Proposed Epidemiological Models and Methods to Improve COVID-19 Contact Tracing and Outbreak Prediction

Jack Hester[a]

[a] *safetraceapi.org*
Email: colab@jackhester.com

Abstract

In the case of pandemics—such as the current COVID-19 pandemic—disease spreads rapidly and testing resources are limited. Contact tracing provides an opportunity to fill some of the gaps left by testing limitations. Individuals with confirmed and presumed or suspected positive cases can be identified via user input of testing status or symptoms, as well as by other data sources. The location and movement of these individuals can then be traced. This location information can be used to create a 4D model of locations over time, and the data can be input into a network model to identify other individuals who were likely exposed as well as identify where or among which groups the next outbreaks are likely to occur. This paper describes a few central models based in epidemiology, graph theory, and population biology that can be used, along with location, case, and symptomatology data to create an effective contact tracing algorithm. It then proposes a SEIR model for tracing infections through a network and discusses limitations and options for further fine-tuning. Such an algorithm has immense potential to drastically improve epidemiological studies, reduce cases, and prioritize interventions during COVID-19 and future epidemics. The algorithm will be implemented via the Safetrace API, part of the MutualAid.world project.

Keywords: contact tracing, COVID-19, SEIR model, network dynamics, SafeTrace API

Goals of Contact Tracing

Contact tracing generally refers to identification of individuals that likely have likely been in close contact with someone infected with a virus or infection, such as SARS-CoV-2.[1] Contact tracing, therefore, allows us to determine individuals that are at a higher risk of becoming infected. These individuals can then monitor for symptoms and isolate themselves to limit secondary spread. Effective contact tracing has many benefits, including:

- slowing and interrupting viral transmission
- better targeting and rationing of test kits
- identifying groups (networks) of individuals that may have a high density of infections at some future time point
- alerting individuals that they have been in contact with the disease
- identifying asymptomatic individuals, including those who have recovered
- suggesting that exposed individuals monitor for symptoms and/or isolate
- prioritizing locations for sanitation
- improving epidemiological models and key parameters of the viral spread

Probabilities in relation to model parameters

Some basic concepts from probability are important to consider as the models are built. This section provides a brief review of these concepts for those without a mathematical background and explains how they might be used in parts of the model designs for contact tracing.

[1] SARS-CoV-2 is the virus that causes the disease, and COVID-19 is the disease itself

Parameters following the binomial distribution

For many purposes, it is useful to consider individuals as either becoming infected or not. While the SEIR model presented later groups individuals into four potential states (see the SIR model section below), the parameters of such a binomial distribution are useful when considering the basis of R_0 values.

The overall distribution of individuals' infection status, given this simplified approach, can be modeled as $X \sim Bin(n, p)$ where X represents the distribution of infected individuals, n represents the total number of individuals in the network (a concept discussed in the network section below), and p represents the probability with which a person will get infected.

The overall probability of being infected is certainly context-specific (an individual who is not in contact with someone infected has $p = 0$ of being infected) and therefore not entirely independent (as a Binomial distribution assumes). However, for each previously uninfected individual that is exposed to an infected individual,[2] each potential infection occurs independently with the probability p.[1]

R_0, the basic reproductive number, is the number of new cases expected to stem from a single infected individual. In a network where each individual (node) interacts with k new people, infecting them with probability p, R_0 can be expressed as $p * k$.[3] In reality, R_0 is context specific and changes over time, thus the "net" reproductive number, R_n is more accurate (though very difficult to accurately predict).

Parameters following the normal distribution

The serial interval (time between successive cases) for COVID-19 is approximately normally distributed based on recent calculations; $X \sim N(\mu, \sigma^2)$ where μ is the mean of the duration of infections, and σ^2 is the variance of duration

[2] the infected individuals' nodes are direct predecessors of the uninfected individuals' nodes in the network

[3] Coincidentally, $R_0 = \frac{\beta}{\gamma}$ in a simple SIR model. These values are described in the SIR model section below.

between infections. While the serial interval for COVID-19 is constantly being re-evaluated, a recent paper suggested that the serial interval for confirmed cases of COVID-19, based on data collected from China as of February 8, 2020 (n=468), is 3.96 days (95% CI 3.53–4.39 days), a period shorter than most previous estimates.[2] This is important when considering how long after exposure an individual may become sick, a value important in the SEIR model proposed later.

Parameters with more complex distributions

Many other parameters that may be considered as the model proposed here is implemented or improved do not follow a binomial or normal distribution, but more complex distributions. An example is the duration that the virus will survive in different mediums and on different surfaces.[3]

Differential probabilities based on symptomatology

The probability that a susceptible individual exposed to an individual with the virus will develop COVID-19 is differential. The context, including environment, behavior and symptomatology of the infected individual, and behavior of the susceptible individual all play a role. Ways of incorporating this into the network and model are described in the Network and SIR sections respectively. Some initial recommendations for addressing this are outlined in the Data Collection section.

Networks

In addition to knowledge of the properties of the SARS-CoV-2 virus and symptomatology of COVID-19, the structure of the population in which a this virus is spreading is critical when estimating disease spread based on contact tracing. Behavior, and therefore contact and network shape, frequently change for many reasons including travel restrictions, quarantining, government-mandated closures, and behavior changes due to illness. It is important to consider network shape dynamics when tracing contact between individuals and estimating the numbers of exposed and infected individuals at future time points. If network topology—as well as the way it changes due to government regulations—can be determined, some information about the future progression of the pandemic can be inferred.[4, 5]

Basics of the network

At its core, the network will involve every individual for whom we have data. Each individual is a node $v \in V$ that will then be set to susceptible, exposed, infected, or recovered based on this information. Each time there is contact between two individuals (say, v, w), an edge will be created between them. The assumption will be made that every pair of nodes is able to infect or receive infection with each connected node, meaning the graph will be a symmetric directed graph $G = (V, A)$. This network will be analyzed at pre-defined time steps (e.g., per hour) and node status and edges will be updated at these time steps. It is also assumed that there are no "original" infections (people cannot become infected without being in some form contact with someone who is sick).

Ideally, a network of all individuals across the world would be created, but this is obviously not realistic. In addition, it would be ideal if all exposures had the same probability of infecting an individual, all individuals recovered at the same rate, all individuals that were ever in contact with an infected individual had the same probability of being exposed, and all methods of transmission were equally likely to expose someone. Of course individuals behave differently and there is no homogeneity of exposure and infection, so these assumptions do not hold in the real world. The initial SEIR model roughly makes these assumptions, and the subsequent section describes potential ways to adjust for this inevitable heterogeneity.

Incorporation of retroactive case identification

If an individual is determined to be infectious, and it seems they were likely infectious before they were tested (due to their symptoms, etc.) then it could be fruitful to "retroactively" set their node in the network to infected at a previous time step and re-run analysis of the network.

With sufficient information, it may also be possible to retroactively suggest individuals that may have spread the virus even though their infection status was unknown at the time if they seem to be a common contact of many subsequently identified cases that do not have a clear alternative source of exposure; the potentially infected user could then be gently alerted. Unfortunately, this would likely require nearly complete information.

Tracking disease flow through the network

SIR models—here a SEIR model—provide the means to track the numbers of susceptible, exposed, infected (infectious), and recovered individuals in the network, and predict trends in infection throughout the network at future time steps. The next section will describe a proposed SEIR model and note its limitations and potential ways to adapt it to more heterogeneous and complex (in a sense more "realistic") networks of individuals.

SIR models and proposed SEIR model

SIR (susceptible, infected/infectious, recovered) models[4] are widely used in epidemiology and population biology to model and predict the number and rate of infections in a population over time. They compartmentalize individuals into each of the categories, S, I, and R. The model dynamics represent the flow of the number of individuals between each of these categories. Flow rates are typically expressed through differential or difference equations.

The SIR model is a basic form that can be expanded and adapted to more complex situations. In addition to adding more compartments to the model, additional flow rates can be added to include elements such as birth, death,

[4] If you would like a visual intuition for the spread of infection in a network based on a SIR model, you are encouraged to look at: `http://systems-sciences.uni-graz.at/etextbook/networks/sirnetwork.html`

vaccination, and re-infection rates. The model proposed in this paper is a variation of the SIR model called the SEIR (susceptible, exposed, infectious, recovered) model. The next section outlines the details and assumptions of this SEIR model.

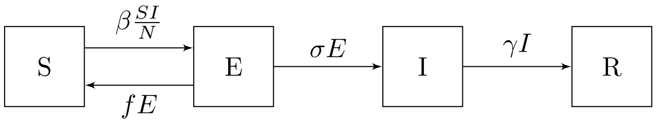
Proposed SEIR model

A SEIR model is chosen over a SIR model here because it is useful to track exposed individuals (not just infectious). Knowledge of exposed individuals allows anticipation of where the next outbreak will be and allow for better resouce allocation. SEIR models are commonly used for the flu, though they typically do not include a flow from exposed back to susceptible as this one does.

Some SIR variations may include other terms such as birth and natural or disease-induced death terms. Because the dynamics of the COVID-19 pandemic are much faster than typical dynamics of birth and death, it is safe to exclude natural birth and death terms. Disease-induced death will be grouped in with the recovered individuals because neither those that died nor those that have recovered can infect others or become re-infected.

The "exposed" group is often considered to be a "pre-infections" group that will eventually become infectious at a certain rate. This model considers the exposed individuals to have been exposed to SARS-CoV-2 and may become infected and infectious at some rate σE (see diagram below), or return to susceptible because they did not actually contract the virus, at some rate fE. The rest of the flow rates and compartments are portrayed in the diagram and differential equations.

The SEIR model variation that this article proposes for use in the contact tracing process is as follows:



where β represents the risk constant[5] from susceptible to exposed individuals, f represents the flow rate constant for individuals that were exposed but not infected, σ represents the flow rate constant of exposed individuals to infected individuals, and γ represents the flow rate constant from infected to recovered individuals.

It then follows that the equations for change per unit time of individuals in each category are:

$$\begin{aligned}
&\frac{dS}{dt} = -\beta\frac{SI}{N} + fE \\
&\frac{dE}{dt} = \beta\frac{SI}{N} - fE - \sigma E \\
&\frac{dI}{dt} = \sigma E - \gamma I \\
&\frac{dR}{dt} = \gamma I \\
&N = S + E + I + R
\end{aligned} \tag{1}$$

The following are a few considerations when interpreting this model:

SI is divided by N to constrain the model to the population size and ensure dimensional analysis checks out. If we add the assumption that, at each time step, an individual in the exposed group will either return to susceptible or become infected (not remain simply exposed), then $f = (1-\sigma)$. It should make little difference whether an individual in R is recovered and alive or dead, as neither will be able to infect other individuals or be infected again.

There are some major assumptions that this SEIR model makes. The next section, therefore, will describe some limitations and provide options for addressing them.

[5] β is sometimes referred to as the "contact rate" and γ is sometimes referred to as the "removal rate."

Limitations and potential improvements on this model

Many assumptions about interaction were described in the network section, and therefore apply when this deterministic SEIR model is used to analyze changes in the network of individuals. This basic SEIR model is a “mean-field model,‘’ which means that all members of the population receive the same mean treatment. Therefore, the baseline model struggles to capture the complex and heterogenous structures of real-life networks and interaction. Heterogeneity of influence between nodes, which is of primary concern, is discussed here.

Dealing with heterogeneity

One crude approach to handling this would involve partitioning the network of all individuals into subnets (by location or by strata of infectiousness), evaluating infection dynamics within each subnet at each time step, and then move the relevant nodes, with their anticipated state at the next time step, to different subnets. This crude process limits the analysis that can be done and makes it more difficult to perform population-level analysis.

A better approach is assigning different values for infectiousness, susceptibility, or other important parameters to each node. The SEIR model can then be adapted to incorporate heterogeneity of effects without having to create a set of (reasonably large) subnets using a set of tools called “pairwise approximation methods.” Much work has been done to apply these pairwise approximation methods to SIR models in recent years.[6, 7, 8] This method allows evaluation of every pair of nodes (two individuals) at each time step t_n, and therefore incorporate the unique parameter values held by each node. This, in turn, addresses the concern about heterogeneity of effects between individuals.

Even with these tools, it is important to be careful about how the model is designed and implemented considering the the frequent movement between locations, and therefore groups of people which can make it difficult to trace infected individuals and subsequent exposures effectively.

Data collection

What's available and what's needed

This approach is certainly constrained by data quality and completeness. Fortunately, much work has gone into estimating some of the parameters needed in a SIR model, especially R_0, and duration of infection (and by extension recovery time) which is useful for estimating γ. There is less information on many of the other parameters, so data collection and continuous updating based on new data for these values is critical.

Beyond estimating model parameters, a large network of individuals whose location is being tracked (as in our 4D model) is critical for effectively tracing exposure, infection, risk of infection if exposed, as well as identifying context-specific differences in the model parameters. It is intuitive that infected individuals who are not traced are missing from the network and have the potential to infect others without warning.

A significant barrier to implementation is identification of the context-specific risks of infection as mentioned in the paragraph above. The following are a few avenues that can be explored and researched that should be incorporated to improve context-specific estimation.

As briefly noted in the Probabilities section, the SARS-CoV-2 virus survives for different durations in (or on) different mediums. It is certainly not realistic to include information on every time of material an infected and subsequently susceptible individual has come into contact with. However, it is worth considering some locational properties such as indoor versus outdoor exposure.

It is also important to consider an individual's symptomatology. It makes sense that an individual who is coughing and sneezing will likely spread more of the virus and spread it further than someone who is breathing normally. There are other nuances such individuals wearing masks, which will typically reduce spread from infectious individuals and sometimes reduce infection if a susceptible individual is wearing one.

In addition, some parameters such as level of infectiousness or how long the virus remains at a given location can be informed through the continuous use of machine learning algorithms on the data that is gathered.

Other needs for data

The two paragraphs above give only a few recommendations. Further recommendations for estimating differences in context-specific risks of infection, as well as estimating other parameters, are important and welcome. It is likely that some research has been published on many such considerations, but those were not the focus of research for this paper.

Brief conclusions

Contact tracing is a powerful tool for fighting pandemics and saving lives, as stated at the beginning of this paper. Opportunities to partner with other research groups, companies, and governmental agencies that will allow us to use their data or resources to improve our contact tracing and prediction algorithms are certainly welcome and beneficial. This could range from cellphone data to information on individuals who have been tested to symptom data from at-home thermometers[9] or similar products. Certainly the needs to maintain patient confidentiality and respect user privacy are central to the MutualAid.world project and must be prioritized. Any opportunity to safely and securely include new data into the network of individuals will improve the algorithm and model prediction.

References

[1] D. Easley, J. Kleinberg, Networks, crowds, and markets: reasoning about a highly connected world, Cambridge University Press, 2010.

[2] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, L. A. Meyers, Serial interval of covid-19 among publicly reported confirmed cases, Emerging Infectious Diseases 26 (6). doi:10.3201/eid2606.200357.
URL http://wwwnc.cdc.gov/eid/article/26/6/20-0357_article.htm

[3] N. van Doremalen, T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, S. I. Gerber, et al., Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1, New England Journal of Medicine (2020) NEJMc2004973doi:10.1056/NEJMc2004973.

[4] M. J. Keeling, K. T. Eames, Networks and epidemic models, Journal of The Royal Society Interface 2 (4) (2005) 295–307. doi:10.1098/rsif.2005.0051.

[5] M. D. Shirley, S. P. Rushton, The impacts of network topology on disease spread, Ecological Complexity 2 (3) (2005) 287–299. doi:10.1016/j.ecocom.2005.04.005.

[6] C. Llensa, D. Juher, J. Saldaña, On the early epidemic dynamics for pairwise models, Journal of Theoretical Biology 352 (2014) 71–81. doi:10.1016/j.jtbi.2014.02.037.

[7] C. T. Bauch, The spread of infectious diseases in spatially structured populations: An invasory pair approximation, Mathematical Biosciences 198 (2) (2005) 217–237. doi:10.1016/j.mbs.2005.06.005.

[8] L. Liu, X. Luo, L. Chang, Vaccination strategies of an sir pair approximation model with demographics on complex networks, Chaos, Solitons & Fractals 104 (2017) 282–290. doi:10.1016/j.chaos.2017.08.019.

[9] [link].
URL https://www.kinsahealth.co/research/