UNIVERSITY OF CAMBRIDGE

Department of Engineering

Granular Modelling of Epidemics Using Belief Propagation

Author Name: Dimitrios Alexandridis

Supervisor: Dr Jossy Sayir

Date: 1/6/2021

I hereby declare that, except where specifically indicated, the work submitted herin is my own original work.

Signed Δημήτριος Αλεξανδρίδης *date* 1/6/2021

Granular Modelling of Epidemics Using Belief Propagation

Engineering Tripos Part IIB Project - Final Report

June 1, 2021

Dimitrios Alexandridis (da460)
Girton College

Supervisor: Dr Jossy Sayir (js851)
Signal Processing Laboratory, CUED

Technical Abstract

Motivation & Objectives

The COVID-19 pandemic has uncovered the fragility of our society in the face of a new highly transmissible virus. Contact tracing and epidemic modelling have played a defining role in tackling the pandemic. The purpose of this project is to develop a technical framework for the next generation of contact tracing and epidemic modelling tools. Our epidemic modelling approach looks at communities at a fine scale, modelling each interaction that each individual within the community participates in. It then generates a probability distribution for the infectious state of each individual within the community, which provides a quantitative measure for their risk of infection. We will utilise Belief Propagation (BP) on the graph of the community to trace the evolution of a dynamic system that models the evolution of the epidemic.

Outline of Report

In section 1 we introduce our proposal and make a brief account of recent work on modelling the COVID-19 pandemic. We also describe the Susceptible-Infectious-Recovered (SIR) model, some features of which have been adapted to our model. Section 2 outlines the theoretical concepts used to develop our graphical model by running BP on factor graphs that describe the community. The section is concluded with a comparison of our BP modelling approach with a Monte Carlo simulation using the example of a Cambridge College, built using synthetic data. Section 3 presents a more advanced model in which Generalised Belief Propagation (GBP), introduced by Yedidia, Freeman and Weiss, is run on region graphs which group individuals into households. A proof of the equivalence between BP and GBP is provided, along with a method to convert a factor graph to a region graph. Section 3 ends with a practical comparison of the GBP modelling approach with the BP approach of section 2 using the example of a nursing home model that is built using synthetic data. Conclusions and implications are discussed in section 4, which ends with the proposal of a privacy-preserving contact-tracing mobile application that utilises the graphical model developed in this project.

Graphical Representation Using Factor Graphs

We model the community using a factor graph in which discrete-state variable nodes represent individual people within the community and factor nodes represent either swab tests or social interactions between individuals. We then prove that if an interaction is a monotonically increasing function with respect to the number of infected states, then the posterior probability of infection upon an interaction is guaranteed to be higher than the prior probability of infection. For the purposes of epidemic modelling, the BP variable-to-factor message passing rule is modified so that a variable only propagates messages that it received from interactions that took place strictly in the past. A practical comparison of BP with a Monte Carlo simulation applied to a synthetic Cambridge College leads to the conclusion that the two methods produce very similar results. This indicates that the BP algorithm can indeed be repurposed to perform simulations using probabilistic graphs.

Graphical Representation Using Region Graphs

A problem with the factor graph model is that the modified BP algorithm works under the assumption that messages going into a node are independent from one another. However, the existence of short cycles in the graph of interactions means that this assumption is poor. Cycles are generally expected to appear between individuals who belong to the same household, friendship group, or work environment because people belonging to these groups interact more than once. By identifying these groups, we can prevent messages from circulating within them. This task is achieved using the Generalised Belief Propagation (GBP) algorithm.

We begin by summarising the proof by Yedidia, Freeman and Weiss which states that BP, which is a special case of GBP, is an iterative algorithm that converges towards a stationary point of the Bethe approximation of free energy. The Bethe approximation of free energy can be visualised using a region graph whose nodes and edges correspond one-to-one to those of the factor graph. We present a general method to modify this region graph into a form that is free of short cycles by merging together the factors that intercept short cycles. The resulting region graph is free of short cycles and the GBP algorithm associated with it involves messages between neighbouring regions only, similar to BP's message-passing between neighbouring nodes only. We apply our GBP algorithm to an example of a synthetic nursing home and compare its results with those of BP. Both methods produce predictions that resemble those of compartmental models of epidemics, indicating that our model is capable of modelling the dynamics of an epidemic successfully. Due to short cycles, the probabilities of infection for BP quickly saturate to 1 whereas those of GBP increase at a slower and more stable rate. This indicates that short cycles lead to BP overestimating the probabilities of infection. Finally, we apply a disease-mitigation policy to the nursing home model and re-run GBP to conclude that the policy's effect is to reduce the expected number of infections by about 17% over the course of a week.

Conclusions

Modelling an epidemic using a probabilistic graphical model allows us to produce a ”simulated simulation” of an epidemic, a term used by Thomas Richardson in 2004 to describe the relationship between density evolution and iterative decoding. Our simulated simulation can be viewed as a level of abstraction above that of a Monte Carlo simulation. The probabilistic model quantifies uncertainty by expressing states in the community with probability distributions instead of simulation outcomes. Whereas in a COVID-19 simulation individuals exchange viral load, in our simulated simulation individuals exchange a probabilistic message that represents the risk of exchanging viral load.

The lack of real fine-scale data about epidemics was a considerable challenge in the development of our model. Instead, we have resorted to simple statistical metrics and heuristics to produce graphical models that are as close to reality as possible. Some suggested next steps for our work would be to test the model using more complex examples, real data and more accurate interaction factor models. It would also be worthwhile to develop a mobile contact tracing application that runs our graphical model in a distributed real-time manner using real data from users' mobile phones.

Acknowledgements

I would like to thank my supervisor, Dr Jossy Sayir, for the continuous encouragement, guidance and inspiration he provided during the development of this work. I also thank my family for working hard to make sure I had a pleasant and healthy environment to work in despite the adversities caused by the ongoing pandemic. Finally, I would like to thank Dr Flavio Toxvaerd for his valuable feedback in infectious disease modelling.

Contents

1 Introduction

This report expands heavily on a paper pending review at the 11th International Symposium for Topics in Coding (ISTC), which was co-authored by Jossy Sayir and myself. The ISTC paper in turn expands on the technical milestone report that was submitted in Lent Term 2021.

1.1 Motivation

The COVID-19 pandemic has uncovered the fragility of our society in the face of a new highly transmissible virus. Healthcare, education, entertainment and many other systems that ensure society's wellbeing have incurred disruption of historic proportions. The considerable changes we have had to make to our everyday lives signifies how vulnerable our lifestyle can be in the face of a pandemic.

1.1.1 The Role of Technology in Tackling the Pandemic

An important difference between the current pandemic and previous ones in history, such as the Spanish flu in 1918, is the defining role that Information and Communication Technology (ICT) has played in present times: Shortly after the new coronavirus was discovered, several inter-personal activities were moved online thanks to fast internet connections and the existence of internet-enabled devices in the majority of households of more economically developed countries [1]. Technology not only mitigated disruption in work, education and entertainment but it also allowed vulnerable people to shield themselves from the virus without isolating from society.

Another set of ICT tools that have been proven useful in tackling the coronavirus is contact tracing and epidemic modelling. The fact that virtually every adult has a mobile phone means that they can be instantly contacted by health officials to be asked to self-isolate if they have been exposed to an infected person, effectively breaking the chain of virus transmission. Smartphones open up to further opportunities: Contact tracing mobile applications gather information about users' inter-personal contacts and automatically notify them if they have come in contact with a positive case. However, such applications have not been able to successfully replace manual contact tracing services yet [2]. Epidemic modelling has been proven useful in guiding government policy by modelling the effect of non-pharmaceutical interventions (NPI) such as lockdowns and mask mandates on the evolution of the pandemic within a country [3]. In summary, contact tracing and epidemic modelling have been key tactics in fighting a virus for which there is no cure yet by preventing infection and breaking the chain of transmission. This is because the primary form of transmission of COVID-19 is person-to-person contact [4] and, as Hippocrates once said, "prevention is better than the cure".

1.1.2 Public Criticism of Current Modelling Tools

However, the above tools have suffered from important pitfalls. Contact tracing applications based on the Apple/Google Exposure Notification protocol [5] such as the NHS Test & Trace App have been criticised for their limited functionality [6]. The contact tracing app notifies the user if their smartphone has, in the past, been in close proximity with the smartphone of an individual who tested positive. However, the app does not produce a quantitative measure of the risk of infection. Moreover, in order to protect users' privacy, the app is unable by design to inform its user of the location where the encounter took place or who the positive contact was. Consequently, when the app notifies a user of a positive contact, the user is unable to evaluate the risk they are in in order to decide whether they will self-isolate or not.

Epidemic modelling on the other hand has been focused primarily on the evolution of coronavirus infections at the country level, rather than at city or community level [3], [7]. Such models have been useful in evaluating the effect of universally enforced disease-prevention policies such as self-isolation rules and mask mandates. However, they cannot provide bespoke guidance for what small communities such as a nursing home or a Cambridge College should do in order to stop infections efficiently. This leads to policy-making within such communities which is based primarily on qualitative arguments.

1.2 Objectives

The purpose of this project is to develop a technical framework for the next generation of contact tracing and epidemic modelling tools. It has the following novel characteristics:

- It looks at communities at a fine scale, modelling each interaction that each individual within the community participates in. This is in contrast to the conventional compartmental models in epidemiology, which rely on a mean-field approximation. The mean-field assumption is that the population under study is large enough that the behaviour of one individual alone has a negligible effect on the model output [8].
- Our approach generates a probability distribution for the infectious state of each individual within the community, which provides a quantitative measure for the risk of infection.

Our proposal is to utilise information-theoretic tools to create a graphical model of an epidemic that looks at individuals within a community and at the interactions between them in order to predict infections. We utilise Belief Propagation (BP) on the graph of the community to trace the evolution of a dynamic system that models the evolution of the epidemic. Our use of BP differs from its use in well-known applications such as LDPC decoding [9]. Our goal is not to approximate static probability marginals but rather to quantify the dynamic behaviour of the epidemic over time. Such a fine-scale model could be used to test and inform bespoke policy-making within communities.

1.2.1 Outline of Report

In section 1, apart from introducing our proposal, we make a brief account of recent work on modelling the COVID-19 epidemic. We also describe the Susceptible-Infectious-Recovered (SIR) model, some features of which have been adapted to our model. Section 2 outlines the theoretical concepts used to develop our graphical model by running BP on factor graphs that describe the community. It is concluded with a comparison of our BP modelling approach and a Monte Carlo simulation using the example of a Cambridge College that is built using synthetic data. Section 3 presents a more advanced model in which Generalised Belief Propagation (GBP), introduced by Yedidia et al. [10], is run on region graphs which group individuals into households. A proof of the equivalence between BP and GBP is provided, along with a method to convert a factor graph model to a region graph model. Section 3 ends with a comparison of our GBP approach with the BP approach of section 2 using the example of a nursing home model that is built using synthetic data. Conclusions and implications are discussed in section 4, which ends with the proposal of a privacy-preserving contact-tracing mobile application that utilises the graphical model developed in this project.

1.3 Previous Work on Epidemic Modelling

Epidemic modelling during the past two centuries has been largely based on compartmental models, which use the mean-field theory to look at a large population in terms of aggregates that obey a series of ordinary differential equations (ODE) [11]. More recently, the vast computational power of modern computer systems has paved the way for individual-based simulations that model the population of a whole country at a fine scale and provide richer and more accurate predictions [3].

1.3.1 SIR Model

The most basic of compartmental models is the SIR model which segments the population in three groups based on their infectious state: The **S**usceptible, **I**nfectious and **R**ecovered/deceased. Let $S(t)$, $I(t)$, $R(t)$ be the number of people in each group. These quantities obey the following ODEs [12]:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \qquad S(0) = S_0, \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \qquad I(0) = I_0, \tag{2}$$

$$\frac{dI}{dt} = \gamma I, \qquad R(0) = R_0, \tag{3}$$

under the equality constraint $S(t) + I(t) + R(t) = N \, \forall \, t \geq 0$, where N is the population size and β, γ are parameters that represent characteristics of the epidemic such as transmissibility and speed of recovery. Figure 1 illustrates the numerical solution of the SIR model for initial conditions $S_0 = N$, $I_0 = R_0 = 0$ and parameters $\beta = 1/2$ and $\gamma = 1/3$.

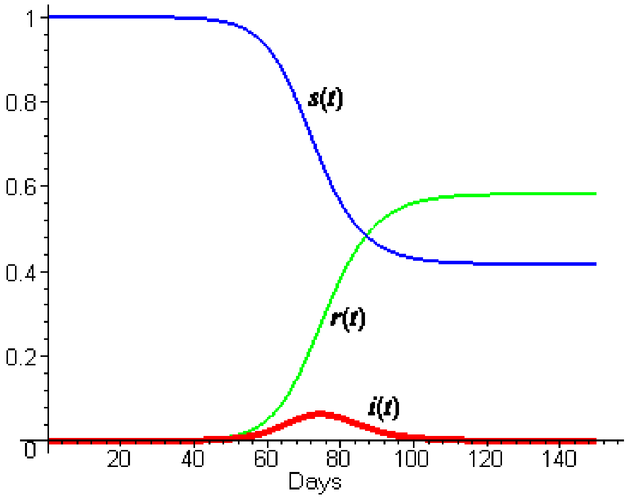


Figure 1: Numerical Solution of SIR model with $\beta = 1/2$, $\gamma = 1/3$ (adapted from [13]). y-axis normalised with respect to N. Notice that $S(t) + I(t) + R(t) = N \, \forall \, t \geq 0$.

The number of infected individuals peaks at day 73. From that day onwards herd immunity is built, leading to an exponential decay and eventually a saturation of the number of infected and recovered because new infections no longer occur.

1.3.2 Modern Approaches Used for COVID-19: Italy

To better explain the novelties of our modelling approach, it is worth summarising the techniques recently employed by the science community in order to model the dynamics of the COVID-19 pandemic. We shall examine the two notable approaches followed by researchers in Italy [7] and the UK [3].

In Italy, researchers used a deterministic compartmental model called SIDARTHE which extends the number of states in the SIR model [7]. The **S**usceptible-**I**infected-**D**iagnosed-**A**iling-**R**ecognised-**T**hreatened-**H**ealed-**E**xtinct model is described by ODEs that include several more parameters than those of the SIR model. Increasing the number of parameters has two important advantages:

- It increases the degrees of freedom, leading to a model that is more complex. A more complex model leads to a better fit on historical data than the SIR model would.
- One can model the effect of different disease-mitigation policies by heuristically adjusting the parameters of the SIDARTHE model in such a way that it describes the modified dynamics of an epidemic after enforcing the disease-mitigation policies.

However, the model has the following important weaknesses:

- The authors of the report emphasise that due to lack of accurate data on disease testing, a large portion of the analysis is based on heuristics drawn from expert knowledge in epidemiology. This means that the model is suboptimal because its predictions depends heavily on heuristics.
- As with all compartmental models, this approach is only applicable for disease-mitigation policies imposed at country-level. Hence it cannot be used to predict the effect of bespoke disease-mitigation policies applied to small communities. For example, it cannot answer the question ”How much staff should a nursing home with 60 residents have in order to avoid a spike in cases?”. In subsection 3.5 we will use our model to provide an answer to this question.

1.3.3 Modern Approaches Used for COVID-19: UK

In the UK, researchers at Imperial College London used an individual-based simulation [3] to produce a model of the pandemic, the results of which played an important role in the UK government's decision to impose the first national lockdown in March 2020 [14]. The simulation draws from demographic data such as population density, age distribution and geographical distribution of school sizes to model the activities that each age group participates in on average.

Modelling the physical contacts that each age group participates in allows us to test the effect of specific, rather than universally imposed, disease-mitigation policies. For example, to test whether high school students should be wearing masks, we can readily incorporate the effect of mask-wearing at schools by simply reducing the assumed transmissibility of the virus for interactions that occur at schools only. By re-running the simulation, we can assess the overall effect that the specific policy has on the pandemic.

However, the individual-based simulation also suffers from weaknesses. The fact that the model is a simulation means that all input data on interactions must be available to a single location (e.g. a researcher's computer) prior to running the simulation. The data are used to construct a model of people's interactions, on which different disease-mitigation policies are tested. This in turn means that practically the model can only be run on synthetic data because it is very hard to gather large amounts of real fine-scale data on people's physical interactions without raising personal privacy concerns.

2 Graphical Representation Using Factor Graphs

The principal mode of COVID-19 transmission is person-to-person contact [4]. This means that a history of social interactions within a community and some information about who is infected and who is not (e.g. through disease testing) could be sufficient information to predict how the epidemic will evolve and who will be infected next.

2.1 Model Description

We model the community using a factor graph in which discrete-state variable nodes represent individual people within the community and factor nodes represent either swab tests, or social interactions between individuals. For the state-space of variable nodes, we borrow the states defined by the SIR model (see subsection 1.3.1). We will constrain ourselves to the simple state-space of the SIR model in all the analyses that will follow. However, our model's state-space can be readily extended to correspond to richer compartmental models such as the SIDARTHE used in [7].

Factor graphs are a suitable candidate to model an epidemic because an epidemic has the following basic characteristics:

- *Discrete state-space* - An individual can be in one of three discrete states: Susceptible to the disease, Infected by it, or Recovered/deceased.

- *Sparse local dependencies between variables* - Assuming that individuals generally socialise with a small subset of the community leads to a sparse graph on which iterative message-passing algorithms can harness sparsity to perform probabilistic inference.

- *Well-understood relationships* - Epidemiology provides insights into how the nature of an interaction (e.g. having a meal at a restaurant) affects disease transmission during that interaction. We can hence design a factor so that it models the effect that the corresponding interaction has on the state of the participants involved.

There are three node types on the graph: Variable nodes (representing individuals), test factors (representing prior evidence on a variable's state) and interaction factors (representing interactions between variables). Inference on the graph can be done efficiently by message-passing, using a modified version of BP.

2.1.1 Variable Nodes & Test Factors

A variable node v_i represents a random variable for the state of individual i in the community (e.g. v_1 for individual 1 in figure 2). It can take 3 values: Susceptible, Infectious and Recovered. We will denote the three states as S, I and R respectively.

For every variable node v_i, we define a factor node $f_i(v_i)$ connected to v_i which summarises any information about v_i prior to inference on the graph. Factors denoted with a numerical subscript always represent test factors. For example, in figure 2, individual v_1 can have a prior probability of infection $f_1(v_1 = I) = 0.8$ and $f_1(v_1 = S) = f_1(v_1 = R) = 0.1$ as a result of a positive covid test result (assuming a test is 80% accurate).

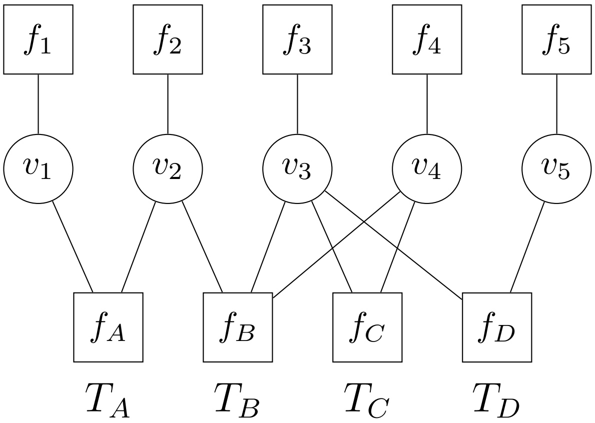


Figure 2: Example graph with a community of 5 individuals participating in 4 interactions.

2.1.2 Interaction Factors

Interaction factors are denoted with an alphabetical subscript to distinguish them from test factors. In figure 2, $f_B(v_2, v_3, v_4)$ represents a social interaction (e.g. a walk at the park) between individuals v_2, v_3 and v_4.

Consider the graph in figure 3 with N individuals participating in one interaction represented by $f_A(v_{1:N})$. The general factor $f_A(v_{1:N})$ connected to N variable nodes can take 3^N possible inputs since each of the N variables can be equal to one of three SIR states. We can constrain the form of $f_A(v_{1:N})$ by assuming the following:

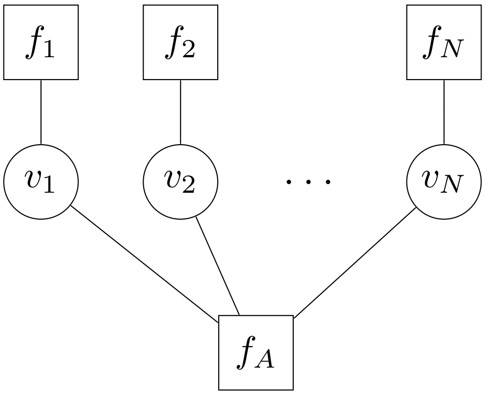


Figure 3: Example graph with one interaction between N individuals.

Assumption 1 *All N individuals participate in interaction A in the same way and they are indistinguishable from each other as far as $f_A(v_{1:N})$ is concerned.*

As a result, the factor output does not depend on which variable is in which state but rather it depends on the number of variables that are in each state, i.e.,

$$f_A(v_{1:N}) = g_A(n_S, n_I, n_R), \tag{4}$$

where n_S, n_I, n_R are the number of variables connected to f_A that are in state S, I, R respectively and g_A is a newly defined function. Making this simplification allows us to design a factor by defining outputs for the $(N+1)N$ possible inputs that $g_A(n_S, n_I, n_R)$ can take instead of the 3^N that a non-symmetric $f_A(v_{1:N})$ would take. This leads to a significant reduction in design parameters and computational complexity. Note that $n_S + n_I + n_R = N$, where N is the total number of individuals in the community. To further simplify the form of f_A, we can derive sufficient (but not necessary) conditions for the form of an interaction factor by making another fundamental assumption about epidemics:

Assumption 2 *Upon an interaction, the probability $p(v_i = S)$ that a participant is susceptible should never increase, i.e., $p(v_i = S) \leq f_i(v_i = S) \,\forall\, i \in \{1, 2, \ldots, N\}$.*

This means that as v_i interacts within the community, the chance $p(v_i = I) + p(v_i = R)$ that they have been exposed to the disease (either having been infected, or having recovered from it) must always increase. Consider the factor graph in figure 3. We will show that if

$$g_A(n_S^{(1)}, n_I^{(1)}, n_R^{(1)}) \leq g_A(n_S^{(2)}, n_I^{(2)}, n_R^{(2)}) \;\forall\; n_S^{(1)} \leq n_S^{(2)}, \tag{5}$$

then f_A will satisfy assumption 2. We start from the fact that the joint distribution of all variables in figure 3 is equal to the product of all factors [10]:

$$p(v_{1:N}) = \frac{1}{Z} f_A(v_{1:N}) \prod_{n=1}^{N} f_n(v_n) \tag{6}$$

where Z is a normalisation constant. Starting from $p(v_i = S) \leq f_i(v_i = S)$ and by marginalising out all variables except for v_i in equation (6), we have:

$$\frac{1}{Z} f_i(v_i = S) \sum_{\{v_{1:N}\}\backslash\{v_i\}} f_A(v_{1:N\backslash i}, v_i = S) \prod_{n \neq i} f_n(v_n) \leq f_i(v_i = S)$$

$$\underbrace{\sum_{\{v_{1:N}\}\backslash\{v_i\}} f_A(v_{1:N\backslash i}, v_i = S) \prod_{n \neq i} f_n(v_n)}_{Q(S)} \leq Z. \tag{7}$$

We also have:

$$Z = \sum_{\{v_{1:N}\}} f_A(v_{1:N}) \prod_{n=1}^{N} f_i(v_i) = \sum_{\alpha \in \{S,I,R\}} f_i(v_i = \alpha) \underbrace{\left[\sum_{\{v_{1:N}\}\backslash\{v_i\}} f_A(v_{1:N\backslash i}, v_i = \alpha) \prod_{n \neq i} f_n(v_n) \right]}_{Q(\alpha)}.$$

The normalisation constant Z is hence a weighted average of $Q(\alpha)$ over the three states if $f_i(v_i)$ is normalised. Inequality (7) is consequently satisfied because of the initial constraint on $f_A(v_{1:N})$, namely,

$$f_A(v_{1:N\setminus i}, v_i = S) \leq f_A(v_{1:N\setminus i}, v_i = \alpha)\, \forall\, \alpha \in \{S, I, R\}.$$

This expression is true because the left-hand side will always have a number of variables equal to S that is greater or equal to that of the right-hand side, satisfying inequality (7). Hence it has been proven that constructing an interaction factor with the form of equation (4) and constraining it with inequality (5) leads to a model that satisfies assumptions 1 and 2 as long as test factors are normalised. Since test factors reflect the prior probability distribution of a variable, they are always normalised.

2.1.3 Example of an Interaction Factor Model

The form of an interaction factor is a model by itself which, as seen in subsection 2.1.2, affects the posterior probability of infection of each participant in the interaction. Its form hence depends on the dynamics that govern disease transmission during person-to-person contact. Epidemiological experiments and heuristics are necessary to design accurate interaction factors. For example, according to [15], the odds ratio of testing positive with COVID-19 after having a meal indoors with a confirmed case compared to not having a meal with a confirmed case is 2.5. In the summer of 2020, the UK population was not vaccinated and the positivity rate for COVID-19 was about 2% [16]. If we apply our model to that period, it is reasonable to assume that the prior probability of being susceptible is very high, i.e., $f(v = I) << 1$ and $f(v = R) << 1$. Let random variable r be equal to S if nobody in the interaction (excluding v) is infectious and I if at least one participant in the interaction (excluding v) is infected. The odds ratio [17] becomes

$$\begin{aligned} O(v,r) = 2.5 &= \frac{P(v=S, r=S)P(v=I, r=I)}{P(v=S, r=I)P(v=I, r=S)} \\ &= \frac{P(v=S|r=S)P(v=I|r=I)}{P(v=S|r=I)P(v=I|r=S)} \\ &= \frac{1-P(v=I|r=S)}{1-P(v=I|r=I)} \times \frac{P(v=I|r=I)}{P(v=I|r=S)} \\ &\approx \frac{P(v=I|r=I)}{P(v=I|r=S)}. \end{aligned} \tag{8}$$

The last step follows because the posterior probabilities $p(v = I|r = I)$, $p(v = I|r = S)$ are both small. This is true to an approximation because the positivity rate for the disease is very low and the chance of getting infected by somebody who is positive is also low. Consequently the odds ratio is weakly affected by $P(v = S|r = S)/P(v = S|r = I)$, with the remaining ratio dominating. We can now use equation (8) to design a factor that models a meal at a restaurant between two friends. Consider again the factor graph of figure 3 with $N = 2$ and let the interaction factor have the form

$$f_A(v_1, v_2) = \alpha^{n_I^2},$$

where α is a parameter and n_I is the number of infected individuals. For example, if $v_1 = S$ and $v_2 = I$ then we would have $n_I = 1$. For simplicity we have neglected the "Recovered" state. For $N = 2$, variable v_2 is equivalent to variable r. We can now use the odds ratio provided by [15] to find α such that the interaction factor f_A produces the same odds ratio as the experiment of two people dining at a restaurant. The conditional probability becomes

$$\begin{aligned} p(v_1 = I|v_2 = I) &= \frac{p(v_1 = I, v_2 = I)}{p(v_1 = I, v_2 = I) + p(v_1 = S, v_2 = I)} \\ &= \frac{Z^{-1} f_1(I) f_2(I) f_A(I, I)}{Z^{-1} f_1(I) f_2(I) f_A(I, I) + Z^{-1} f_1(S) f_2(I) f_A(S, I)} \\ &= \frac{Z^{-1} f_1(I) f_2(I) \alpha^4}{Z^{-1} f_1(I) f_2(I) \alpha^4 + Z^{-1} f_1(S) f_2(I) \alpha} \\ &= \frac{f_1(I)\alpha^3}{f_1(I)\alpha^3 + f_1(S)} \\ &\approx \alpha^3 \frac{f_1(I)}{f_1(S)}, \end{aligned} \tag{9}$$

where for the last step we have again made the realistic assumption that $f_1(S) >> f_1(I)$ since for a positivity rate of 2%, $f_1(S) \approx 0.98$. We also constrain α to be small enough that $\alpha^3 f_1(I) << f_1(S)$. By a similar argument we have

$$p(v_1 = I|v_2 = S) = \frac{f_1(I)\alpha}{f_1(I)\alpha + f_1(S)} \approx \alpha \frac{f_1(I)}{f_1(S)}. \tag{10}$$

By plugging equations (9) and (10) into the approximate odds ratio of equation (8), we get $O(v, r) \approx \alpha^2$ and hence $\alpha = \sqrt{2.5}$. Starting from experimental results on the infectiousness of eating a meal indoors and by heuristically choosing an appropriate model for the interaction factor, we have derived the value that parameter α should take in order for our modelled interaction to produce the same odds ratio. A similar process can be followed for different types of interactions until a library of realistic interaction factors are created for a variety of events, e.g. students sitting next to each other at school or a nurse assisting the resident of a nursing home. These "building blocks" can then be used to construct a factor graph that accurately models the dynamics of an epidemic within the community.

2.2 Adaptation of Belief Propagation to Epidemic Modelling

Before deriving the modified BP algorithm that can perform inference on the graph of interactions, it is worth investigating the working principles behind the standard BP algorithm. BP was first formulated by Judea Pearl in 1982 [18], although an earlier version of it was described by Gallager in 1962 as a decoding algorithm for the Low-Density Parity Check (LDPC) Codes he developed [19].

2.2.1 Overview of Standard BP

When applied to a factor graph, the iterative message-passing rules and the equation of the belief for a variable take the following form [10]:

$$m_{v_j \to f_i}(v_j) = \prod_{f \in n(v_j) \setminus f_i} m_{f \to v_j}(v_j), \tag{11}$$

$$m_{f_i \to v_j}(v_j) = \sum_{v \in n(f_i) \setminus v_j} f_i(\{v : v \in n(f_i)\}) \prod_{v \in n(f_i) \setminus v_j} m_{v \to f_i}(v), \tag{12}$$

$$b(v_i) = \prod_{f \in n(v_j)} m_{f \to v_j}(v_j), \tag{13}$$

where $n(x)$ is the set of neighbours of x. The algorithm allows us to efficiently marginalise otherwise intractable joint probability distributions with a large number of variables by taking advantage of the sparsity of connections in the underlying graph, i.e., the sparsity in the dependencies between variables. In his original paper [18], Pearl describes the iterations of message-passing as a diffusion of beliefs from the leaves to the root (when the factor graph is a tree) and vice-versa.

It can be easily proven that BP yields the exact marginals for trees. However, theoretical bounds for the algorithm's surprising accuracy on sparse cyclic graphs have not yet been found [20]. The reason why BP's performance is unpredictable in graphs with cycles is because the algorithm assumes that messages going into a node are independent from each other, a condition which is only true for trees. BP's inability to capture the dependence between messages is the reason why it is only an approximate inference method. In the context of LDPC decoding, Gallager notes that inference is accurate as long as the number of message-passing iterations is smaller than the length of the shortest cycle, a condition which ensures that messages do not circulate back into the nodes that generated them [19]. By avoiding graphs with short cycles, we hence have some assurance that the BP results will be accurate.

In the context of our epidemic model, we considered BP to be a suitable algorithm for inference because its basic principle of local message-passing bears a strong resemblance with the locality of human interactions and the exchanges of infection during those interactions (which could be seen as "infection messages").

In section 2, by running BP on the graph of interactions we are neglecting the potential existence of short cycles in the graph of interactions and their adverse effect on inference. However, this is an unrealistic assumption to make because within a community individuals tend to interact repeatedly with certain people (e.g. those within their household), leading to short cycles in the factor graph. In section 3, we will use GBP to mitigate these short cycles by exchanging messages between households rather than individuals.

2.2.2 Modified BP Algorithm

In order to apply BP on the model, we need to modify the scheduling of message-passing. Because each interaction occurs at a specific time instance T, we want messages to only propagate forward in time. For example, in figure 2 if interaction A occurred before interaction B, i.e., $T_A < T_B$, then messages received by f_A should be independent of messages sent by f_B but not vice-versa. This ensures that the state of infection of v_2 at time $T_A < t < T_B$, which is summarised by message $m_{v_2 \to f_B}$, is only dependent on $f_1(v_1)$,

$f_2(v_2)$ and $f_A(v_1, v_2)$, i.e., all the information that is relevant to the state of v_2 at that time.

Assumption 3 *Incoming messages to an interaction at time $t = T$ are independent of all interactions occurring at time $t > T$.*

We hence introduce a timestamp T_X for the time instance each interaction X occurred. We then modify the variable-to-factor message so that a message $m_{v_j \to f_X}$ from v_j to f_X is the product of every factor K neighbouring v_j that has $T_K < T_X$, i.e,

$$m_{v_j \to f_X} = \prod_{\{f_K \in n(v_j) : T_K < T_X\}} m_{f_K \to v_j}, \tag{14}$$

where $n(v_j)$ is the set of neighbours of v_j. The factor-to-variable message and the belief of a variable remain the same as for standard BP, i.e., equal to equations (12) and (13) respectively. Figure 4 is a copy of figure 2 in which the process of calculating $m_{v_3 \to f_C}$ is illustrated, assuming that $T_A < T_B < T_C < T_D$. Blue arrows represent messages that were calculated in previous iterations and the red arrow represents message $m_{v_3 \to f_C}$. Dashed edges represents dependencies/messages which are ignored in the calculation of $m_{v_3 \to f_C}$, in accordance with the modified variable-to-factor message (equation (14)).

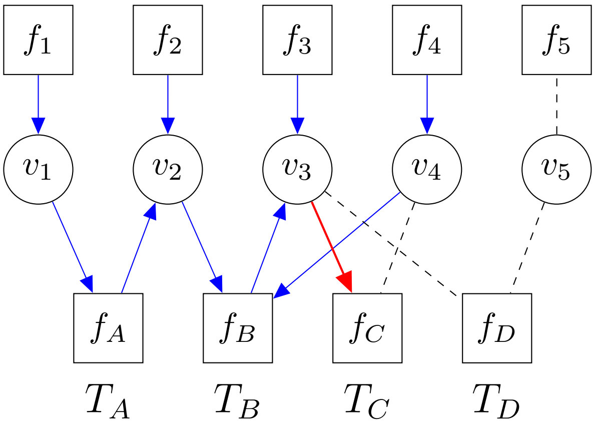


Figure 4: Copy of factor graph in figure 2 illustrating the calculation of $m_{v_3 \to f_C}$ with BP.

Notice that the number of BP iterations will determine how recent an interaction needs to be in order to affect the marginal belief. Assuming that all messages in the factor graph of figure 4 are initialised with the value $m = 1$, a single BP iteration will yield a marginal belief $b(v_3) = f_3$. With 3 BP iterations, the belief $b(v_3)$ depends on f_2, f_3, f_4, f_5, f_B, f_C and f_D. Finally, with 5 BP iterations the belief $b(v_3)$ will depend on all factors in the factor graph of figure 4.

2.3 Comparison of BP with Monte Carlo: A Cambridge College

The lack of real fine-scale epidemiological data for interactions means that it is difficult to evaluate the performance of the modified BP algorithm. Instead, we can compare marginal probabilities of infection $p(v_i = I)$ generated by BP against a simulation of the interactions that are described by the graphical model. A Monte Carlo (MC) simulation on a graph of a synthetic community is described by the following steps:

1. We assume N_0 randomly picked individuals to be the "patients zero" (the first ones to carry the disease). The "patients zero" are the same for all MC runs. All other individuals are assumed susceptible (free from the disease). We then consider the interaction(s) that occurred first, at time $T = 1$.

2. For interaction X that happened at time T_X, we sample states as follows:

   - To participants that were in state S (susceptible) before interaction X occurred, we assign states sampled from probability distribution

   $$p_S(v) = \frac{1}{Z_S} \begin{cases} N_S(X) & \text{if } v = S \\ N_I(X) + \gamma & \text{if } v = I \end{cases}. \tag{15}$$

   - To participants that were in state I (infected) before interaction X occured, we assign states sampled from probability distribution

   $$p_I(v) = \frac{1}{Z_I} \begin{cases} N_I(X) + \gamma & \text{if } v = I \\ N_R(X) + \delta & \text{if } v = R \end{cases}, \tag{16}$$

   where $N_S(X)$, $N_I(X)$, $N_R(X)$ are the number of variables which were in states S, I, R before interaction X happened. With this process, upon an interaction a previously susceptible individual may become infected or remain susceptible. Similarly, a previously infected individual may remain infected or recover from the disease. Parameters γ and δ represent pseudo-counts of infected and recovered participants. They control the dynamics of the interaction. Several assumptions about epidemics have been made in order to come up with this step. These are discussed at the end of this subsection.

3. We repeat Step 2 for the interaction(s) that happened after X until all interactions have been "visited". When the MC run is complete, we store the final states of all variables.

After all MC runs are complete, we normalise the counts for the final states of each variable to come up with a marginal probability distribution for each variable. We now construct a probabilistic model that is equivalent to the above simulation:

- The individuals that were assumed to be "patients zero" for Monte Carlo are given priors $f_i(v_i = S) = 0.15$, $f_i(v_i = I) = 0.8$, $f_i(v_i = R) = 0.05$. All other individuals are assigned priors $f_i(v_i = S) = 0.8$ and $f_i(v_i = I) = f_i(v_i = R) = 0.1$ to reflect the fact that they are most likely susceptible a-priori. This is in line with the simulation above.

- Each interaction factor takes the same universal form

$$f_X(\{v_i\}) = \alpha^{n_I(\{v_i\})} \times \beta^{n_R(\{v_i\})}, \tag{17}$$

  where α, β are parameters and $n_I(\{v_i\})$, $n_R(\{v_i\})$ is the number of participants in interaction X that are in state I and R respectively. We have proven that if $\beta \geq \alpha > 1$ then the graph will satisfy assumption 2. In a more sophisticated graph, separate parameters α_X, β_X could be defined for each factor to reflect the nature of each interaction.

In step 2 of the MC simulation, we made a number of assumptions about the inner workings of an epidemic. The assumptions made for the MC simulation where chosen to be consistent with the assumptions behind the simple form of interaction factor used in equation (17). More specifically, we have assumed the following:

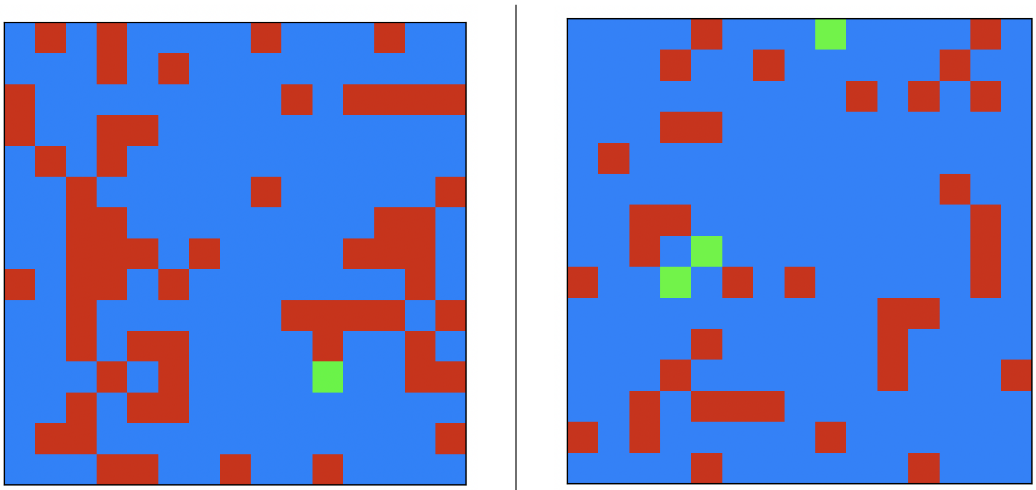
- The sampled states from probability distribution (15) are only assigned to individuals who were susceptible. Similarly, sampled states from probability distribution (16) are only assigned to individuals who were infected. This ensures that a susceptible individual can become infected during the MC run but an infected individual cannot become susceptible. Similarly, an infected individual can become recovered but not vice versa.

- The probability of a susceptible person becoming infected during an interaction depends on the number of infected people participating in it. This is in line with the interaction factor model of equation (17).

- The mechanism with which people recover from the disease is the same as that with which people become infected. In other words, the probability that an infected person recovers from the disease after an interaction depends on the number of recovered participants in it. This is in line with the introduction of parameter β in equation (17) but it is an unrealistic assumption to make because recovery from a disease does not depend on who one interacts with. This design choice was made in order to keep the factor form simple. However, one could argue that if an infected person interacts with a large number of recovered individuals, then this is a sign that the disease has existed long enough in the individual's social circle that they will soon recover from it with high probability. For a more realistic model of recovery from a disease, more complex factors would be designed.

2.4 Results & Comparison

We apply the two modelling methods described above on a synthetic community that resembles a Cambridge College. The community has 225 students participating in 400 interactions that occur within 200 time units (i.e., 2 interaction occur on average at every time instance). There are $N_0 = 4$ ”patients zero” and 3 students on average per interaction. Figure 5a illustrates the maximum a-posteriori (MAP) states after 10 iterations of BP (when convergence was reached). Figure 5b illustrates the final states of individuals averaged over the outcomes of 80 Monte Carlo simulations. Boxes in blue, red and green represent individuals who are a-posteriori Susceptible, Infected or Recovered respectively.



(a) MAP states calculated with 10 iterations of BP.

(b) Final states averaged over 80 Monte Carlo runs.

Figure 5: Posterior states on a synthetic community of 225 individuals with model parameters $\alpha = \beta = 1.4$, $\gamma = 0$, $\delta = 0.1$). Each box on the chessboard corresponds to a unique individual and the color on each box represents the maximum a-posteriori state of the individual (blue: Susceptible, red: Infected, green: Recovered).

The two methods produce fairly similar results, yielding the same posterior state for 76% of individuals in the community. Configuring parameters α and γ changes the rate at which individuals are infected in the BP model and Monte Carlo respectively. Similarly, β and δ control the rate of recovery from the disease with each method respectively. With the chosen parameters, the BP model predicts that after 400 interactions 27% of the population become infected (Table 1), while recovery in the community is at its onset. Starting from simple heuristics and using synthetic data, BP has yielded results that exhibit the basic dynamics of an epidemic and its performance is equivalent to a Monte Carlo simulation.

%	Prior	BP averaged marginals	MC outcomes
Susceptible	98.2	72.4	77.3
Infected	1.8	27.1	21.3
Recovered	0	0.5	1.3

Table 1: Percentage of population in each state.

3 Graphical Representation Using Region Graphs

A problem with the factor graph model outlined above is that the modified BP algorithm works under the assumption that messages going into a node are independent from one another. In the case of the synthetic community of section 2.4, we had the luxury of purposefully designing a factor graph that does not have short cycles. In reality this is a poor assumption to make because an individual within a community tends to interact with the same people over and over again, thereby creating short cycles in the graph of interactions. Consequently, it is fair to expect that a realistic graphical model of the community will have plenty of short cycles that impede inference due to strong dependencies between messages that are assumed independent. Thus if c is the length of the shortest cycle in which nodes x and y participate in, message $m^{(i)}_{x \to y}$ sent out by node x to y at iteration i depends on the same message $m^{(j)}_{x \to y}$ that the node sent out at iteration $j = i - c$. Thus information is not "diffused" within the graph as Gallager described [19] but rather it is "recycled", leading to inaccurate beliefs.

Figure 6 illustrates message $m_{v_3 \to f_D}$ which is problematic as a result of the length-4 cycle in the factor graph of figure 2. Notice that $m_{v_3 \to f_D}$ has been fed with message $m_{f_4 \to v_4}$ twice, meaning that $m_{v_3 \to f_D}$ is more sensitive towards the values of $f_4(v_4)$ than it should due to a "recycling" of the $m_{f_4 \to v_4}$ message.

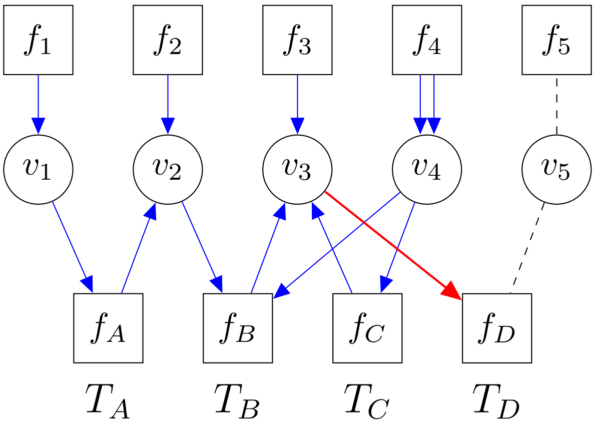


Figure 6: Copy of factor graph in figure 2 in which the calculation of $m_{v_3 \to f_D}$ is illustrated.

Given that we do not know what the correct results should be for the posterior probabilities of infection on our synthetic data, it is difficult to quantify the effect that cycles have on the accuracy of the model. Nevertheless, the effect of cycles can be mitigated by considering the reason why they appear in the graph of interactions. Cycles are generally expected to appear between individuals who belong to the same household, friendship group, or work environment because people belonging to these groups interact more than once, leading to cycles in the graph. By identifying these groups, we can prevent messages from circulating within them. This task can be achieved using the Generalised Belief Propagation (GBP) algorithm introduced by Yedidia, Freeman and Weiss [10].

By adapting GBP to the epidemic modelling problem, we will group nodes that exhibit short cycles into regions of nodes and derive a graph of regions from the factor graph. We can then propagate messages between regions rather than nodes. Such inter-region messages have weaker dependencies because an appropriately designed region graph only has long cycles. In subsection 3.1 we will formally introduce region graphs and justify their equivalence with the factor graphs they are derived from using the concept of free energy from physical systems. We will then proceed with outlining a method to construct such region graphs and an algorithm to perform inference on them, justifying why this modified GBP algorithm is equivalent to the modified BP algorithm.

3.1 Introduction to Region Graphs

This subsection summarises the definitions and fundamentals of region graphs, adapted from [10]. It was deemed necessary to include an introduction to region graphs so that the reader has the appropriate context to understand the construction of a region graph from the graph of interactions in subsection 3.2.

3.1.1 Gibbs Free Energy

Consider a factor graph (V, F, E) where V is the set of variable nodes, F the set of factor nodes and E the set of edges. The probabilistic graph can be viewed as a description of a physical system that is in state $\boldsymbol{s}$ with probability

$$p(\boldsymbol{v} = \boldsymbol{s}) = p(v_1 = s_1, ..., v_{|V|} = s_{|V|}) = \frac{1}{Z} \prod_{f_i \in F} f_i(\boldsymbol{v}_i = \boldsymbol{s}_i)$$

where $\boldsymbol{v}_i$ is the subset of variables in V that are inputs to f_i and $\boldsymbol{s}_i$ the corresponding states. The energy of the system is defined using the Boltzmann distribution, i.e.,

$$p(\boldsymbol{v} = \boldsymbol{s}) = \frac{1}{Q} \exp\left[-\frac{E(\boldsymbol{v} = \boldsymbol{s})}{kT}\right]$$

where we can assume that $kT = 1$ since temperature T has no meaning in this context and can hence take any arbitrary value that will simply scale the energy. We can also assume $Q = Z$ because the value of the normalisation constant Q can be chosen arbitrarily as long as a constant offset is correspondingly added to $E(\boldsymbol{s})$ to make the probability distribution proper. Hence we have

$$p(\boldsymbol{s}) = \frac{1}{Z} \exp\left[-E(\boldsymbol{s})\right] \Leftrightarrow \tag{18}$$

$$E(\boldsymbol{s}) = -\ln\left[Zp(\boldsymbol{s})\right] = -\ln\left[\prod_{f_i \in F} f_i(\boldsymbol{s}_i)\right] = -\sum_{f_i \in F} \ln f_i(\boldsymbol{s}_i). \tag{19}$$

We now define the **Gibbs Free Energy** $G(b)$ with respect to a trial probability distribution $b(\boldsymbol{s})$. It is equal to the difference between the variational average energy $U(b)$ and the variational entropy $H(b)$ which are in turn defined below:

$$U(b) = -\sum_{\boldsymbol{s}} b(\boldsymbol{s}) E(\boldsymbol{s}) \tag{20}$$

$$H(b) = -\sum_{\boldsymbol{s}} b(\boldsymbol{s}) \ln(b(\boldsymbol{s})) \tag{21}$$

The origin of these quantities is beyond the scope of this project and it lies in statistical physics, where $G(b)$ describes the maximum work that can be extracted from an isolated system [21]. The property of $G(b)$ that we are interested in is that it is equal to the KL divergence of $b(\boldsymbol{s})$ and $p(\boldsymbol{s})$ plus a constant that is independent of $b(\boldsymbol{s})$: Starting from the definition of $G(b)$ and using equations (20) and (21) we have

$$
\begin{aligned}
G(b) &= U(b) - H(b) \\
&= -\sum_{\boldsymbol{s}} b(\boldsymbol{s})E(\boldsymbol{s}) - \left(-\sum_{\boldsymbol{s}} b(\boldsymbol{s}) \ln(b(\boldsymbol{s})) \right) \xRightarrow{\text{eq. (19)}} \\
&= -\ln(Z) + \sum_{\boldsymbol{s}} b(\boldsymbol{s})(\ln b(\boldsymbol{s}) - \ln p(\boldsymbol{s})) \\
&= -\ln(Z) + D(b||p).
\end{aligned}
\tag{22}
$$

Hence at the global minimum of $G(b)$, $b = p$. We will proceed to show that every region graph corresponds to an approximation of $G(b)$. We will then show that GBP converges to a local minimum of this approximation with respect to beliefs $b(\boldsymbol{s})$. If the approximation of $G(b)$ is accurate, then the resulting beliefs $\hat{b}(\boldsymbol{s})$ generated by GBP will approximately minimise $D(b||p)$ and $\hat{b}(\boldsymbol{s})$ will be a good approximation of the true distribution $p(\boldsymbol{s})$.

3.1.2 Regions and Valid Region Graphs

We begin with the definition of a region from [10]: A **region** R is a set F_R of factors and a set V_R of all the variables that are neighbouring at least one factor in F_R. We can now define the probability distribution $p_R(\boldsymbol{v}_R)$ of the region, the region energy $E_R(\boldsymbol{v}_R)$, the region average energy $U_R(b_R)$ and the region entropy $H_R(b_R)$, where $b_R(\boldsymbol{v}_R)$ is the corresponding region belief that attempts to approximate $p_R(\boldsymbol{v}_R)$:

$$p_R(\boldsymbol{v}_R) = \frac{1}{Z_R} \prod_{f_i \in F_R} f_i(\boldsymbol{v}_i), \tag{23}$$

$$E_R(\boldsymbol{v}_R) = -\sum_{f_i \in F_R} \ln f_i(\boldsymbol{v}_i), \tag{24}$$

$$U_R(b_R) = \sum_{\boldsymbol{s}_R} b_R(\boldsymbol{s}_R) E_R(\boldsymbol{s}_R) \tag{25}$$

$$H_R(b_R) = -\sum_{\boldsymbol{s}_R} b_R(\boldsymbol{s}_R) \ln b_R(\boldsymbol{s}_R) \tag{26}$$

These definitions follow naturally from the corresponding definitions for the whole factor graph (V, F, E) if one considers a region to be simply a subgraph (V_R, F_R, E_R). Having defined a region, we can now define a region graph.

A **region graph** $(\mathcal{V}, \mathcal{E}, \{c_{1:|\mathcal{V}|}\})$ derived from a factor graph (V, F, E) is a graph whose set of vertices $\mathcal{V}$ consists of regions that are derived from the factor graph. The set of edges $\mathcal{E}$ of the region graph consists of undirected links between regions such that if two regions $R_1 \in \mathcal{V}$, $R_2 \in \mathcal{V}$ have $(V_{R_2} \cup F_{R_2}) \subseteq (V_{R_1} \cup F_{R_1})$, then $\mathcal{E}$ includes a undirected link between R_1 and R_2. Each region R_i is associated with a counting number c_i, which is an integer and is chosen by design. A region graph corresponds to a particular approximation of the free energy $G(b)$ of the factor graph it was derived from. The region-based approximate entropy $H_{\mathcal{R}}$, region-based average energy $U_{\mathcal{R}}$ and region-based approximate free energy

$G_{\mathcal{R}}$ are defined in terms of of the set of beliefs $\{b_R\}$ associated with regions $R \in \mathcal{V}$:

$$H_{\mathcal{R}}(\{b_R\}) = \sum_{R \in \mathcal{V}} c_R H_R, \tag{27}$$

$$U_{\mathcal{R}}(\{b_R\}) = \sum_{R \in \mathcal{V}} c_R U_R, \tag{28}$$

$$G_{\mathcal{R}}(\{b_R\}) = U_{\mathcal{R}}(\{b_R\}) - H_{\mathcal{R}}(\{b_R\}). \tag{29}$$

In order for $G_{\mathcal{R}}(\{b_R\})$ to approximate the factor graph's free energy $G(b)$, the contribution to average energy $U_{\mathcal{R}}(\{b_R\})$ and entropy $H_{\mathcal{R}}(\{b_R\})$ of each node in the factor graph must be balanced. In other words, in calculating free energy each $v \in V$ and each $f \in F$ must be counted once. However, each variable and factor node appears in multiple regions and in different numbers. The counting numbers $\{c_R\}$ ensure that every node in the factor graph is counted only once in the calculation of $H_{\mathcal{R}}(\{b_R\})$, $U_{\mathcal{R}}(\{b_R\})$ and $G_{\mathcal{R}}(\{b_R\})$. For this to happen, the counting numbers must satisfy the following condition:

$$\sum_{\{R \,:\, x \,\in\, V_R \cup F_R\}} c_R = 1 \quad \forall \quad x \in V \cup F, \tag{30}$$

i.e., for every variable or factor x in the factor graph, the sum of the counting numbers of regions that include x must be 1, so that the effect of x is only counted once in the region-based free energy $G_{\mathcal{R}}(\{b_R\})$ given by equation (29). A region graph that satisfies equation (30) is called a valid region graph. To see why it is crucial that a region graph is valid, consider the region-based average energy $U_{\mathcal{R}}(\{b_R\})$ of a valid region graph:

$$\begin{aligned} U_{\mathcal{R}}(\{b_R\}) &= \sum_{R \in \mathcal{V}} c_R U_R \\ &= \sum_{R \in \mathcal{V}} c_R \sum_{\boldsymbol{s}_R} b_R(\boldsymbol{s}_R) E_R(\boldsymbol{s}_R) \\ &= -\sum_{R \in \mathcal{V}} c_R \sum_{f_i \in F_R} \sum_{\boldsymbol{s}_i} b_i(\boldsymbol{s}_i) \ln f_i(\boldsymbol{s}_i) \\ &= -\sum_{f_i \in F} \sum_{\boldsymbol{s}_i} b_i(\boldsymbol{s}_i) \ln f_i(\boldsymbol{s}_i) \sum_{\{R: f_i \in F_R\}} c_R \\ &= -\sum_{f_i \in F} \sum_{\boldsymbol{s}_i} b_i(\boldsymbol{s}_i) \ln f_i(\boldsymbol{s}_i) \\ &= -\sum_{\boldsymbol{s}} b(\boldsymbol{s}) \sum_{f_i \in F} \ln f_i(\boldsymbol{s}_i) \\ &= U(b), \end{aligned} \tag{31}$$

where it has been assumed that region-based beliefs $b_R(\boldsymbol{s}_R)$ are consistent, i.e., there exists a universal joint belief $b(\boldsymbol{s})$ from which every region belief $b_R(\boldsymbol{s}_R)$ was marginalised:

$$\sum_{\boldsymbol{s}_{\bar{R}}} b(\boldsymbol{s}) = b_R(\boldsymbol{s}_R) \qquad \forall \qquad R \in \mathcal{V}, \tag{32}$$

where $\boldsymbol{s}_{\bar{R}}$ is defined such that $\boldsymbol{s} = [\boldsymbol{s}_R, \boldsymbol{s}_{\bar{R}}]^T$. Consequently, we have shown that a valid region graph has the same average energy as the factor graph it was derived from. The same cannot be said about the entropy $H_{\mathcal{R}}(\{b_R\})$. Nevertheless, it is reasonable to expect that $H_{\mathcal{R}}(\{b_R\}) \approx H(b)$ because H is a measure of uncertainty of the beliefs $\{b_R\}$ which are marginalisations of a global belief $b(\boldsymbol{s})$. Hence $G_{\mathcal{R}}(\{b_R\}) \approx G(b)$.

3.2 Construction of Region Graph for Epidemics

We have shown that the free energy $G_{\mathcal{R}}(\{b_R\})$ of a valid region graph is an approximation of the free energy $G(b)$ of the factor graph, which is in turn equal to the sum of the KL divergence $D(b||p)$ and a constant. Consequently, if we minimise $G_{\mathcal{R}}(\{b_R\})$ with respect to b, we expect to reduce $D(b||p)$ and produce a belief that approximates p well. In [10] it is proven that the GBP algorithm converges to a stationary point of $G_{\mathcal{R}}(\{b_R\})$.

Our goal is to find a general method to convert the graph of interactions to a valid region graph whose corresponding free energy $G_{\mathcal{R}}(\{b_R\})$ is a good approximation of $G(b)$. This ensures that stationary points of $G_{\mathcal{R}}(\{b_R\})$ are close to local minima of $G(b)$. We will then derive a simple form of GBP which converges to stationary points of the region graph's free energy $G_{\mathcal{R}}(\{b_R\})$. Depending on how good of an approximation $G_{\mathcal{R}}(\{b_R\}) \approx G(b)$ is, GBP will converge to accurate beliefs. A useful starting point for these tasks is to investigate the structure of the region graph whose free energy is minimised by BP.

3.2.1 BP as an Approximate Free Energy Minimisation Algorithm

In [10], Yedidia et al. reach the insightful conclusion that BP can be interpreted as an iterative algorithm that converges to a stationary point of the Bethe free energy $G_{Bethe}(\{b_R\})$, which is an approximation of free energy that is well known in the physics literature [22]. We elaborate on the corresponding proof introduced in [10] and append a few contributions of our own, which are signified at the end of this subsection.

For a factor graph (V, F, E), $G_{Bethe}(\{b_R\})$ corresponds to a region graph (from now on referred to as the Bethe region graph) that consists of a set $\mathcal{V}_L$ of $|F|$ large regions and a set $\mathcal{V}_S$ of $|V|$ small regions. Each large region encapsulates a factor node $f \in F$ and its neighbouring variable nodes $n(f)$. Each small region encapsulates a variable node $v \in V$. If a small region corresponds to variable v and variable v neighbours factor f, then the small region is connected with an undirected link to the large region that encapsulates f. Figure 7 illustrates the Bethe region graph that was derived from the factor graph of figure 2.

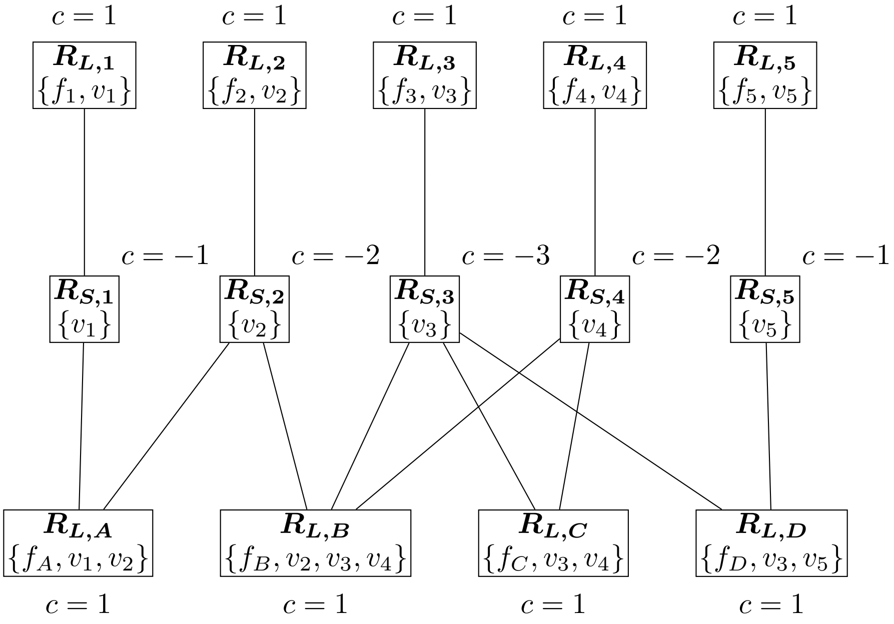


Figure 7: Bethe region graph for the community with 5 individuals.

Notice that the form of the region graph is in direct correspondence with the original factor graph. Factors are replaced by large regions and variables are replaced by small regions. To find the counting numbers $\{c_R\}$, we solve the system of simultaneous equations (30). For the Bethe region graph, it is easy to see that the counting number for all large regions is $c_R = 1$ because each factor appears in exactly one large region. For small regions the counting number c_R is simply $1 - d_R$ where d_R is the number of edges of R. Using equations (27), (28), we calculate the Bethe average energy and the Bethe entropy as follows:

$$\begin{aligned} H_{Bethe}(\{b_R\}) &= \sum_{R \in \mathcal{V}} c_R H_R \\ &= \sum_{R_S \in \mathcal{V}_S} c_{R_S} H_{R_S} + \sum_{R_L \in \mathcal{V}_L} c_{R_L} H_{R_L} \\ &= - \sum_{R_S \in \mathcal{V}_S} (1 - d_{R_S}) \sum_{s_{R_S}} b_{R_S}(s_{R_S}) \ln b_{R_S}(s_{R_S}) \\ &\quad - \sum_{R_L \in \mathcal{V}_L} b_{R_L}(\boldsymbol{s}_{R_L}) \ln b_{R_L}(\boldsymbol{s}_{R_L}), \end{aligned} \tag{33}$$

$$\begin{aligned} U_{Bethe}(\{b_R\}) &= \sum_{R \in \mathcal{V}} c_R U_R \\ &= \sum_{R_L \in \mathcal{V}_L} U_{R_L} \\ &= - \sum_{R_L \in \mathcal{V}_L} b_{R_L}(\boldsymbol{s}_{R_L}) \ln f_{R_L}(\boldsymbol{s}_{R_L}), \end{aligned} \tag{34}$$

where it should be noted that $b_{R_S}(s_{R_S})$ has a scalar input because it is the belief for the unique variable node that resides in R_S. Also, $f_{R_L}(\boldsymbol{s}_{R_L}) = f_i(\boldsymbol{s}_i)$ is simply the function of the unique factor node f_i inside R_L. We can now show that the BP belief (equation (13)) is the solution that minimises $G_{Bethe}(\{b_R\})$ under the following constraints for each region R:

- $b_R(\boldsymbol{s}_R) \geq 0 \; \forall \; \boldsymbol{s}_R$ because the belief is a probability distribution;
- $\sum_{\boldsymbol{s}_R} b_R(\boldsymbol{s}_R) = 1$ because the belief of a region must be a proper distribution;

Moreover, as it was noted in the definition of the region graph (equation (32)), the belief b_R of a region must be a marginalisation of one single belief over all variables, or alternatively for each large region R_L and each of its neighbouring small regions R_S

$$\sum_{\boldsymbol{s}_{R_L} \setminus s_{R_S}} b_{R_L}(\boldsymbol{s}_{R_L}) = b_{R_S}(s_{R_S}). \tag{35}$$

We can hence define a Lagrangian $L(\{b_R\})$ for this optimisation problem:

$$\begin{aligned} L(\{b_R\}) = G_{Bethe}(\{b_R\}) & \\ + \sum_R \sum_{\boldsymbol{s}_R} \alpha_R(\boldsymbol{s}_R) b_R(\boldsymbol{s}_R) & \\ + \sum_R \beta_R (\sum_{\boldsymbol{s}_R} b_R(\boldsymbol{s}_R) - 1) & \qquad (36) \\ + \sum_{R_S \in \mathcal{V}_S} \sum_{R_L \in n(R_S)} \sum_{s_{R_S}} \gamma_{R_S,R_L}(s_{R_S}) \left(b_{R_S}(s_{R_S}) - \sum_{\boldsymbol{s}_{R_L} \setminus s_{R_S}} b_{R_L}(\boldsymbol{s}_{R_L}) \right), & \end{aligned}$$

where $n(R_S)$ is the set of neighbours of R_S (which exclusively consists of large regions). Parameters $\alpha_R(\boldsymbol{s}_R)$, β_R are Lagrange multipliers for the inequality and equality constraint of each region R. They ensure that $b_R(\boldsymbol{s}_R)$ is a valid and proper distribution. Parameters $\gamma_{R_S,R_L}(s_{R_S})$ are equality constraints for each neighbouring pair of regions (R_S, R_L) and each possible state s_{R_S} of the small region.

By differentiating (36) with respect to the belief $b_{R_S}(s_{R_S})$ of a small region and setting it to zero to find a stationary point of the Lagrangian, we have:

$$\begin{aligned} \frac{dL}{db_{R_S}(s_{R_S})} &= \frac{dU_{Bethe}}{db_{R_S}(s_{R_S})} - \frac{dH_{Bethe}}{db_{R_S}(s_{R_S})} + \alpha_{R_S}(s_{R_S}) + \beta_{R_S} + \sum_{R_L \in n(R_S)} \gamma_{R_S,R_L}(s_{R_S}) \\ &= 0 + (1 - d_{R_S})(1 + \ln b_{R_S}(s_{R_S})) + \alpha_{R_S}(s_{R_S}) + \beta_{R_S} + \sum_{R_L \in n(R_S)} \gamma_{R_S,R_L}(s_{R_S}). \end{aligned} \qquad (37)$$

We will constrain our model to always have non-zero beliefs and probabilities because we assume that we can never be completely certain about the state of an individual. Consequently, the inequality constraint is inactive and $\alpha_R(\boldsymbol{s}_R) = 0 \quad \forall \quad R, \boldsymbol{s}_R$. Setting the derivative above to zero, we have:

$$b_{R_S}(s_{R_S}) = \exp\left[\frac{1}{d_{R_S} - 1} \left(\beta_{R_S} + \sum_{R_L \in n(R_S)} \gamma_{R_S,R_L}(s_{R_S}) \right) - 1 \right]. \qquad (38)$$

By repeating the process for a large region R_L, for which $dU_{Bethe}/db_{R_L}(\boldsymbol{s}_{R_L}) = -\ln f_{R_L}(\boldsymbol{s}_{R_L})$, we have

$$b_{R_L}(\boldsymbol{s}_{R_L}) = f_{R_L}(\boldsymbol{s}_{R_L}) \exp\left(-\beta_{R_L} + \sum_{R_S \in n(R_L)} \gamma_{R_S,R_L}(s_{R_S}) - 1 \right). \qquad (39)$$

We can now show that standard BP belief satisfies equations (38), (39) and the equality constraints of the optimisation problem. In other words, the BP belief converges to a stationary point of $G_{Bethe}(\{b_R\})$. Define the variable (small region) to factor (large region) message as $m_{R_S \to R_L}(s_{R_S}) = \exp(\gamma_{R_S,R_L}(s_{R_S}))$ and define also the factor-to-variable message $m_{R_L \to R_S}(s_{R_S})$ using the message-passing rule (11). Consequently equation (38) takes the form

$$\begin{aligned} b_{R_S}(s_{R_S}) &= C_{R_S}\left(\prod_{R_L\in n(R_S)} m_{R_S\to R_L}(s_{R_S})\right)^{\frac{1}{d_{R_S}-1}} \\ &= C_{R_S}\left[\prod_{R_L\in n(R_S)}\left(\prod_{R'_L\in n(R_S)\setminus R_L} m_{R'_L\to R_S}(s_{R_S})\right)\right]^{\frac{1}{d_{R_S}-1}} \\ &= C_{R_S}\left[\prod_{R_L\in n(R_S)}\left(m_{R_L\to R_S}(s_{R_S})\right)^{d_{R_S}-1}\right]^{\frac{1}{d_{R_S}-1}} \\ &= C_{R_S}\prod_{R_L\in n(R_S)} m_{R_L\to R_S}(s_{R_S}), \end{aligned} \tag{40}$$

where for the second step we have used the relationship between variable-to-factor and factor-to-variable messages as defined by BP (equation (11)). For the third step of the derivation, note that the double product counts $d_{R_S}-1$ times each message $m_{R_L\to R_S}$ to small region R_S. Finally, C_{R_S}, C_{R_L} are normalisation constants since beliefs have been constrained to be proper distributions. We can use C_R to calculate the region-specific Lagrange multiplier β_R that ensures $b_R(\boldsymbol{s}_R)$ is proper. Equation (40) is clearly equal to the belief of a variable node as defined by BP (equation (13)). Starting from equation (39), substituting for $\gamma_{R_S,R_L}(s_{R_S})$ and normalising, the belief of a large region becomes

$$b_{R_L}(\boldsymbol{s}_{R_L}) = C_{R_L} f_{R_L}(\boldsymbol{s}_{R_L}) \prod_{R_S\in n(R_L)} m_{R_S\to R_L}(s_{R_S}). \tag{41}$$

Starting from the equality constraint (35) that ensures beliefs between a large region R_L and a neighbouring small region R_S are consistent, we expand its left hand side using equation (41) to get

$$\begin{aligned} \sum_{\boldsymbol{s}_{R_L}\setminus s_{R_S}} b_{R_L}(\boldsymbol{s}_{R_L}) &= C_{R_L}\sum_{\boldsymbol{s}_{R_L}\setminus s_{R_S}} f_{R_L}(\boldsymbol{s}_{R_L})\prod_{R'_S\in n(R_L)} m_{R'_S\to R_L}(s_{R'_S}) \\ &= C_{R_L} m_{R_S\to R_L}(s_{R_S})\sum_{\boldsymbol{s}_{R_L}\setminus s_{R_S}} f_{R_L}(\boldsymbol{s}_{R_L})\prod_{R'_S\in n(R_L)\setminus R_S} m_{R'_S\to R_L}(s_{R'_S}) \\ &= C_{R_L}\left(\prod_{R'_L\in n(R_S)\setminus R_L} m_{R'_L\to R_S}(s_{R_S})\right)\times \\ &\quad \sum_{\boldsymbol{s}_{R_L}\setminus s_{R_S}} f_{R_L}(\boldsymbol{s}_{R_L})\prod_{R'_S\in n(R_L)\setminus R_S} m_{R'_S\to R_L}(s_{R'_S}), \end{aligned} \tag{42}$$

where in the third step we have used the message-passing rule of equation (11). If we now equate (40) and (42) in accordance with the equality constraint described by (35), we come up with a formula for the factor-to-variable message $m_{R_L\to R_S}(s_{R_S})$:

$$\begin{aligned} C_{R_S} \prod_{R'_L \in n(R_S)} m_{R'_L \to R_S}(s_{R_S}) &= C_{R_L} \left(\prod_{R'_L \in n(R_S) \setminus R_L} m_{R'_L \to R_S}(s_{R_S}) \right) \\ &\quad \times \sum_{\boldsymbol{s}_{R_L} \setminus s_{R_S}} f_{R_L}(\boldsymbol{s}_{R_L}) \prod_{R'_S \in n(R_L) \setminus R_S} m_{R'_S \to R_L}(s_{R'_S}) \\ m_{R_L \to R_S}(s_{R_S}) &= \frac{C_{R_L}}{C_{R_S}} \sum_{\boldsymbol{s}_{R_L} \setminus s_{R_S}} f_{R_L}(\boldsymbol{s}_{R_L}) \prod_{R'_S \in n(R_L) \setminus R_S} m_{R'_S \to R_L}(s_{R'_S}), \end{aligned} \tag{43}$$

which is indeed the standard BP factor-to-variable message (equation (12)). We have hence shown that the BP equations yield beliefs $\{b_R\}$ that are at a stationary point of the Bethe free energy $G_{Bethe}(\{b_R\})$ under appropriate equality constraints that ensure the beliefs are valid and consistent probability distributions. We achieved this by defining the variable-to-factor message as the exponential of the Lagrangian multipliers γ and by implicitly defining the factor-to-variable message using the BP variable-to-factor message-passing rule of equation (11).

Earlier it was shown that the factor graph's free energy $G(b)$ is equal to the KL divergence of the true probability distribution p and the belief b. If $G_{Bethe}(\{b_R\})$ is a good approximation of $G(b)$, then the beliefs given by BP will point close to a stationary point of $G(b)$. If this stationary point is a local minimum, then the beliefs b will approximate the true probability distribution p. Consequently, the accuracy of the beliefs depends on how good of an approximation $G_{Bethe}(\{b_R\})$ is to $G(b)$ and on whether the stationary point is the global minimum. In equation (31) we showed that the region-based average energy $U_{\mathcal{R}}(\{b_R\})$ is always equal to the true average energy $U(b)$. However, the region-based entropy $H_{\mathcal{R}}(\{b_R\})$ is not equal to $H(b)$. Hence the accuracy of the beliefs yielded by BP depends on the accuracy of the region-based entropy, which in turn must depend on the design choices of the underlying problem, namely:

- The structure of the factor graph and, in particular, the length of cycles in it.
- The structure of the chosen region graph, which uniquely corresponds to a specific message-passing algorithm.

Moreover, the initialisation of the beliefs before BP is run will determine which stationary point of region-based free energy the algorithm will converge to. In the analysis above, we chose a structure for the region graph that corresponds to the Bethe free energy approximation and the standard BP algorithm. A different region graph structure will correspond to a different message-passing algorithm and free energy approximation. What the analysis above has shown is that choosing a region graph whose region-based free energy accurately approximates the factor graph's free energy guarantees that with proper initialisation the resulting message-passing algorithm will converge to accurate beliefs.

The proof that BP minimises the Bethe approximation of free energy is an adaptation of the proof in [10], with the most notable contributions on our part being the construction of the Bethe region graph in figure 7 from the factor graph in figure 2 and the detailed derivations of equations (40), (41), (42), (43). Finally, the insight drawn in the paragraphs below these equations are also our own contributions that supplement the theory in [10].

3.2.2 Converting the Factor Graph to a Region Graph

We have justified the merit of region graphs as a graphical method to construct an approximation of free energy. We have also seen that BP corresponds to the Bethe region graph, which has the same structure as the factor graph. As the graph of individuals is expected to have short cycles, so will the Bethe region graph, leading to a poor region-based approximation of free energy and consequently inaccurate beliefs. To justify this claim, we return to Gallager's view of BP as an algorithm that diffuses local information, with short cycles "recycling" messages which impede the accuracy of beliefs. With the Bethe graph as our starting point, we will construct a region graph that is free of short cycles. We will then derive the corresponding message-passing equations, which are expected to perform well because they operate on a region graph without short cycles.

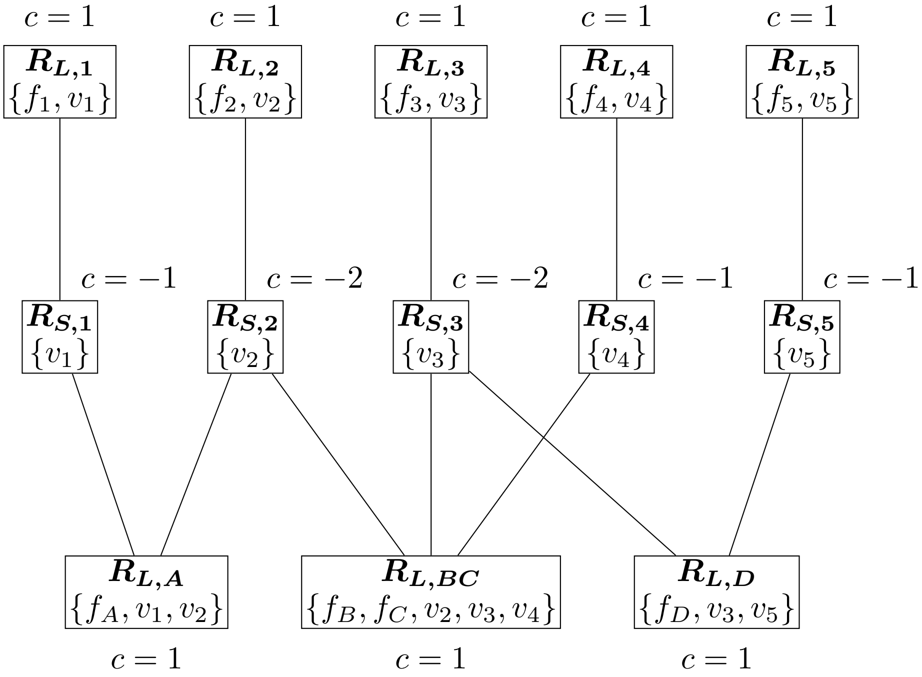
Our goal is to convert the factor graph to a region graph whose regions encapsulate interactions between individuals who interact with one another frequently (thereby causing short cycles in the factor graph). The large regions of this region graph may, for example, represent households, friendship groups or groups of people working in the same office space. Below is an outline of a general method to convert the factor graph (graph of interactions between individuals) to a region graph (graph of interactions among groups) that is free of short cycles:

- **Step 1: Identify short cycles.**
  We will constrain ourselves to cycles of length smaller or equal to 8 because heuristically such cycle lengths are the most problematic. Short cycles may be located by calculating all possible paths of length 4, 6, or 8 in the factor graph and keeping those that begin from and end in the same variable node v_i. Such a naive search may be computationally expensive but not prohibitively so for a graph that models a community of a few hundreds of people.

- **Step 2: Merge large regions.**
  Starting from the Bethe region graph of our epidemic model, the first step is to merge large regions whose factors intercept a short cycle. This is done by replacing the corresponding large regions with a single large region whose set of nodes is equal to the union of the merged regions' sets of nodes. The counting numbers of the new large regions remain $c = 1$. Small regions are unchanged.

- **Step 3: Re-calculate counting numbers $\{c_R\}$ of small regions.**
  Since the number of neighbours of small regions has changed as a result of large regions merging together, we re-calculate the counting numbers for small regions as $c_{R_S} = 1 - d_{R_S}$, where d_{R_S} is the number of edges of R_S in the modified region graph. In this way, we ensure that the region graph is valid.

Figure 8 illustrates the cycle-free region graph that was constructed from the factor graph of figure 2 using the method described above. The factor graph has a cycle of length 4 whose path is $v_3 - f_C - v_4 - f_B - v_3$. Since the cycle is intercepted by factor nodes f_C and f_B in the factor graph of figure 2, we have replaced the large regions $R_{L,B}$ and $R_{L,C}$ in the Bethe region graph (figure 7) with a single large region $R_{L,BC}$ that contains both factors. We hence come up with the region graph of figure 8, which happens to be cycle-free.



$c = 1$
$\boldsymbol{R_{L,1}}$ $\{f_1, v_1\}$
$c = 1$
$\boldsymbol{R_{L,2}}$ $\{f_2, v_2\}$
$c = 1$
$\boldsymbol{R_{L,3}}$ $\{f_3, v_3\}$
$c = 1$
$\boldsymbol{R_{L,4}}$ $\{f_4, v_4\}$
$c = 1$
$\boldsymbol{R_{L,5}}$ $\{f_5, v_5\}$
$c = -1$
$\boldsymbol{R_{S,1}}$ $\{v_1\}$
$c = -2$
$\boldsymbol{R_{S,2}}$ $\{v_2\}$
$c = -2$
$\boldsymbol{R_{S,3}}$ $\{v_3\}$
$c = -1$
$\boldsymbol{R_{S,4}}$ $\{v_4\}$
$c = -1$
$\boldsymbol{R_{S,5}}$ $\{v_5\}$
$\boldsymbol{R_{L,A}}$ $\{f_A, v_1, v_2\}$
$c = 1$
$\boldsymbol{R_{L,BC}}$ $\{f_B, f_C, v_2, v_3, v_4\}$
$c = 1$
$\boldsymbol{R_{L,D}}$ $\{f_D, v_3, v_5\}$
$c = 1$

Figure 8: Cycle-free region graph for the community with 5 individuals.

When applying the method above to a more realistic factor graph that includes hundreds of individuals and interactions, the resulting region graph will be free of short cycles by design but it may still suffer from longer cycles. The chosen threshold L_{max} for the maximum length of cycles we wish to eliminate involves a tradeoff: As L_{max} increases, the computational cost of identifying and eliminating cycles increases. Meanwhile, the additional accuracy in the generated beliefs as a result of eliminating cycles of length L_{max} is small because "recycled" messages have a small impact when the cycle that facilitates them is long.

3.2.3 Applying GBP to the Region Graph

The GBP algorithm, introduced by Yedidia et al. [10], is a general message-passing algorithm that can be applied to any valid region graph. It is the solution of the constrained optimisation of a general region-based approximation of free energy. We start by adapting the GBP equations from [10] to region graphs constructed with the method outlined in subsection 3.2.2 (e.g. figure 8). We come up with the message passing equation $m_{R_S \to R_L}(v_{R_S})$ from a small region to a large region

$$m_{R_S \to R_L}(v_{R_S}) = \prod_{R'_L \in n(R_S) \setminus R_L} m_{R'_L \to R_S}(v_{R_S}) \tag{44}$$

and message $m_{R_L \to R_S}(v_{R_S})$ from a large region to a small region

$$\begin{aligned} m_{R_L \to R_S}(v_{R_S}) &= \sum_{\boldsymbol{v}_{R_L} \setminus v_{R_S}} f_{R_L}(\boldsymbol{v}_{R_L}) \prod_{R'_S \in n(R_L) \setminus R_S} m_{R'_S \to R_L}(v_{R'_S}) \\ &= \sum_{\boldsymbol{v}_{R_L} \setminus v_{R_S}} \prod_{a \in F_{R_L}} f_a(\boldsymbol{v}_a) \prod_{R'_S \in n(R_L) \setminus R_S} m_{R'_S \to R_L}(v_{R'_S}), \end{aligned} \tag{45}$$

where F_{R_L} is the set of factor nodes in region R_L. The belief for a small region is

$$b_{R_S}(v_{R_S}) = \prod_{R_L \in n(R_S)} m_{R_L \to R_S}(v_{R_S}). \tag{46}$$

The message-passing equations are very similar to BP with the only difference being that message $m_{R_L \to R_S}(v_{R_S})$ marginalises the product of multiple factors that reside within the large region. In contrast, when we run BP we only marginalise over one factor because only one factor resides in each large region of the Bethe graph. It can be easily proven that the above message-passing equations converge to a belief that resides at a stationary point of the region-based free energy. The steps of the proof are almost identical to those followed in subsection 3.2.1. Since the region graph generated by the method of subsection 3.2.2 is guaranteed to be free of short cycles, the beliefs generated by equations (44), (45) and (46) will be more accurate than those generated by BP.

This message-passing algorithm may seem like a trivial improvement compared to BP. However, the region graph it operates on is the only region graph that, by design, has the following two properties:

- It is free of short cycles, which are well known to be the cause of inaccurate inference [19].
- It involves message-passing rules that only involve neighbouring regions.

One could construct valid region graphs that are more complex by, for example, merging small regions together or introducing additional regions in such a a way that short cycles are eliminated. Such region graphs would meet the first property but not the second one because the resulting message-passing algorithm would involve messages between non-neighbouring regions, owing to the higher complexity of the region-based free energy optimisation problem. In such a case, Gallager's interpretation of BP as a local propagation of probability collapses because messages are not local anymore. Apart from increasing the computational complexity of the algorithm, non-local messages are of little use to us because one of our primary objectives is to run the message-passing algorithm in a distributed manner to achieve probabilistic contact tracing. This application strictly requires local messages as it will be observed in subsection 4.1.

3.3 Modified GBP Algorithm

We have come up with a systematic method to group together factors which exhibit short cycles into a modified region graph. Moreover, with GBP as our starting point we have derived a message-passing algorithm for this region graph. In order to apply the message-passing algorithm to the problem of epidemic modelling, we make the following two modifications:

- The first modification is identical to that made for BP in subsection 2.2.2: In order for messages to only propagate forward in time, a variable must not propagate messages from future to past interactions. Consequently equation (44) must be modified to the following form:

  $$m_{R_S \to R_L}(v_{R_S}) = \prod_{R'_L \in \{n(R_S): T_{R'_L} < T_{R_L}\}} m_{R'_L \to R_S}(v_{R_S}).$$

  where T_{R_L} can be defined as the timestamp of the latest interaction $f \in F_{R_L}$.

- In order to reflect the close dependence between variables that exhibit a short cycle, the product $f_{R_L}(\boldsymbol{v}_{R_L})$ of factors within a large region must be replaced by a single new factor that is appropriately designed to demonstrate the desired dependence. For the example of figure 8, $f_B(v_2, v_3, v_4)f_C(v_3, v_4)$ is replaced with a new function $f_{BC}(v_2, v_3, v_4)$.

The latter modification eliminates the initial assumption of independence between interactions B and C which was made by defining separate factors f_B and f_C in the factor graph. The form of $f_{BC}(v_2, v_3, v_4)$ must be modelled with a similar process to that followed in subsection 2.1.3. For example, suppose that v_2, v_3, v_4 are coworkers and f_B, f_C represent two occasions in which the coworkers had lunch together in the office. The factor f_{BC} that replaces f_B, f_C can be a more abstract model of the relationship between the three people. For example, it could represent the dependencies between three people who work in the same office space. In other words, what is going to determine the probability of infection of each individual is merely the fact that they work in the same office space and not the number of interactions that they have within it.

3.4 Practical Comparison of BP and GBP: A Nursing Home

We will demonstrate the benefits of the GBP approach by applying it to an example and comparing its results against those yielded by BP. The example will be based on synthetic data that are more complex than those used in subsection 2.3 to compare BP with Monte Carlo. More specifically, the model will have different types of interactions and it will include short cycles. We will model individual interactions using the factor form of equation (17). Due to the increased complexity of this example, it was practically difficult to construct a Monte Carlo simulation that would be equivalent to the graph of interactions. For this reason, we will limit ourselves to a comparison of GBP with BP only.

Consider a nursing home that has 60 residents denoted by variables $\{v_1, ..., v_{60}\}$ and 10 staff members denoted by variables $\{s_1, ..., s_{10}\}$. Each resident participates in three interactions every day:

- **Sharing of room facilities with a roommate ($f_A(v_{2q-1}, v_{2q})$):** Each of the nursing home's 30 rooms has 2 residents, with the q-th room being occupied by v_{2q-1} and v_{2q}. f_A represents the interaction that roommates have as a result of sharing the same living space and bathroom. We will assume that this interaction is the most infectious and describe it with parameters $\alpha_A = 1.3$, $\beta_A = 1$.

- **Assistance by member of staff ($f_B(v_i, s_j)$):** We assume that the nursing home has 10 assistants/nurses, one for every 6 residents. Nurse s_j is responsible for residents $v_{6j-5}, v_{6j-4}, v_{6j-3}, v_{6j-2}, v_{6j-1}, v_{6j}$ and interacts with each resident individually every day. The factor is configured with parameters $\alpha_B = 1.25$,$\beta_B = 1$, making the assumption that interactions with staff are less risky than interactions between roommmates.

- **Recreational activities ($f_C(v_k, v_{k+10}, v_{k+20}, v_{k+30}, v_{k+40}, v_{k+50})$):** We assume that if we pick every 10th resident, we come up with a group of 6 residents that interact with each other during recreational activities, such as exercise. The factor is configured with parameters $\alpha_C = 1.1$, $\beta_C = 1$ because this interaction is assumed to be the least infectious.

The above interactions occur every day for 7 days. We run BP using the method outlined in subsection 2.3. To run GBP, we first use the method outlined in subsection 3.2.2 to convert the factor graph of the community to a region graph that is free of short cycles. Figure 9 illustrates a portion of the factor graph for the first 6 residents in the nursing home. f_{Vi} is the test factor for resident v_i and f_{S1} is the test factor for nurse s_1, who assists the 6 residents. For clarity we have omitted recreational activities f_C from the illustration.

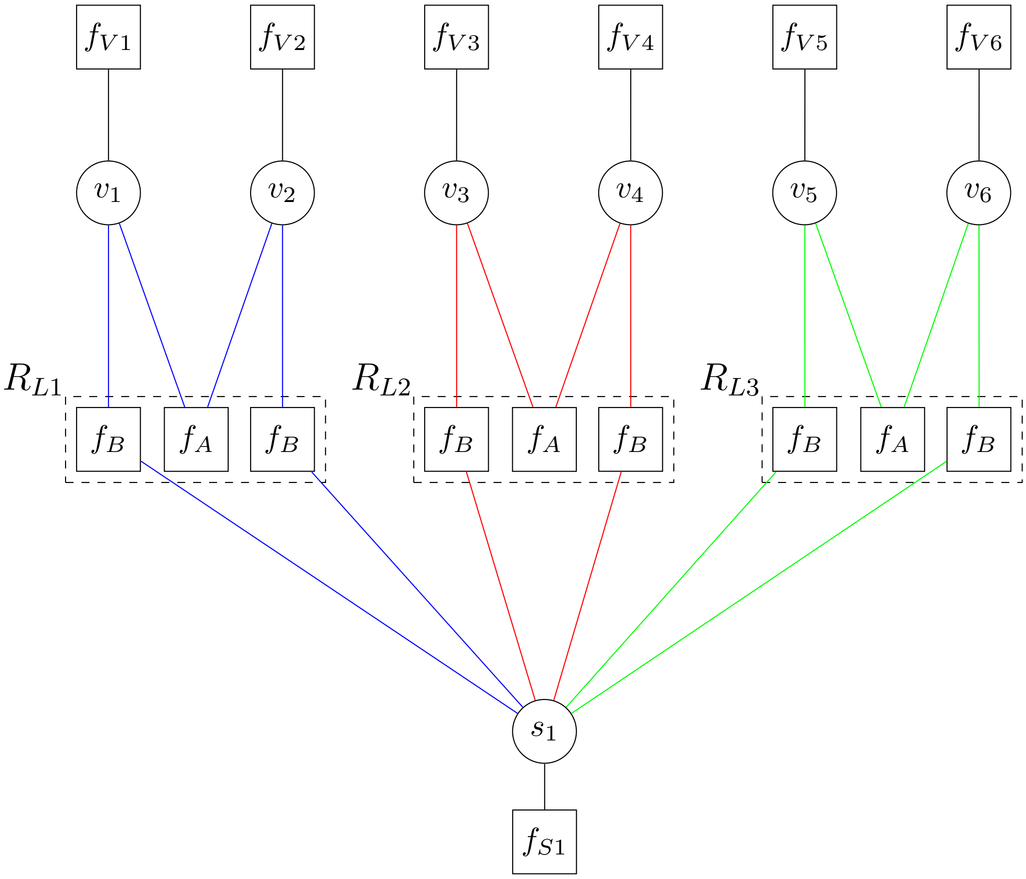
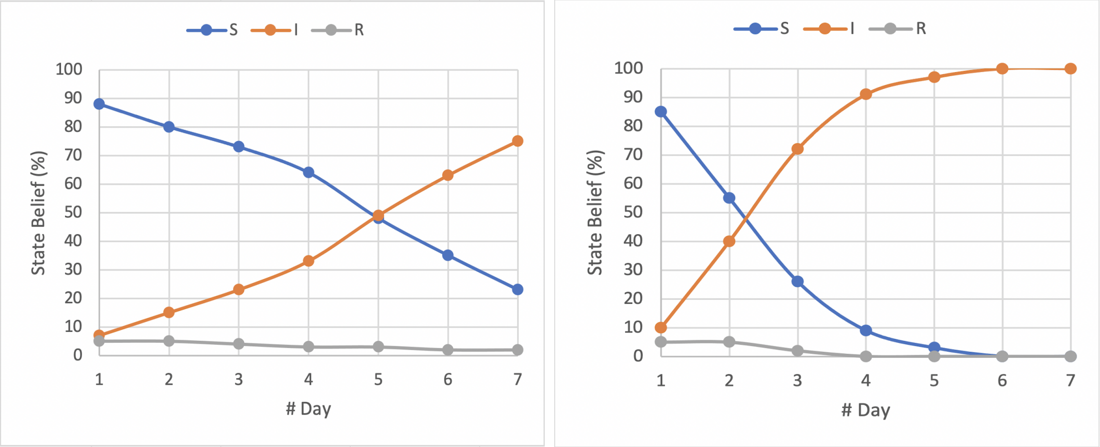


Figure 9: Portion of factor graph that models the nursing home. The first 6 residents are shown, with factors representing the interactions between them and the member of staff that assists them. Recreational activities (factor f_C) are omitted for clarity.

In the factor graph, every consecutive pair of resident nodes suffers from length-6 cycles (with each cycle drawn with a different colour in figure 9). We can eliminate these cycles by encapsulating the interactions that each pair of roommates has with their assigned member of staff and between them. The resulting large regions for the first 6 residents are denoted R_{L1}, R_{L2}, R_{L3} in figure 9. The resulting region graph is free of cycles of length less than or equal to 8. The product $f_A(v_{2i-1}, s_j)f_B(v_{2i-1}, v_{2i})f_A(v_{2i}, s_j)$ in each consolidated region is replaced by a single factor $f_A(v_{2i-1}, s_j, v_{2i})$ which takes as inputs all three variable nodes. Factors f_C representing recreational activities were designed to not cause any short cycles in order to simplify the demonstration of the method. We can now run GBP on the region graph using the message-passing equations of subsection 3.2.3.

We initialise the variables with prior probabilities $[0.9, 0.05, 0.05]$ of being susceptible, infected and recovered respectively. However, 3 randomly picked nursing home residents are set to be ”patients zero” with priors $[0.15, 0.8, 0.05]$, i.e., most likely infected. The message-passing algorithms are run for 10 iterations.

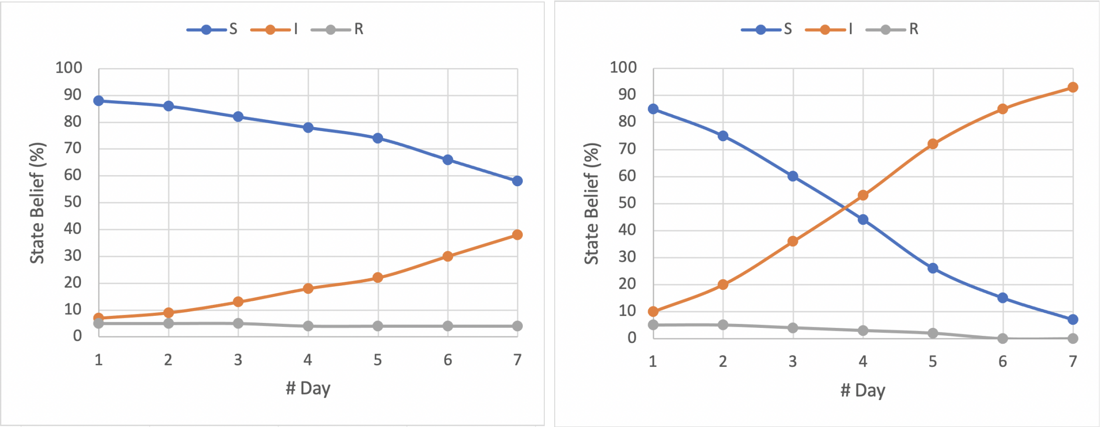
Figure 10a illustrates the marginal beliefs of residents averaged among all residents over time. These are generated by BP and they are calculated at the end of each day for 7 days. Similarly, figure 10b illustrates the marginal beliefs for staff members as generated by BP. Figure 11a, 11b illustrate the same quantities respectively but these are calculated using GBP.



(a) Marginal belief averaged over residents.

(b) Marginal belief averaged over members of staff.

Figure 10: Evolution of marginal beliefs at a nursing home over the course of a week (generated by BP).



(a) Marginal belief averaged over residents.

(b) Marginal belief averaged over members of staff.

Figure 11: Evolution of marginal beliefs at a nursing home over the course of a week (generated by GBP).

As seen in figure 10b, BP predicts that by day 5 with very high probability all staff will be infected. The infection probability among residents has risen at a slower rate. This could be attributed to the fact that a resident interacts regularly with 2 people only, their roommate and their assigned member of staff. In contrast, a member of staff interacts regularly with 6 residents, exposing themselves to thrice the number of people. Notice that the evolution of beliefs for each state bears great resemblance with the evolution of states predicted by the SIR model (figure 1), i.e., both models resemble a transformation of an inverse exponential e^{-t}. However, in our model the population is small (60 residents + 10 members of staff) and the timescale is short (1 week). Consequently, the number of recovered individuals remains negligible during the whole period and the virus spreads throughout the whole population of the nursing home because there is no physical boundary to stop it. This leads to increasing probabilities of infection for residents and staff.

The solution yielded by BP can be directly compared with that of GBP (figure 11) in order to understand how short cycles affect beliefs generated by BP in figure 10. GBP predicts a slower rate of increase in infection probability. On day 7, GBP predicts a probability of infection among residents equal to 0.39 (figure 11a) whereas the equivalent BP prediction is 0.75 (figure 10a). This is expected because GBP does not treat the interaction f_A between roommates and the interactions f_B between a resident and a member of staff as independent events. The interactions between v_1, v_2 and s_1 (figure 9) are treated as one big interaction because the two roommates and their member of staff are connected together via short cycles. Consequently, GBP mitigates the effect of short cycles, which have prompted BP to generate excessively high beliefs of infection. BP assumes that v_1 and s_1 are connected by two paths: v_1-f_B-s_1 and v_1-f_A-v_2-f_B-s_1. It has hence assumed that there are two "pathways" through which v_1 could be infected by s_1. In reality, v_1, v_2 and s_1 interact with each other so often that any additional "pathways" connecting any two of them together should have a negligible effect on their state. GBP accounts for this principle by merging "pathways" together and hence provides a more accurate description of the dependence between v_1, v_2 and s_1.

3.5 Using GBP to Evaluate Disease-Mitigation Policies: Doubling the Nursing Home Staff

To demonstrate the usefulness of modelling a nursing home using GBP, we will now evaluate the effect of a disease-mitigation policy. Suppose that in response to the COVID-19 pandemic, the nursing home doubles the number of staff by introducing volunteers. Each member of staff s_i is now assigned to residents v_{3i-2}, v_{3i-1} and v_{3i}. The factor graph of figure 9 is consequently modified to the form in figure 12.

In figure 12, coloured edges represent the new sets of short cycles in the factor graph. Notice that for the first 6 residents, doubling the staff has reduced the short cycles from 3 to 2. The consolidated large regions R_{L1}, R_{L2} are also shown. Applying GBP on the modified region graph produces the results shown in figures 13a and 13b for residents and staff respectively.

Doubling the staff reduces the number of residents that each member of staff is responsible for. As a result, on day 7 the probability that a resident is infected is 0.31 (figure 13a), which is lower than the previous probability of infection of 0.39 (figure 11a). Doubling staff also means that each member of staff comes into contact with fewer people, leading to fewer paths through which they could become infected. As a result, the probability of infection of a member of staff is 0.51 on day 7 (figure 13b) whereas previously it was 0.93 (figure 11b). According to our model, doubling the number of staff hence reduces the expected number of infections among residents and staff combined by roughly 17%. This is by no means a conclusion we can draw with confidence because our model is based on a series of heuristics and simplifications. Even so, the use of factor graphs to visualise and quantify the interactions within a nursing home provides a useful framework to think of a community as a system of interactions and to evaluate the effect that changes in the system's structure can have on the dynamics of an epidemic within it.

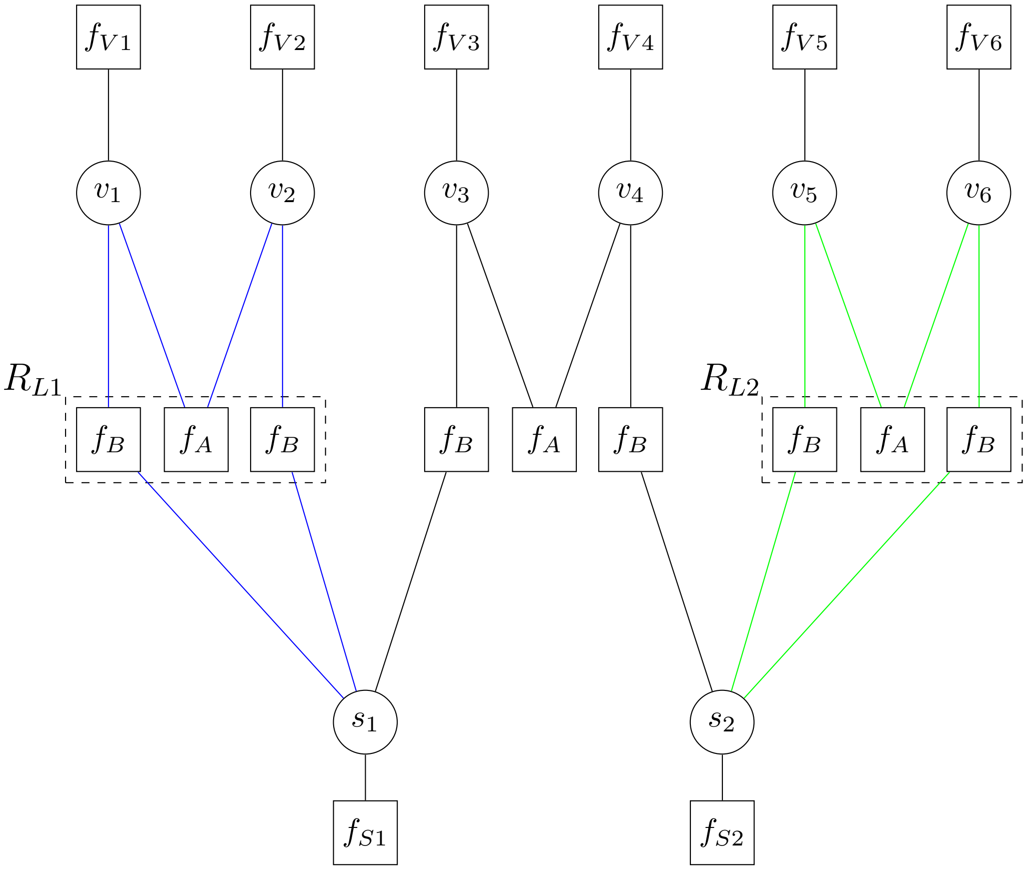
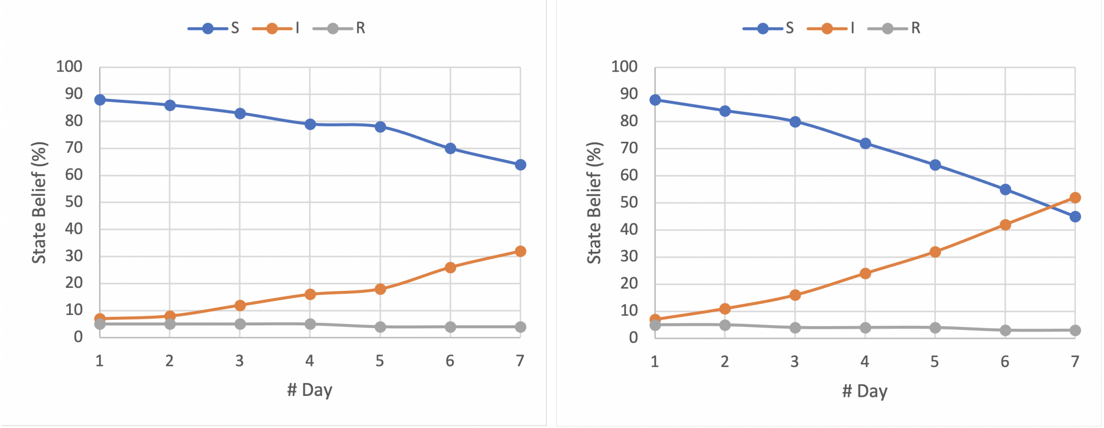


Figure 12: Portion of factor graph modelling the nursing home after doubling the number of staff as part of a disease-mitigation policy.



(a) Marginal belief averaged over residents.

(b) Marginal belief averaged over members of staff.

Figure 13: Evolution of marginal beliefs at a nursing home over the course of a week after doubling the number of staff (generated by GBP).

4 Conclusion

The crucial role technology has played in tackling the COVID-19 pandemic has signified the need for advanced modelling tools that can harness the vast computational power available to trace epidemics and to evaluate the effectiveness of policies that attempt to mitigate them. We strongly believe that the future of epidemic modelling relies on looking at communities at a fine scale and considering the effect each individual interaction has on the evolution of an epidemic.

Starting from the theory behind Belief Propagation on factor graphs, we have constructed a graphical model of interactions within a small community. For diseases that are transmitted through person-to-person contact, we have shown that the graphical model can be used to trace the evolution of an epidemic within the community. Probabilistic inference on the model is achieved using a modified version of BP. In the same way that BP can be interpreted as a local diffusion of probability within a sparse probabilistic model [19], our message-passing algorithm can be interpreted as a local diffusion of probability of disease transmission from one individual to the next.

Through the example of a synthetic Cambridge College, we showed that BP can make predictions about infections within the community that are similar to those of a Monte Carlo simulation, which is the current modelling approach used by epidemiologists [3]. However, it is important to understand the difference between the two modelling approaches: Modelling an epidemic using a probabilistic graphical model allows us to produce a ”simulated simulation” [23] of an epidemic, a term used by Thomas Richardson in 2004 to describe the relationship between density evolution and iterative decoding. Our simulated simulation can be viewed as a level of abstraction above that of a Monte Carlo simulation. The probabilistic model quantifies uncertainty by expressing states in the community with probability distributions instead of simulation outcomes. Whereas in a COVID-19 simulation individuals exchange viral load, in our simulated simulation individuals exchange a probabilistic message that represents the risk of exchanging viral load. A simulated simulation has the following important advantages:

- Expressing inference results in terms of probabilities allows us to directly quantify the risk of infection of each individual in the community using the belief $b(v = I)$. In this way, we predict not only who in the community will become infected but also with what probability this infection will occur. For example, the nursing home model in subsection 3.4 signified with high confidence that the staff are at a much higher risk of infection compared to residents.

- An inference algorithm that is executed with local message-passing can be readily executed in a distributed manner, i.e., using multiple processing units that can communicate with one another. This property is used to design an advanced contact tracing mobile application which we would call Probabilistic Contact Tracing (see subsection 4.1 below).

The GBP theory introduced in [10] motivated an alternative interpretation of Belief Propagation as a free energy optimisation algorithm. It also provided the theoretical tools to create region graphs that do not suffer from the short cycles that factor graphs do. We used the GBP theory to come up with a methodology to convert a factor graph to an equivalent region graph that does not suffer from short cycles. We also derived the message-passing algorithm that this particular region graph is uniquely associated with. The Nursing Home example allowed us to practically compare the new message-passing algorithm with BP. The comparison revealed that the "recycled" messages that BP suffers from lead to overly high probabilities of infection. The region graph representation motivated us to reformulate the problem and remove short cycles. The resulting GBP beliefs are more stable compared to those of BP.

In applications of BP such as LDPC decoding, message-passing is used to approximate static marginals of a factorised joint distribution such as equation (6). For our graphical model, the modification in scheduling so that messages only propagate forward in time means that the message-passing algorithm traces the behaviour of a dynamic system instead of converging towards static marginals. If T_X is the timestamp of the last interaction X in the model, then BP yields beliefs that summarise the epidemiological state of the community (the state of the dynamic system) at $t = T_X$. If we remove from the graph all interactions that have timestamp T such that $T_X \geq T > T_Y$ and then run BP on the pruned graph, the resulting beliefs summarise the system state at $t = T_Y$. Consequently, time evolution is modelled with the help of interaction timestamps and appropriate scheduling of messages.

The lack of real fine-scale data about epidemics was a considerable challenge in the development of our model. Instead, we have resorted to simple statistical metrics and heuristics to produce graphical models that are as close to reality as possible.

Some suggested next steps in the development of this work would be the following:

- Using expert knowledge in epidemiology and sociology, more complex forms of interaction factors could be developed, so that factors are realistic representations of the interactions they correspond to. This requires a good understanding of how different types of interactions affect disease transmission, which can be acquired through experimental work.

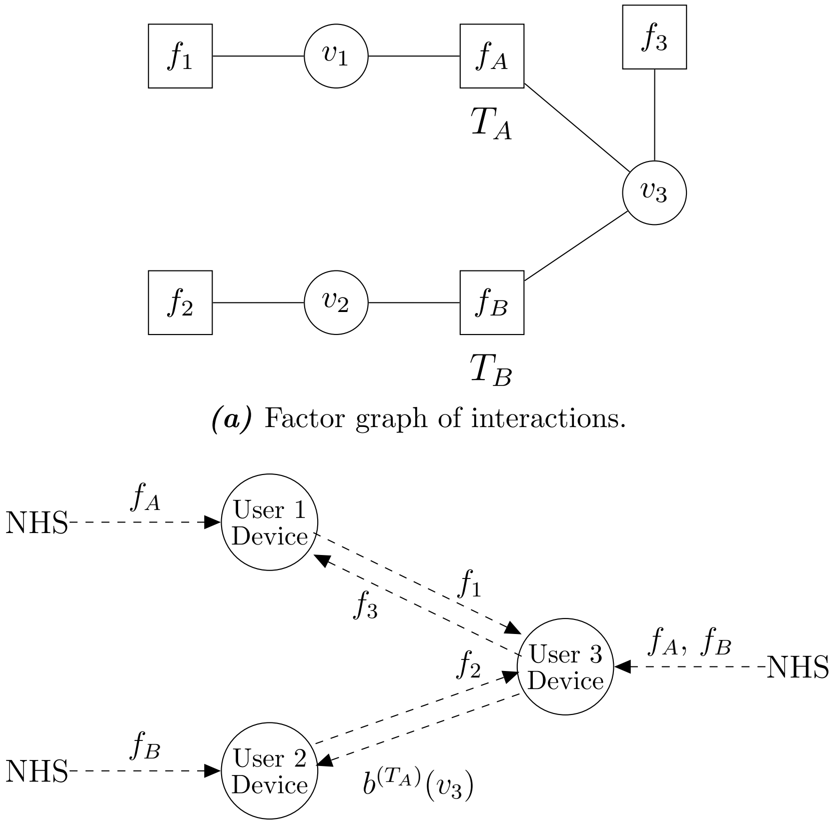
- The method we have presented for converting a factor graph to a region graph that is free of short cycles is general. One could encapsulate interactions in different ways and to different degrees. It is important to further investigate how the choice of region graph structure affects inference on the model of interactions.

- Finally, it would be worthwhile to develop and test a mobile software application that implements our graphical model in a real-time distributed manner, as outlined in subsection 4.1 below. The mobile application would use real data from people's interactions to predict infections, while preserving personal privacy to the same degree that current contact tracing applications do.

4.1 Distributed BP for Probabilistic Contact Tracing

As discussed in subsection 1.1.2, national apps based on the Apple/Google Exposure Notification protocol, such as the NHS Test & Trace App, have only been partially useful in tackling the pandemic [6]. One reason for their limited usefulness is that the apps are inherently unable to quantify a person's risk of exposure. The NHS Test & Trace App locally records the interactions a person participates in and, if in the future someone they interacted with tests positive, the application will notify the user that they may have been exposed. To preserve personal privacy, the app cannot provide information as to when and where the exposure took place. If it could, the user would be able to better evaluate their risk of exposure. For example, if the user was exposed to an infected individual in a closed crowded space, then the risk of being infected is higher compared to having been exposed to an infected individual while walking at the park.

Using the same technological infrastructure as current contact tracing apps do, i.e., mobile phones wirelessly connected at short range, a software application can implement our probabilistic model using real interactions, while preserving a similar degree of privacy as current contact tracing applications do. Upon running BP locally, the mobile application can display a probability of infection for its user, given all their previous interactions. The working principles of this application, which could be named Probabilistic Contact Tracing, will be demonstrated with the help of an example. Consider the factor graph of figure 14a, in which v_1 interacts with v_3 at time T_A and v_2 interacts with v_3 at a subsequent time $T_B > T_A$. Factors f_1, f_2, f_3 are the priors that represent the probability of infection of each individual at the start of the day. Factors f_A, f_B are interaction factors.



(a) Factor graph of interactions.

(b) Equivalent message-passing of mobile devices via Bluetooth.

Figure 14: Demonstration of Probabilistic Contact Tracing using mobile phones.

Figure 14b illustrates the implementation of distributed BP using remote message-passing between mobile phones over Bluetooth. At the start of the day, each device has knowledge of its user's prior. At time T_A:

- Individuals v_1 and v_3 meet. Device 1 owned by individual v_1 receives f_3 from device 3 which is owned by v_3. Similarly, v_3 receives f_1 from device 1. In this way, both devices know both priors.

- Device 1 gathers information on what type of interaction took place between v_1 and v_3. This can be inferred from the current GPS location, the duration for which device 1 was close to device 3 and/or user input. Using this information, device

3 accesses an online NHS database of interaction factor models and downloads the factor f_A that best models the interaction that just took place. With knowledge of f_1, f_3 and f_A, device 1 calculates the belief

$$b^{(T_A)}(v_1) = f_1(v_1) \sum_{v_3} f_A(v_1, v_3) f_3(v_3)$$

at time T_A and informs its user of the current probability that they are infected. Device 3 independently follows the same steps and calculates

$$b^{(T_A)}(v_3) = f_3(v_3) \sum_{v_1} f_A(v_1, v_3) f_1(v_1)$$

using the BP equations (12) and (13).

At the subsequent time T_B:

- Individuals v_2 and v_3 meet. Device 3 sends its latest marginal belief $b^{(T_A)}(v_3)$ to device 2 and device 2 sends factor f_2 to device 3.
- Device 3 gathers information on what type of interaction took place between v_2 and v_3. It then accesses the online NHS database of interaction factor models and downloads the factor f_B that best models the interaction that just took place. Device 2 independently acquires knowledge of f_B by following the same steps. With knowledge of f_1, f_2, f_3, f_A and f_B, device 3 calculates the updated belief at time T_B

$$\begin{aligned} b^{(T_B)}(v_3) &= b^{(T_A)}(v_3) \sum_{v_2} f_B(v_2, v_3) f_2(v_2) \\ &= f_3(v_3) \left(\sum_{v_1} f_A(v_1, v_3) f_1(v_1) \right) \left(\sum_{v_2} f_B(v_2, v_3) f_2(v_2) \right). \end{aligned}$$

It then displays it to its user. Device 2 uses its knowledge of f_2, f_B and $b^{(T_A)}(v_3)$ to calculate

$$b^{(T_B)}(v_2) = f_2(v_2) \sum_{v_3} f_B(v_2, v_3) m^{(T_B)}_{3 \to B}(v_3) = f_2(v_2) \sum_{v_3} f_B(v_2, v_3) b^{(T_A)}(v_3).$$

A similar process is followed for subsequent interactions. In summary, during interaction X, the device of participant v_i transmits its latest belief $b^{(T_X - 1)}(v_i)$ to all neighbouring devices. This belief is equal to the variable-to-factor message $m^{(T_X)}_{i \to X}(v_i)$ at time T_X. Since the form of the interaction factor f_X is known by all participants in X, a participant v_j can combine all the received messages and its knowledge of the interaction factor f_X to update their belief $b^{(T_X)}(v_j)$ using the BP equations (12) and (13).

Notice that the BP algorithm is executed in a fully de-centralised manner. The role of the NHS public library is simply to provide an accurate epidemiological description of different types of interactions. Notice also the privacy-preserving properties of the distributed system: An individual shares information about their infectious state only with the people they interact with. Moreover, each individual is only aware of the interactions they have participated in and no central entity can infer the structure of the complete factor graph. Probabilistic Contact Tracing uses interaction data more effectively compared to the current test and trace applications, while preserving a similar degree of personal privacy. The framework could be used in conjunction with the Apple/Google Exposure Notification system to give individuals a holistic view of their risk of exposure which includes a continuously updated probability of infection, along with notifications of positive contacts.

References

[1] *Access to computers from home.* Nov. 2017. DOI: 10.1787/a70b8a9f-en. URL: https://doi.org/10.1787/a70b8a9f-en.

[2] Rory Cellan-Jones. *Coronavirus: What went wrong with the UK's contact tracing app?* June 2020. URL: https://www.bbc.com/news/technology-53114251.

[3] Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, et al. *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand.* Tech. rep. London: Imperial College COVID-19 Response Team, Mar. 2020. DOI: 10.25561/77482.

[4] European Centre for Disease Prevention and Control. *Latest Evidence: Transmission of COVID-19.* Sept. 2020. URL: https://www.ecdc.europa.eu/en/covid-19/latest-evidence/transmission.

[5] Apple Inc. and Google Inc. *Exposure Notifications Frequently Asked Questions Preliminary-Subject to Modification and Extension.* Tech. rep. Sept. 2020.

[6] Cat Ferguson. *Do digital contact tracing apps work? Here's what you need to know. — MIT Technology Review.* Nov. 2020. URL: https://www.technologyreview.com/2020/11/20/1012325/do-digital-contact-tracing-apps-work-heres-what-you-need-to-know/ (visited on 05/28/2021).

[7] Giulia Giordano et al. "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy". In: *Nature Medicine* 26.6 (June 2020), pp. 855–860. ISSN: 1546170X. DOI: 10.1038/s41591-020-0883-7. URL: https://doi.org/10.1038/s41591-020-0883-7.

[8] M. S. Bartlett. "Measles Periodicity and Community Size". In: *Journal of the Royal Statistical Society. Series A (General)* 120.1 (1957), p. 48. ISSN: 00359238. DOI: 10.2307/2342553.

[9] Gottfried Lechner, Jossy Sayir, et al. "Efficient DSP implementation of an LDPC decoder". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing 2004* (2004). DOI: 10.1109/ICASSP.2004.1326914. URL: https://www.researchgate.net/publication/224750910.

[10] Jonathan S Yedidia, William T Freeman, and Yair Weiss. *Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms.* Tech. rep. Mitsubishi Electric Research Laboratories, 2004.

[11] D. J. Daley and J. Gani. *Epidemic Modelling.* Cambridge University Press, Feb. 1984. DOI: 10.1017/cbo9780511608834. URL: https://www-cambridge-org.ezp.lib.cam.ac.uk/core/books/epidemic-modelling/6F7376322E00A98D6801B97D9429A0CF.

[12] Herbert W Hethcote. *The Mathematics of Infectious Diseases.* Tech. rep. 4. 2000, pp. 599–653. URL: https://epubs.siam.org/page/terms.

[13] David Smith and Lang Moore. *The SIR Model for Spread of Disease - The Differential Equation Model — Mathematical Association of America*. Dec. 2004. URL: https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model (visited on 05/13/2021).

[14] Dr Sabine L. van Elsland and Ryan O'Hare. *COVID-19: Imperial researchers model likely impact of public health measures — Imperial News — Imperial College London*. Mar. 2020. URL: https://www.imperial.ac.uk/news/196234/covid-19-imperial-researchers-model-likely-impact/ (visited on 05/14/2021).

[15] Kiva A. Fisher et al. "Community and Close Contact Exposures Associated with COVID-19 Among Symptomatic Adults 18 Years in 11 Outpatient Health Care Facilities — United States, July 2020". In: *MMWR. Morbidity and Mortality Weekly Report* 69.36 (Sept. 2020), pp. 1258–1264. ISSN: 0149-2195. DOI: 10.15585/mmwr.mm6936a5. URL: https://www.cdc.gov/mmwr/volumes/69/wr/mm6936a5.htm.

[16] Public Health England. *Testing — Coronavirus in the UK*. May 2021. URL: https://coronavirus.data.gov.uk/details/testing?areaType=nation%26areaName=England (visited on 05/15/2021).

[17] Magdalena Szumilas. *Information Management for the Busy Practitioner Explaining Odds Ratios*. Tech. rep. 3. 2010.

[18] Judea Pearl. "Reverend Bayes on Inference Engines: a Distributed Hierarchical Approach." In: *Proceedings of the Second National Conference on Artificial Intelligence* (1982), pp. 133–136.

[19] R. Gallager. "Low-density parity-check codes". In: *IRE Transactions on Information Theory* 8.1 (1962), pp. 21–28. DOI: 10.1109/TIT.1962.1057683.

[20] Yair Weiss. "Correctness of Local Probability Propagation in Graphical Models with Loops". In: *Neural Computation* 12.1 (2000), pp. 1–41. DOI: 10.1162/089976600300015880.

[21] P. Perrot. *A to Z of Thermodynamics*. Supplementary Series; 27. Oxford University Press, 1998. ISBN: 9780198565529. URL: https://books.google.co.uk/books?id=n3hKx7u71%5C_MC.

[22] H A Bethe, H H Wills, and W L Bragg. "Statistical theory of superlattices". In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 150.871 (July 1935), pp. 552–575. ISSN: 0080-4630. DOI: 10.1098/rspa.1935.0122. URL: https://royalsocietypublishing.org/.

[23] Tom Richardson. *Talk at ISIT 2004*. Chicago, 2004.

[24] Aleksei Krasikov. *krashkov/Belief-Propagation*. URL: https://github.com/krashkov/Belief-Propagation.

A Software Implementation of Examples

The starting point for the computer code used to produce the results of the Cambridge College model (subsection 2.4) and the Nursing Home model (subsection 3.4) was an open-source python repository retrieved from GitHub [24], which implements factor graphs and belief propagation. The python library `igraph` is used to define factors, variables and the factor graph as class instances with properties and methods. The original repository was heavily modified so that it implements the modified interaction factors described in subsection 2.1.2 and the modified BP algorithm described in subsection 2.2. Since our GBP algorithm introduced in subsection 3.2.3 has the same formulation as the BP algorithm of subsection 2.2, no additional computer code was required to implement GBP.

B Risk Assessment Retrospective

The initial risk assessment identified eye strain, back problems and loneliness as the three safety hazards associated with this project. Indeed these were the major challenges faced. All three were more intense than anticipated due to the coronavirus restrictions, which led to all university work moving online. This meant that during the 2020-2021 academic year I spent about 8 hours on average doing computer work.

Eye strain and back problems were mitigated with frequent breaks from screen time, ensuring that these breaks were spent in standing position or doing some form of exercise outside when possible. A standing desk was also used during computer work to reduce sitting time. Loneliness was mitigated by scheduling social activities every day while at university or spending time with family while at home.

If this project was to start again, a lot more emphasis would be put on the risk involved with back problems and back pain. I would have bought a second monitor to reduce slouching and would have scheduled daily physical exercises focused on keeping my back, neck and shoulder muscles in good shape.

C COVID-19 Disruption

The project exclusively involved computer work and hence a contingency plan was not necessary. In order to ensure that work progress and communication between myself and my supervisor remained continuous, weekly meetings were set up and online collaboration tools such as Microsoft OneNote, Microsoft OneDrive and Zoom were used during the meetings.

The coronavirus pandemic affected my project work in the same way that it affected all of my other university work. In Michaelmas Term, I was anxious about the health and safety of the older members of my family in Greece, whom I would have been unable to visit if they fell ill. On two occasions my household in Cambridge received a false positive pooled test, which led to considerable disruption for the 5 days we had to self-isolate each time until it was eventually confirmed that the test was indeed a false positive. In Lent Term, national lockdowns in the UK and Greece (my home country) forced me to stay in Greece until May 2021. Being at home, I took an active role in housework and family matters, which reduced the amount of time I could spend on project work and university work in general.