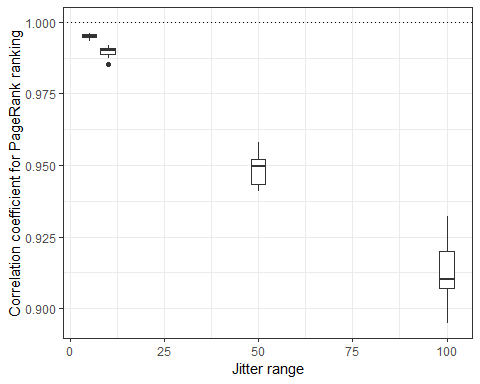


Figure 5: Effects of modifying movement batch sizes on the ranking of holdings according to PageRank.  
Each individual box represents the distribution of Kendall’s tau correlation coefficients, comparing the ranking according to PageRank of holdings in a dataset with modified batch sizes to the ranking in the true data, for the 13 periodic sub-networks created from a single modified dataset. In the figure on the left, batch sizes have been jittered, with the jitter range increasing from left to right along the x-axis. In the figure on the right, batch sizes have been rounded to the nearest multiple of a specified unit, with this rounding unit increasing from left to right. In the rounded datasets, each data point represents the correlation coefficient for a single periodic sub-network; in the jittered datasets, each data point is an average across correlation coefficients for 3 sub-networks for the same period, as obtained by 3 simulations of jitter\_weight(). The dotted line at y = 1 represents perfect agreement between the ranking in the true data and the ranking in the modified data.

Where batch size modifications result in correlation coefficients that are much lower than 1, this indicates that the ranking of holdings according to PageRank is very different than in the true data. Using these particular modifications could lead to inaccurate identification of the most (and least) important holdings in the network based on their PageRank.

## Methods

### Initial data processing

Entries with weight (batch size) 0 or representing moves from a holding to itself (“loops”) were removed, as they were considered irrelevant for disease transmission and as potentially complicating the interpretability of analyses. Additionally, all repeated moves from and to the same holdings on the same day were aggregated into a single entry per day, with weights (batch sizes) summed up.

### Modification of weights for privacy enhancement

Movement weights were modified with the following functions:

* [movenet::jitter\_weights()](https://digivet-consortium.github.io/movenet/reference/jitter_weights.html): This adds random noise between -range and range to movement weights, while ensuring that resulting weights remain greater than 0. jitter\_weights() was applied with various range arguments: 5, 10, 50, 100 (until the order of magnitude of the mean weight in the data). To take into account the effects of random sampling, 3 simulations were run with each range argument.
* [movenet::round\_weights()](https://digivet-consortium.github.io/movenet/reference/round_weights.html): This rounds movement weights to multiples of unit, and also sets unit as the minimum possible value for the resulting weights. round\_weights() was applied with various unit arguments: 5, 10, 50, 100 (until the order of magnitude of the largest weight in the data).

### Creation of network representations

Static networks were created from true and privacy-enhanced datasets, for each subsequent 28 days in the data.

During the creation of static network representations (snapshots), all repeated moves from and to the same holdings were aggregated into a single network edge per snapshot, with batch sizes summed up.

### Static network analyses (weighted)

*How much to move here from the individual result sections? I know they’re mostly methods but I need some intro there too…*

Average shortest path length, weighting 1/batch\_size, assortativity between in-strength and out-strength, weighting batch\_size

Ranking of holdings according to local network properties: Betweenness, weighting 1/batch\_size Strength and PageRank, weighting batch\_size For strength used geometric mean between in-strength and out-strength. Holdings were ranked according to their centrality, in decreasing order. The ranking in each privacy-enhanced dataset is compared to the ranking in the true dataset using Kendall’s tau rank correlation coefficient.

## Bibliography

Brin, Sergey, and Lawrence Page. 1998. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” *Computer Networks and ISDN Systems* 30 (1): 107–17. <https://doi.org/10.1016/S0169-7552(98)00110-X>.

Dijkstra, Edsger W. 1959. “A Note on Two Problems in Connexion with Graphs.” *Numerische Mathematik* 1: 269–71. <https://doi.org/10.1007/BF01386390>.

Newman, Mark E. J. 2001. “Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality.” *Physical Review E* 64 (June): 016132. <https://doi.org/10.1103/PhysRevE.64.016132>.

Opsahl, Tore. 2011. “Shortest Paths in Weighted Networks.” <https://toreopsahl.com/tnet/weighted-networks/shortest-paths/>.

Opsahl, Tore, Filip Agneessens, and John Skvoretz. 2010. “Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths.” *Social Networks* 32 (3): 245–51. <https://doi.org/10.1016/j.socnet.2010.03.006>.