Estimating the covariance structure of SIS for general infection matrices

Eric Cator

Joint work with Henk Don and Piet Van Mieghem

Radboud Universiteit Nijmegen

Contact process

The model: contact process

Consider N nodes that can either be infected or healthy. An infected node i heals (becomes healthy) with rate $\delta_i \geq 0$.

Furthermore, if i is infected, it infects a node j with rate A_{ij}; this means that if i is infected and j is healthy, j can become infected at rate $A_{ij} \geq 0$.

If $I \subset \{1, \ldots, N\}$ represents the infected nodes (I is the "state" of the process), then

$$I \to I \cup \{j\} \text{ at rate } \sum_{i \in I} A_{ij}$$

$$I \to I \setminus \{i\} \text{ at rate } \delta_i.$$

Mean Field Approximation

Mean flow

If we consider a state $X(t) \in \{0,1\}^N$ at some time t, we can calculate the expected jump in a small time period h:

$$\mathbf{E}(X_i(t+h) - X_i(t) \mid X(t)) = -X_i(t)\delta_i h + (1 - X_i(t)) \sum_{j=1}^{n} A_{ji} X_j(t) h.$$

Stability

If we have meta-stability, we should find that

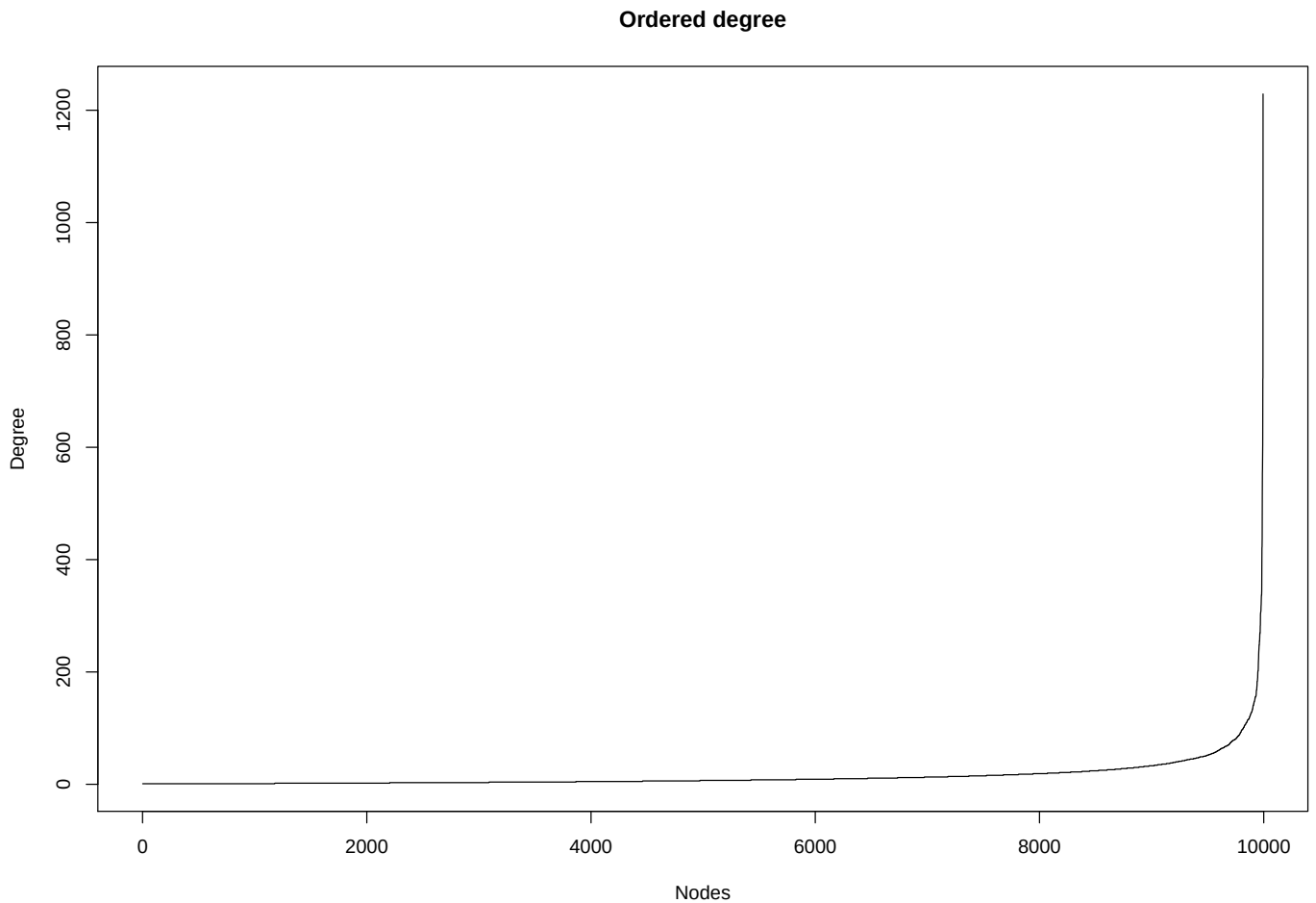
$$0 = \sum_{j=1}^{n} (A_{ji} - \delta_i)\mathbf{E}(X_i) - \sum_{j=1}^{n} A_{ji}\mathbf{E}(X_i X_j).$$

In MFA we approximate $\mathbf{E}(X_i X_j) = \mathbf{E}(X_i)\mathbf{E}(X_j)$.

Mean Field Approximation

Example Simulated network

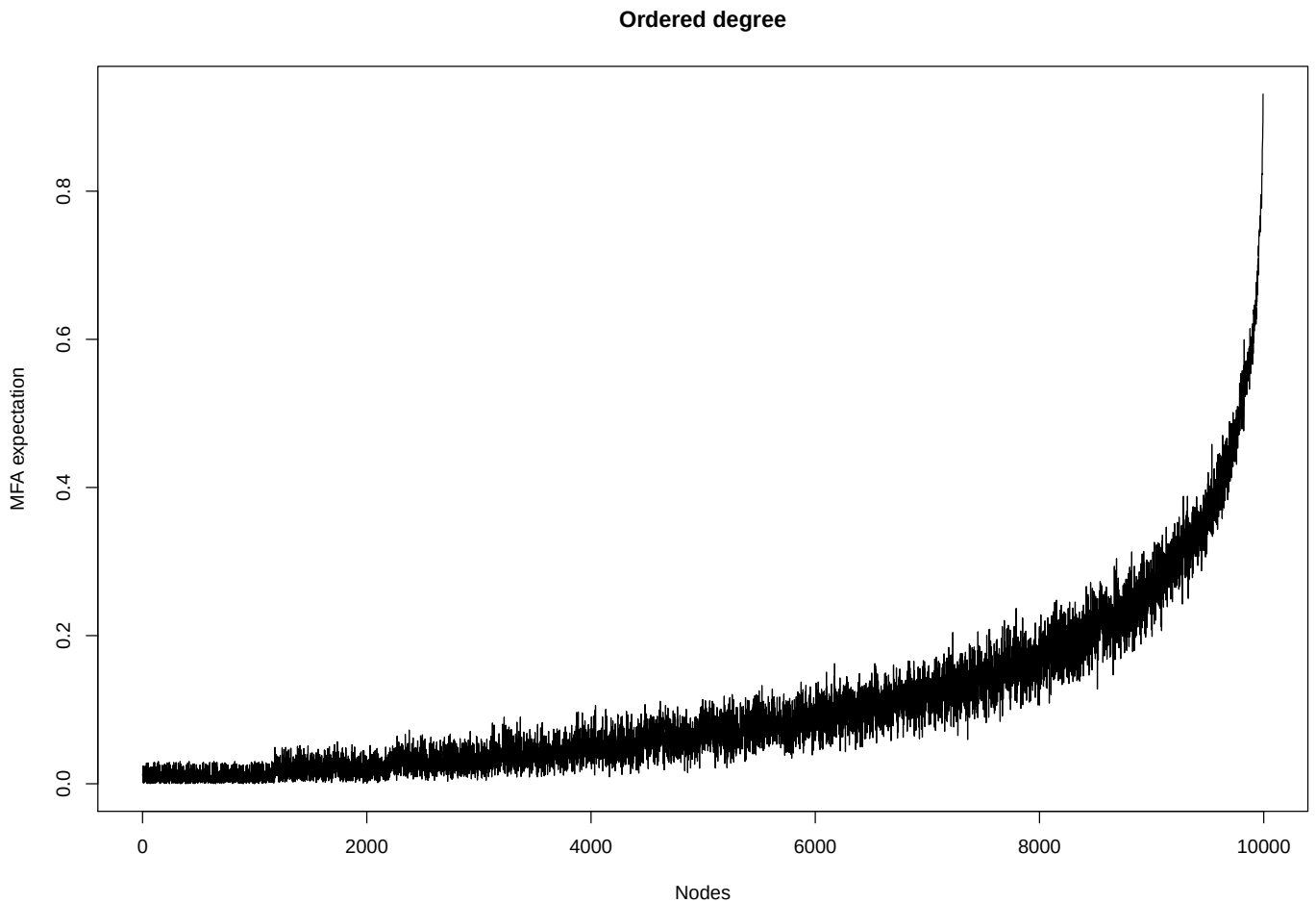
$N = 9994$ nodes, heavy tailed degree-distribution.



Mean Field Approximation

Example Simulated network

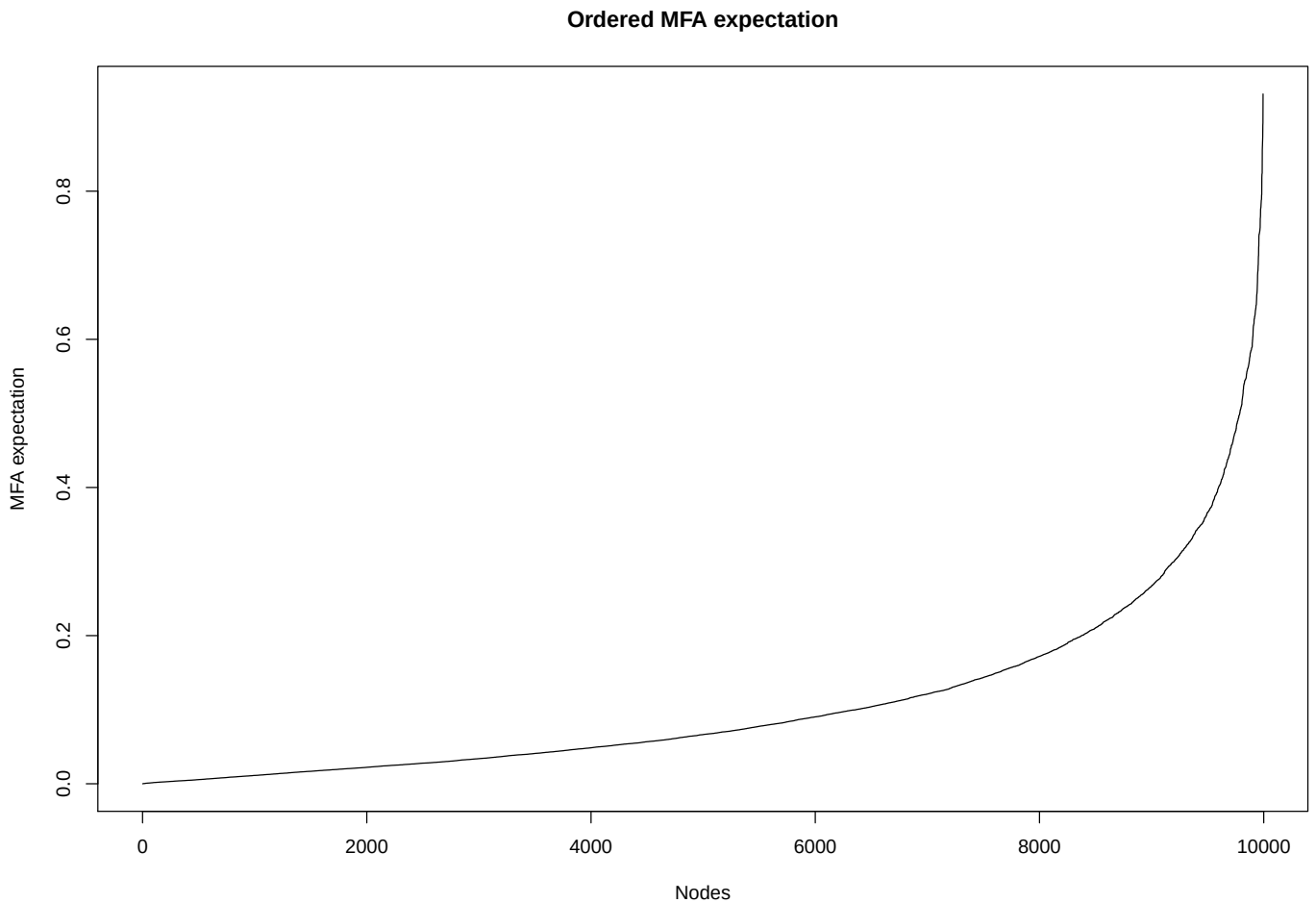
$N = 9994$ nodes, heavy tailed degree-distribution.



Mean Field Approximation

Example

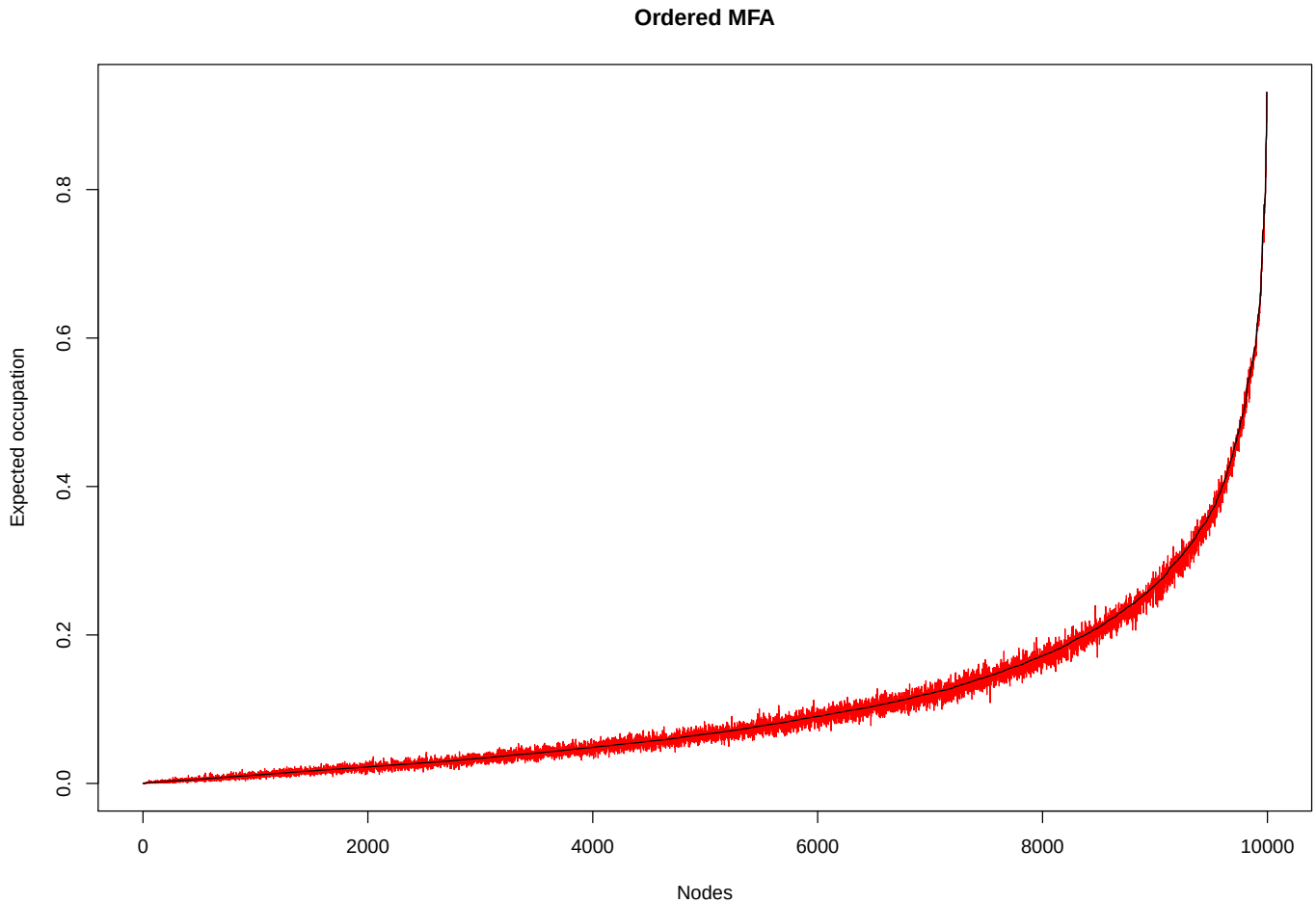
$N = 9994$ nodes, heavy tailed degree-distribution.



Mean Field Approximation

Example

Simulation of about $6.5 \cdot 10^6$ events. Average occupation given.



Mean Field Approximation

Shortcomings of MFA

- Fluctuations?
- Correlations?
- No possibility to improve approximation (at the cost of extra computations).

Idea: approximate A by structured matrix

We suggest to approximate A by writing

$$A \approx W^T H,$$

with W and H $k \times N$-dimensional non-negative matrices. This is known as Non-negative Matrix Factorization (NMF).

Non-negative Matrix Factorization

Equilibrium

When $A \approx W^T H$ and healing rates are given by Δ, apply MFA:

$$\Delta_i \mathbf{E}(X)_i = (W\mathbf{E}(X))^T H_i - (W\mathbf{E}(X))^T H_i \mathbf{E}(X)_i.$$

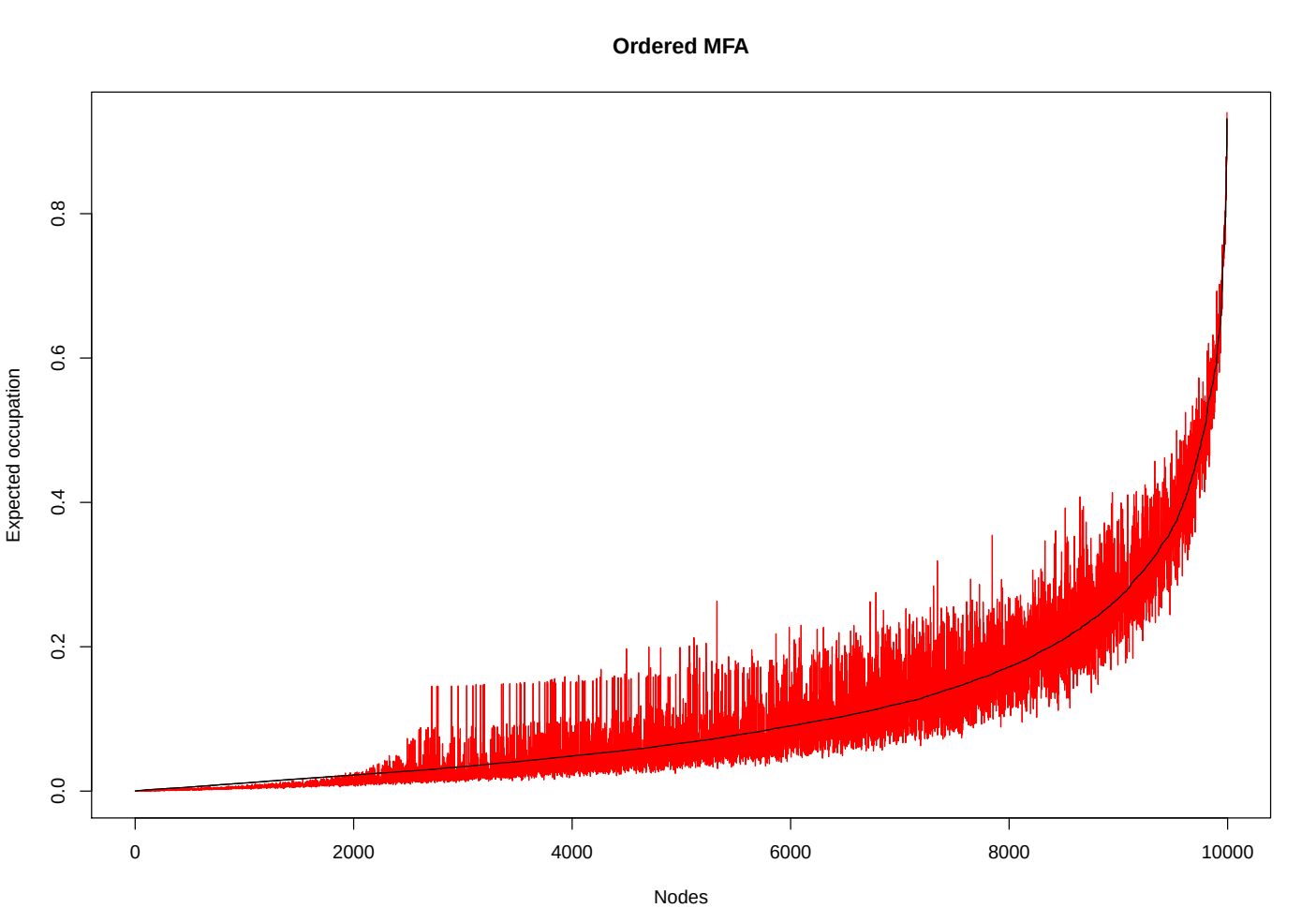
Define $\tilde{C} = W\mathbf{E}(X) \in \mathbb{R}^k$. We get

$$\mathbf{E}(X)_i = \frac{\tilde{C}^T H_i}{\Delta_i + \tilde{C}^T H_i} \text{ and } \tilde{C} = \sum_{i=1}^{N} \frac{(\tilde{C}^T H_i) W_i}{\Delta_i + \tilde{C}^T H_i}.$$

How does this compare to original MFA?

Non-negative Matrix Factorization

Compare MFA for simulated network



Non-negative Matrix Factorization

Feature space

When $A = W^T H$, and the healing rates are given by the vector Δ, each node has a $2k + 1$ dimensional feature:

$$Z_i = (W_i, H_i, \Delta_i).$$

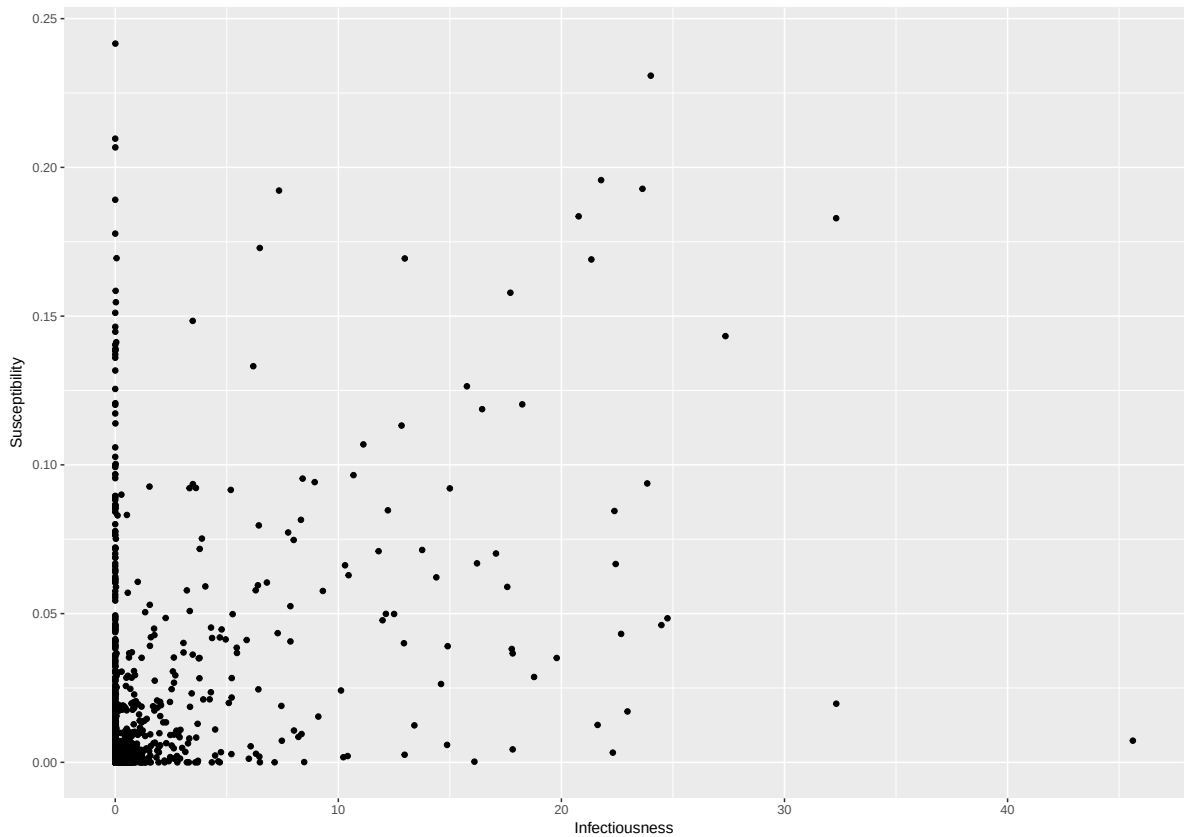
We say that W_i is the *infectiousness*, H_i is the *susceptibility* and Δ_i the healing rate. Node i infects node j with rate $W_i^T H_j$.

Now we could define clusters on the basis of these features: two nodes are almost indistinguishable if they have almost the same features.

Factorized infection matrix

Indistinguishability

When $A = W^T H$, a set of nodes $G \subset \{1, \ldots, N\}$ is indistinguishable, precisely when $\forall i, j \in G : \; Z_i = Z_j$.



The process $(N_1(t), \ldots, N_r(t))$

Clustering the nodes

We form r clusters of nodes that have Z-values close together: $B_1, \ldots, B_r$ is a partition of $\{1, \ldots, N\}$. Define

$$N_j = N_j(t) = \#\{i \in B_j \mid X_i(t) = 1\}.$$

Set $m_j = \#B_j$. Define Y_j as the mean of the Z-values in cluster j:

$$Y_j = \frac{1}{m_j} \sum_{i \in B_j} Z_i.$$

In reasonable approximation, the vector $(N_1, \ldots, N_r)$ is now a Markov process, with transition rates determined by $Y_1, \ldots, Y_r$.

The process $(N_1(t), \ldots, N_r(t))$

Transition rates

Write $Y_j = (Y_{w,j}, Y_{h,j}, Y_{\delta,j})$. The rates are given by:

$$N_j \to N_j + 1 \text{ at rate } (m_j - N_j) \sum_{k=1}^{r} N_k Y_{w,k}^T Y_{h,j}$$

$$N_j \to N_j - 1 \text{ at rate } Y_{\delta,j} N_j.$$

Equilibrium: $Y_{\delta,j} N_j = (m_j - N_j) \left(\sum_{k=1}^{r} N_k Y_{w,k}^T \right) Y_{h,j}$.

Define $C = \sum_{k=1}^{r} N_k Y_{w,k} \in \mathbb{R}^k$. We get

$$N_j = \frac{C^T Y_{h,j}}{Y_{\delta,j} + C^T Y_{h,j}} \cdot m_j \text{ and } C = \sum_{j=1}^{r} \frac{m_j (C^T Y_{h,j}) Y_{w,j}}{Y_{\delta,j} + C^T Y_{h,j}}.$$

The process $(N_1(t), \ldots, N_r(t))$

Equilibrium

$$N_j = \frac{C^T Y_{h,j}}{Y_{\delta,j} + C^T Y_{h,j}} \cdot m_j \text{ and } C = \sum_{j=1}^{r} \frac{m_j (C^T Y_{h,j}) Y_{w,j}}{Y_{\delta,j} + C^T Y_{h,j}}.$$

Compare this to MFA when $A = W^T H$:

$$\mathbf{E}(X)_i = \frac{\tilde{C}^T H_i}{\Delta_i + \tilde{C}^T H_i} \text{ and } \tilde{C} = \sum_{i=1}^{N} \frac{(\tilde{C}^T H_i) W_i}{\Delta_i + \tilde{C}^T H_i}.$$

This shows that with properly chosen clusters, $N_j^\infty \approx \sum_{i \in B_j} \mathbf{E}(X)_i$.

The process $(N_1, \ldots, N_r)$

Fluctuations

$$N_j \to N_j + 1 \text{ at rate } (m_j - N_j) \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} N_k$$

$$N_j \to N_j - 1 \text{ at rate } Y_{\delta,j} N_j.$$

Define the fluctuations away from equilibrium:

$$D_j = N_j - N_j^\infty.$$

Infections: $I_j \sim \text{Pois}\left(h(m_j - N_j^\infty - D_j) \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} (N_k^\infty + D_k) \right).$

Healings: $H_j \sim \text{Pois}\left(h Y_{\delta,j} (N_j^\infty + D_j) \right).$

The process $(N_1, \ldots, N_r)$

Normal approximation

$\{I_j\}$ and $\{H_j\}$ are all independent. When clusters are large enough, Poisson variables are well approximated by normal random variables. Define $\Delta D_j = I_j - H_j$. Up to main order, we get:

$$\begin{aligned}\mathbf{E}(\Delta D_j) \approx h(m_j - N_j^\infty) \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} D_k - hD_j \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} N_k^\infty \\ - hY_{\delta,j} D_j \\ \mathrm{Var}(\Delta D_j) \approx 2hY_{\delta,j} N_j^\infty.\end{aligned}$$

Define $B(t)$ to be an r-dimensional Brownian motion. We get

$$dD(t) = KD(t)dt + \mathrm{diag}(\sqrt{2\mathrm{diag}(N^\infty) Y_\delta})dB(t),$$

$$K = \mathrm{diag}(m - N^\infty) Y_h^T Y_w - \mathrm{diag}(Y_h^T Y_w N^\infty + Y_\delta).$$

The process $(N_1, \ldots, N_r)$

Explicit solution

Define $\Sigma_0 = \mathrm{diag}(2\mathrm{diag}(N^\infty)Y_\delta)$. Then

$$D(t) = e^{Kt}D(0) + \int_0^t e^{K(t-s)}\Sigma_0^{1/2}dB(s).$$

Since K only has negative eigenvalues when MFA solution exists, there exists a stationary solution. Covariance matrix Σ is given by:

$$\Sigma = \int_0^\infty e^{Ks}\Sigma_0 e^{K^T s}\, ds.$$

This also solves the matrix equation

$$K\Sigma + \Sigma K^T = -\Sigma_0.$$

The process $(N_1, \ldots, N_r)$

Explicit solution

$$K\Sigma + \Sigma K^T = -\Sigma_0.$$

This matrix equation has an explicit solution if K is diagonalizable:

$$K = V\Lambda V^{-1}.$$

We get $\Lambda V^{-1}\Sigma V^{-T} + V^{-1}\Sigma V^{-T}\Lambda = -V^{-1}\Sigma_0 V^{-T}$, so

$$(\Lambda_{ii} + \Lambda_{jj})(V^{-1}\Sigma V^{-T})_{ij} = -(V^{-1}\Sigma_0 V^{-T})_{ij}.$$

Define J to be the all ones matrix, and we see that

$$\Sigma = -V\frac{V^{-1}\Sigma_0 V^{-T}}{\Lambda J + J\Lambda}V^T.$$

The process $(N_1, \ldots, N_r)$

Conlusion

We found that the vector-values proces $N(t)$ has an approximating stationary distribution, given by

$$N(t) \sim \mathcal{N}_r(N^\infty, \Sigma).$$

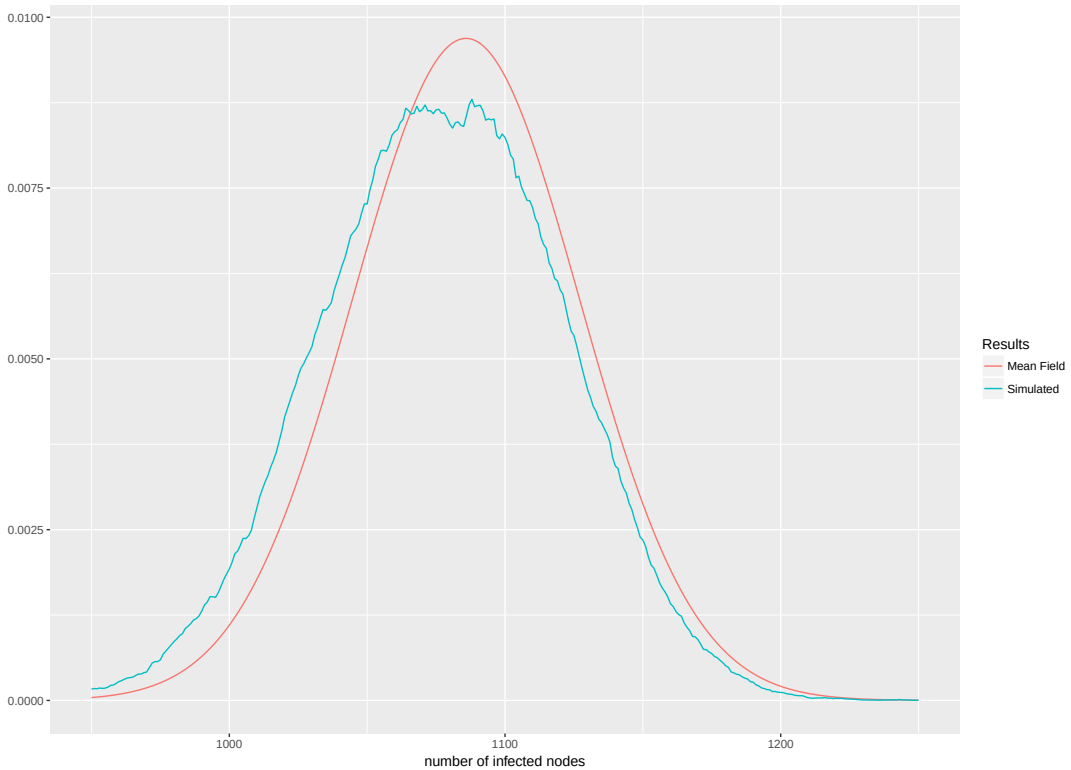
We have also linked the time-evolution to the eigenvalues of the matrix K. We used a string of approximations:

- First approximate A by $W^T H$.
- Choose r clusters, and use average infectiousness, susceptibilty and healing rate for all nodes within a cluster. This way, $N(t) = (N_1(t), \ldots, N_r(t))$ becomes a Markov process.
- Approximate $N(t)$ by a non-linear SDE.
- Only consider highest order terms, and solve linear SDE.

Example: simulated network

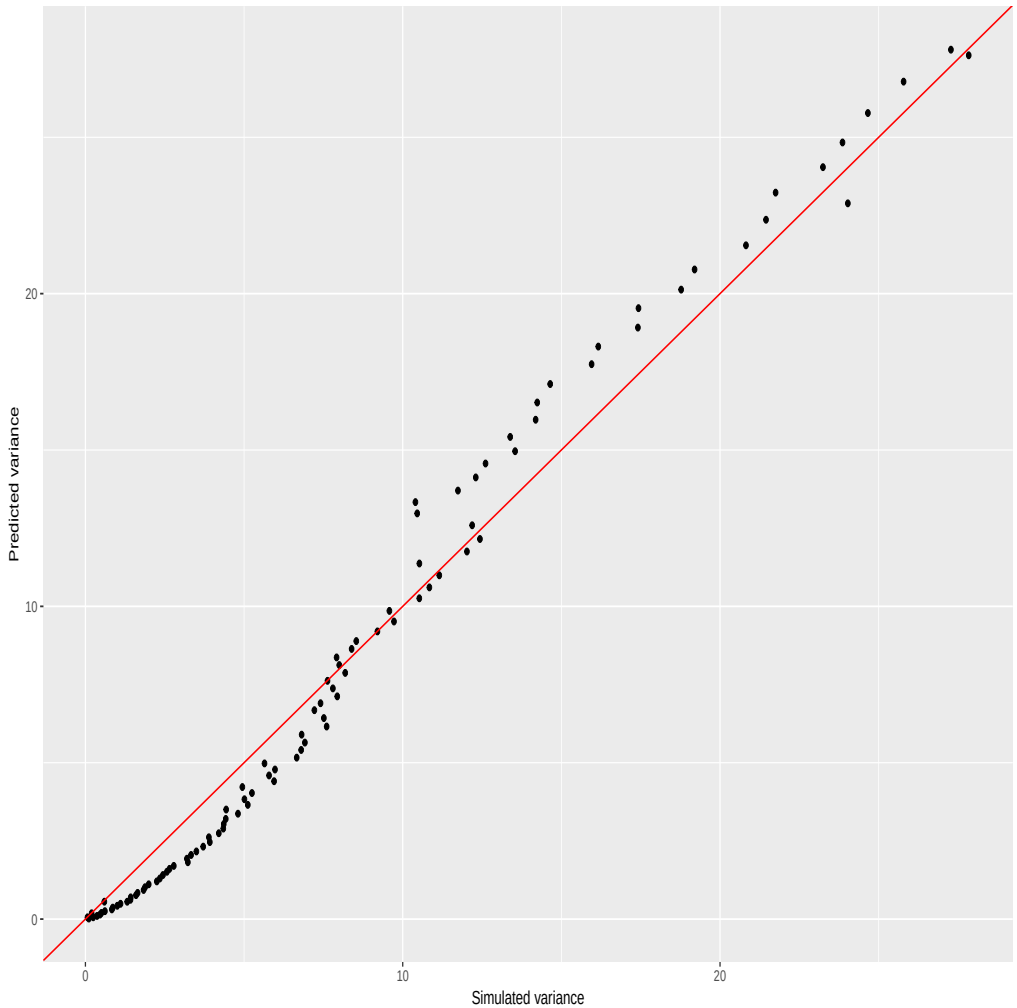
Total number infected

Total number of infected: $\mathcal{N}(\sum_{j=1}^{r} N_j^{\infty}, \sum_{j=1}^{r} \sum_{j'=1}^{r} \Sigma_{jj'})$.



Example: simulated network

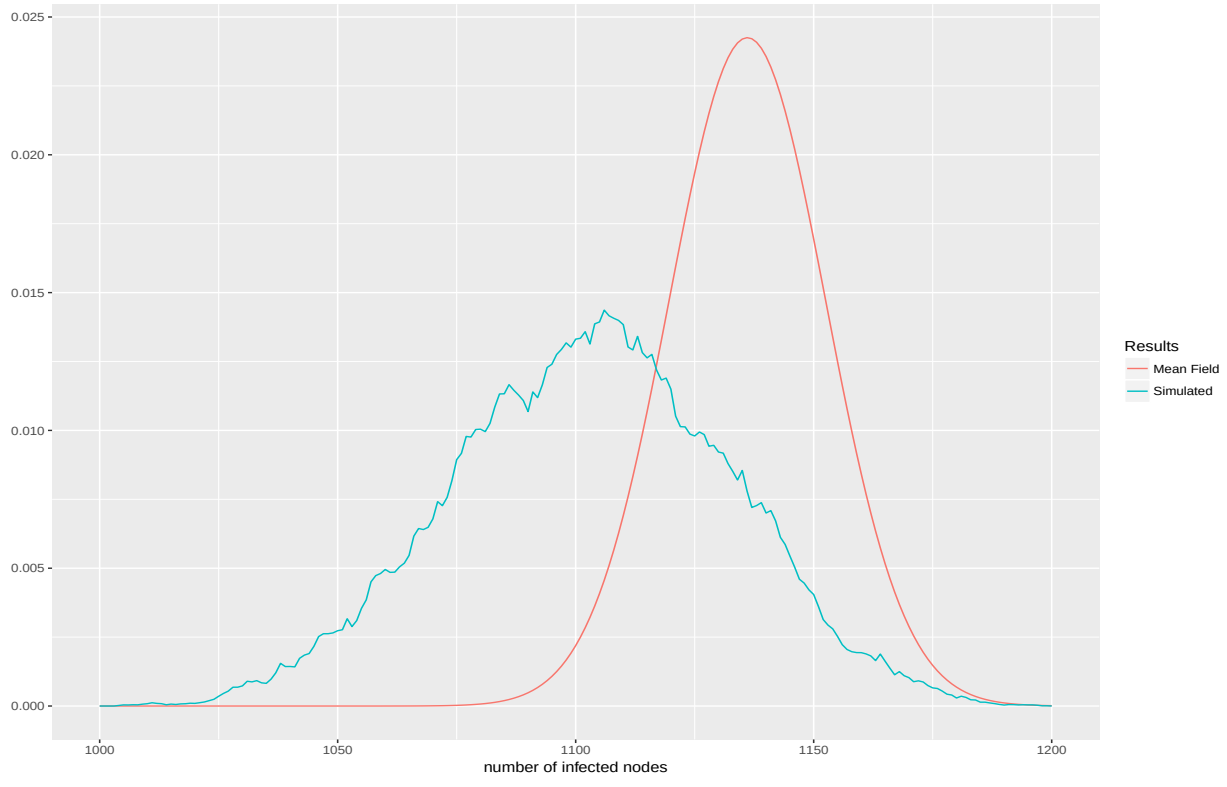
Simulated and predicted variance of the clusters.



Example: airport network

Airport network

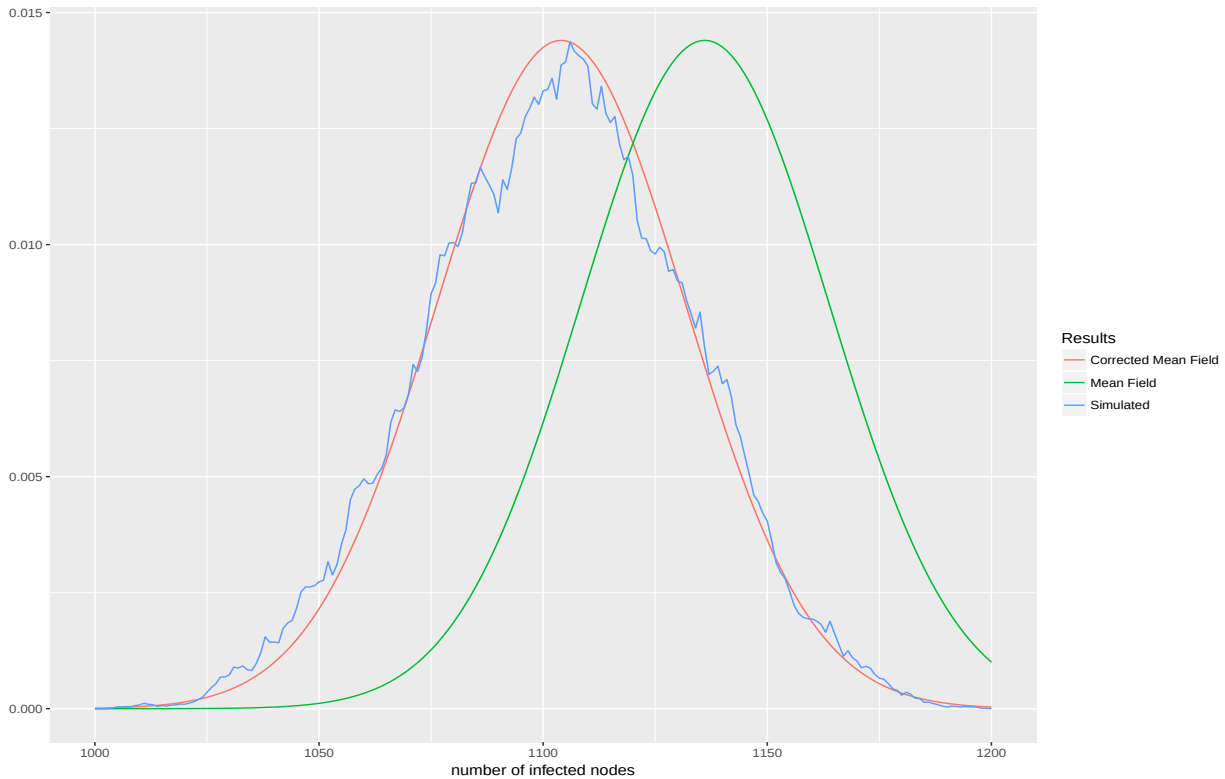
Matrix A is asymmetric and infections rates vary; 3425 nodes. We try 1 dimensional factorisation, with 3425 clusters.



Example: airport network

Airport network

Now no factorisation ($W = I, H = A$), with 3425 clusters. We also correct MFA.



Correct the mean of MFA

Use covariance prediction

Rate equation for expectations:

$$\frac{d\mathbf{E}(N_j)}{dt} = \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} \mathbf{E}(N_k(m_j - N_j)) - Y_{\delta,j}\mathbf{E}(N_j)$$

$$= \sum_{k=1}^{r} m_j Y_{h,j}^T Y_{w,k} \mathbf{E}(N_k) - Y_{\delta,j}\mathbf{E}(N_j) - \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} \mathbf{E}(N_k N_j)$$

$$= \sum_{k=1}^{r} m_j Y_{h,j}^T Y_{w,k} \mathbf{E}(N_k) - Y_{\delta,j}\mathbf{E}(N_j)$$

$$- \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} \mathbf{E}(N_k)\mathbf{E}(N_j) - \sum_{k=1}^{r} Y_{h,j}^T Y_{w,k} \mathrm{Cov}(N_k, N_j)$$

Correct the mean of MFA

Use covariance prediction

Use the estimate for the covariance ($\mathrm{Cov}(N_k, N_j) = \Sigma_{kj}$) and put derivative to 0:

$$\mathrm{diag}(Y_h^T Y_w \Sigma) \approx \mathrm{diag}(m) Y_h^T Y_w \mathbf{E}(N) - \mathrm{diag}(Y_\delta)\mathbf{E}(N) - \mathrm{diag}(\mathbf{E}(N)) Y_h^T Y_w \mathbf{E}(N).$$

This gives a corrected estimate for the expected infection of each cluster. This new value may be (slightly) negative, in which case we put it to 0.

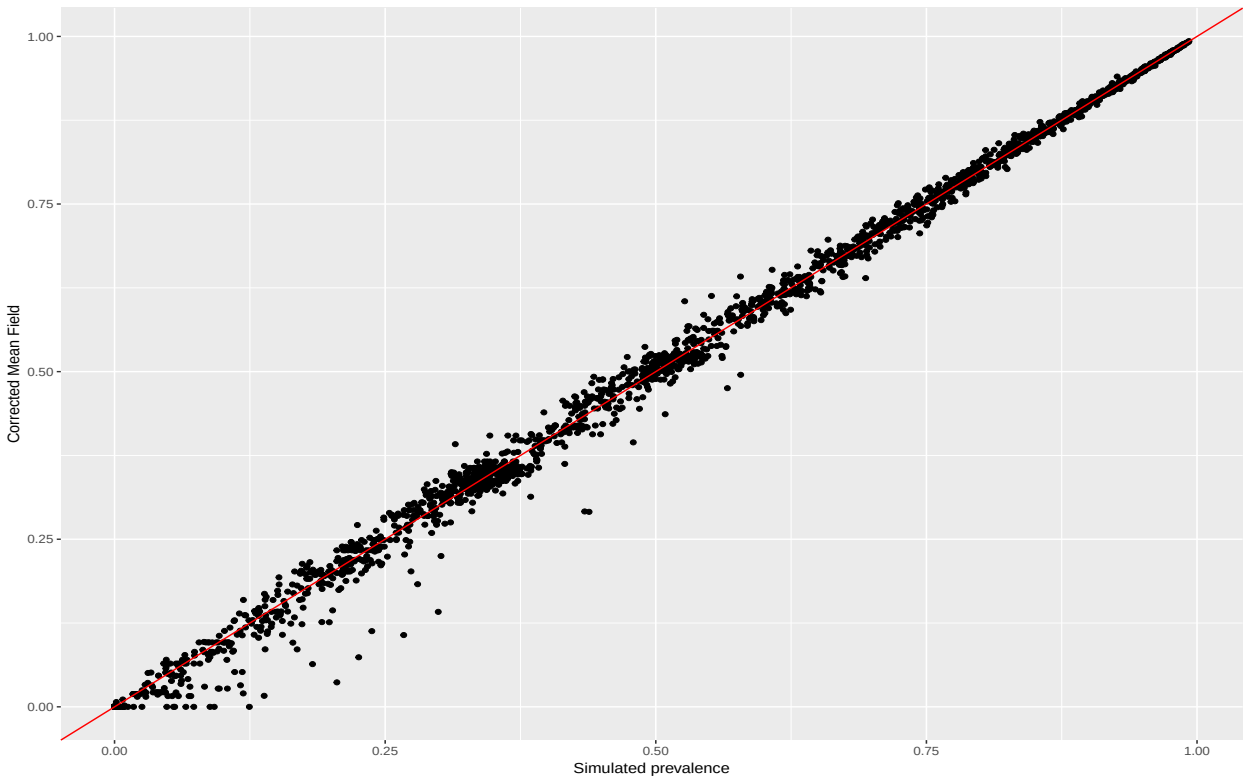
Not always effective

We found that this correction is small when using low dimensions or few clusters.

Example: airport network

Airport network

Corrected MFA for each node.



Work in progress

Limit theorems?

- If we know that the features of all the nodes have a reasonable distribution, can we prove that the contact process converges to a normal process, as the number of nodes increases?
- As we increase the dimension of the feature space, will the approximation to the true contact process get better? Under what conditions?

Non-negative Matrix Factorization

- What should we optimise when trying to determine W and H? For example, the diagonal is irrelevant for us.
- If A and $W^T H$ are close, what does this mean for the contact process? Can we control the difference in meta-stable distribution?