Statistical Modelling of Interaction Lengths

Jianan Rui

02 August 2022

Supervisor: Dr Riccardo Rastelli and Dr Michael Fop

School of Mathematics and Statistics
University College Dublin

1 Introduction

Network data usually contains a collection of nodes and edge list, or even the weight of the edges, which allows us to observe clearly the internal connections of the network. Node represents each different entity and edge is the interaction between the entities, as for the weight of the edge it can usually be seen as the amount of effort needed to interact from one entity to another.

In analysis, network data is often thought of as a graph, a directed graph or an undirected graph. Consider two nodes i and j in a graph, and $(i\ ,\ j)$ in the edge list. For undirected graphs, the connection between i and j is unpointed, it means $i{\rightarrow}j = i{\leftarrow}j$, so it can be written as $i{\sim}j$. For directed graph, $i{\rightarrow}j \neq i{\leftarrow}j$.

Based on this framework, we can apply it to a wide range of research and analysis in different fields, such as finance, epidemiology, Internet and transportation. But beyond that, more and more interest is being put into modelling network data, that is, modelling network data so that we can interpret the data in a better way and use the models to make forecasts and so forth.

This work has been ongoing since the latter half of the last century and has resulted in a number of profound models, one of which is the Latent Position Model (Hoff et al., 2002), the theory on which this article is based. After the Latent Position Model, new improvements are constantly being proposed, either to create a more sophisticated model by taking more factors into account, or to improve it in a certain direction depending on the characteristics of the network data objects being analysed. An important statistical modeling approach presented for specific network data is Interaction Lengths (Rastelli and Fop,2020), which is based on the Stochastic Block Model for the analysis of dynamic networks with time continuity.

In this project we will attempt to apply the analysis of Interaction Lengths to the Latent Position Model, and due to time constraints and the complexity of the model, our model will be heavily simplified and will focus on modelling network data using the given parameters.

2 Latent Position Model

2.1 Concept

According to (Hoff et al., 2002)'s summary of the previous findings, If we observe $i{\rightarrow}j$ and $j{\rightarrow}k$, this means that the distance between i and k in latent space is not too long and that there is a high probability that they are also connected. And this latent space in which they are located refers to an abstract multi-dimensional space containing feature values that we cannot interpret directly, but which encodes a meaningful internal representation of externally observed events, for example, potential transitive tendencies in network relations.

In summary, the Latent Position Model assumes that each node i has an unknown position Z_i in Latent Space, and we can calculate the probability of a connection between two nodes based on the Euclidean distance between them, the shorter the Euclidean distance, the more likely they are to be in connect with each other.

2.2 Modelling

N is the number of nodes in the network data, Y is the N x N adjacency matrix with diagonal zero, the value of Y_{ij} represents the connection between node i and node j. If $Y_{ij} = 1$, a connection is established between the two; If $Y_{ij} = 0$, no connection between the two. Let Z is a N × D matrix where each row is composed by $Z_i = (Z_{i1}\ , \ .\ .\ .\ .\ , \ Z_{iD})$, the vector in each row represents the position of each node in the latent space. Also reintroduce an unknown parameter α.

After setting these conditions, we take a conditional independence approach to modeling by assuming that the presence or absence of a connection between two individuals is independent of all other connection in the system.

Thus we can get Latent Position Model:

$$P(Y|Z,\alpha) = \prod_{i \neq j}^{N} P(y_{ij}|z_i, z_j, \alpha)$$

Then use logistic regression to parameterise the above model to obtain a linear equation:

$$\eta_{ij} = logodds(y_{ij} = 1|z_i, z_j, \alpha) = \alpha - |z_i - z_j|$$

And so the probability of a connection between two nodes is:

$$\pi_{ij} = \frac{e^{\eta ij}}{1-e^{\eta ij}}$$

Finally, the probability of the whole graph is obtained (which is also equal to Likelihood Function):

$$\begin{aligned} P(Y|Z,\alpha) &= \prod_{i \neq j}^{N} P(y_{ij}|, z_i, z_j, \alpha) && (1) \\ &= \prod_{i \neq j}^{N} {\pi_{ij}}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} && (2) \end{aligned}$$

which gives us the model we desire.

2.3 Estimation

An important step in the whole process is estimation, i.e. how do we get the latent space position Z of the nodes and the value of α to maximize the probability of the whole graph, or we can say maximizing the Likelihood Function. Because only maximizing the Likelihood Function can we find the probability distribution and parameters that best explain the observed network data.

In the Latent Position Model, Markov Chain Monte Carlo inference is used for estimation, which will not be described in detail here as it is not the focus of the project. Given prior information on α and Z, then sampling from the posterior distribution as follows:

- Identify an MLE $\hat{Z}$ of Z, centered at the origin, by direct maximization of the Likelihood Function.
- Using $Z_0 = \hat{Z}$ as a starting value, construct a Markov Chain over model parameters as follows:

1. Sample a proposal $\check{Z}$ from J($Z|Z_k$), a symmetric proposal distribution.
2. Accept $\check{Z}$ as Z_{k+1} with probability $\frac{p(Y|\check{Z},\alpha_k)}{p(Y|Z_k,\alpha_k)} \frac{\pi(\check{Z})}{\pi(Z_k)}$; otherwise,set $Z_{k+1} = Z_k$.
3. Store $\tilde{Z}_{k+1} = argmin_{TZ_{k+1}} tr(\hat{Z} - TZ_{k+1})'(\hat{Z} - TZ_{k+1})$.

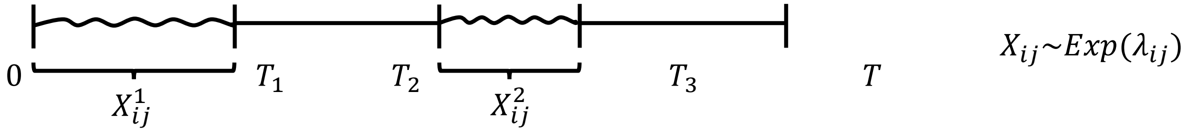
In addition, Markov Chain Monte Carlo inference is indeed a relatively good approach for small networks, but when it comes to large networks, such as thousands of nodes, the computational cost increases significantly. Therefore, there are many faster and more efficient inference methods developed during the years to solve the high computational cost.

3 Modelling of Interaction Lengths

3.1 Concept

In the original Latent Position Model, the probability of a connection between two nodes depends on their Euclidean distance in latent space. But with the discovery of (Rastelli and Fop,2020), we can use the new model to analyse the Interaction Lengths between nodes directly on the basis of the Latent Position Model.

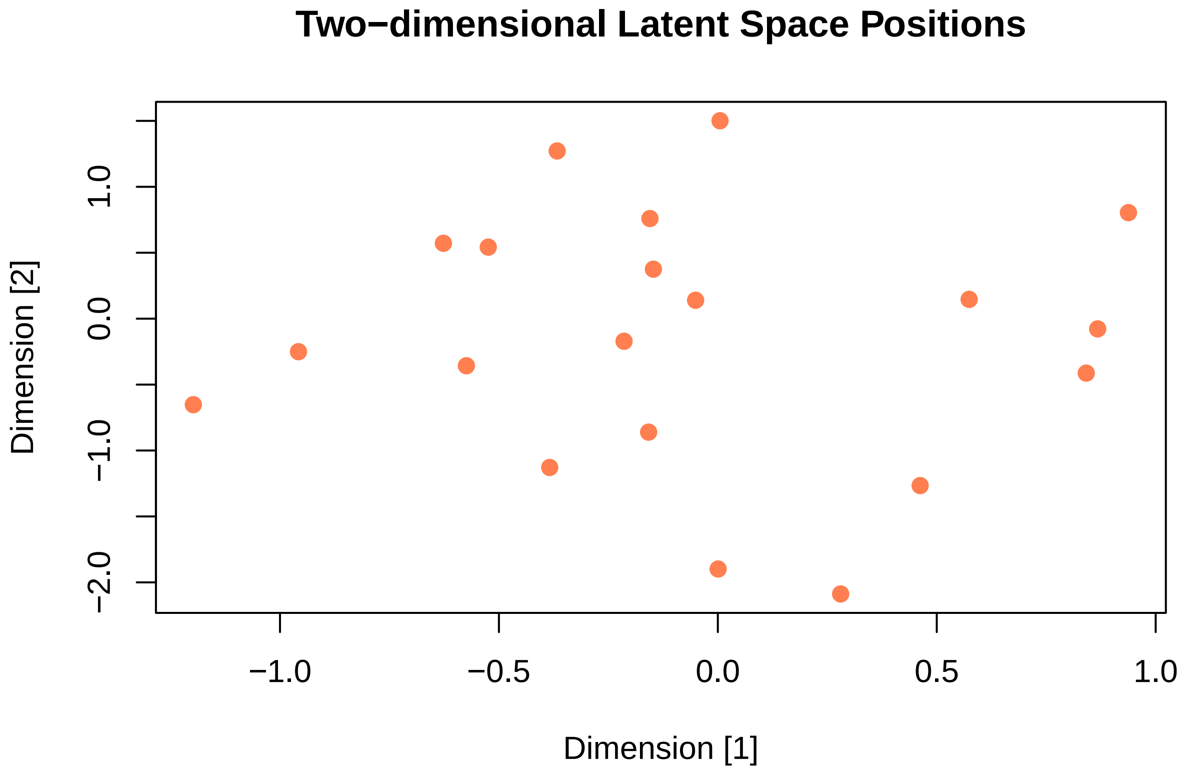
At the same time, the application of Interaction Lengths is based on a continuous period of time, which we separate into small sections, in which Interaction and Non-Interaction always alternate. Once we know the Interaction Length and whether it is an Interaction or not we can reconstruct the whole network. In this project, due to the complexity of Interaction Lengths in continuous time, we only consider the variable that we model is the sum of all the Interaction Lengths X_{ij}^T.



3.2 Generating parameters

The first step for the complete model is to obtain the parameters by estimation, and since estimation is not the focus of this project, here we first generate the parameters randomly.

We aim to generate a network of 20 nodes, all positions Z in the latent space obey the normal distribution: $Z \sim N(0, 1)$:



Next we set the value of parameter α to 2, this gives us all the parameters that should have been obtained by estimation.

3.3 Generating simulated network

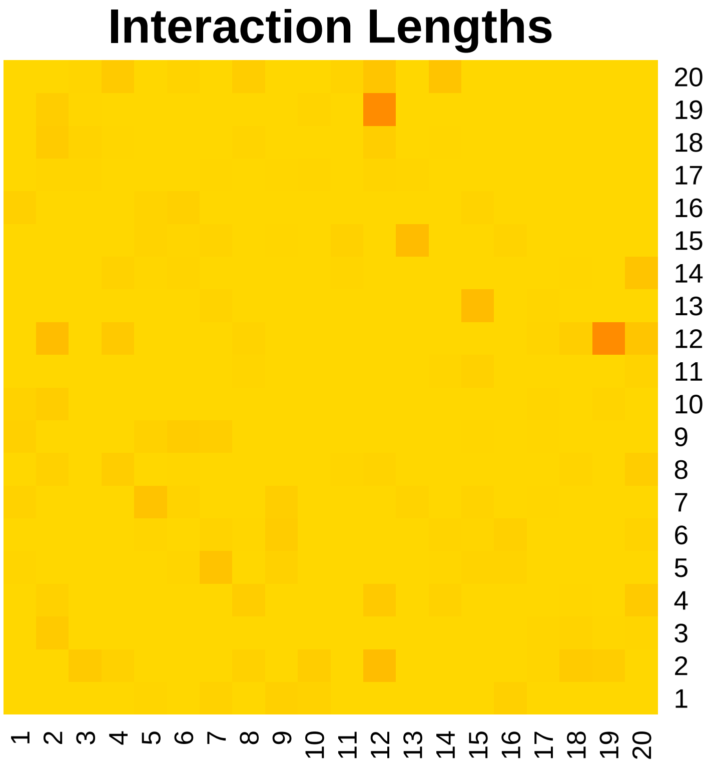
Based on the known latent space position of the nodes and the value of the α, we can first calculate the rate λ of the exponential distribution that the Interaction Lengths follow:

$$\lambda_{ij} = exp(-\alpha - Z_i \cdot Z_j)$$

For example, for the Interaction Length between node 1 and 2:

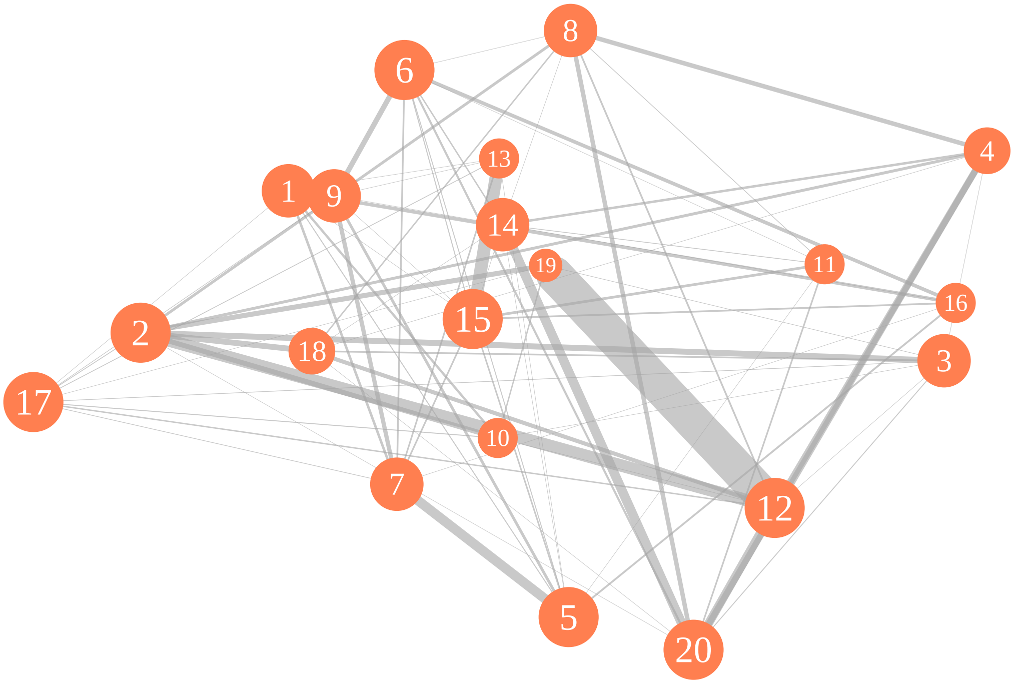
$$\begin{aligned} \lambda_{12} &= exp(-2 - Z_1 \cdot Z_2) \qquad (3) \\ &= 0.202054 \qquad (4) \end{aligned}$$

Also, to make the generating network simpler and easier to observe, some of the larger exponential rates ($\lambda_{ij} > 0.13$) are ignored. With the exponentially distributed rate in place, we can then generate Interaction Lengths ($X_{ij} \sim Exp(\lambda_{ij})$) between the various nodes:



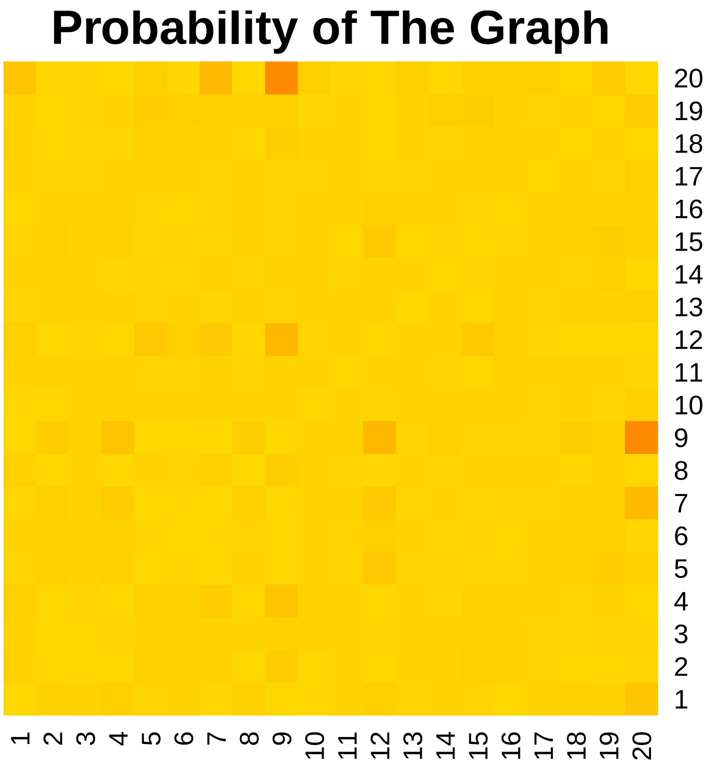
Which in turn, resulting in a simulated undirected network that strictly follows our assumptions.

Simulated Network Based on Interaction Lengths

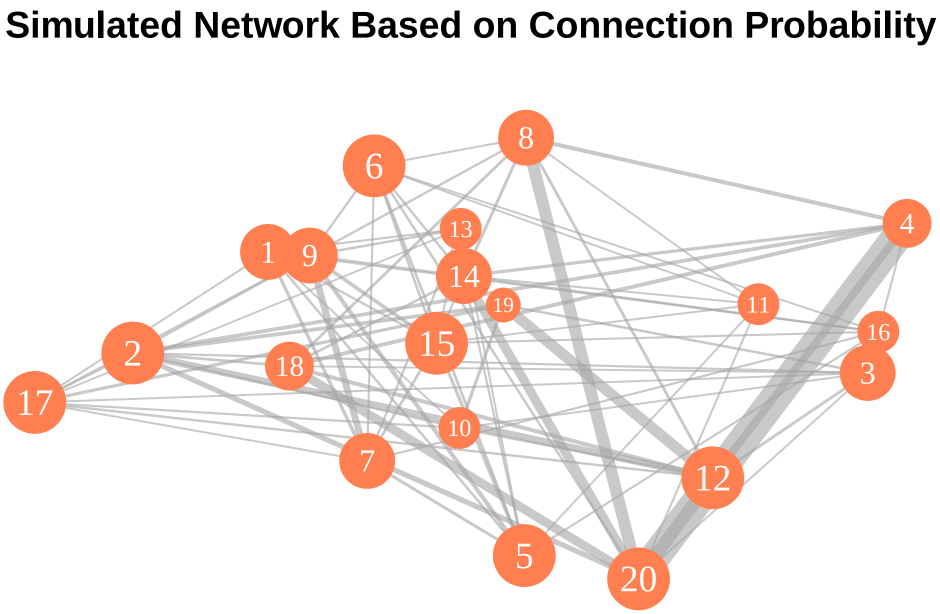


3.4 Transform the network

Since $P(Y_{ij}|Z,\theta) = \lambda_{ij}Exp(-\lambda_{ij}X_{i,j})$, and we know the exponential distribution rate λ and Interaction Lengths X, the data can be represented in a different way, where we transform it into binary to check which pairs of nodes have the longest interactions.



Based on the probability of the whole graph, we can set a probability value as a threshold for filtering, and only those with a probability value greater than this will be considered to have a connection between nodes, this results in our simulated network, which has expected Interaction Lengths equal to $\frac{1}{\lambda}$.



By transforming the network we can see that the probability-based model presents a network and its Interaction Lengths that are almost identical to the original network, indicating that our model is feasible.

4 Estimation

In this section we will try our hand at estimation. Due to the complexity and time consuming nature of the estimation itself, what we do will be significantly simplified and more familiar with the concepts and calculation process.

4.1 Statistical Inference

Statistical inference is a process of extending to whole populations by analysing data on samples to obtain probability distributions. In the Latent Position Model, we make inference by observing the raw network data and assuming that these network data and the parameters in our Likelihood Function follow certain statistical distributions, so as to obtain approximate values that maximise the Likelihood Function to give our model a high degree of accuracy.

In fact, from the earliest Markov Chain Monte Carlo inference to Expectation Maximization inference to Variational Inference, new algorithms have been developed to apply to various models based on the Latent Position Model. Each of these algorithms has its own advantages and disadvantages. As we described earlier, Markov Chain Monte Carlo is computationally expensive for large networks, so here we try to use Variational Inference, a statistical inference method that works well for large networks.

4.2 Variational Inference

4.2.1 Preliminaries

Bayes' rule states:

$$\underbrace{P(\theta|Y)}_{Posterior} = \underbrace{P(Z)}_{Prior} \times \frac{\overbrace{P(Y|\theta)}^{Likelihood}}{\underbrace{\int P(Y|\theta)P(\theta)d\theta}_{Evidence}}$$

where θ is the parameter we want to infer and Y is the observed data. But in general it is difficult to obtain Marginal Likelihood because of the difficulty of integration, so we have tried to convert this process into a derivative which greatly simplifies the steps required.

We want to get at the posterior distribution, this can be done by finding density $q^*(\theta)$ from a family of densities Q that best approximates the posterior distribution:

$$q^*(\theta) = \underbrace{argmin}_{q^*(\theta)\in Q} KL(q^*(\theta)||p(\theta|Y))$$

where $KL(.||.)$ represents the Kullback-Leibler divergence, which is to measure the difference between two probability distributions over the same variable. So our task is to obtain the minimum KL divergence, which has this property:

$$\begin{aligned}
KL(q^*(\theta)||p(\theta|Y)) = \int_\theta q(\theta)log[\frac{q(\theta)}{p(\theta|Y)}]d\theta \quad (5)\\
= \int_\theta [q(\theta)log(q(\theta))]d\theta - \int_\theta [q(\theta)log(p(\theta|Y)]d\theta \quad (6)\\
= E_q[log(q(\theta))] - E_q[log(p(\theta|Y)] \quad (7)\\
= E_q[log(q(\theta))] - E_q[log[\frac{p(Y,\theta)}{p(Y)}]] \quad (8)\\
= E_q[log(q(\theta))] - E_q[log(p(Y,\theta))] + E_q[log(p(Y))] \quad (9)\\
= E_q[log(q(\theta))] - E_q[log(p(Y,\theta))] + log(p(x)) \quad (10)
\end{aligned}$$

because $log(p(x))$ is a constant, so we could just ignore it during optimization process. Moreover, minimizing KL divergence means maximizing its negative, so we can define the evidence lower bound ($ELBO(q)$):

$$\begin{aligned}
ELBO(q) = -KL(q(\theta)||p(\theta|Y)) - log(p(x) \quad (11)\\
= E_q[log(p(Y,\theta))] - E_q[log(q(\theta))] \quad (12)
\end{aligned}$$

where our ultimate goal is to maximise $ELBO(q)$ as much as possible.

4.2.2 Inference Process

Due to the complexity of Variational Inference itself, completing a full estimation in the remaining time of the project is a very difficult task. Therefore, we focus on the estimation of Latent Space Position Z, which we assume its initial position in each dimension follows the Noraml distribution: $N(0,1)$. And we will estimate the value of μ and σ for each position subject to $N(\mu,\sigma)$ by minimizing $ELBO(q)$.

Assume N is the number of nodes, i and j are different nodes,X_{ij} is the interaction lengths, we can obtain the following conditions by calculation.

Prior:

$$Z \sim N(0,1)$$

Posterior:

$$p = p(Z|Y) \tag{13}$$

$$\propto p(Y|Z)p(Z) \tag{14}$$

$$= \prod_{i=1}^{N}\prod_{i<j}^{N} exp[-\alpha - Z_i \cdot Z_j - exp(-\alpha - Z_i \cdot Z_j)X_{ij}] \times (2\pi)^{-\frac{n}{2}} exp(-\frac{1}{2}\sum_{i=1}^{N} Z_i \cdot Z_i) \tag{15}$$

Closed form distribution density:

$$q = q(Z|Y) \tag{16}$$

$$= q(Z|\tilde{\mu}, \tilde{\sigma^2}) \tag{17}$$

$$= (2\pi\tilde{\sigma^2})^{-\frac{n}{2}} exp(-\frac{1}{2\tilde{\sigma^2}}\sum_{i=1}^{N}(Z_i - \tilde{\mu})^2) \tag{18}$$

From these we can calculate KL divergence and got the following:

$$E_q[log(p(Y|Z))] = \textstyle\sum_{i=1}^{N}\sum_{i<j}^{N} -\tilde{\mu_i}\tilde{\mu_j} - exp(-\tilde{\mu_i}\tilde{\mu_j})X_{ij}$$

and

$$E_q[log(p(Z))] = E_q[-\frac{n}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} Z_i \cdot Z_i] \tag{19}$$

$$\approx -\frac{1}{2}\sum_{i=1}^{N}(\tilde{\mu_i} \cdot \tilde{\mu_i} + \tilde{\sigma_i} \cdot \tilde{\sigma_i}) \tag{20}$$

$$\propto -\sum_{i=1}^{N}(\tilde{\mu_i} \cdot \tilde{\mu_i} + \tilde{\sigma_i} \cdot \tilde{\sigma_i}) \tag{21}$$

and

$$E_q[log(q(Z))] \approx -\frac{1}{2}\sum_{i=1}^{N} log(2\pi\tilde{\sigma_i} \cdot \tilde{\sigma_i}) \tag{22}$$

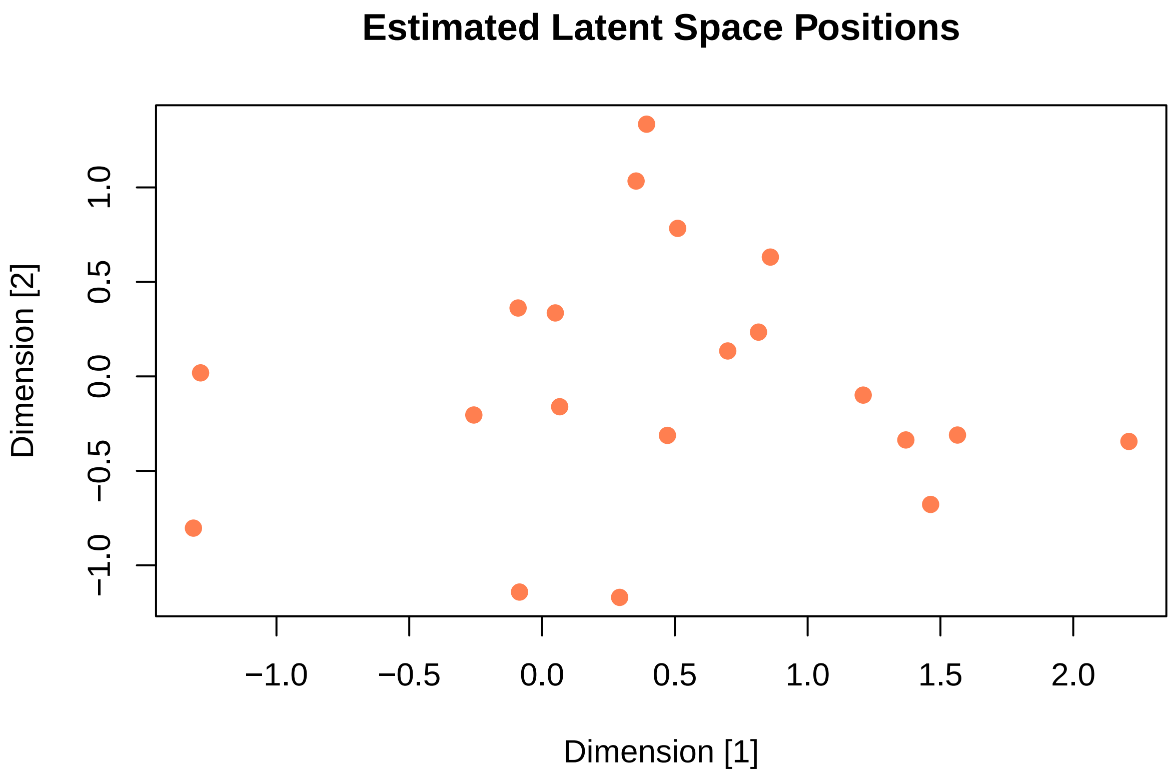
$$\propto -\sum_{i=1}^{N} log(2\pi\tilde{\sigma_i} \cdot \tilde{\sigma_i}) \tag{23}$$

With these we can get ELBO,

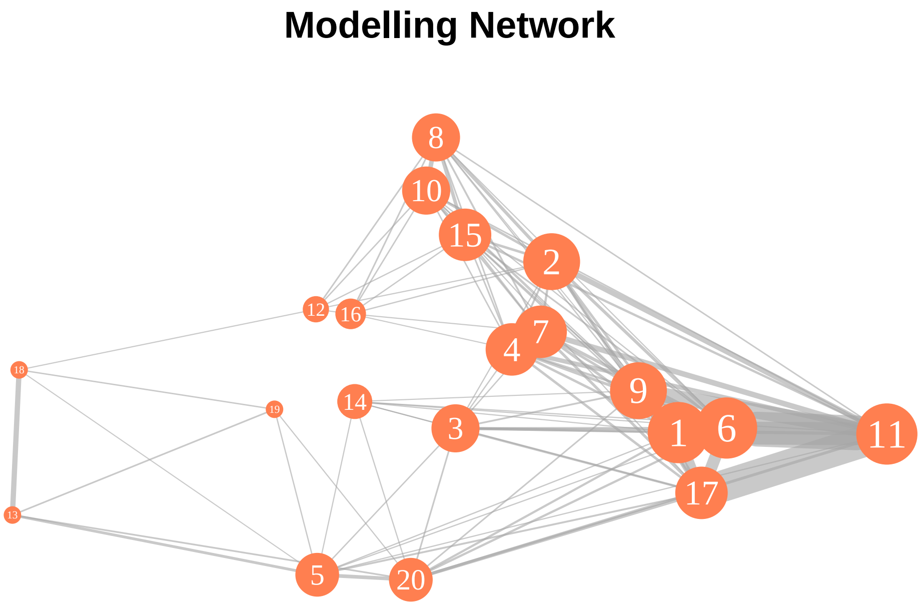
$$ELBO(q) \approx [\sum_{i=1}^{N}\sum_{i<j}^{N} -\tilde{\mu}_i\tilde{\mu}_j - exp(-\alpha - \tilde{\mu}_i\tilde{\mu}_j)X_{ij}] - [\sum_{i=1}^{N}(\tilde{\mu}_i \cdot \tilde{\mu}_i + \tilde{\sigma}_i \cdot \tilde{\sigma}_i)] + [\sum_{i=1}^{N} log(2\pi\tilde{\sigma}_i \cdot \tilde{\sigma}_i)]$$

4.2.3 Estimated parameters

By optimizing $ELBO(q)$ for maximum value, we can finally obtain the position of each node in each dimension of the two-dimensional latent space, which follows a normal distribution with certain μ and σ values.



Once we have the estimated position, we can perform the previous operations to obtain the modelling network.



5 Application in Real Dataset

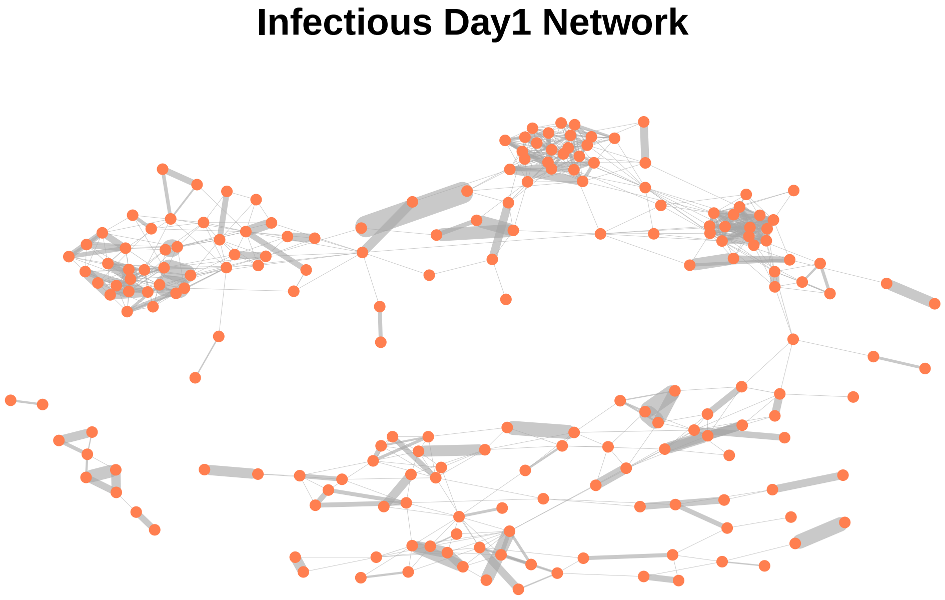
5.1 Infectious dataset

This dataset contains the daily dynamic contact networks collected during the Infectious SocioPatterns event that took place at the Science Gallery in Dublin, Ireland, during the artscience exhibition “INFECTIOUS: STAY AWAY”.

The nodes represent visitors of the Science Gallery while the edges represent close-range face-to-face proximity between the concerned persons. The weights associated with the edges are the number of 20 seconds intervals during which close-range face-to-face proximity has been detected. The dataset contains daily records for up to three months; here, to simplify the work, we only analyse records for day1.

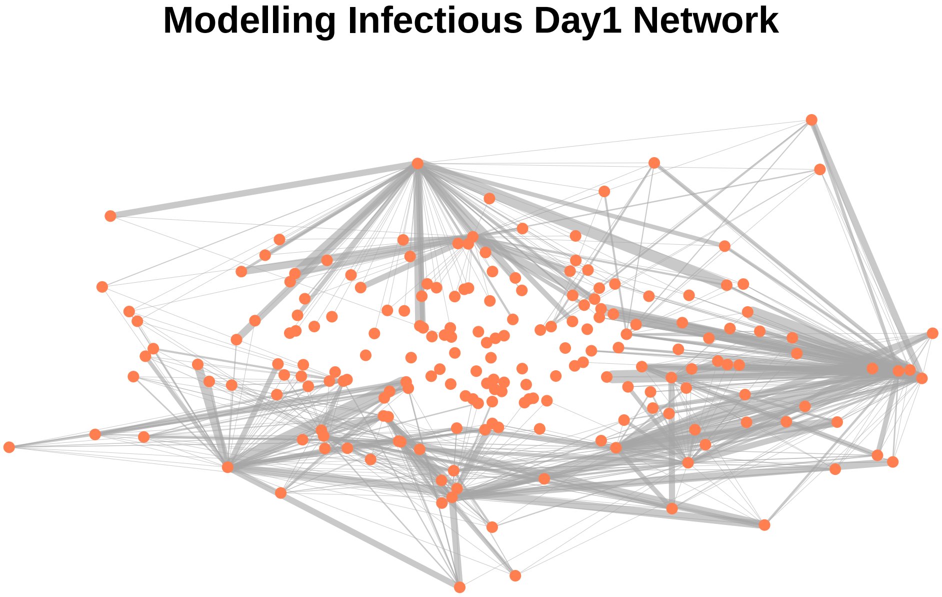
5.2 Analysis and modelling

The network below shows the raw data, which indicates that due to the specific nature of Science Gallery visits, visitors tend to follow a fixed path and interact more with people in adjacent visitor groups.



Infectious Day1 Network

We then repeated the previous steps to analyse and re-model the network, and we can find that our model indeed capture some of the information from the original dataset, but the results of the modelling still lacked precision. Because the estimated parameters were only simple estimates of the latent space position of the nodes, and more optimisation could be done to improve the precision of the model. For example, the estimation of α and further evaluation of the position distribution.



6 Conclusion

This statistical model differs from the previous analysis of binar interactions and takes into account the analysis of Interaction Lengths within a continuous period of time. It is therefore well placed to analyse data on, for example, epidemic types and to reconstruct the original network as far as possible, which is of great interest in a world that is currently plagued by viruses. My work can be further refined in the future, as mentioned above, by further improving the accuracy of the model, especially the estimation part, and thus refining it to make it a complete model. I look forward to the continuation of this part of the work. And I am very grateful to my supervisors for their patience and guidance in helping me to solve a lot of puzzles and difficulties in this project.

Reference

Friel, N., Rastelli, R., Wyse, J., & Raftery, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences, 113*(24), 6629-6634.

Gollini, I., & Murphy, T. B. (2016). Joint modeling of multiple network views. *Journal of Computational and Graphical Statistics, 25*(1), 246-265.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(2), 301-354.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association, 97*(460), 1090-1098.

Mao, L. (2019). Introduction to Variational Inference. [Blog] leimao.github.io, Available at: <*https://leimao.github.io/article/Introduction-to-Variational-Inference/#Variational-Inference*>.

Rastelli, R., & Fop, M. (2020). A stochastic block model for interaction lengths. *Advances in Data Analysis and Classification, 14*(2), 485-512.

Rastelli, R., Friel, N., & Raftery, A. E. (2016). Properties of latent variable network models. *Network Science, 4*(4), 407-432.

Salter-Townshend, M., & Murphy, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis, 57*(1), 661-671.

Shortreed, S., Handcock, M. S., & Hoff, P. (2006). Positional estimation within a latent space model for networks. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences, 2*(1), 24.

Sosa, J., & Buitrago, L. (2021). A review of latent space models for social networks. *Revista Colombiana de Estadística, 44*(1), 171-200.