EPIDEMIC STUDY ON SOCIAL NETWORK VIA BAYESIAN INFERENCE

Sibghat Ullah

1772576 Sapienza University of Rome

Table of Content

Abstract

The project is related to the study of Stochastic Epidemiology spreading across a social network. The social Network here can be thought of a realization of Exponential Random Graphs. The epidemic model here has exponentially distributed transmission times. Specifically, a Node in the Network (i.e Node is a Person in this case) can be in four different periods which are

- A Person or Node is in Susceptible state if that Person is not yet infected, but in future that could happen.
- A Person or Node is in Exposed state if he/she has caught the virus, but can't transfer that virus to other people.
- A Person or Node is in Infected state if he/she can infect other people.
- After a specific time, an Infected person or Node is categorized as to be in Removal state which means he/she has no further role in spreading of the disease. It could mean that Person has been either cured or found to be dead.

I would be describing it as an SEIR epidemic model which refers to the concept.

Susceptible → Exposed → Infectious → Removed

I'll be studying this Model in pure Bayesian Statistics using Monte Carlo Markov Chain. The most important Question in this inference would be to study the properties of the spread disease and the Network on which it happened. Furthermore, the important aspect of this study is the effect of parameters that control the epidemic. While the Network parameters are also important, their roles have been discussed slightly less due to the complex analytical expressions and the lack of Software that can be used to study that in a wonderful way.

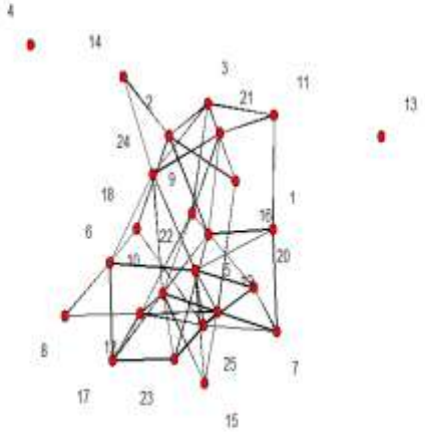
Epidemic and Network Model

The epidemic starts once a single Node or individual becomes infected. And this infection can be transferred to the other Nodes. The waiting time for spreading this disease/virus across a specific edge is exponentially distributed with mean $1/\beta$. This simply means that the time a Person spends in Susceptible state is exponentially distributed with mean given above.

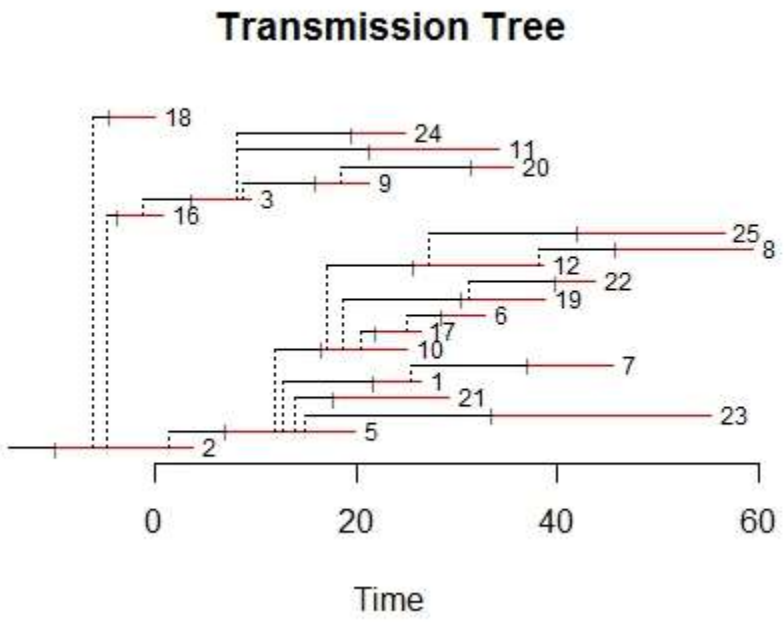
The time for a specific Node to be in Exposed and Infectious state is modeled as gamma distribution with parameters (θ_E, k_E) and (θ_I, k_I) respectively. I could use here the exponential distribution but studies have shown that Gamma is ideal for Bayesian Inference in this setting.

Finally, at the end of disease spreading we have ourselves a subgraph which shows all the infected Nodes during the epidemic and the edges across which it happened. This graph is called the Transmission Tree by Researchers and Scientists and is usually denoted by ₱. Further explanation of this new graph shows that this can also be categorized as a Tree since it doesn't have any cycles. In the figure below, a very general Exponential Random Graph is shown. The red points indicate the Node whereas the undirected arrows indicate the edges between those Nodes. Note that this graph has only 25 Nodes and is undirected whereas the actual Data that I

would study (Hagelloch Data) would have 187 Nodes and would be considered as a directed graph.



(an undirected exponential random graph with 25 Nodes)



(an epidemic Transmission Tree for above Graph)

As for the Network structure, we would have a population of finite size N in which the social Network would be modeled as an Exponential random graph denoted by G. The vertex set would be

$$V=\{1,2,3,\ldots\ldots\ldots\ldots,N\}$$

Also, for any distinct vertices i and j the edge {i,j} would exist with a probability p, and it would be independent of all the other edges in the network. A reminder here is again that these are the most General Notations and symbols, and that our actual Hagelloch Data can only be considered a realization of this Network and Epidemic structure.

Bayesian Inference

In this section, I'll try to be more precise, and specific with the Notations and the Statistical model. Please note that this is a complete Bayesian Inference So I'd have to choose the Statistical Model, the Prior and the Monte Carlo Markov Chain algorithm which would guarantee in a sensible posterior distribution.

Data

The fundamental data used in this case would be the times at which each individual Node enters in Infectious and Removal states. The data at which a Node enters in exposed state is not given, and I would have to estimate that as part of the Bayesian process.

The exposure, infectious and removal times for a specific Node i are denoted by $E_{i,}$ $I_{i,}$ and $R_{i.}$ The sets $E=\{E_1, E_2, \ldots\ldots, E_N\}$ $I=\{I_1, I_2, \ldots\ldots, I_N\}$ and $R=\{R_1, R_2, \ldots\ldots, R_N\}$ indicate the exposure, infectious and removal times for all nodes in the Network. A more common Notation for (E,I,R) in the articles is T. Also, let k be the initially exposed node and in this case, I would also try to estimate it since it's not given as part of the input. Also, let E_{-k} be the set of exposure times except for the individual k. Finally, all the nodes that were infected during the epidemic are denoted as 1,.,m.

Likelihood/Statistical Model

Remember that I'd be mentioning the data with a notation of T where T stands for set of times. The parameter of interests would be

$$\beta,\theta_E,k_{E,}\theta_I,k_{I,}G,₱$$

and T would be

$$I, R$$

So, I can write the likelihood as

$$L(I,R|\ \beta,\theta_E,k_{E,}\theta_I,k_{I,}G,₱)$$

For simplicity, I'd break down the Likelihood as L1, L2, L3, and L4 whereas L1 means the contribution of Likelihood because of ₽. L2 is the contribution of the Likelihood of those edges over which the epidemic did not pass, and it can be denoted as G \ ₽. Furthermore, L3 and L4 are the contribution of the likelihood which occurred due to the Exposure to infectious and infectious to removal processes.

A brief description of the parameters is given below

- θ_E, k_E correspond to the process which transforms a specific edge from exposed state to infectious state and together they are parameters of a gamma distribution.
- θ_i, k_i correspond to the process which transforms a specific edge from infectious state to removal state and together they are parameters of a gamma distribution.
- B correspond to the process which transforms a specific edge from susceptible state to exposed state. It is the parameter of exponential distribution.
- G correspond to the structure of Network topology which I initially have. It characterizes how many and which Nodes are connected to each specific node initially. Note that just having an edge between two nodes doesn't necessarily mean that the transition (virus/epidemic) between them has occurred. It could be the case that it would never occur in some cases.
- ₽ correspond to the subgraph of the original graph G in which there has been successful transitions. So, it can also be considered as Transmission tree whose root node is k.

One change of the Notation that I'll have with the practical implementation/Code is that I would not use the symbols such as G and ₽ to measure the effect of Network. Rather I'd have Parameters Classroom 1, Classroom 2, and Household distance which I would use through a linear Model. More specifically,

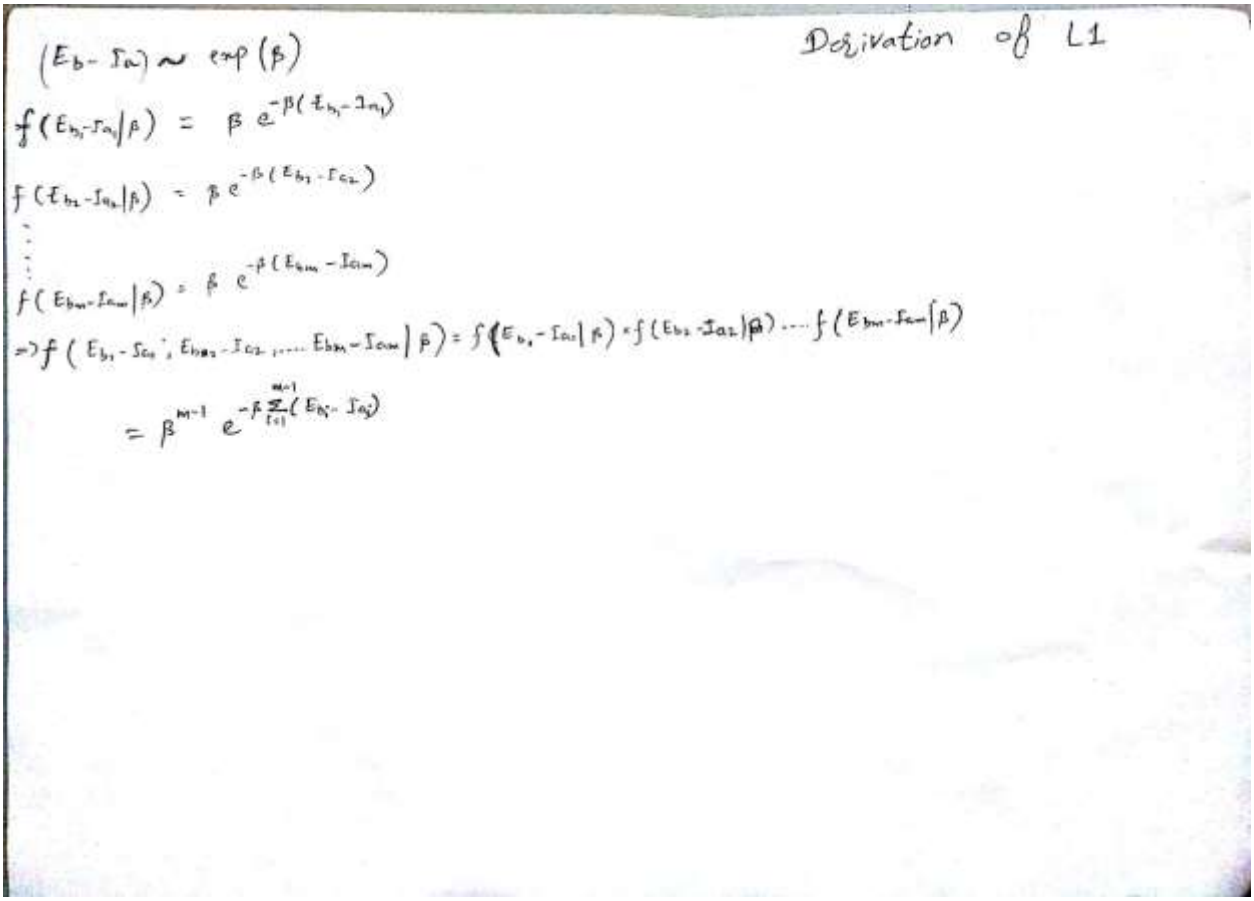
Linear model (~ Classroom1, Classroom2, House distance) alongside an intercept would be What it means to measure the effects of G and ₽ in the Hagelloch Data, I'd be using the parameters which if combined in a linear model are the logical equivalent of a General random graph G and the subgraph ₽. These parameters are described as:

- Classroom1 is a logical value, which shows if the two Nodes/Persons have studied in the same Class for kids. Since this could play a significant role in spreading the epidemic.
- Classroom2 is same as Classroom1, only difference is it is for the kids whose age are above six years. So perhaps two nodes who happen to be above six years old are studying in the same class.
- House Distance is the physical distance between two houses in which two Nodes reside. If the distance is very low, there is a strong chance of transition.

Now, let me come again towards L1, L2, L3 and L4 and let's find it's derivation which is quite trivial.

$$L1 = \beta^{m-1} \exp[-\beta \sum_{(a,b)\in P} E_b - I_a]$$

The derivation of L1 is given in the figure below:



And for L2, I have

$$L2 = \exp\{-\beta \sum_{(a,b)\in G\setminus P}[\{(E_b \wedge R_a) - I_a\} \vee 0]\}$$

So then, L1 and L2 can be combined written as

$$L1L2 = \beta^{m-1} \exp -\{\beta A\}$$

Where

$$A = \left[\sum_{(a,b\in \mathbb{P})}(Eb - Ia)\right] + \left[\sum_{(a,b\in G\setminus \mathbb{P})}\{(Eb \wedge Ra) - Ia\} \vee 0\right]$$

In previous articles, A has been referred to as the overall Infectious pressure throughout the Network. It simply means that A is the amount of infectious pressure applied by every node in Network to spread the epidemic, and since its fundamentally happening because of the Global Network structure, this can be said that it is Network effect or Global Network effect.

Then the contribution to likelihood from (exposed to infectious) process for each single node in the network is given below. Remember that this contribution is only measured for those nodes only which I know are exposed and can be infected in future. Thus,

$$L3 = \frac{[\prod_{i=1}^{m}(I_i - E_i)]^{k_E-1}\ \theta_E{}^{-mk_E} \exp\{-\frac{\sum_{i=}^{m}(I_i - E_i)}{\theta_E}\}}{constant}$$

Where constant is simply a constant for gamma distribution with parameters above. It can be ignored for simplicity.

Similarly, the contribution to likelihood from (infectious to removal) process for each single node in the network is given below. Again, remember that this is only measured for those nodes which I know are infectious and can be removed in future. Thus,

$$L4 = \frac{[\prod_{i=1}^{m}(R_i - I_i)]^{k_I-1}\ \theta_I{}^{-mk_I} \exp\{-\frac{\sum_{i=}^{m}(R_i - I_i)}{\theta_I}\}}{constant}$$

Where constant is simply a constant for gamma distribution with parameters above. It can be ignored for simplicity.

The derivation of L3 and L4 is given below. Except for L2, the other contributions of likelihood L1, L3 and L4 are very trivial to compute. Overall, the derivation of L2 is missing which was impossible for me to derive but the result for L2 is mentioned above with the help of some previous studies.

(Derivation of L_3)

$I_1 - E_1 \sim \text{Gamma}(k_E, 1/\theta_E)$

$$f(I_1 - E_1 \mid k_E, \theta_E) = \frac{1}{\Gamma(k_E)} \times \frac{(I_1 - E_1)^{k_E - 1} \, e^{-(I_1 - E_1)/\theta_E}}{(\theta_E)^{k_E}}$$

$$f(I_2 - E_2 \mid k_E, \theta_E) = \frac{1}{\Gamma(k_E)(\theta_E)^{k_E}} \times (I_2 - E_2)^{k_E - 1} \, e^{-(I_2 - E_2)/\theta_E}$$

$$\vdots$$

$$f(I_m - E_m \mid k_E, \theta) = \frac{1}{\Gamma(k_E)(\theta_E)^{k_E}} \times (I_m - E_m)^{k_E - 1} \, e^{-(I_m - E_m)/\theta_E}$$

$$f(I_1 - E_1, I_2 - E_2 \ldots I_m - E_m \mid \theta_E, k_E) = f(I_1 - E_1 \mid k_E, \theta_E) \times f(I_2 - E_2 \mid k_E, \theta) \ldots \times f(I_m - E_m \mid k_E, \theta_E)$$

$$= \frac{1}{\left[\Gamma(k_E)(\theta_E)^{k_E}\right]^m} \times \left[(I_1 - E_1) \times (I_2 - E_2) \ldots (I_m - E_m)\right]^{k_E - 1} \times e^{-\left[(I_1 - E_1) + (I_2 - E_2) + \ldots (I_m - E_m)\right]/\theta_E}$$

$$= \frac{1}{\left[\Gamma(k_E)\right]^m} \times \theta_E^{-mk_E} \left[\prod_{i=1}^{m} (I_i - E_i)\right]^{k_E - 1} \times e^{-\left[\sum_{i=1}^{m} (I_i - E_i)\right]/\theta_E}$$

(Derivation of L_4)

$R_1 - I_1 \sim \text{Gamma}(k_2, \theta_2)$

$$f(R_1-I_1 \mid k_2, \theta_2) = \frac{1}{\Gamma(k_2) \times (\theta_2)^{k_2}} \times (R_1-I_1)^{k_2-1} \; e^{-(R_1-I_1)/\theta_2}.$$

$$\vdots$$

$$f(R_m-I_m \mid k_2, \theta_2) = \frac{1}{\Gamma(k_2) \times (\theta_2)^{k_2}} \times (R_m-I_m)^{k_2-1} \; e^{-(R_m-I_m)/\theta_2}.$$

$$f(R_1-I_1, R_2-I_2 \ldots R_m-I_m \mid k_2, \theta_2) = f(R_1-I_1 \mid k_2, \theta_2) \times f(R_2-I_2 \mid k_2, \theta_2) \ldots\ldots f(R_m-I_m \mid k_2, \theta_2)$$

$$= \frac{1}{\left[\Gamma(k_2)(\theta_2)^{k_2}\right]^m} \times \left[(R_1-I_1) \times (R_2-I_2) \times (R_3-I_3) \ldots (R_m-I_m)\right]^{k_2-1} \times e^{-[(R_1-I_1)+(R_2-I_2)\ldots R_m-I_m)]/\theta_2}$$

$$= \frac{1}{\left[\Gamma(k_2)\right]^m} \frac{\theta_2^{-mk_2}}{} \times \left[\prod_{i=1}^{m} (R_i-I_i)\right]^{k_2-1} \times e^{-\left[\sum_{i=1}^{m} (R_i-I_i)\right]/\theta_2}$$

Prior Distributions

In the previous articles done on this case, studies have shown that it's better to use flat or uniform prior if we also must estimate the exposure time, which is the case now with the Hagelloch data since exposure time is not given. But, I'd be using Inverse gamma as prior for all the epidemic parameter since its conjugate with our Likelihood (not for all the epidemic parameters but still) and is, thus easy to compute/derive.

So, from now, the results that would follow would be classified as using Inverse Gamma prior distribution for all epidemic parameters and gamma prior distribution for the Network parameters.

MCMC Algorithm

MCMC algorithm would give us a posterior distribution of all the parameters of interest discussed above through the help of the Ergodic Markov Chain for each one of those parameters. That Markov Chain would then be used for point estimation, confidence interval, standard errors and much more. Remember here that the actual implementation of this MCMC algorithm is provided through the R package called 'epinet' whose link is provided in the reference section.

Functionality

The prior distribution of each epidemic parameter alongside Likelihood and Posterior is given below. Note that where possible, I used conjugate prior distribution as it speeds up the computation. So, for those prior distribution that would result in a conjugate posterior, the Algorithm would use full conditional to update the MCMC through the implementation of Gibbs sampling. As for the other case where the posterior is not recognizable, the Algorithm would implement Metropolis Hastings Algorithm to get MCMC.

Parameter	**Prior**	**Likelihood**	**Posterior**
B	gamma (a,b)	Exponential Distribution (mean=1/ B)	Gamma (m+a-1, A-b) Note: m and A have been discussed in likelihood L1L2
θ_I	IG (a,b)	L4	IG (a+m k_I , -(b+sigma(Ri-Ii))
θ_E	IG (c,d)	L3	IG (c+m k_I , -(d+sigma(Ii-Ei))
k_I	IG(a,b)	L4	Not Recognizeable
k_E	IG(c,d)	L3	Not Recognizeable

Note that for the Network Parameters Classroom1, Classroom2 and House Distance, the prior distribution (No Matter what it is), doesn't result in a recognizable posterior let alone conjugate posterior, so for the sake of simplicity they're not given in the above table since writing the complex symbols and expression that they have is of no benefit. Furthermore, it is obvious that the MCMC Algorithm must use Hastings Algorithm to update these Network parameters. Rest assured, the posterior distribution is achieved. The complete diagnostics of MCMC for all the parameters (Network or Epidemic) is discussed below. Furthermore, the results, inferences, Standard Errors and Confidence Interval are also discussed in the next section. The derivation of Posterior for the Epidemic parameters however is given and can also be verified using any table for conjugate priors on Internet.

$$L_{joint}(E, I, R, k_E, k_I, \theta_E, \theta_I, \beta) = L_1 L_2 L_3 L_4 \, p(\theta_E, k_E, \theta_I, \theta_I)$$

$$L_1 = \beta^{m-1} e^{-\beta \sum_{i=1}^{m-1} (E_{t_i} - I_{a_i})}$$

$$L_2 = \beta^{\cdot} \exp\left[-\beta \sum \left[\{(E_{t_i} \wedge R_a) - I_a\} V_0\right]\right]$$

$$L_1 L_2 = \beta^{m-1} \exp[-\beta A] \quad \{\text{where } A = \sum (E_{t_i} - I_{t_i}) + \sum [\quad]\}$$

$$L_3 = \frac{1}{(\Gamma(k_E))^m} \theta_E^{-m k_E} \left[\prod_{i=1}^{m} (I_i - E_i)\right]^{k_E - 1} \times e^{-\left[\sum_{i=1}^{m} (I_i - L_i)\right]/\theta_E}$$

$$L_4 = \frac{1}{(\Gamma(k_I))^m} \theta_I^{-m k_I} \times \left[\prod_{i=1}^{m} (R_i - I_i)\right]^{k_I - 1} \times e^{-\left[\sum_{i=1}^{m} (R_I - I_I)\right]/\theta_I}$$

$$\pi(k_E, k_I, \theta_E, \theta_I, \beta) = \pi(k_E)\, \pi(k_I)\, \pi(\theta_E)\, \pi(\theta_I)\, \pi(\beta)$$

$k_E \sim \text{Gamma}(8, 20)$

$k_I \sim \text{Gamma}(15, 25)$

$\theta_I \sim \text{Gamma}(0.25, 0.75) = \dfrac{1}{\Gamma(a)\Gamma(b)} \theta_I^{a-1} e^{-\theta_I/b}$

$\theta_E \sim \text{Gamma}(0.25, 1)$

$\beta \sim \text{Gamma}(0, 4)$

1. Conditional for θ_I :

$$\pi(\theta_I \mid E, I, R, k_E, k_I, \theta_E, \beta) = \frac{\pi(E, I, R, k_E, k_I, \theta_E, \theta_I, \beta)}{\pi(E, I, R, k_E, k_I, \theta_E, \beta)}$$

$$= L_4 \times \pi(\theta_I)$$

$$= \theta_I^{-m k_I} \times e^{-\left[\sum_{i=1}^{m} (I_i - I_i)\right]/\theta_I} \times$$

$\pi(\theta_I) \propto \theta_I^{a-1} \times e^{-\theta_I/b}$

$$= \theta_I^{-mk_I + a - 1} \exp\left(-\frac{[\Sigma(R_i - I_i)] - B}{\theta_I} \right)$$

$$= \theta_I^{(a - mk_I) - 1} \exp\left(\frac{-[\Sigma(R_i - I_i)] - B}{\theta_I} \right)$$

Not Recognizable: $\theta_I^{(a-1)} \exp\left(-\frac{(B + [\Sigma( )]}{\theta_I} \right)$

$$\sim IG\left((a + mk_I), -(B + [\sum_{i=1}^{n}(R_i - I_i)]) \right)$$

2. Conditional for θ_E:

$$= L_3( ) \times \pi(\theta_E)$$

$$= \theta_E^{-mk_E} e^{-[\sum_{i=1}^{m}(I_i - E_i)]/\theta_E} \times \pi(\theta_E)$$

$$= \theta^{-mk_I + c - 1} e^{-[(\sum_{i=1}^{m}(I_i - E_i)]/\theta_E + d/\theta_I]}$$

$\pi_1(\theta_E) \approx \theta_I^{-c-1} e^{-d/\theta_E}$

$$= \theta_I^{(c + mk_I) - 1} e^{-[(\sum_{i=1}^{m}(I_i - E_i) + d)]/\theta_I}$$

$$\approx IG\left((c + mk_I), \left(\sum_{i=1}^{m}(I_i - E_i) + d\right) \right)$$

3 Conditional for β:

$$= L_1 L_2[ ) \pi(\beta)$$

4. Conditional of k_E:

$$= L_3(\,)\ \pi(k_E) \qquad \pi(k_E) \approx (k_E)^{-a-1} e^{-b/k_E}$$

$$= \frac{1}{(\Gamma(k_E))^m} \theta_E^{-mk_E} \times \left[\prod_{i=1}^{m} (I_i - E_i)\right]^{k_E} (k_E)^{-a-1} e^{-b/k_E}.$$

$$= \frac{\left((\theta_E)^{-m} \left(\prod_{i=1}^{m} (I_i - E_i)\right)\right)^{k_E} (k_E)^{-a-1} e^{-b/k_E}}{(\Gamma(k_E))^m}$$

$$= \frac{(a)^{k_E} (k_E)^{-a-1} e^{-b/k_E}}{(\Gamma(k_E))^m} \quad \text{(Not Recognizable)}$$

$$= \frac{\left((\theta_E)^{-m} \prod_{i=1}^{m} (I_i - E_i)\right)^{k_E} (k_E)^{-a-1} e^{-b/k_E}}{(\Gamma(k_E))^m}$$

5. Conditional for k_I

$$= L_H(\ ) \times \pi(k_I)$$

$$= \frac{1}{(\Gamma(k_I))^m} (\theta_I)^{-m k_I} \left[\pi(R_i - I_i)\right]^{k_I - 1} \times (k_I)^{-a-1} e^{-b/k_I}$$

$\pi(k_I) = (k_I)^{-a-1} e^{-b/k_I}$.

$$= \frac{(\theta_I)^{-m} \left[\prod_{i=1}^{m} (R_i - I_i)\right]^{k_I - 1}}{(\Gamma(k_I))^m} \times (k_I)^{-a-1} e^{-b/k_I}$$

Not Recognizable.

MCMC Diagnostic

The MCMC Diagnostic is done using R package 'Coda'. The MCMC diagnostic on epidemic Parameters

$\beta, \theta_E, k_E, \theta_I, k_I, E$

suggests that they have achieved Stationary Distribution. The Stationary distribution test and Halfwidth Mean test were accomplished through "Hiedel.Diag testing". The results can be verified with the help of R Code.

The Network Parameters however, haven't got impressive results when it comes to MCMC Diagnostics. This is obvious since there is Multicollinearity between the Independent Network parameters which are

Intercept, Classroom1, Classroom2, House Distance

So, it suggests that we shouldn't use all the Network parameters for the Inference. Since to handle Multicollinearity, we must select a subset of Network parameters out of all the given parameters. Given the logic of the domain that this problem belongs to, and looking at the results of MCMC diagnostics, I'd be using only House hold distance as a serious candidate from Network parameters which has a lot of impact on spreading the Disease. The results of MCMC diagnostics for all Parameters, however are shared here:

Parameter Type	**Parameter**	**Stationary Distribution Test**	**Halfwidth Mean Test**
Epidemic	β	Passed	Passed
Epidemic	θ_E	Passed	Passed
Epidemic	k_E	Passed	Passed
Epidemic	θ_I	Passed	Passed
Epidemic	k_I	Passed	Passed
Epidemic	E (exposure time)	Passed	Passed
Network	Intercept	Passed	Failed
Network	Classroom1	Passed	Failed
Network	Classroom2	Passed	Failed
Network	House Distance	Passed	Passed

Autocorrelation in Markov Chain

Auto correlation is one of the main problems that Markov Chains face. If there is a lot of Autocorrelation, then the mixing of Markov Chains tends to be slow. So, more iterations are required for the Convergence of MCMC. There're also theories alongside mathematical proofs that a higher correlation means less accurate estimates like Standard Error and Mean. That of course leads to relatively poor Confidence Intervals.

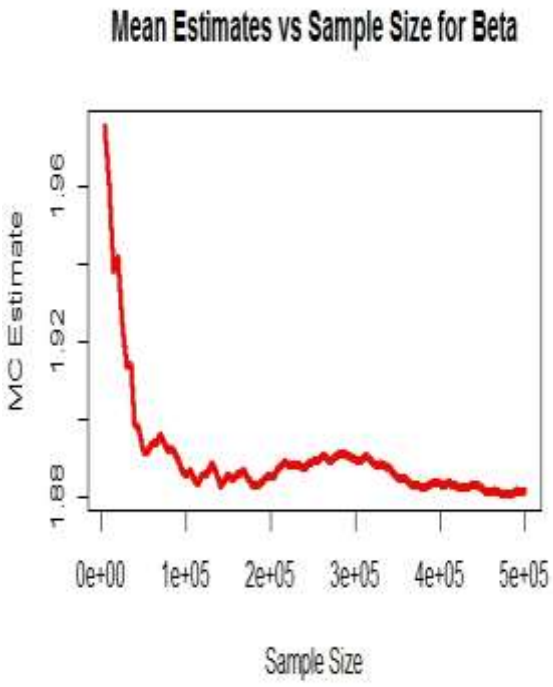
Considering all that in mind, I think it's fair to say that Autocorrelation in MCMC, for all the parameters needs to be discussed. As such, the table below gives us Autocorrelation results in MCMC for our model.

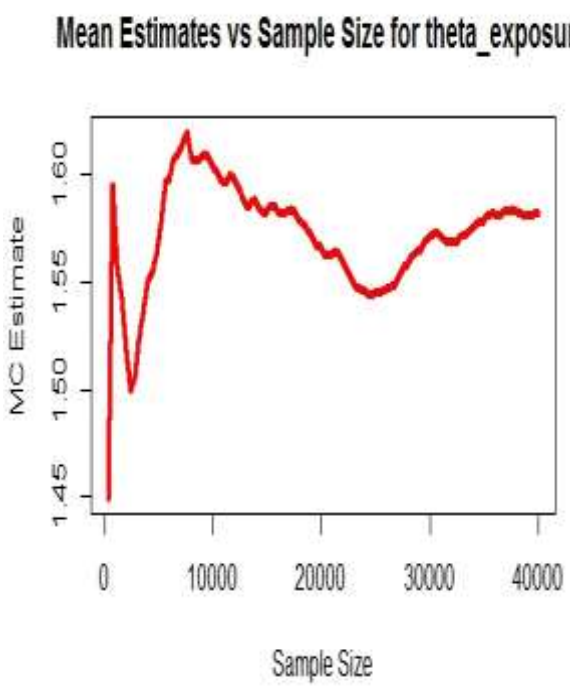
Parameter Type	**Parameter**	**Lag 50 Autocorrelation**	**Lag 100 Autocorrelation**	**Lag 500 Autocorrelation**	**Comment**
Epidemic	B	0.008	0.010	-0.006	Good Mixing, Easily Converged
Epidemic	θ_E	0.34	0.28	0.11	Slow Mixing, Still Easily Converged
Epidemic	k_E	0.34	0.31	0.15	Slow Mixing, Still Easily Converged
Epidemic	θ_I	-0.0003	0.0004	-0.001	Good Mixing, Easily Converged
Epidemic	k_I	0.0005	-0.002	-0.0001	Good Mixing, Easily Converged
Epidemic	E	0.096	0.052	0.012	Good Mixing, Easily Converged
Network	Intercept	0.83	0.79	0.68	Very Poor Mixing, Somehow Converged

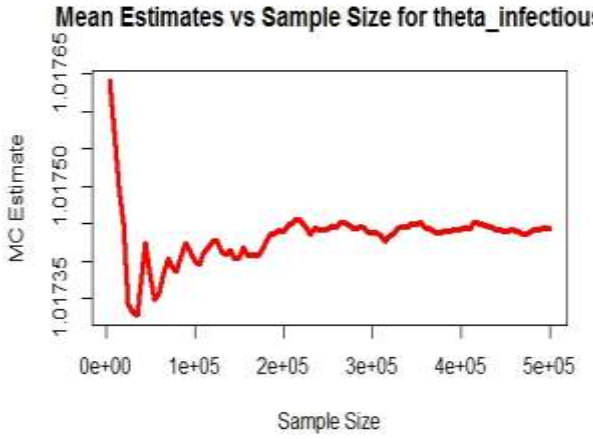
Network	Classroom 1	0.99	0.99	0.98	Extremely Poor Mixing, Somehow Converged
Network	Classroom 2	0.62	0.53	0.33	Slow Mixing, Converged
Network	House Distance	0.79	0.75	0.65	Slow Mixing, Still Satisfactory results

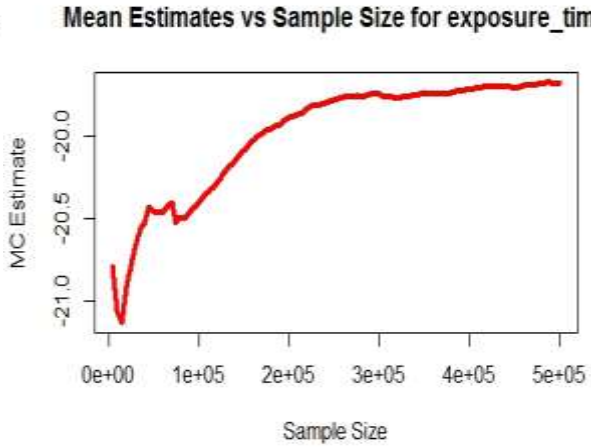
Estimates and the effects of Sample Size

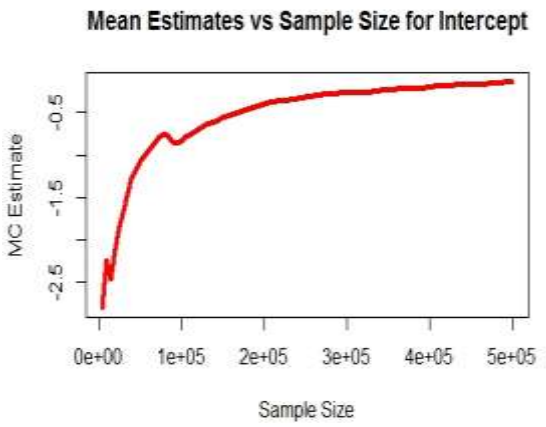
In this portion, I'd like to show, with the help of graphical figures, the effects of sample size on the Mean estimates. This would help me understand, whether the increasing sample size is helping to improve the convergence of MCMC or not. Please note that this is an informal Technique and can be easily challenged. Yet, I'd like to stress that it's always important to look at every aspect of Diagnostics. And While this technique may not be 100 percent reliable or even accurate, it certainly gives a General situation of convergence and is, somehow a decent criterion to check MCMC diagnostics.

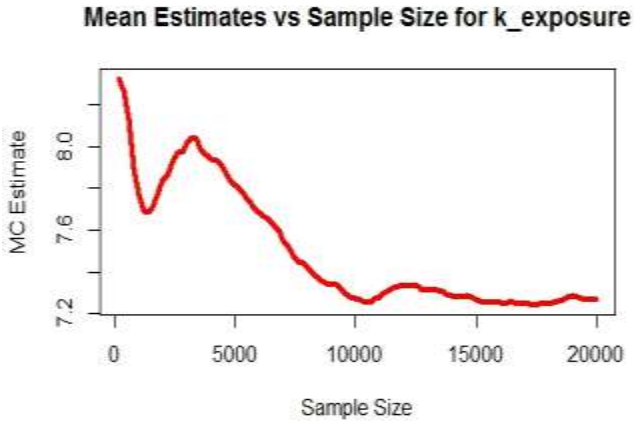

Mean Estimates vs Sample Size for Beta
MC Estimate
1.88
1.92
1.96
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Sample Size

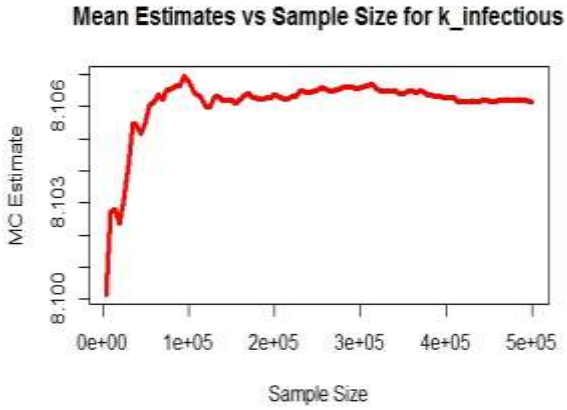

Mean Estimates vs Sample Size for theta_exposure
MC Estimate
1.45
1.50
1.55
1.60
0
10000
20000
30000
40000
Sample Size

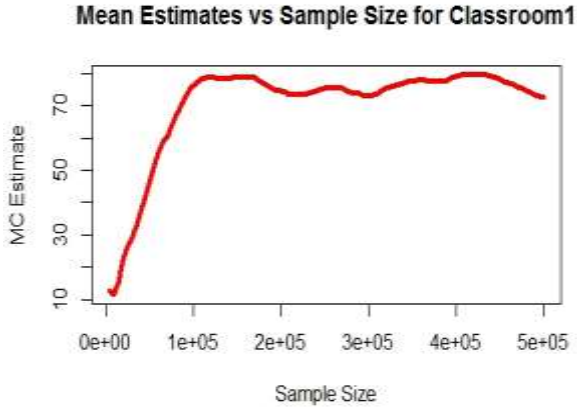

Mean Estimates vs Sample Size for theta_infectious
MC Estimate
1.01735
1.01750
1.01765
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Sample Size

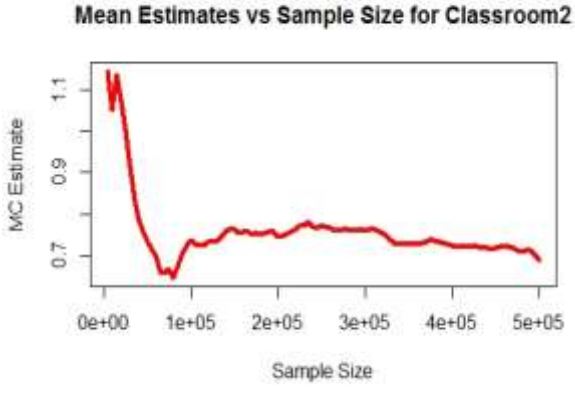

Mean Estimates vs Sample Size for exposure_time
MC Estimate
-21.0
-20.5
-20.0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Sample Size

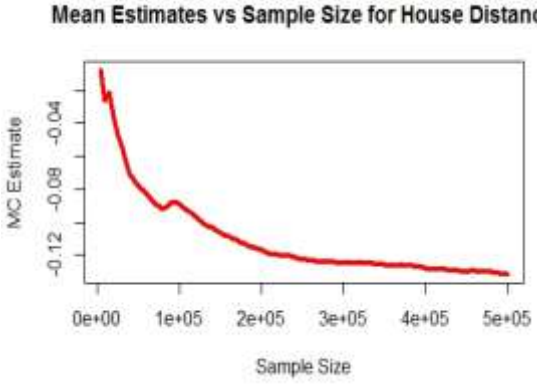












These figures suggest that while, in some cases, it is slow, the increasing sample size certainly is playing a role in MCMC convergence since the estimator (in this mean) is becoming centered

around. Obviously, theoretically I can always prove from the properties of Stationary, Ergodic MC that it would converge.

Effective Sample Sizes

Effective Sample Sizes is used in statistics when the correlation in sample of a specific Distribution is a genuine issue. Below is table given which gives us the effective sample size for each MCMC.

Parameter	Effective Sample Size
B	23708
θ_E	576
k_E	357
θ_I	412447
K_i	462620
E	12287
Intercept	916
Classroom1	710
Classroom2	441
House Distance	966

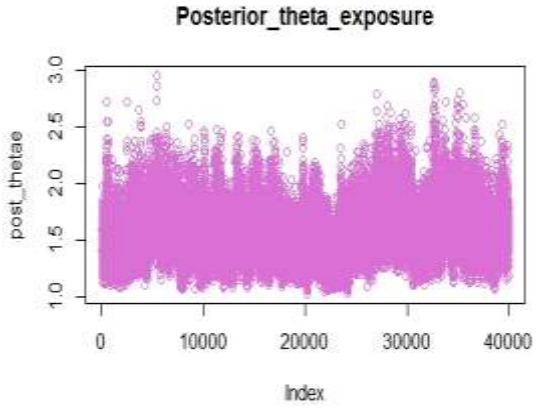
Conclusion about Diagnostics

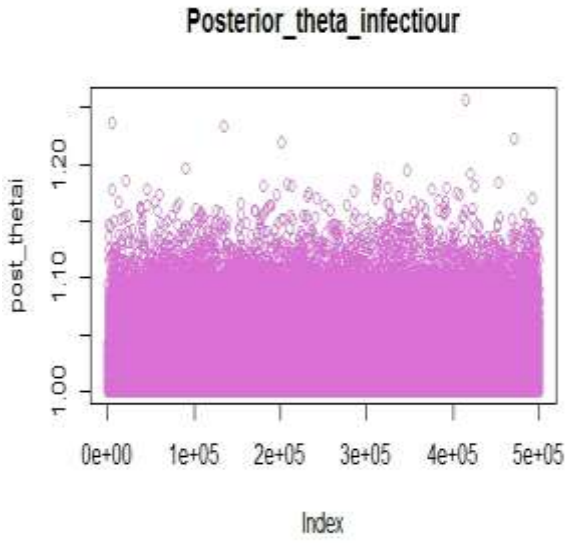
After the aforementioned results, I think it's quite reasonable to suggest that all the Epidemic parameters have achieved stationary distribution. And thus, can be used in Inference thereafter. As for the Network parameters, which if combined in a linear fashion were to give us the Network effects on the epidemic, have failed to give impressive results in MCMC diagnostic. Those Network parameters, even though, have achieved Stationary distribution, are not ideal to be used in the Inference. The reason for that is Multicollinearity, and the Numeric stability associated with that Multicollinearity. Thus, I find it fair to suggest that only the "House distance" parameter can be used in Inference when it comes to Network parameters.

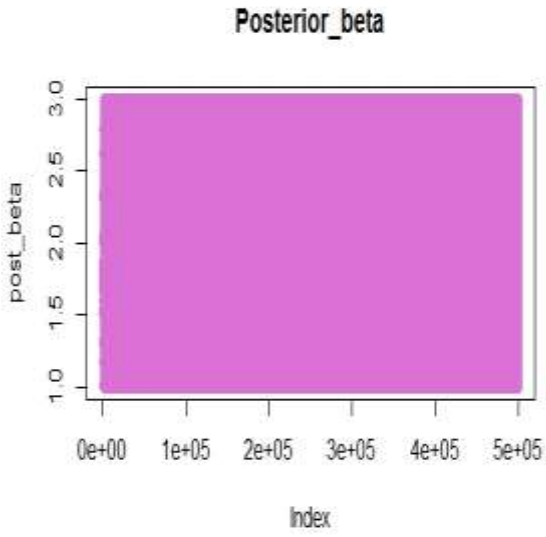
Inference and Results

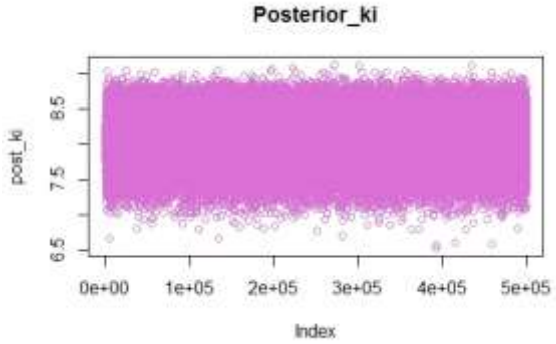
In this section, I'd discuss in detail the results, the inference and the conclusion on epidemic. Noe that this is based on the MCMC and I have already discussed the convergence and Diagnostics. So, it's safe to say that the parameters discussed thereafter have achieved posterior distribution (through full conditionals, Gibbs sampling and Metropolis Hastings Algorithm).

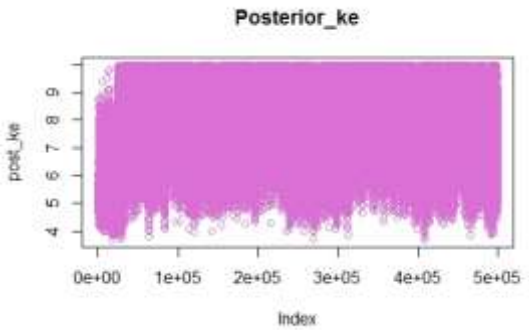
Trace plots of Parameters

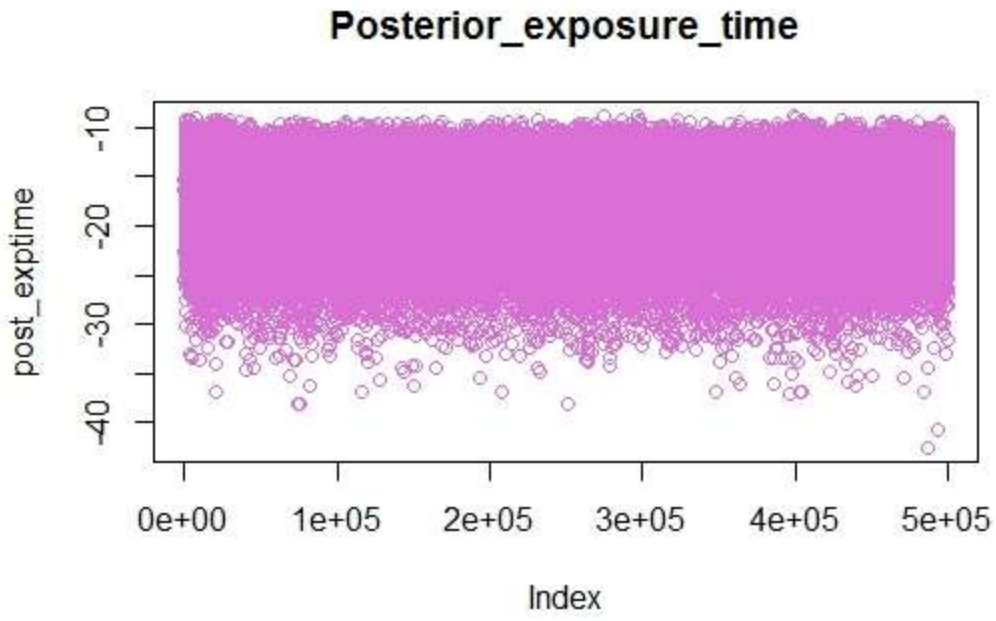






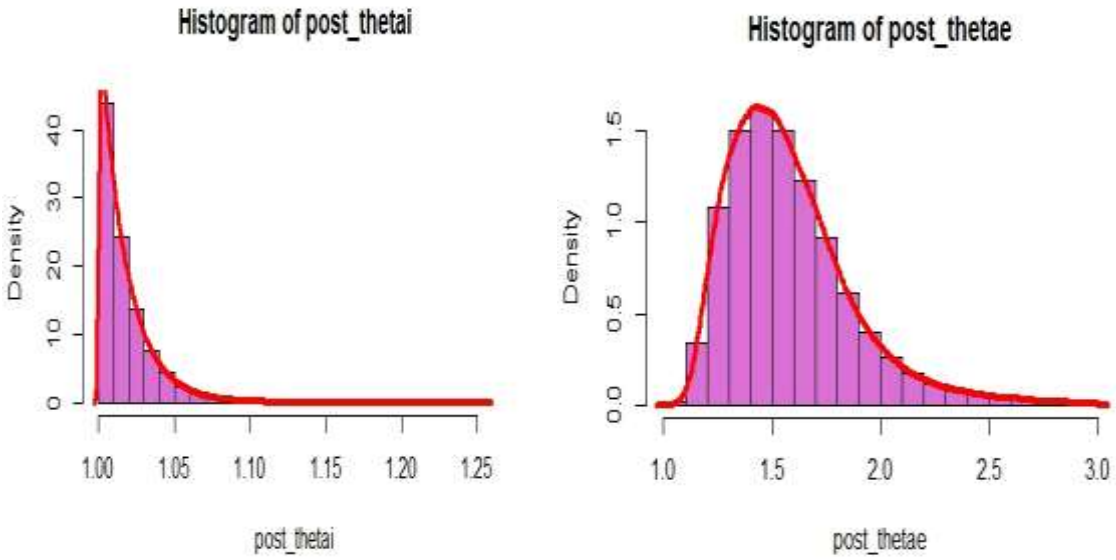


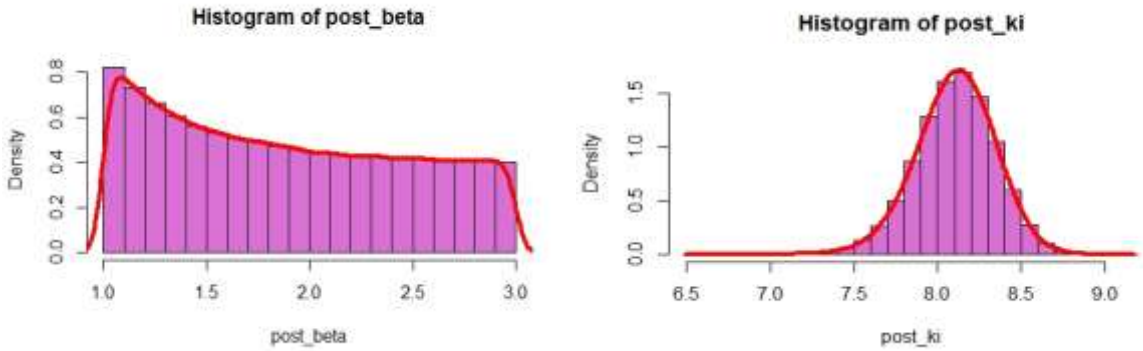


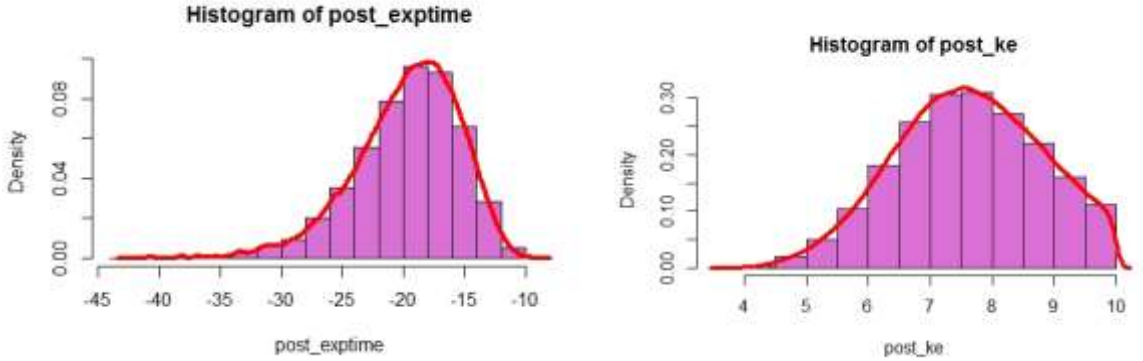


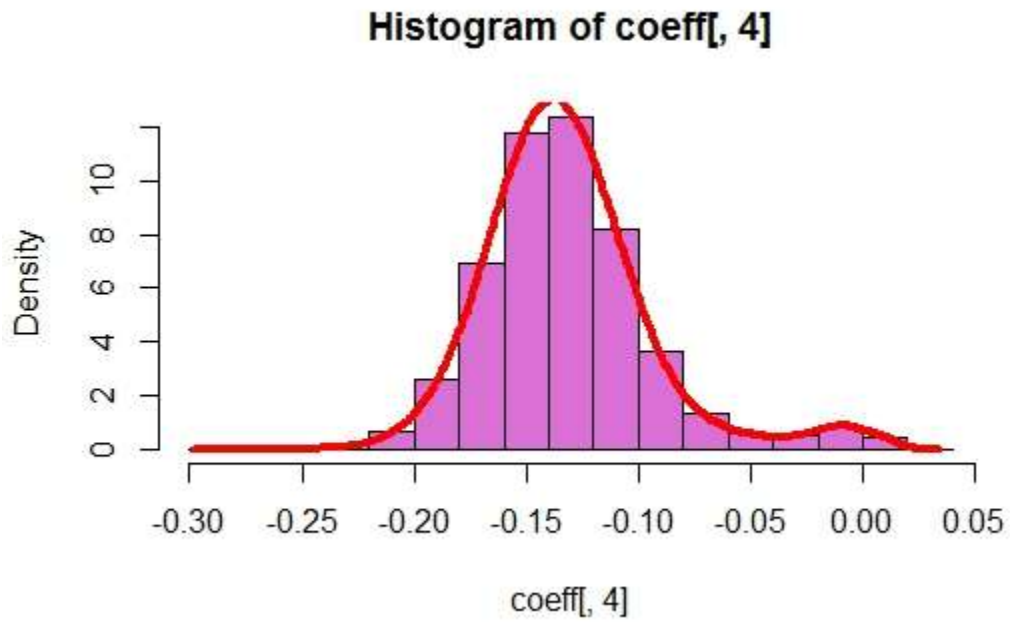
Histograms and Densities

Histograms and Densities of all the Parameters of interest (all epidemic parameters and one Network parameters) are given below.









Remember here that the posterior densities of only B, θ_E and θ_i are recognized since they were chosen from conjugate models and they are respectively Gamma, Inverse Gamma, and Inverse Gamma. The actual parameters can be easily specified through the help of the table discussed in

MCMC. The other posterior densities are not recognized except for House distance which has a Normal posterior density. Also, please note that there's a lot of gap in theory and practice. The R package 'epinet' doesn't allow me to have the likelihood measured for Hagelloch data. I can only specify the priors, and get the posterior. So, even with the best of efforts, I'm unable to get the posterior parameters for Network specifically. However, looking at the histogram and density, I can guess the parameters for 'House distance'. Rest assured, the posterior distribution of epidemic parameters B, θ_E and θ_i can be specified easily. As for all the other parameters, either the posterior distribution isn't recognizable or because of the limitation of the Software (this problem arises only for the Network parameters), I can't mention it. That's why I mentioned at the start of this report that Network effects can't be measured correctly for this Data. However, it can be done for a Generic exponential random Graph which was the core reason this R package was developed.

Inferences

In this section, I will talk about the Inference related to each parameter in the model. This would include Mean, Standard Error, Confidence Intervals and Point Estimation. I used different packages for that**.** The R package 'mcmcse' was used. Also, some manual coding was used as well.

So, first I'd talk about point estimator for each individual parameter and error associated with it. For this, I used the aforementioned R package.

Parameter	**Estimator of Expected Value**	**SE**
B	1.88	0.003
θ_E	1.57	0.007
θ_i	1.01	2.733422e-05
K_i	8.10	0.0003
K_E	7.61	0.02
E	-19.61	0.03
House distance	-0.13	0.001

Remember here that the Standard Error is calculated through batch means method where the size of each batch is the square root of the sample size. Furthermore, the Estimator is the estimate of expected value of that parameter.

I also calculated the Estimator of expected value manually (without a built in Package) and Standard deviation. I also built Highest Posterior Density Confidence Intervals for each parameter with the help of R package 'Coda'. In the table below, they are presented.

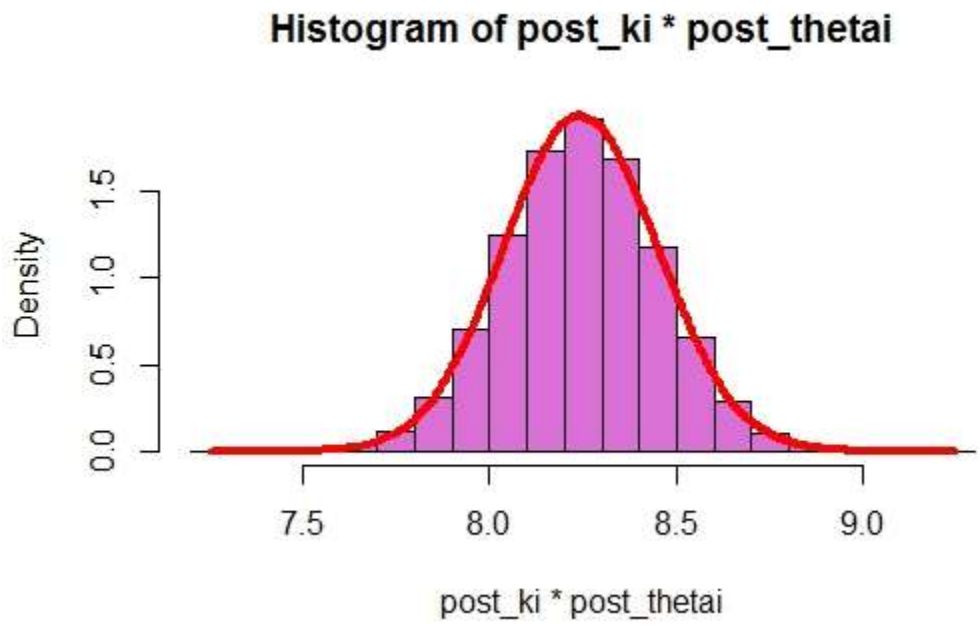
Parameter	Estimator of Expected Value	Standard Deviation	Highest Posterior Density (0.95) Confidence Intervals
B	1.88	0.59	(1.0006,2.87)
θ_E	1.57	0.28	(1.12,2.14)
Θ_i	1.01	0.017	(1,1.05)
K_i	8.10	0.23	(7.63,8.56)
K_E	7.61	1.17	(5.64,9.98)
E	-19.67	4.31	(-28.03, -11.9065)
House Distance	-0.13	0.038	(-0.21, -0.054)

Further Inference

The graph below sows the posterior estimate of mean for Infectious process. In other words, this graph shows how the mean of

Infectious-> Removal

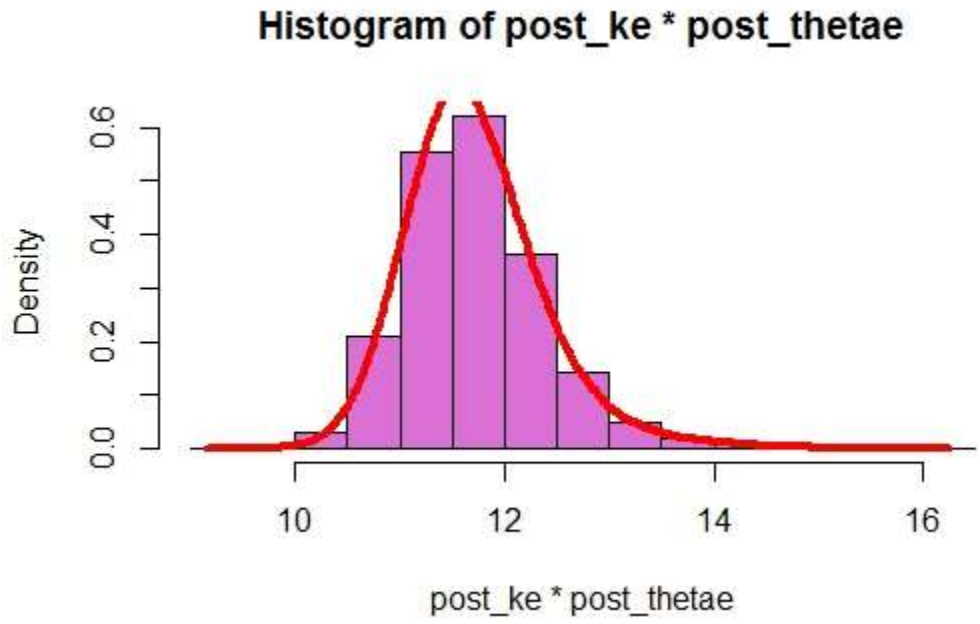
Process would be distributed given the data and the statistical model.



In the same way, I'll have another graph which would show how the estimate of mean for the

Exposed-> Infectious

Process would be distributed given the statistical model and the data.



It is also important in the case of an epidemic study to find out the number of infectious people infected by an individual. This has been calculated through the help of an equation in the

https://online.stat.psu.edu/~dhunter/talks/samsiBayesNetworks2010.pdf

and it uses the posterior distribution of epidemic parameters. It can also be seen in the code.

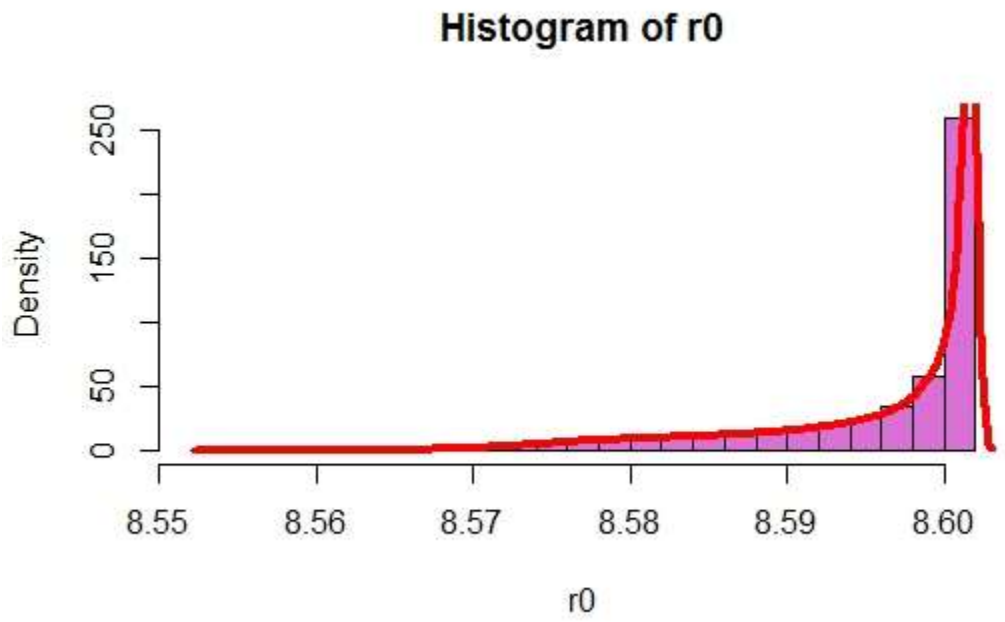
Remember that in the specified article, they had used the Network parameter as part of the Bayesian Inference. I, however haven't been successful in complete Bayesian Inference on all the Network parameters. Therefore, to calculate the average reproduction number, I must take a big assumption about the probability of Network contacts. I've assumed the probability of a contact in the network to be 0.046 which btw is inferenced in the article. After specifying the probability, I can use the posterior epidemic parameters and find out the average reproduction number. The equation for average reproduction number is given as:

$$R = (NP)\{1 - \left(\frac{1}{1 + \beta\theta i}\right)^{Ki}\}$$

Where N is the total number of nodes in the Network, in Hagelloch data, it is 187. P however, is assumed (not been able to successfully infer it as part of the Bayesian inference due to limitation of Software). As for the other parameters β, θi and Ki I will use their posterior distribution. So now, I'm in perfect position to analyze this parameter R. The inference is given below.

Parameter	Inferential function	Value
R	Highest posterior density (0.95) Confidence Interval	(8.57,8.60)
R	Estimate for expected value	8.59
R	Standard Deviation	0.007
R	Point Estimate and SE	8.59, 4.1177 * e^-05

The histogram and probability density is given as:



Acknowledgement, References and Further Study

In analyzing the epidemic data, especially the Hagelloch data through the only R package available to study this, I must say that role of Network parameters is obviously unclear and quite Non-trivial. Even the research articles and past studies support this claim of mine. However, that obviously shouldn't stop me from doing my best. I must say, that even the best studies on Network parameters weren't satisfactory or intuitive to the best of my understanding. This was, and has been my greatest weakness in this project. Apart from that, I think I have been successful in complete Bayesian inference for all the epidemic parameters. The choice of priors and the hyperparameters has been done through a comprehensive study on epidemic models. Nonetheless, there's always need for improvement.

Overall, I feel quite satisfied and hopeful for an excellent grade in this subject. I must acknowledge all the studies previously done on this subject specifically:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.587.5607&rep=rep1&type=pdf

https://online.stat.psu.edu/~dhunter/talks/samsiBayesNetworks2010.pdf

http://www.stat.yale.edu/Conferences/ICSS2010/abstracts/David%20R%20Hunter.pdf

I remain hopeful to do more study on this topic in future Statistics courses.