Corona Contact Tracing: Optimal contact inhibition for decelerating the pandemic

Leonard Salewski,* Thomas Hoffmann,* Matthias Blaschke,* and Martin Lellep*

(Dated: 03/22/2020)

Scope: Created during the #WirVSVirus Hackathon of the German government as initiative against the spreading of the COVID-19 pandemic in early 2020.

Open source: Available under `https://github.com/PellelNitram/corona_contact_tracing`.

In this work we develop a graph convolutional approach to predict the health status of all agents in this simulation. It uses past contacts as well as observed health information to derive this prediction. It is able to deal with partial data such as missing locations and missing health status, as it is a probabilistic approach.

Lastly, this prediction can be used to derive measures to reduce the diseases' spread. We propose to find a optimal trade of between removing the least amount of edges in said graph (e.g. through quarantine, social distancing, etc.) and limiting the spread of the disease. Opposed to classical non-pharmaceutical intervention methods such as contact tracing, our approach directly identifies the nodes with the greatest potential to accelerate the diseases' spread in the network.

This technical report was created within the #WirVsVirus Hackathon of the German government and is Work in Progress!

Keywords: COVID-19, #WirVsVirus hackathon, epidemiology, graph theory, mathematical modeling

I. INTRODUCTION

The outbreak of the SARS-COV-2 virus and the associated COVID-19 illness sweep rapidly across the world. Some patients need ventilation support to survive and the exponential growth in infected persons quickly overwhelms any available medical resources [1].

Thus the identification of contact persons is of great importance to control the spread of the disease. Some governments, such as the Singapurian, use location trace data from mobile phone providers. Other approaches are more user centric and build apps that utilise GPS data of individuals.

It is known from past outbreaks and epidemiologic research that such contact tracing and non pharmaceutical interventions (NPI) like school cancellations are important tools to reduce the impact of crisis like the ongoing one. However, both are not particularly directed interventions might come at the cost of an increased social or economic cost.

We formulate a mathematical framework that is suitable to be used for subsequent optimisations which contacts should be avoided while on the other hand decreasing the social and economic cost.

The section II will explain the mathematical foundations that are necessary to understand our main approach presented in section III. Sections IV and V present our working state at the end of the #WirVsVirus hackathon and provide extensive outlook, respectively. In the outlooks, additionally, we propose to use mathematical optimisation to compute very targeted NPIs (e.g. only cancellation of large events) and optimal placement of limited tests (e.g. prioritize potential super spreaders). These perspectives have the potential to control the disease with minimal effects on daily life.

II. MATHEMATICAL BASICS

A. Graphs and graph convolutions

Graphs are sets of nodes connected by edges, as shown in Fig. 1

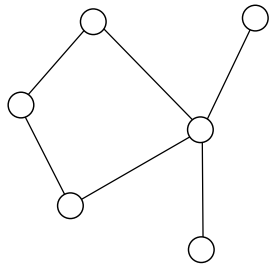

FIG. 1. Example graph where nodes are individuals and edges are their contacts at time step t.

We model partially infected populations as a graph, where each individual (interchangeably called agent) is a node. Edges of this graph model contacts between two agents. The dynamics of infections throughout the population is described by graph convolutions. The definition of a graph convolution [2] for here is

$$h_{v_i}^{(l+1)} = \sum_{j \in A(i)} h_{v_j}^{(l)} \tag{1}$$

* Contributed equally.

with $h_{v_i}^{(l)}$ denoting the feature vector of node i of iteration (l) and $A(i)$ as all neighbours of node i as described by the adjacency matrix A. This formulation is equivalent to the matrix formulation $h^{(l+1)} = Ah^{(l)}$ with A as adjacency matrix as shown in A. We consider here adjacency matrices without diagonal elements.

Each agent i is modeled by D features, $h_{v_i} \in \mathbb{R}^D$. Therefore, the feature matrix, $h^{(l)}$, consists of all agents' features at time (l) and is thereby of dimension $N \times D$ where there are N agents in the population and each agent is described by D features. A three dimensional feature space is used in this work, $D = 3$, modeling three possible health states. The unit vectors of this space are interpreted as following:

- $\vec{e}_0$: susceptible state
- $\vec{e}_1$: infected state
- $\vec{e}_2$: recovered state

A uniform distribution over these possible states expresses complete uncertainty of the health state of an agent.

B. SIR Model

Our basic stochastic SIR model relies on the assumptions that every person in an environment can be modeled as a point value on a grid, which has a location (i.e. GPS coordinates) and an infection state. These states can be either *susceptible* (S), *infected* (I), *recovered* (R) or in advanced models also *under quarantine* (Q) or *dead* (D). All individuals, here called agents, have a probability (here called diffusion rate d) to make a step on the grid per time step on a predefined grid. The movement is a random walk pattern. In the case that some agents meet at the same location, disease spreading can occur. An infected agent spreads the disease with the probability β to all the agents in its close vicinity (same location on the grid). Furthermore, recovery is covered by taking a recovery rate into account, i.e. a probability γ to recover from the disease per time step. If an infected agent recovers from the disease, the state of the agent changes from *infected* to *recovered*, which is definite (no double infections). The process ends, when no infected agents are left. For more background the reader is reffered to [3] and [4].

III. GRAPH-BASED FRAMEWORK

Our novel graph-based approach to model the dynamics of the agents' health states incorporates communal effects and temporal decease effects through a graph contribution and an individual health contribution term, respectively.

Our main propagation rule is based on the definition of a graph convolution shown in Eq. (1) and reads as follows in component notation

$$h_{v_i,m}^{(l+1)} = \underbrace{\sum_k \frac{\hat{A}_{v_i,k}^{(l)}}{\sum_j \hat{A}_{v_i,j}^{(l)}} h_{k,m}^{(l)} \delta_{m,1}}_{\text{Graph}} + \underbrace{(h_{v_i}^{(l)} \cdot T)_m}_{\text{Temporal}} \tag{2}$$

with m being the health state index ranging from 0 to 2, $\hat{A}_{v_i,k}^{(l)}$ infection-adjusted adjacency matrix component, δ the Kronecker delta and T as temporal transition matrix.

The propagation consists of two parts, first the graph contribution and second the temporal contribution. While the former captures the dynamics of infections based on the social contacts between agents, the former ensures that an infected agent heals over time and becomes resistant against the Corona virus. Figure 2 visualises the propagation rule and the two subsequent sections explain the terms in greater detail.

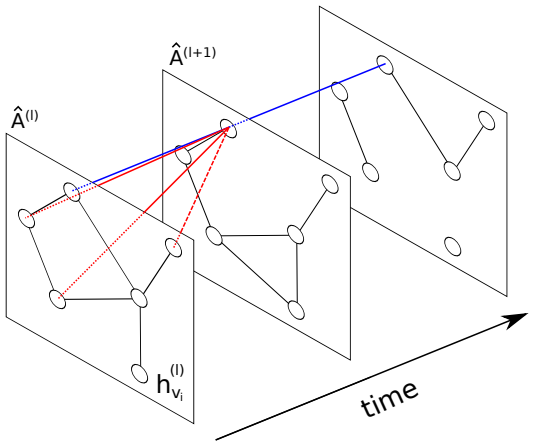

FIG. 2. Visualisation of propagation rule from equation (2). The blue connection visualises the temporal term and the red connections visualise the graph term. The dashed red line is does not contribute to the graph term as the two nodes are not connected according to $\hat{A}(l)$.

A. Explanation of the graph contribution term

The graph contribution models how infected agents spread the disease through contacts with susceptible agents. This is modeled by the term

$$(h_{v_i,m}^{(l+1)})_{\text{Graph}} = \sum_k \frac{\hat{A}_{v_i,k}^{(l)}}{\sum_j \hat{A}_{v_i,j}^{(l)}} h_{k,m}^{(l)} \delta_{m,1}. \tag{3}$$

A sum over all neighbouring agents' features $h_{k,m}^{(l)}$ is weighted by the normalised infection-adjusted graph connections as shown in cyan. The Kronecker delta, as shown in orange, ensures that only the I feature is added

as this is the only one that matters during social contacts between agents.

The infection-adjusted adjacency matrix $\hat{A}$ is constructed from A and I which are the regular continuous adjacency matrix and the infection matrix, respectively. These three quantities are explained in the following:

- The adjacency matrix A is time dependent, $A^{(l)}$, and inferred from data. In our use case, $A_{ij} = \frac{1}{dist(v_i,v_j)+\epsilon}$, hence A_{ij} is large when persons i and j have been in contact. ϵ serves as regularization for small distances.

- The infection matrix is constructed as

$$I = \begin{pmatrix} 0 & 0 & 0 \\ \beta & 0 & \alpha \\ 0 & 0 & 0 \end{pmatrix} = (I_{ij})_{i,j} \tag{4}$$

  with i as the index of the host state and j is the index of the contact person state. The states that we consider here are ordered as follows: susceptible, infected, recovered. β denotes the probability of infection after contact (also known as attack rate). α models the probability of being reinfected, which we assume to be zero ($\alpha = 0$) based upon current medical [5].

- $\hat{A}$, with $\hat{A}_{ij} \in [0, 1]$, is the infection-adjusted adjacency matrix that takes the infection interactions into account and is computed as follows

$$\hat{A}_{ij} = A_{ij} \cdot \frac{h_{v_1}^T I h_{v_2} + h_{v_2}^T I h_{v_1}}{\beta}. \tag{5}$$

  The weighted scalar product of the health states of agents i and j is used to evaluate whether the edge is relevant for the infection dynamics. Only when an infected person and a susceptible have contact, the edge A_{ij} should be considered, otherwise it should be dropped. The sum in the denominator comes from the fact that both, agent i and j, can act as host during a contact. The division by β normalises the factor to one to ensure $\hat{A}_{ij} \in [0, 1]$. Since I is not symmetric, p_a is a proper normalization because the sum is in $\{0, p_a\}$. Note that the fraction has the desired properties for pure S-, I- and R-persons.

Figure 2 visualises the influence of the infection-adjusted adjacency matrix to the agents' states at the next time step. The two solid red lines contribute directly while the dashed red lines does not.

1. *Explanation of the temporal contribution term*

The transition of a person's health state $h_{v_i}^{(l)}$ is determined by three rules that are stated in the following:

- A susceptible person always stays susceptible.

- An infected person has a probability γ, called recovery rate, to recover. The complementary probability $1 - \gamma$ denotes that the person remains sick.

- A recovered person could have a probability to be re-infected, but we assume this to be zero throughout this work. Thus a recovered person always stays recovered [5].

These three rules are combined into a temporal transition matrix T, which describes the health state of an agent as time passes. This matrix reads as

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-\gamma & \gamma \\ 0 & 0 & 1 \end{pmatrix}. \tag{6}$$

The temporal update rule based on the health status thus becomes

$$H^{(l+1)} = H^{(l)} T. \tag{7}$$

Figure 2 visualises the influence of the temporal component as red line connecting agents at subsequent time steps.

IV. CURRENT WORKING STATE

A. SIR Model

For testing and data generation we have implemented a SIR Model using Python. The code can be found in the given Github repository.
An example of an early model state is shown in Figure 3.

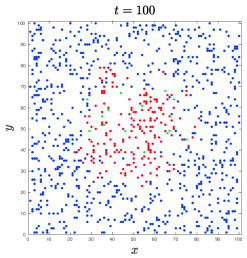

FIG. 3. Early state of SIR model, blue dots = susceptibles, red dots = infected agents, green dots = recovered agents.

In this case, 1000 agents were initialized on a 100 by 100 grid, where a certain amount of infected agents were introduced as a seed. Letting the agents perform random walks on the grid (maximally one step on the grid each time step with probability d) and letting the model converge, the results in Figure 4 can be observed.

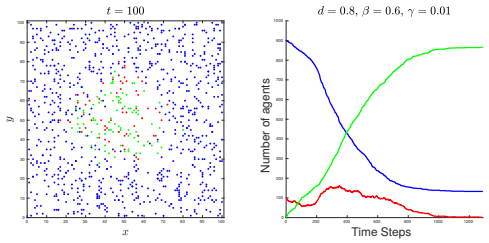

FIG. 4. Plot of the proportions of susceptible (blue), infected (red) and recovered (green) individuals in each state over time.

With the above stated parameters ($d = 0.8$, $\beta = 0.6$, $\gamma = 0.01$), the disease does not spread over the whole population. However more than 80% were infected over time, which could be a very likely scenario of the corona outbrake.

B. Improving the Simulation

Random walk is commonly used for simple Simulations. However, this movement pattern does not represent the daily routine of the majority of citizens. It is more likely that the major part of the population moves around their hometown and travels small distances which leads to clustering of the whole population. To improve the simulation, the agents are split into two groups. One group moves random walk like. This group represents e.g. service providers and delivery services. The other group moves around given points. This movement is implemented using polar coordinates with random radius and random angle. The movement pattern of the two groups is illustrated in Figure 5.

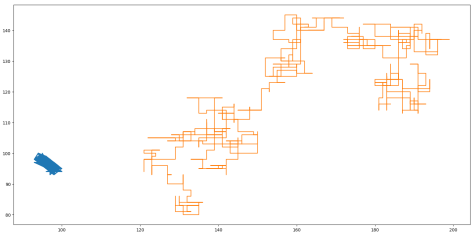

FIG. 5. Movement pattern of the two groups. Random walk (orange) and localized movement (blue).

With these two groups, new features such as multiple outbreaks in different clusters can be observed. An example of the improved model is shown in Figure 6. In this case, 3000 agents were initialized on a 200 by 200 grid with the infection parameters $\beta = 0.6$ and $\gamma = 0.01$. 70% of the agents move in clusters. The infection waves at $t = 40$, $t = 210$ and $t = 650$ can be explained by the infection of a new cluster. The second infection wave is illustrated in Figure 7. The whole simulation can be found as a video `https://youtu.be/c7ehtN-1n9w`.

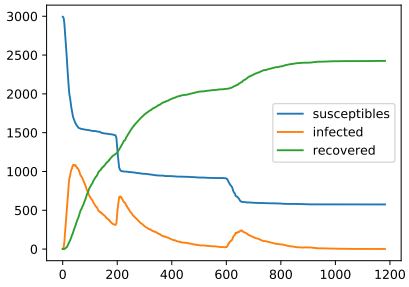

FIG. 6. Plot of the proportions of susceptible (blue), infected (orange) and recovered (green) individuals in each state over time.

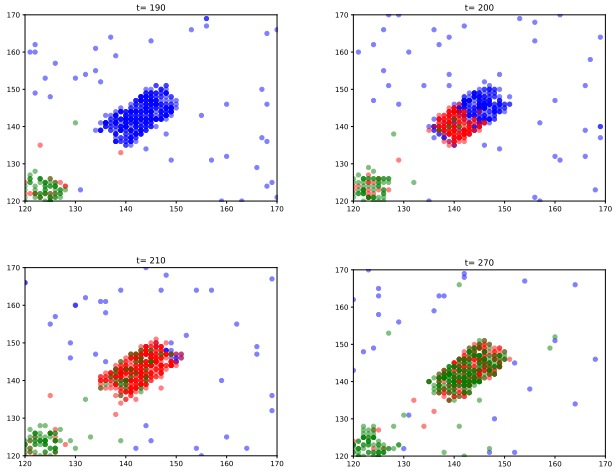

FIG. 7. Second infection wave from figure 6. Blue: susceptibles, red: infected, green: recovered. $t = 190$ cluster gets infected (top left), $t = 200$ disease spreads (top right), $t = 210$ most of the cluster is infected (bottom left) and $t = 270$ cluster recovers (bottom right).

V. OUTLOOK

In the following we will present two extensions to the above methods. First we demonstrate a baysian approach to estimate the populations' health stte. Second we show how the above methods could be utilized to derive disease containment policies.

A. Baysian way of estimating the Probabilities

A major drawback of our analysis is that we only take fixed disease states of individuals into account, i.e either

susceptible, infected or recovered. However, in real live, due to the lack of data, these states are mostly unknown and have to be estimated by a probability distributions. If we model these probabilities based on detailed assumptions, we can use Bayes' posterior probability to update these distributions based on the individual interaction of people. Thus, we define X as the random variable that represents the state of a single individual. It can be assumed, that X is well approximated by the multinomail distribution ($n \gg 1$):

$$X = (X_S, X_I, X_R) \sim \mathrm{Mn}(n; p_S^{(t)}, p_I^{(t)}, p_R^{(t)}) \tag{8}$$

where n is the population size, and $p_S^{(t)}$, $p_I^{(t)}$, and $p_R^{(t)}$ are the probability for an individual to be either in susceptible, an infected or a recovered state. These probabilities change over time as the disease progresses, while

$$p_S^{(t)} + p_I^{(t)} + p_R^{(t)} = 1 \tag{9}$$

As we can only estimate the probabilities $p_S^{(t)}$, $p_I^{(t)}$, and $p_R^{(t)}$ from samples of our population. Therefore, we assume that the probabilities are taken from a prior distribution, in this case the Dirichlet distribution:

$$\left(p_{S,i}^{(t)}, p_{I,i}^{(t)}, p_{R,i}^{(t)}\right) \sim \mathrm{Dir}\left(\alpha_{S,i}^{(t)}, \alpha_{I,i}^{(t)}, \alpha_{R,i}^{(t)}\right). \tag{10}$$

The parameters $\alpha_{1,i}^{(0)}$, $\alpha_{2,i}^{(0)}$, and $\alpha_{3,i}^{(0)}$ can be estimated from $p_S^{(0)}$, $p_I^{(0)}$, and $p_R^{(0)}$ and are specific for each individual i.

If we now assume that two individuals with unknown state meet at the same location, an interaction occurs and we cannot neglect, that at least one of the individuals may carry the disease and may potentially infect the other. Therefore, we can assume, that each individual is present in an infectious superstate $\langle x \rangle$ where

$$\langle x \rangle = \left(\langle x_{S,i} \rangle, \langle x_{I,i} \rangle, \langle x_{R,i} \rangle\right) = \left(p_S^{(t)}, p_I^{(t)}, p_R^{(t)}\right) \tag{11}$$

If the state of an individual in this group known, due to a previous test for example, the state collapses into the right state (e.g $(0, 1, 0)$ for an infected individual).

Based on this state, we can define the graph based Bayesian update rules, where the infection rate β is taken into account, as followed [6]:

$$\alpha_{S,i}^{(t+1)\prime} = \alpha_{1,i}^{(t)\prime} \tag{12}$$

$$\alpha_{I,i}^{(t+1)\prime} = \alpha_{2,i}^{(t)\prime} + \beta \cdot \sum_{j \in A_{v_i}} x_{I,j} \tag{13}$$

$$\alpha_{R,i}^{(t+1)\prime} = \alpha_{3,i}^{(t)\prime} \tag{14}$$

$$\alpha_{0,i}^{(t+1)\prime} = \alpha_{S,i}^{(t+1)\prime} + \alpha_{I,i}^{(t+1)\prime} + \alpha_{R,i}^{(t+1)\prime} \tag{15}$$

The update is only done for the α_I as only the possible infection of an individual can change the infection state of another individual.

However, this model still does not take the potential recovery into account. We can add this by looking at the estimated Values of $p_{S,i}^{(t+1)}$, $p_{I,i}^{(t+1)}$, and $p_{R,i}^{(t+1)}$

$$\mathrm{E}[p_{S,i}^{(t+1)}] = \frac{\alpha_{S,i}^{(t+1)\prime}}{\alpha_{0,i}^{(t+1)\prime}} = \frac{\alpha_{S,i}^{(t+1)}}{\alpha_{0,i}^{(t+1)\prime}} \tag{16}$$

$$\mathrm{E}[p_{I,i}^{(t+1)}] = \frac{\alpha_{I,i}^{(t+1)\prime}}{\alpha_{0,i}^{(t+1)\prime}} - \gamma = \frac{\alpha_{I,i}^{(t+1)}}{\alpha_{0,i}^{(t+1)}} \tag{17}$$

$$\mathrm{E}[p_{R,i}^{(t+1)}] = \frac{\alpha_{R,i}^{(t+1)\prime}}{\alpha_{0,i}^{(t+1)\prime}} + \gamma = \frac{\alpha_{R,i}^{(t+1)}}{\alpha_{0,i}^{(t+1)}} \tag{18}$$

Based on this equation system, we can calculate $\alpha_{k,i}^{(t+1)}$ for $k \in S, I, R$. Needless to say, that if the expectation value of $p_{I,i}^{(t+1)}$ or $p_{R,i}^{(t+1)}$ return a value smaller than 0 or larger than 1, we assume $\gamma = 0$ and therefore $\alpha_{k,i}^{(t+1)\prime} = \alpha_{k,i}^{(t+1)}$. Something similar may be found in [7]

B. Policy Design

In the previous sections we have shown how predict the health state of all individuals in a location tracked population. The next step is to use this information to compute optimal policies. These policies guide a governments' and societies' response to the spread of COVID-19 and modify how the disease is able to spread in a population.

An optimal policy always keeps the infection counts below the medical systems' capacity and the total number of infections as small as possible. Mathematically speaking we want to minimize the sum of all future infections

$$\min \sum_{\forall t} N_i^{(t)} \tag{19}$$

whilst

$$N_i^{(l)} \leq N_{limit}, \forall t \tag{20}$$

A policy may consist out of two different kinds of actions, non pharmaceutical interventions (NPI) and test prioritization (TP).

1. *Non Pharmaceutical Interventions (NPI)*

All non pharmaceutical interventions can be understood as some kind of edge removal in our graph-based approach:

- Isolation of an infected individual removes all of its edges with very high probability.

- Quarantine of a contact person removes all of its edges with high probability.
- Social distancing removes some edges of many individuals.
- Cancellation of large events remove many edges of many individuals.

Formally speaking, the square matrix $C \in \mathbb{B}$ with dimensions $N \times N$ models desirable edge cancellations. This is the policy, which will be optimized. To avoid the trivial solution of $C = 0$, the cancellation of all edges, we also want to minimize the the number of cancellations

$$\min_C -\sum_i \sum_j C_{ij} \tag{21}$$

Note, that this matrix does not have to know the edges of a future time step, it only expresses which edges must not exist. It is multiplied element-wise onto the adjacency matrix A to obtain the adjacency matrix with applied cancellations $\bar{A} = A \odot C$.

To obtain an optimal cancellation policy one thus must jointly minimize eqs. (19) and (21) whilst fulfilling eq. (20).

2. *Test Prioritization (TP)*

When tests are limited, we argue that they should be used to discover as much as possible about the health state of the overall population. This in turn allows non pharmaceutical interventions such as school cancellations to become more efficient. Currently there are only rough medical-based guidelines who should be tested and who should not.

Lets assume there are $t_{\max}$ tests per time step. A test reveals the true health state of an individual (ignoring false negatives and false positives)

$$h_{v_i}^{(t)} \xrightarrow{\text{test}} h_{v_i}^{(t+1)} \in \{\vec{e}_0, \vec{e}_1, \vec{e}_2\} \tag{22}$$

The test assignment T with dimension N is a binary variable describing which individuals should be tested.

Appendix

Appendix A: Consistency of notations

Using the notation from the blog and the paper, including $H' \in \mathbb{R}^{N\times D}$, $A \in \mathbb{R}^{N\times N}$, $H \in \mathbb{R}^{N\times D}$, the propagation rule is:

$$H' = AH \tag{A1}$$

is equivalent to

$$\begin{aligned} H'_{v_i,m} &= \sum_l \overbrace{A_{v_i,l}}^{\in\{0,1\}} H_{l,m} \\ &= \sum_{l'} H_{l',m} \end{aligned} \tag{A2}$$

where l' takes all $A_{v_i,l} = 1$, i.e. neighbours, into account.

Hence, I do understand that the notations are equal.

Appendix B: Formulate normalisation

Use from ICLR 2017 paper the normalisation $D_{ii} = \sum_j \hat{A}_{ij}$.

Use the knowledge from the blog post to write:

$$\begin{aligned} \overbrace{H'}^{\in\mathbb{R}^{N\times D}} &= \overbrace{D^{-1}}^{\in\mathbb{R}^{N\times N}} \overbrace{\hat{A}}^{\in\mathbb{R}^{N\times N}} \overbrace{H}^{\in\mathbb{R}^{N\times D}} \\ \Leftrightarrow H'_{v_i,m} &= \sum_k \overbrace{D^{-1}_{v_i,k}}^{=D_{v_i,v_i},\delta_{v_i,k}} (\hat{A}H)_{km} \\ &= D^{-1}_{v_i,v_i} (\hat{A}H)_{v_i,m} \\ &= D^{-1}_{v_i,v_i} \sum_k \hat{A}_{v_i,k} H_{k,m} \\ &= \sum_k \underbrace{\frac{\hat{A}_{v_i,k}}{\sum_j \hat{A}_{v_i,j}}}_{\text{normalised}} H_{k,m}. \end{aligned} \tag{B1}$$

The weighting by the adjacency matrix is, indeed, normalised to its column sums.

Note that this part is also normalised:

$$\begin{aligned} \sum_m H'_{v_i,m} &= \sum_m \sum_k \frac{\hat{A}_{v_i,k}}{\sum_j \hat{A}_{v_i,j}} H_{k,m} \\ &= \sum_k \frac{\hat{A}_{v_i,k}}{\sum_j \hat{A}_{v_i,j}} \underbrace{\sum_m H_{k,m}}_{=1,\text{ per construction}} \\ &= 1 \end{aligned} \tag{B2}$$

[1] Z. Wu and J. M. McGoogan, JAMA (2020), 10.1001/jama.2020.2648, https://jamanetwork.com/journals/jama/articlepdf/2762130/jama_wu_2020_vp_200028_1.pdf.

[2] T. N. Kipf and M. Welling, CoRR **abs/1609.02907** (2017).

[3] H. H. Weiss, Materials matematics , 0001 (2013).

[4] J. M. Epstein, Nature **460**, 687 (2009).

[5] L. Bao, W. Deng, H. Gao, C. Xiao, J. Liu, J. Xue, Q. Lv, J. Liu, P. Yu, Y. Xu, F. Qi, Y. Qu, F. Li, Z. Xiang, H. Yu, S. Gong, M. Liu, G. Wang, S. Wang, Z. Song, W. Zhao, Y. Han, L. Zhao, X. Liu, Q. Wei, and C. Qin, bioRxiv (2020), 10.1101/2020.03.13.990226, https://www.biorxiv.org/content/early/2020/03/14/2020.03.13.990226

[6] J. A. Rice, *Mathematical statistics and data analysis* (Cengage Learning, 2006).

[7] O. Stojanović, J. Leugering, G. Pipa, S. Ghozzi, and A. Ullrich, PloS one **14** (2019).