TemporalPrestige

Nicholas Winsley

2025-03-19

Contents

Introduction

Epidemiology is a subject of much contemporary relevance. The recent Covid-19 pandemic highlighted the importance of effective methods for combating the spread of disease. With the emergence of automated contact tracing technology, prophylactic identification and isolation

of high-risk individuals could be practicable in the future. This study investigates methods for the identification of high-risk individuals in a temporal network.

We will begin by defining what is meant by "high-risk" in this context. Government's wish to intervene for individuals who are most instrumental in the spread of a disease. Therefore, risk is a combination of the likelihood of a particular individual becoming contagious due to prior contacts, and the likelihood of the individual spreading the illness via future contacts. In practice, future contacts are not known, therefore this study will primarily focus on the risk of infection.

Social networks can be modeled as a number of close personal contacts, each paired with a measure of proximity. Each contact is written as (i, j, t), where i and j are the interacting individuals, and t is the time of the interaction (note that (i, j, t) is semantically no different from (j, i, t)). This representation is called a temporal network, and it is frequently used in the study of epidemics. In practice, contact events are identified by using contact tracing.

Contact tracing is a methodology for combating the spread of infectious diseases transmitted by close personal contacts. By uncovering close contact events, contact tracing can be used to identify people at high risk of infection, and foresee future growth or contraction of the epidemic. This information can inform further interventions (e.g. quarantining, disease tests, vaccination). Traditionally, contact tracing focused on confirmed cases, which were reported to authorities. When a new case is reported, an official asks the patient to recall all recent contacts. Although this method has been shown to be effective (Fetzer and Graeber (2021)), it is labor-intensive and does not give a complete picture. Recently, digital contact tracing has emerged as a cost-effective (albeit unreliable) alternative to conventional contact tracing methods. In it's latest incarnations, digital contact tracing uses portable bluetooth devices which detect close contacts between carriers of the device. In addition, digital contact tracing can provide indicators of proximity between the interacting individuals. Statistical methods are often used (in combination with contact tracing) to identify high-risk individuals in a social network. This study investigates two common methods for the analysis of contact tracing data: Simulation and Social Network Analysis (SNA).

Social Network Analysis (SNA) is an interdisciplinary approach for the study of entities

and their relationships with each other. SNA involves constructing a network to model a real-world situation of interest, and calculating metrics related to the network structure. These metrics can be broadly categorized into two types: Population-level measures and Individual-level measures. This study is primarily concerned with individual-level measures. Centrality is a commonly-used individual-level measure which is defined as the influence of a particular individual within a network. Prestige (sometimes called status or rank) is a similar individual-level measure for directional graphs, which only considers incoming dyads. Although SNA has been used in many fields since it's creation in the 1930s, it's value in epidemiology only became apparent in 1985, when Klovdahl (1985) applied SNA to AIDS data.

Simulation is an effective method for ascertaining properties of a temporal network. Some classical models have been deterministic: the differential equation model of Kermack and McKendrick (1927) is a notable example. However, deterministic models typically rely on simplifying assumptions, and thus do not capture the full granularity of the network. Compartmental models are a popular simulation approach in which the population is divided into groups, and individuals transition between groups over time. The Susceptible-Infected-Recovered (SIR) model is a quintessential compartmental model in which all individuals are initially susceptible, and individuals may become infected due to a contact with a contagious individual. Once infected, individuals transition to the recovered state at a predictable recovery rate, where they stay for the remainder of the simulation. Recovery rates are typically sampled from a probability distribution, which may be estimatable by exogenous information (e.g. medical knowledge, recovery rates for similar diseases). Key metrics are averaged over many simulations to approximate a true underlying distribution. A plethora of summary metrics have been applied to simulated epidemics. Macdonald (1952) introduced the reproductive number, which is defined as the number of cases resulting from a single infection. Holme (2018) used the time for the disease to go extinct (i.e. no new cases can occur).

This study aims to address several key research questions:

1. Can we make simplifying assumptions to reduce computation time of simulations?

2. Which centrality measures are most effective when applied to epidemiology?

3. How can we extend these measures to cases where contact risk varies?

Simulation is important for answering questions 2. and 3. as it provides a sort of "ground-truth" against which centrality measures can be compared.

Methods

Background

In 2020, the New Zealand Ministry of Health (MoH) commissioned a pilot study of the "CovidCard", a portable device which used bluetooth technology to record contacts between carriers of the device. Adults 19 years of age or older who live in Ngontotahā West and East were recruited to participate in a seven-day study. Additionally, people who live outside these boundaries but work within the Ngontotahā Village were also permitted to take part in the trial. Ngontotahā was chosen because it met several key criteria, namely compactness, geographical isolation, small population size and high sociodemographic diversity. In total, 1,191 people participated in the study. At the end of the trial period, a subset of 158 participants from the main trial were contacted by MoH case investigators to establish contacts that they had over the trial period using a modified version of the MoH case investigation protocol. The study compared the CovidCard to conventional case investigation methods, and found a greater rate of reciprocal interactions identified by the CovidCard. In short, the study concluded that the CovidCard is a highly effective contact tracing approach. We use this data to compare existing centrality metrics, and develop novel alternatives.

Data Description

The CovidCard is a bluetooth device developed for detecting close-contact events between carriers. Each card advertises it's presence and detects signals from other cards. Algorithms evaluate the radio signal strength indicator (RSSI) of close-contacts in real-time, and the signal strength is aggregated over 15 minute time intervals. The raw RSSI values are transformed

into distance estimates by the path loss model, proposed by Seidel and Rappaport (1992), for which

$$RSSI \propto -20 \log_{10}(distance)$$

Noise in distance estimates is subsequently reduced by signal processing methods, most commonly Kalman Filters.

Each interval was classified as either < 1 meter, < 2 meter and < 4 meter proximity, and the total number of intervals belonging to each class was summed over a two-hour period. The cards can hold up to 128 contact events in short-term cache memory at any given time, of which some are recorded in long-term flash memory. An interaction was recorded in flash memory if it was longer than 2 minutes in duration, and the RSSI exceeded -62dBm (roughly corresponding to a distance of less than 4 meters). For more details on the CovidCard, see Admiraal et al. (2022).

Data Preparation

The last day of the trial period saw an anomalously high number of close-contacts, most likely because participants congregated at a single location for card collection. For this reason, contact events which occurred on the last day of the trial were omitted. Participants who could not be cross-verified by case investigation were removed. If two cards registered the same contact event, and gave conflicting proximity values, one of the proximity values was arbitrarily removed. Contact dates were converted to numeric times by calculating the time elapsed (in hours) between the contact date and the start of the trial. Some cards were collected before the last day of the trial, resulting in an anomalous number of contacts during card collection. For this reason, all contact events which occurred on the day in which they were uploaded were removed. Data on the proximity classes was processed to form non-overlapping categories. For instance, the number of 15-intervals with a distance less than 4 metres, $n_{<4}$, was transformed by subtracting $n_{<2}$ and $n_{<1}$ to get $n_{\geq 3,<4}$; the total number of 15-minute intervals between 3 and 4 metres. By doing this, we get a categorical variable

on which further statistical models are based.

Centrality

Centrality is roughly defined as the influence of a given node in a graph. The exact meaning of a "central" node varies depending on the specific context.

Degree Centrality

Degree centrality is a simple metric for static networks where we only consider the number of neighbors of a given node. Let A be the adjacency matrix for a static network i.e. $A_{ij} = 1$ if i and j are neighbours and 0 otherwise. The degree centrality is defined as:

$$C_D(i) = \sum_{j=1}^{N} A_{ij} \tag{1}$$

Degree centrality is simple and easily calculable, however it does not incorporate knowledge of the entire network structure. Improving this point is the motivation for our next centrality measure.

Closeness Centrality

For undirected graphs, Bavelas (1950) proposed closeness centrality

$$C_c(u) = \frac{N-1}{\sum_{v \neq u} d(u,v)} \tag{2}$$

where N is the number of nodes and $d(u, v)$ is the distance of the shortest path between v and u. Thus, referring to Figure 1, the closeness centrality of node E is $\frac{4}{1+1+2+3} = \frac{4}{7}$. Closeness centrality can be interpreted as the efficiency with which a node can access all other nodes in a network.

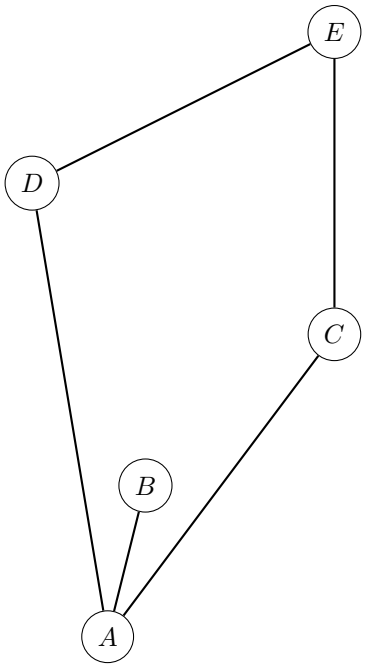


Figure 1: A simple undirected graph

Betweenness Centrality

In some situations, the effect of removing a node on transmission through a network may be highly important (for instance, when quarantining individuals during a pandemic). This is the primary motivation behind Betweenness Centrality (Freeman (1977)), which is defined as

$$C_B(v) = \frac{1}{(N-1)(N-2)} \sum_{s \neq v \neq r} \frac{\sigma_{s,r}(v)}{\sigma_{s,r}} \tag{3}$$

where $\sigma_{s,r}$ is the number of shortest paths (geodesics) between s and r, and $\sigma_{s,r}(v)$ is the number of such paths that pass through v. The denominator term, $(N-1)(N-2)$, ensures that the value is normalized between 0 and 1.

Percolation Centrality

In practice, additional information around the percolation state of nodes may be known. For instance, in epidemiology we may know that certain individuals are infected. To incorporate knowledge of percolation state into centrality metrics, Piraveenan, Prokopenko, and Hossain

(2013) proposed percolation centrality:

$$C_P^t(v) = \frac{1}{(N-1)(N-2)} \sum_{s \neq v \neq r} \frac{\sigma_{s,r}(v)}{\sigma_{s,r}} \frac{x_s^t}{[\sum x_i^t] - x_v^t} \tag{4}$$

where x_i^t is the percolation state of node i at time t. The percolation state ranges from 0 to 1, where 1 means the individual is certainly infected, and 0 means the individual is healthy. A decimal value (say 0.6) could, for instance, represent a probability of infection or a proportion of a township which is infected.

Adjusted Percolation Centrality

I propose a novel variant of percolation centrality, which I will call Adjusted Percolation Centrality. Adjusted Percolation Centrality only considers paths which do not pass through any percolated nodes. By doing this, it ensures that redundant paths are not considered. I will define an unpercolated path as a path where no incident nodes are percolated, except for the start and end nodes, which may have any percolation state. The mathematical definition is

$$C_P^t(v) = \frac{1}{M_P} \sum_{s \neq v \neq r} \frac{\sigma_{s,r}^P(v)}{\sigma_{s,r}^P} \frac{x_s^t}{[\sum x_i^t] - x_v^t} \tag{5}$$

where M_P is the number of pairs (i, j) where there is an unpercolated path between i and j, and $\sigma_{s,r}^P$ is the number of shortest unpercolated paths between s and r.

Katz Centrality

Betweenness and Closeness centrality focus on shortest paths, which potentially do not give the full picture. Katz centrality (Katz (1953)) is an alternative approach which considers all paths between nodes. For the i'th node in a network, Katz centrality is defined as:

$$C_K(i) = \sum_{j=1}^{n} (I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + ...)_{ij} \tag{6}$$

where A is the adjacency matrix and $0 <= \alpha < 1$. A^n is simply the n-step adjacency matrix for the network, and α^n decreases as n grows, therefore longer paths are down weighted. For a technical reason, the infinite sum $(I + \alpha A + ...)$ converges when $\alpha <= \frac{1}{\rho(A)}$ where $\rho(A)$ is the spectral radius of A i.e. the largest absolute value of any of it's eigenvalues. Under this condition, the sum converges to $(I - \alpha A)^{-1}$. If we are only interested in the first n steps, we adjust the formula accordingly to get $(I - \alpha^{n+1}A^{n+1})(I - \alpha A)^{-1}$. This result is convenient for temporal networks, where a finite number of steps in each snapshot is often assumed.

Epidemic Simulations for Static Networks

This section describes a novel approach to epidemic modelling using static networks, which I use as a "ground truth" to validate Adjusted Percolation Centrality. A social network is modeled by a static graph, where each node represents an individual, and each edge represents an ongoing, time-independent relationship between two individuals. It is assumed that contacts between any two individuals i and j follows a Poisson Process with a rate parameter of λ_{ij}, and each contact has a constant probability, β, of being infectious. The sojourn times (period between contacts) of contacts between i and j follow an exponential distribution with a rate of λ.

Lemma 1: Let Y follow a Poisson Process with rate λ, and X be the Markov process where for each arrival in Y, X includes the arrival with probability p. Then X is a Poisson Process with a rate parameter of $p\lambda$.

The proof of Lemma 1 is a well-known result which is not discussed here. Lemma 1 implies that the sojourn times of infectious contacts between i and j follow independent exponential distributions with a fixed rate parameter of $\beta\lambda_{ij}$. Due to the memorylessness property of the exponential distribution, the time until the next infectious contact, from any starting time, follows the same distribution. By these assumptions, epidemics can be efficiently simulated in the most general case, without prior knowledge of contacts. Consider a simulation ending at time T, which starts with a set of infected nodes. We iterate through all neighbours of each infected node, and sample a time until the next infectious contact, adding it to a sorted list of infection times as we go. At each subsequent step, the smallest infection time is selected

from the list. If this time is greater than T, the simulation stops. Otherwise we sample an infection time for each neighbour of the infected node. If the sampled infection time for a given node is greater than it's current infection time, it is ignored. Figure 2 shows a static graph representing a small social network, where the edge weights correspond to the rate parameter, $\beta\lambda_{ij}$. Consider a simulation on this graph which terminates at the time T = 100. Initially, only A is infected. We sample t = 60, 40, 50 for neighbours C, B and D respectively. In the next step, the infection time of C (t = 40) is removed from the list, and the time 70 is sampled for node E. Thus, the infection time for E is 40 + 70 = 110. Then, the infection time of node D (t = 50) is removed from the list, and each neighbour is considered in turn. Suppose we sample a sojourn time of 70 for E, hence the new infection time is 50 + 70 = 120 which is greater than 110, thus it is ignored. Likewise, suppose the time 40 is sampled for the infection process between D and B. The new infection time for B is 50 + 40 = 90 which is greater than 60, thus it is ignored.

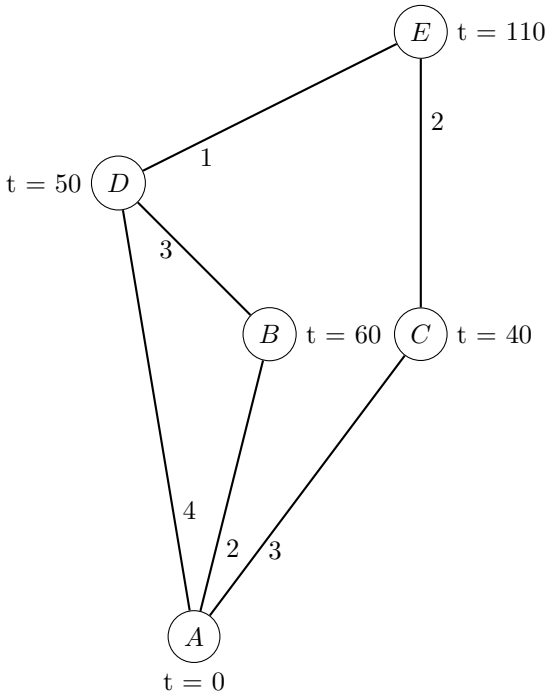


Figure 2: A simple social network represented as a static graph

In general, contact rate parameters may be estimable from contact tracing data or theoretical values. It should be clear that this method is not a replacement for simulations on contact tracing data, and should only be used as a crude alternative when the dataset extremely

large.

The importance of a given node was tested empirically by calculating the difference in average reproductive number (over 100 simulations) between graphs which include and exclude the node in question. A high difference in reproductive numbers indicates that the node is instrumental in the spread of the disease.

The Erdos-Renyi model (Erdos, Rényi, et al. (1960)) was used to generate random graphs of 40 nodes (four of which are initially infected), and the connectedness property (an underlying assumption of betweenness centrality) was verified for each graph. We start with an edge between every vertex, and each edge is included with a constant probability, p. Figure 3 shows the results for various edge inclusion probabilities. Both variants of percolation centrality showed substantial correlations with the empirical measure of importance.

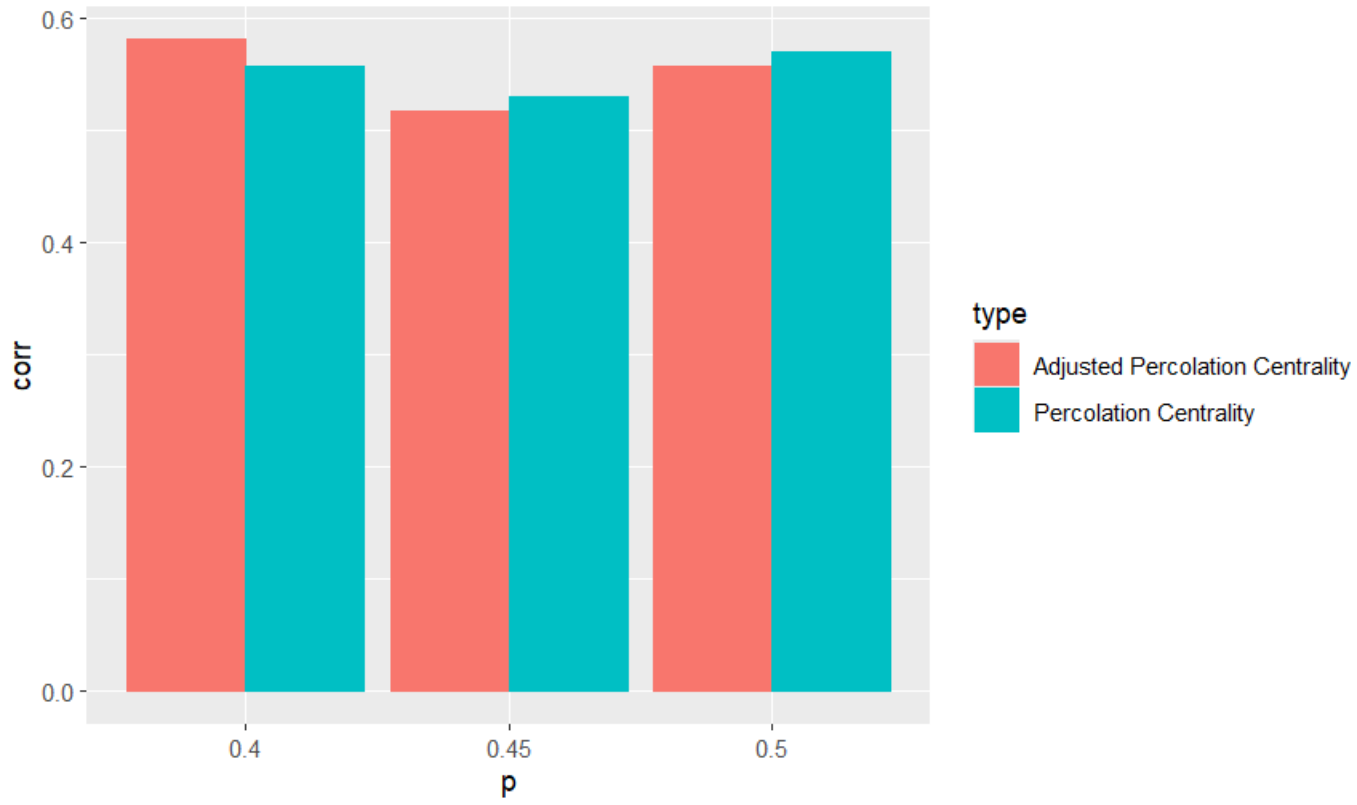


Figure 3: Correlation between the difference of reproductive numbers and the two variants of percolation centrality for different edge inclusion probabilities

Temporal Katz Centrality

Temporal Katz Centrality, proposed by Grindrod et al. (2011), follows fairly straightforwardly from @ref(eq:katz). Denote by A_t the adjacency matrix for the snapshot at time t. For a temporal network with snapshots at times $1, 2, 3, ..., T-1, T$, the Temporal Katz Centrality is defined by the matrix product

$$(I - \alpha A_1)^{-1}(I - \alpha A_2)^{-1}...(I - \alpha A_{T-1})^{-1}(I - \alpha A_T)^{-1} \tag{7}$$

The centrality for a given node is likewise calculated by the row sum of this matrix product.

Time-Ordered Networks

Generalization of conventional centrality measures to temporal networks requires a high-granularity representation of the network as a graph. Kempe, Kleinberg, and Kumar (2000) proposed a graph where the edge weights are contact times, however this model fails to account for differential rates of transmission. Kim and Anderson (2012) proposed a more general solution using a time-ordered directed graph. Consider a network of N nodes for which M edges are observed over T time points. Without loss of generality, discretize the contact times to get a list, $t = (0, 1, 2, \cdots, T)$. We can construct a time-ordered graph where each node appears T + 1 times. Denote by υ_t the node υ at time t. In this graph, a directed edge from υ_t to u_{t+1} only exists if $\upsilon = u$ or there is a contact between υ and u at time t. We can construct this graph for any temporal network without loss of information. In practice, computational constraints may require aggregation of contact times and thus loss of information.

To illustrate this idea, consider a simple temporal network with five individuals, as shown in Table 1.

Table 1

Figures 2 and 3 show snapshots of the network at the two time points

Figure 4 shows the temporal network represented as a directed graph.

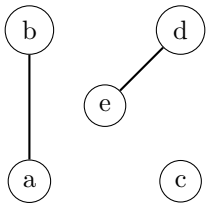


Figure 4: Snapshot of the network when t = 1

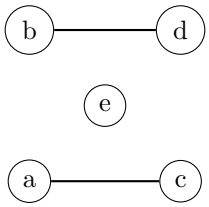


Figure 5: Snapshot of the network when t = 2

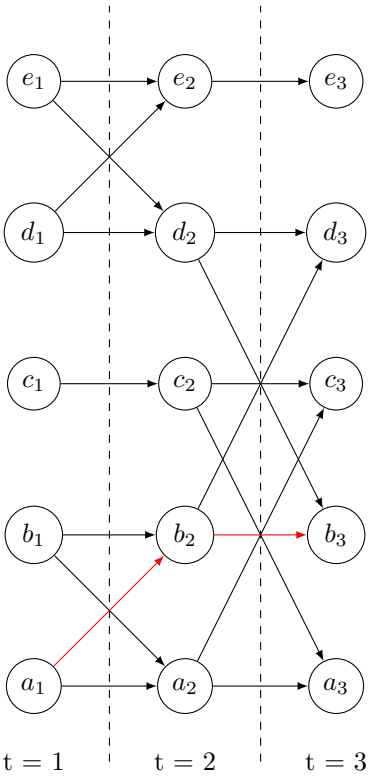


Figure 6: A simple temporal network represented as a digraph. The temporal shortest path from a to b is shown in red.

First Individual	Second Individual	Time of contact
a	b	1
b	d	2
a	c	2
d	e	1

Define the distance of the temporal shortest path length over time interval [i, j], denoted by $d_{i,j}(\upsilon, u)$, as the smallest $d = j - n$, where $i \leq n \leq j$ and there is a path from υ_n to u_j. Thus, in Figure 1, the shortest path distance $d_{1,3}(a, b)$ is two, with the temporal shortest path being a_1-b_2-b_3. By representing a temporal network as a high-granularity digraph, we can now generalize conventional measures of prestige and centrality to temporal networks.

Temporal Closeness Rank

This study is primarily concerned with the likelihood of infection, as estimating transmission requires future contact tracing data, which is usually not known. Likelihood of infection is roughly analogous to the idea of prestige (also known as rank). I will use the terms prestige and rank interchangeably throughout this paper. In a directional network, a prestigious node is the object of many ties i.e. has many incoming connections. This is a distinct concept from centrality, which is also concerned with outgoing connections. Many conventional measures of centrality are ill-defined in directional graphs, owing to the fact that directional graphs are not necessarily strongly-connected. Due to this limitation, we usually only consider nodes in the influence domain of node i i.e. the set of all nodes from whom n_i is reachable. Lin (1976) proposed the following measure for directional relations, called the proximity prestige:

$$P_p(n_i) = \frac{I_i/(g-1)}{\sum d(n_j, n_i)/I_i} \tag{1.}$$

where I_i is the size of the influence domain of node i, and g is the total number of nodes in the network.Intuitively, this is the proportion of the network covered by the influence domain, divided by the average temporal distance over the influence domain. When the influence domain is empty, the proximity prestige is defined to be 0.

For temporal networks, I propose a modified version

$$P_p(u_i) = \sum_{t=0}^{i-1} \frac{I_{t,i,u}/(g-1)}{\sum_{v \in I_{t,i,u}} d_{t,i}(v,u)/I_{t,i,u}} \tag{5.}$$

where $I_{t,i,u}$ is the influence domain of u over the time interval [t, i]. We will call this the temporal proximity prestige, or TPP for short. The temporal proximity prestige can be normalized by dividing by i.

Kim and Anderson [2012] proposed temporal closeness, a similar metric which considers all time intervals $[t, i], t \in [0, i-1]$.

$$C_{i,j} = \sum_{i \leq t < j} \sum_{u \in V} \frac{1}{d_{j,t}(v,u)} \tag{3.}$$

When u is unreachable from v over [t, j], $d_{t,j}(v,u) = \infty$. We cover cases where the denominator is infinite by assuming that $\frac{1}{\infty} = 0$. Note that as we are considering a directional graph, $d_{t,j}(v,u)$ is not equivalent to $d_{t,j}(u,v)$. To turn @ref(eq:tc) into a prestige measure, we simply reverse the direction of the paths to get:

$$C^P_{i,j} = \sum_{i \leq t < j} \sum_{u \in V} \frac{1}{d_{t,j}(u,v)} \tag{4.}$$

We will call this the temporal prestige. The temporal prestige can be normalized by dividing by $(|V| - 1)(j - i)$.

Multiplicative Temporal Closeness Rank

When each edge is associated with a probability of transmission, as may be the case in epidemiology models, the probability of a path may be of greater interest than the temporal length. In this case, we can generalize existing methods by considering a digraph where the edge weights are the natural log of the probability. Figure 2 shows a graph of this kind.

In accordance with the Susceptible-Infected (SI) framework, the probability of transmission from an individual to them self is assumed to be one, and hence the natural log becomes 0. This representation has several useful mathematical properties. Consider a path, P, starting

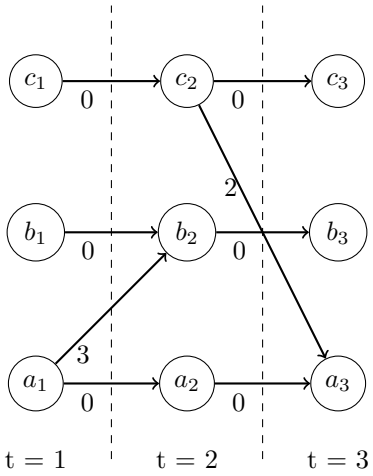


Figure 7: A simple temporal network represented as a digraph

at v_i and ending at u_j. The probability of this path is equal to $\prod_{k=i}^{j} E_k$, where E is the list of transmission probabilities of path P. The probability of path P can be calculated by:

$$e^{\sum_{w \in E} log(w)}$$

It can be shown that for the representation in Figure 2, all highest-probability paths to u_i can be calculated in $O(i|V|^2)$ time using a modified version of the Reversed Evolution Network (abbreviated as REN) algorithm proposed by Hanke and Foraita (2017). Algorithm 1 shows the psuedocode for this algorithm.

Absorption Rank

Rocha and Masuda (2014) proposed TempoRank, an extension of the illustrious Google PageRank algorithm to temporal networks. TempoRank considers the stationary distribution of a random walk through the temporal network. However, TempoRank does not generalize well to epidemiology modelling, where infection may be a permanent state. Intuitively, some sort of aggregation over all previous contacts would be preferred. In this section, I propose an aggregate metric for temporal networks. Without loss of generality, we will consider a temporal network consisting of a set of nodes, N, and positive integer contact times, $t = (1, 2, 3, \cdots, T)$. Let $p_{ijk}(t)$ denote the probability of transmission for the k'th contact between individual i and individual j, at time t. Let $n_{ij}(t)$ denote the number of

Result: Multiplicative Closeness Rank
Input : Data file with each row representing a contact and a probability of transmission
Output: Temporal Prestige for a given node

```
contacts <- List of contact times sorted in increasing order. Each contact time is a data
  structure with a map of nodes to out-neighbours.
while Data file has next line do
    Read line
    Add line to contacts using bisection search
end
tcp <- Temporal closeness prestige
reachable <- Set of reachable nodes
Add target node to reachable
sums <- Map of node id to log of the highest probability path (obtained by summing
  edge weights)
back <- Pointer to final contact time in contacts list
while back is not null do
    temp <- Map of updated path lengths for this iteration
    foreach node in reachable do
        out-neighbours <- back.out-neighbours[node]
        foreach neighbour in out-neighbours do
            Add neighbour to reachable set
            weight <- Edge weight of connection between node and neighbour
            temp[neighbour] = Max(temp[neighbour], sums[node] + weight)
        end
    end
    foreach node in reachable do
        sums[node] = Max(temp[node], sums[node])
        mcr += expsums[node]
    end
    Decrement back pointer
end
Return tcp
```

Algorithm 1: Modified version of the REN algorithm.

contacts between i and j at time t. Define the transition probability matrix for each contact time as:

$$\mathbf{B}_{ij}(t) = \begin{cases} 1 & i = j, s_i(t) = 0 \\ 0 & i \neq j, s_{ij}(t) = 0 \\ \prod_{m \in N, m \neq i} \prod_{k=1}^{n_{im}(t)} (1 - p_{imk}(t)) & i = j, s_i(t) > 0 \\ (1 - \mathbf{B}_{ii}(t))(s_{ij}(t)/s_i(t)) & i \neq j, s_{ij}(t) > 0 \end{cases} \tag{8}$$

where $s_{ij}(t)$ and $s_i(t)$ are defined as:

$$s_{ij}(t) = 1 - \prod_{k=1}^{n_{ij}(t)} (1 - p_{ijk}(t)) \tag{9}$$

$$s_i(t) = \sum_{j \in N, j \neq i} s_{ij}(t) \tag{10}$$

Denote by $\mathbf{B}_i(t)$ the transition matrix obtained by taking $\mathbf{B}(t)$, and setting all entries in the i'th row to zero, except the diagonal entry (which is necessarily one). The walk $\mathbf{B}_i = (\mathbf{B}_i(1), \mathbf{B}_i(2), \cdots, \mathbf{B}_i(T))$ is an absorbing random walk, and the product $C_i^A(t) = \mathbf{B}_i(1)\mathbf{B}_i(2)\cdots\mathbf{B}_i(t)$ will be called the absorption rank for individual i, at time t. The absorption rank of individual i is interpreted as the probability that a random walk through the temporal network passes through i.

If we assume a constant transmission probability, p, for all contacts, @ref(eq:absorb) reduces to:

$$\mathbf{B}_{ij}(t) = \begin{cases} 1 & i = j, s_i(t) = 0 \\ 0 & i \neq j, s_{ij}(t) = 0 \\ \prod_{m \in N, m \neq i} (1 - p)^{n_{ij}(t)} & i = j, s_i(t) > 0 \\ (1 - \mathbf{B}_{ii}(t))(s_{ij}(t)/s_i(t)) & i \neq j, s_{ij}(t) > 0 \end{cases} \tag{11}$$

where $s_{ij}(t)$ is defined as:

$$s_{ij}(t) = 1 - (1 - p)^{n_{ij}(t)} \tag{12}$$

and $s_i(t)$ is similarly defined as:

$$s_i(t) = \sum_{j \in N, j \neq i} s_{ij}(t) \tag{13}$$

In practice, p may be estimable from domain-specific knowledge or simply guessed. The value of p is not particularly important, however values close to 0 or 1 may cause the output values to cluster together, making differences difficult to detect.

Simulation

Stochastic simulations typically follow the standard Markovian framework, in which we assume transmission depends only on the current state of the network. The contact times can be thought of as discrete snapshots of the network, and it is assumed that during each snapshot, only one "hop" can occur. In other words, if we have two contacts (i, k, t) and (k, j, t), i cannot infect j (via k) at time t. The standard Markovian framework typically employs simplifying assumptions to ensure ease of implementation. The Susceptible-Infected-Recovered (SIR) model is a common implementation which assumes that all individuals are susceptible at the beginning, and once infected they remain infectious for a recovery period. Once recovered, individuals cannot be infected again. Other common implementations include the Susceptible-Infected (SI) model and the Susceptible-Infected-Susceptible (SIS) model.

Algorithms

For simulation, we use the event-based algorithm first proposed by Kiss et al. (2017), and described in great detail by Holme (2021). In this algorithm, contacts are conveniently stored an adjacency list format (each node is represented by a data structure with a list of neighbours). For each neighbour, a sorted list of contact times (and associated transmission probabilities) is stored. All infection events are stored in a min-heap ordered by infection

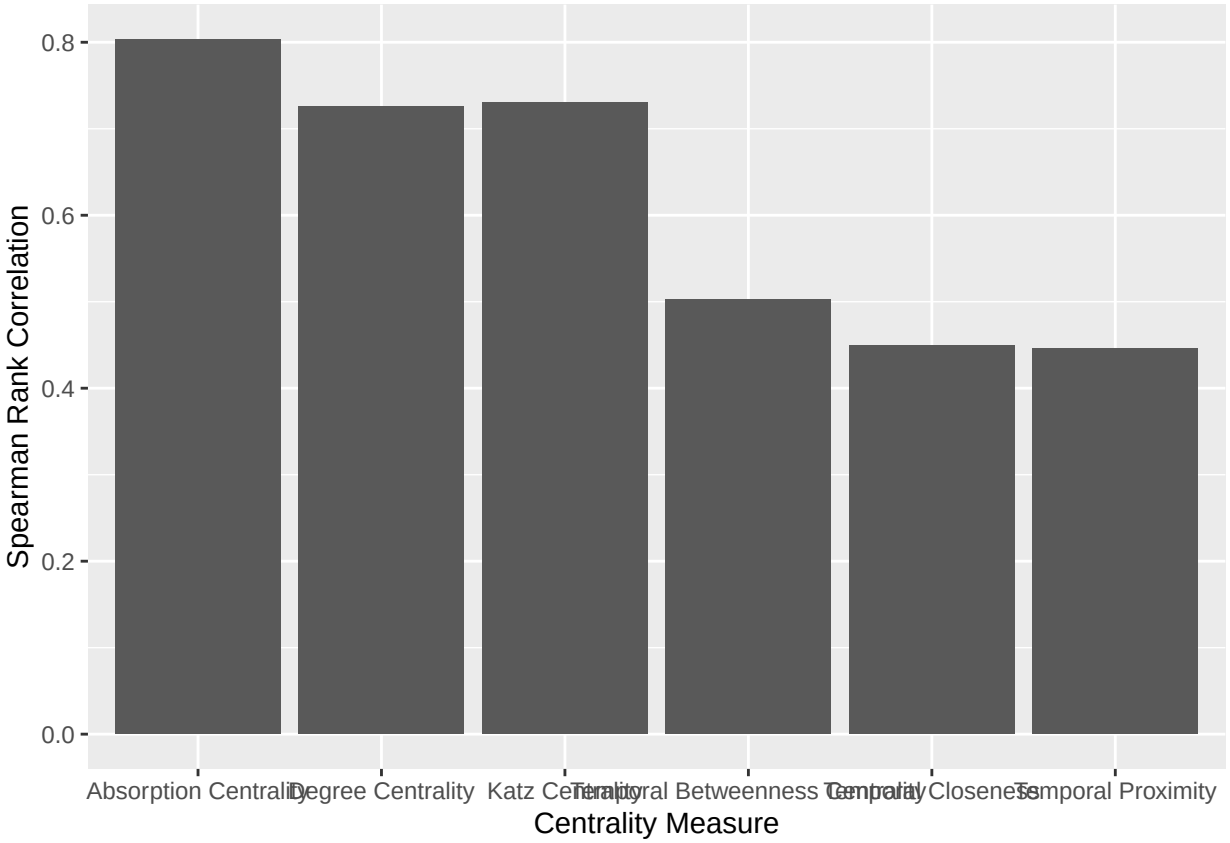
time. At each step, the earliest infection is removed and processed. This continues until no infection events remain in the heap. When node i is infected, we iterate through the list of neighbours and sample an infection time for each neighbour, adding the event to the heap as we go. If we assume a constant transmission probability, the index of the first infectious contact follows a geometric distribution. We can sample this index by

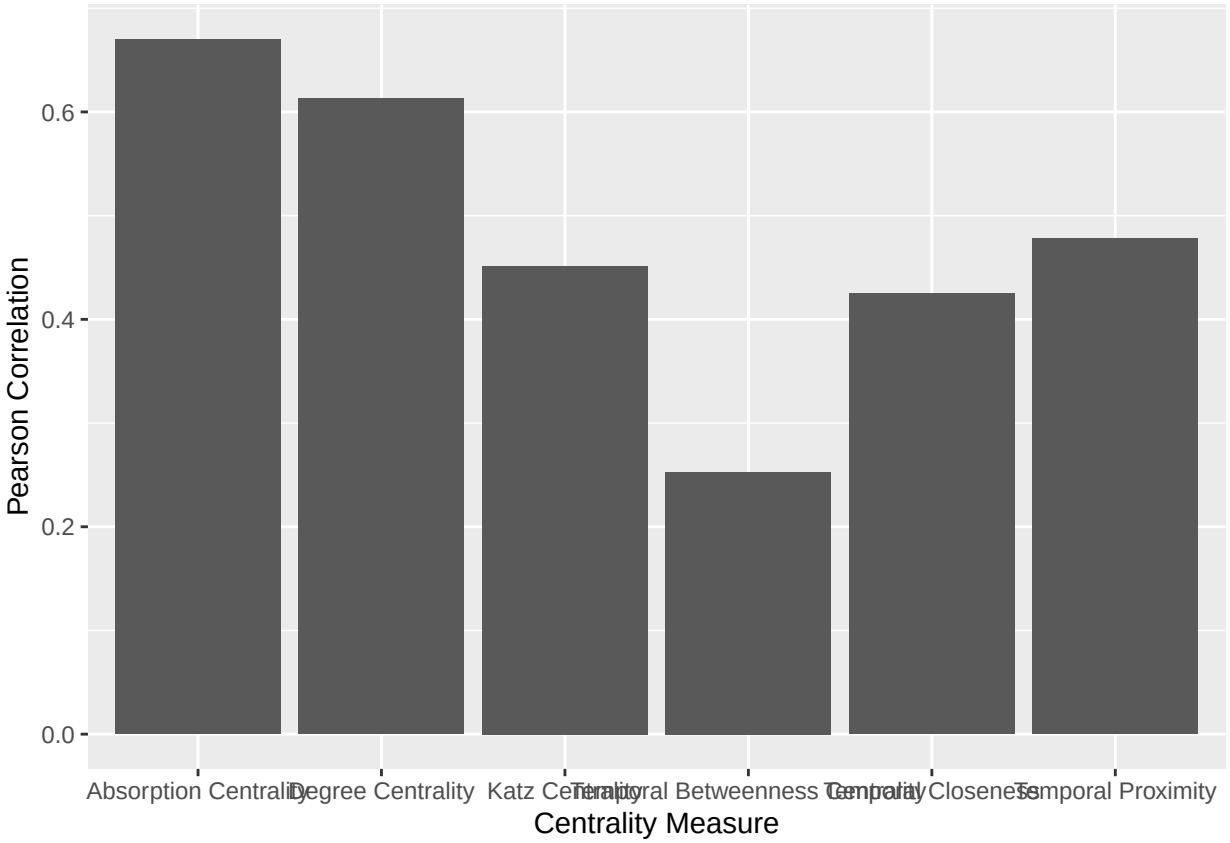
$$\lceil \frac{log(1-X)}{log(1-\beta)} \rceil$$

where β is the fixed transmission probability and $X \sim Uniform(0, 1)$.

Results

Five measures were calculated for the CovidCard dataset, namely Temporal Closeness (denoted by TP) and Proximity Rank (denoted by TPP), Absorption Rank, Temporal Katz Centrality and Temporal Degree Centrality. For a fair comparison, all centrality measures were calculated by assuming constant transmission probabilities for every contact event. This was done because Temporal Katz Centrality does not generalize to varying transmission probabilities. These measures were correlated to the observed number of times each individual was infected over 751,000 simulations (shown in Figure 8). The dataset contained 751 individuals in total, and 1000 simulations for each of the 751 possible starting node were carried out.


0.8
0.6
0.4
0.2
0.0
Spearman Rank Correlation
Absorption Centrality
Degree Centrality
Katz Centrality
Temporal Betweenness Centrality
Temporal Closeness
Temporal Proximity
Centrality Measure



The Spearman Correlations are generally large, with Absorption Rank showing a larger correlation ($\rho \approx 0.8$) compared to other metrics. Conversely, the metrics generally showed lower Pearson correlations, especially Katz centrality. This suggests that the metrics should not be used for prediction, rather they should be used for ranking the relative importance of individuals within the same network. TP and TPP performed similarly with regards to ranking individuals, however TPP outperformed TP in terms of Pearson Correlation.

Figure 8 shows the Pearson correlations between the five metrics. With the exception of Proximity Rank and Closeness Rank, the metrics are mostly independent, with many weak or negative correlations. This suggests that the metrics are tapping into different effects, and our rankings could be improved by combining many metrics as predictors in a regression model.

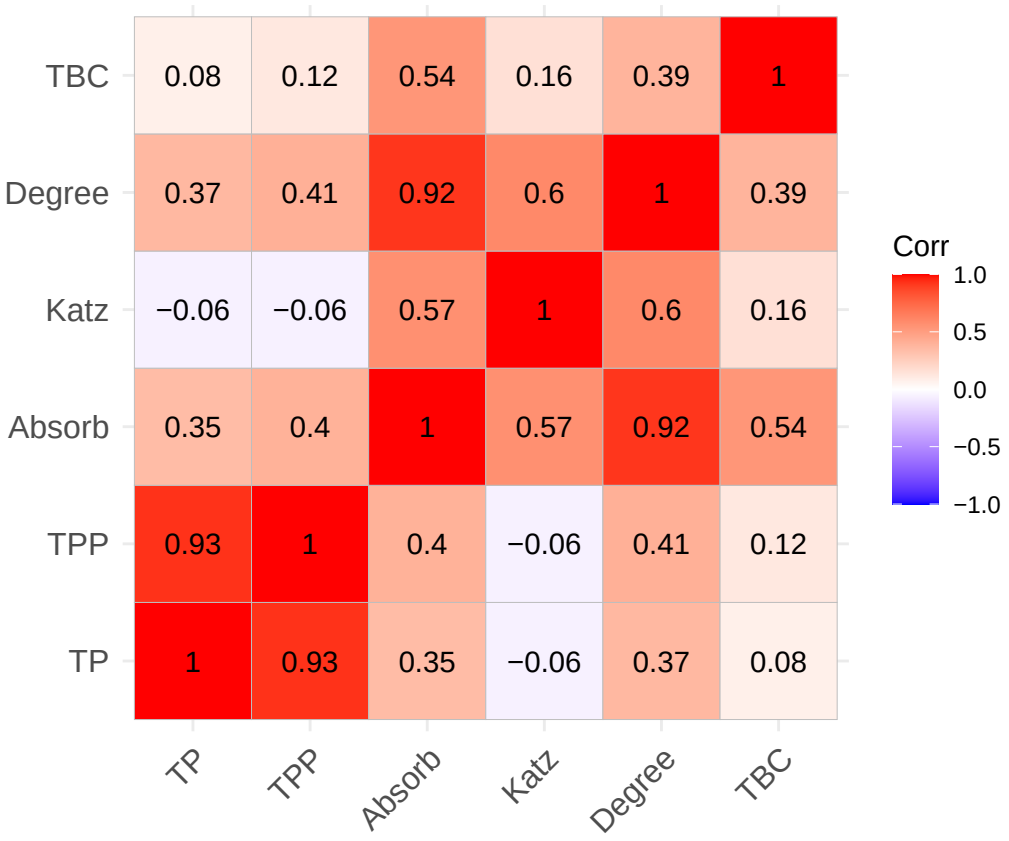


Figure 8: Correlations between the five centrality measures.

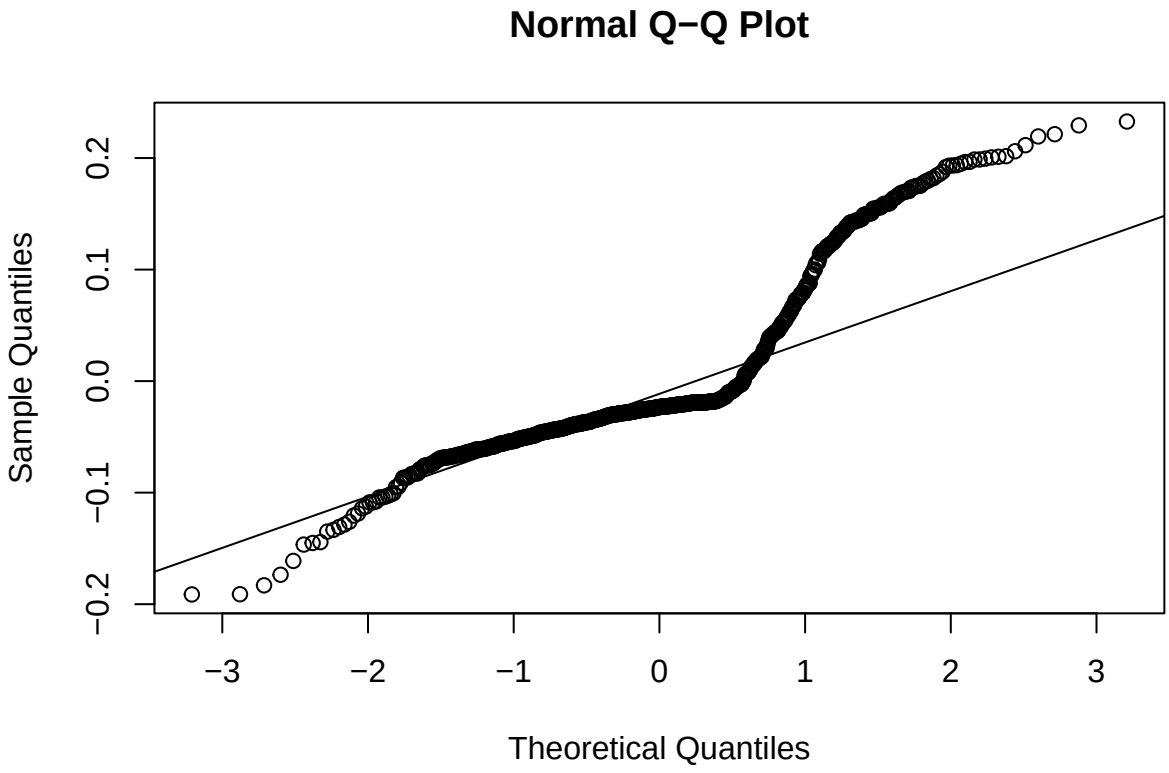


Figure 9: Standardized residuals of the model plotted against theoretical quantiles of the normal distribution.

Model Selection

Figure 9 shows the Q-Q plot for a linear regression model predicting the observed likelihood of infection, with the five centrality metrics as predictors. The relationship is clearly non-linear, with a large number of extreme values for the residuals. A transformation of the output variable may satisfy the linearity assumption, particularly if the distribution is skewed.

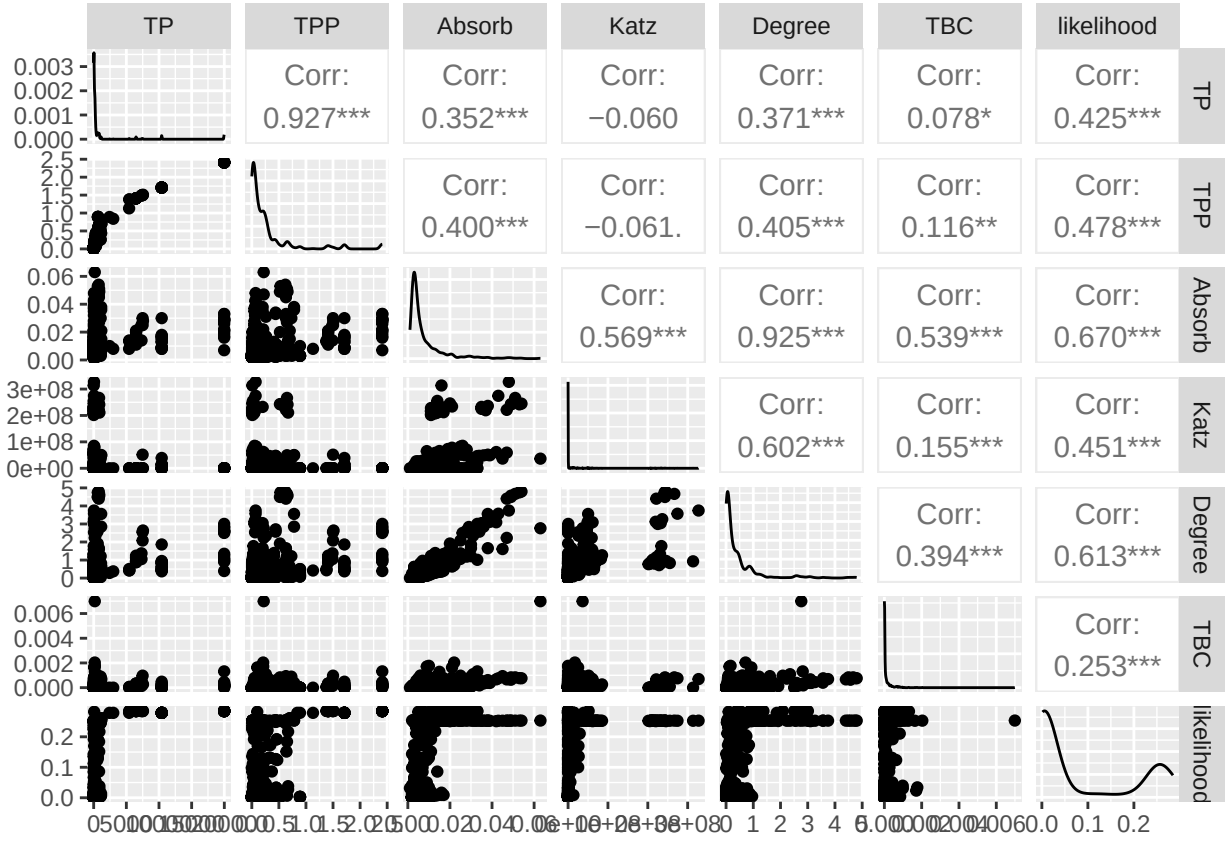
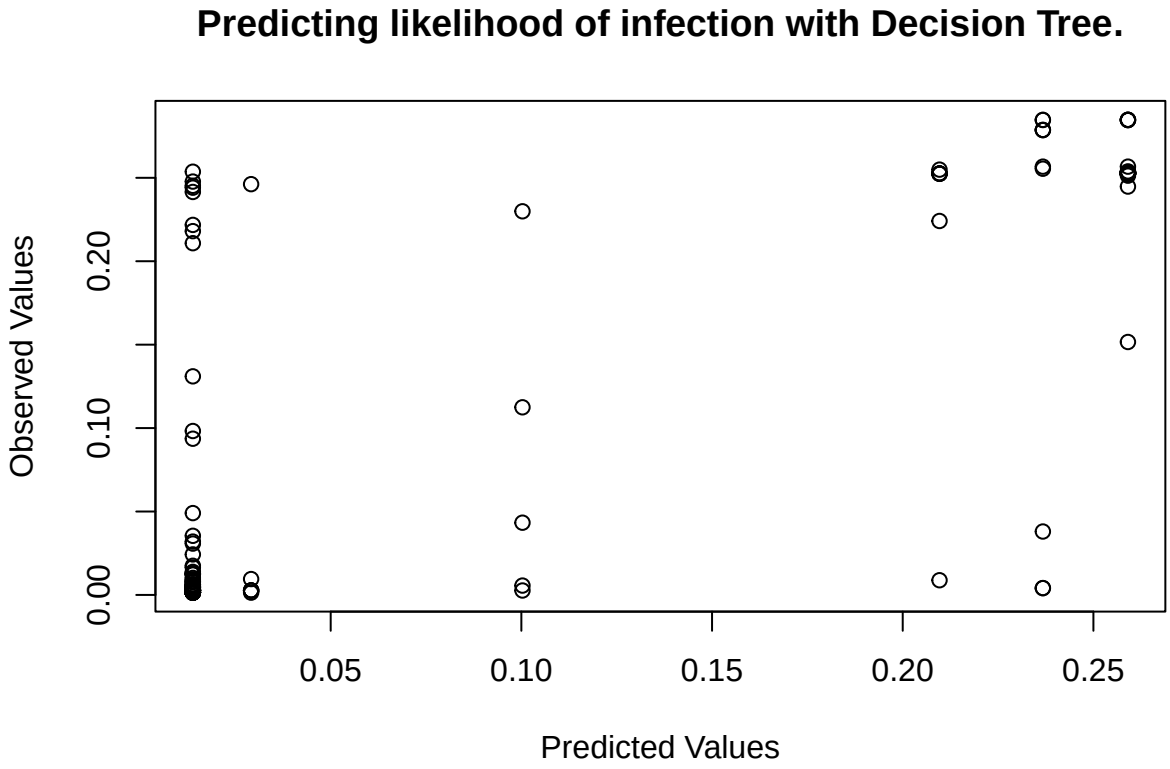


Figure 10: Pairwise relationships between the centrality metrics and the output variable (likelihood)

Looking at the pairwise relationships (shown in Figure 10), the assumptions of linear regression clearly cannot be satisfied. The output variable (plotted in the bottom-right corner) is bimodal and not amenable to common transformations (e.g. log, square root, polynomial etc). Machine Learning and Artifical Intelligence are often used for regression when the distribution of the data is non-standard. We use three machine learning methods - Gradient Boosting Regression, Decision Trees and Polynomial Regression - to predict the likelihood of infection from the six centrality metrics.

Polynomial Regression is based on a similar concept to linear regression, however it allows for powers greater than one in the regression equation (e.g. $y = \beta_1 X + \beta_2 X^2$). Closed-form solutions are rarely known, so the coefficients must be "learned" by numerical optimization methods. Decision Trees split the data into progressively smaller subgroups using decision rules (e.g. X $>$ 40), and each group (sometimes called a leaf node) is assigned a predicted value. The prediction for a given observation is then the predicted value of it's respective leaf node, which is determined by the decision rules. At each branch of the tree, the optimal decision rule is found by minimizing a splitting criteria (e.g. Sum of Squared Errors (SSE), Mean Squared Error (MSE) etc). Gradient Boosting Regression takes a weighted average of predictions for many decision trees. This may result in better predictions, especially when the number of predictors is large or the data is very noisy.

The data was split into training and test sets, comprising 75% and 25% of the data respectively. The models were trained on the training set and evaluated on the test set. Hyper parameters were optimized by 10-fold cross validation.



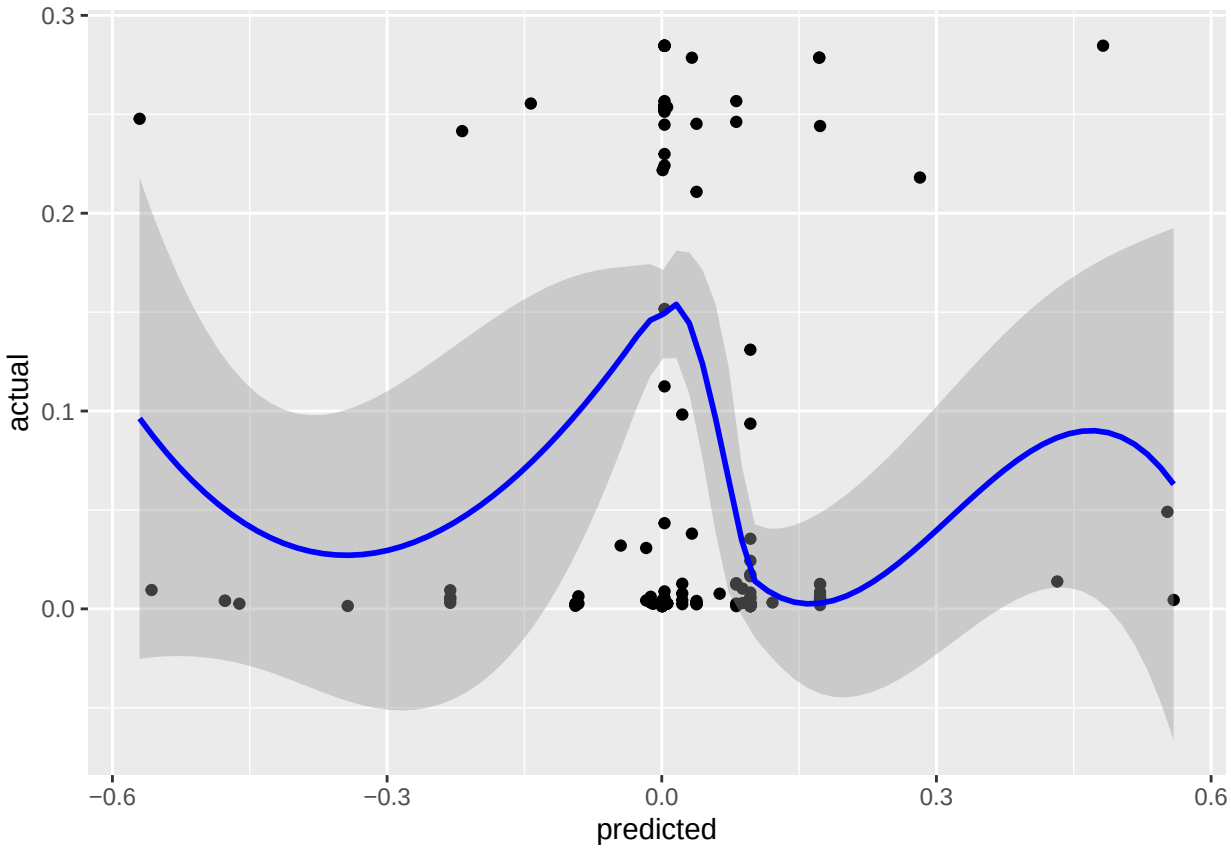
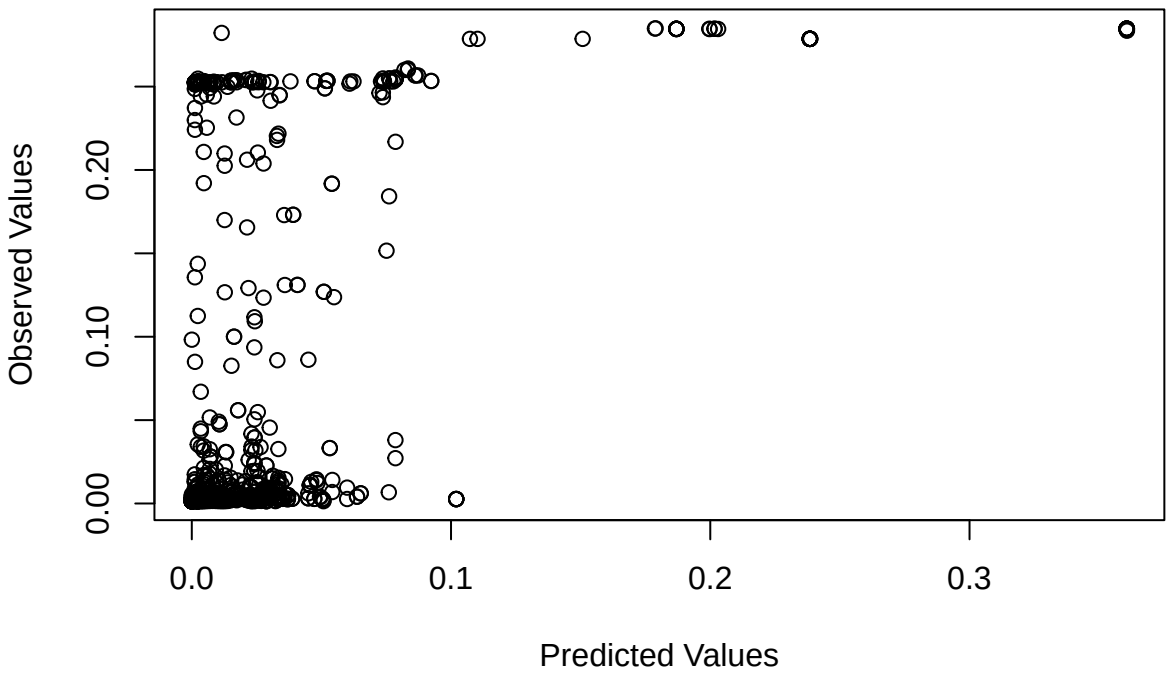
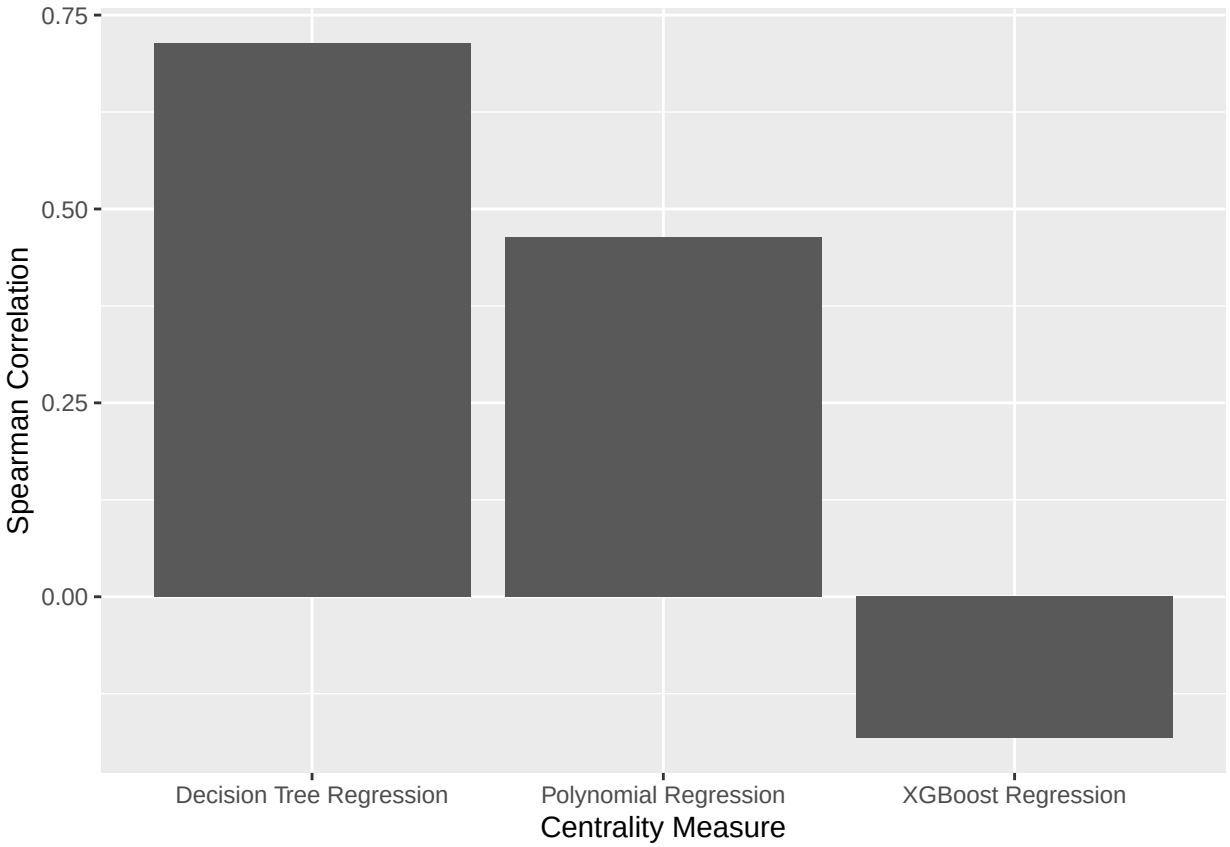
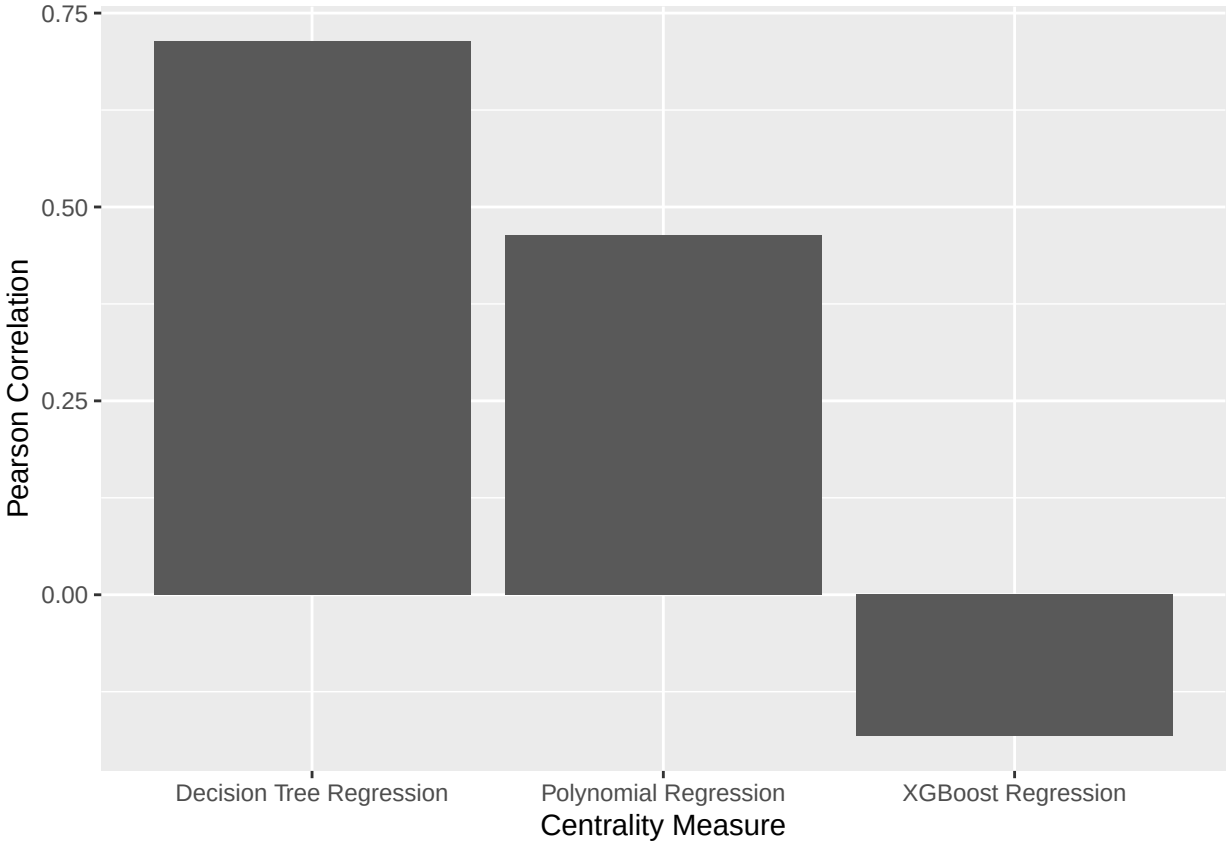


Figure 11: Predicted vs Actual values after training Gradient Boosting model and applying to test set.


Predicting likelihood of infection with Polynomial Regression
Observed Values
0.00
0.10
0.20
0.0
0.1
0.2
0.3
Predicted Values

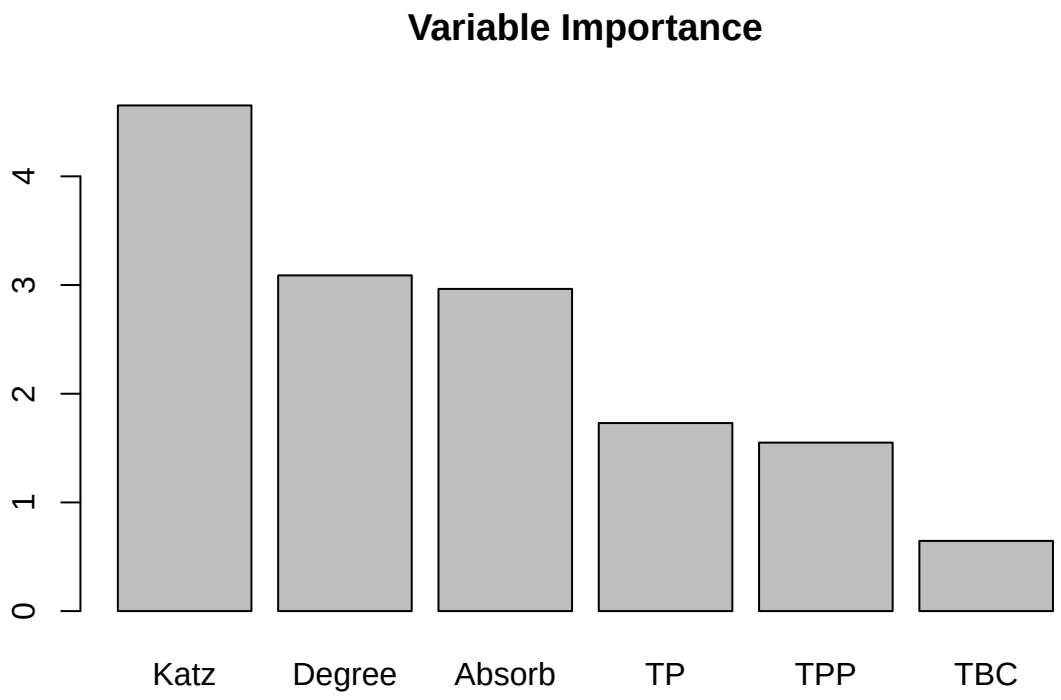

0.75
0.50
0.25
0.00
Spearman Correlation
Decision Tree Regression
Polynomial Regression
XGBoost Regression
Centrality Measure



Figures 11-13 show the predicted values plotted against observed values for the three algorithms. Figures 14 and 15 show the Spearman and Pearson correlations. All algorithms performed worse than Absortion Rank on both metrics. XGBoost and polynomial regression both performed poorly, whereas the decision tree performed similarly to Absorption Rank.

Feature Selection

If some predictors in a model are unimportant, we risk losing model performance due to the "Curse of Dimensionality". Large numbers of unimportant predictors may result in increased overfitting and difficulty with finding meaningful patterns in the data. The process of selecting a good set of predictors is called feature selection.



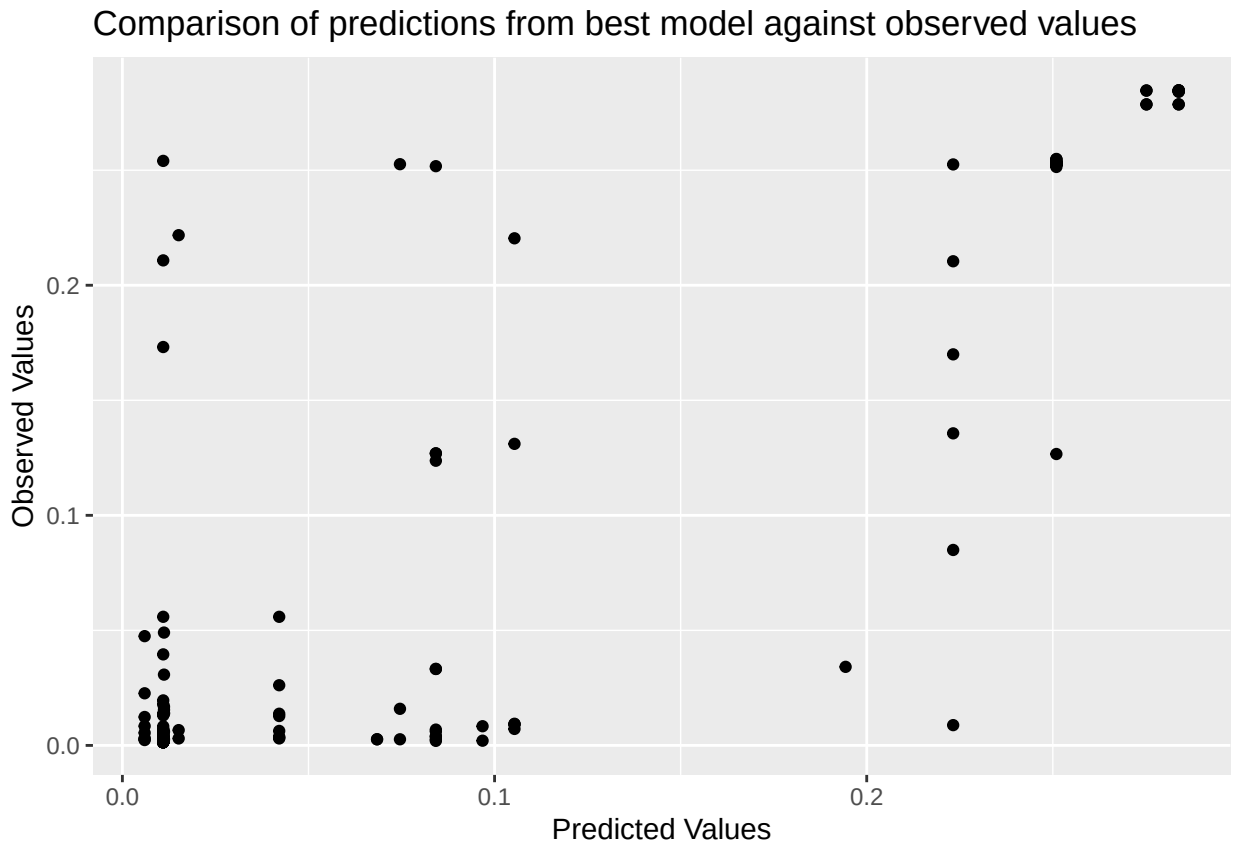
Katz centrality, Degree Centrality and Absorption Centrality had significantly higher importance rankings than other predictors. However, correlated predictors (for instance, TP and TPP) may have attenuated importance rankings due to multicollinearity (high correlation between features). For this reason, iterative feature selection methods (i.e. testing many combinations of features systematically) are often used in lieu of feature importance scores.

We test all possible combinations of two or more features by 5-fold cross validation. Each model was tested with four different values of the complexity parameter, a regularization criterion which limits the complexity of the tree in order to prevent over-fitting. Models were evaluated by the Mean Squared Error (MSE) metric, and the best-performing model was tested against a holdout set.

Table 1: Five best performing feature combinations and their MSE

MSE	TP	TPP	Absorb	Katz	Degree	TBC
0.002228	✗	✗	✓	✓	✓	✗
0.002441	✓	✓	✓	✗	✓	✗
0.002537	✓	✓	✓	✗	✓	✓
0.002538	✓	✓	✓	✓	✗	✗
0.002552	✗	✓	✓	✗	✓	✓

The best-performing model used Absorption Rank, Temporal Katz Centrality and Degree Centrality. As expected, Absorption Centrality was included in all five models, however the inclusion of Temporal Proximity Prestige did not exclude Temporal Prestige. Figure 12 compares the Root Mean Squared Error (RMSE) of the final model to that of the individual predictors. Katz Centrality and Temporal Prestige were omitted due to extremely large RMSE values.



Clearly, the Decision Tree model has the lowest RMSE, followed closely by Absorption Rank and Temporal Betweenness Centrality. It should be clear that RMSE rankings are only meaningful in the context of prediction, and these results should be interpreted accordingly. When ranking the relative importance of nodes in a network, other metrics should be given greater importance (e.g. Spearman Rank Correlation).

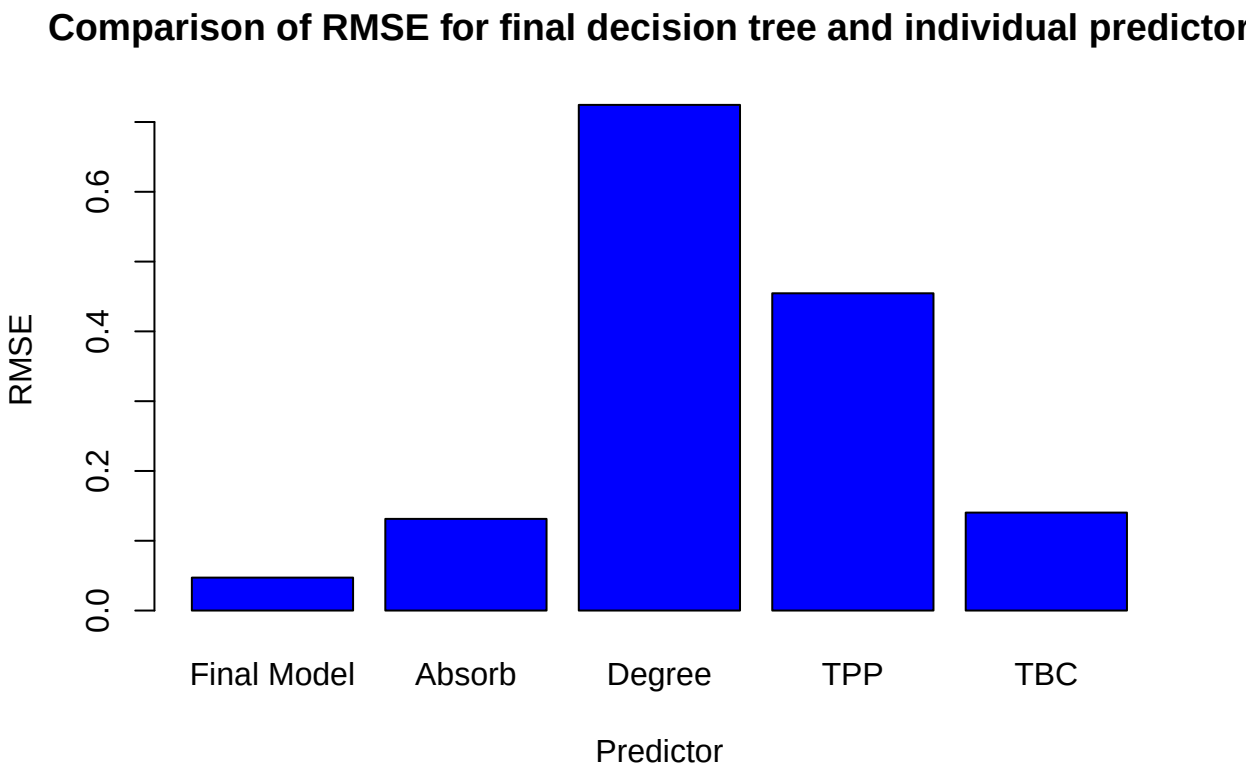


Figure 12: Comparison of RMSE for final decision tree and individual predictors

Constant Transmission Probability

The constant transmission probability assumption is commonplace in the literature, most likely because it greatly speeds up the simulation algorithm. Probabilities were estimated by using proximity categories, which we will call classes, as predictors in a generalized linear model. The number of 15-time intervals spent in each class is given over the two hour observation period. Transmission probabilities were then calculated by a logistic regression model of the form:

$$log\left(\frac{\pi(x)}{(1-\pi(x))}\right) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3$$

where $\pi(x)$ is the transmission probability of the contact and c_i is the number of 15-minute time intervals spent in class i during the two hour observation period. The classes are indexed in increasing order of proximity (i.e. class 3 indicates greater proximity than class 2), therefore

the coefficients were chosen to satisfy the constraint $\beta_3 > \beta_2 > \beta_1$. One may treat the classes as levels of an ordinal variable and assume a proportional relationship, which reduces the model to:

$$log\left(\frac{\pi(x)}{(1-\pi(x))}\right) = \alpha + \beta_1(c_1 + 2c_2 + 3c_3)$$

However, this simplified model was rejected due to insufficient evidence to suggest a proportional relationship.

To test the constant transmission probability assumption, we compare the observed probability of infection for 751 individuals for two different algorithms (run on the same temporal network). In the first algorithm, the probability of transmission is calculated separately for every contact. In the second algorithm, we assume that for two nodes i and j, the probability of transmission is a constant value p_{ij}, regardless of the duration and proximity of the contact event, and this probability is estimated by averaging the event-specific probabilities p_{ijt} (where t is a time index) over all time points. As in the first algorithm, event-specific probabilities are calculated by the same logistic regression model.

In total, 1000 simulations were run for each algorithm. For each run, the indicator variable, I_i, is 1 if node i was infected, and 0 otherwise. The observed probability of infection for node i is the average of I_i over all 1000 runs.

Denote by $\hat{P^{\mathbf{1}}}$ the vector of observed probabilities for algorithm 1 (and likewise for algorithm 2). Thus, $\hat{P_i^{\mathbf{1}}} - \hat{P_i^{\mathbf{2}}}$ is the observed difference between the two algorithms for node i. Suppose we want to test whether this difference is equal to 0. The test statistic is

$$Z = \frac{\hat{P_i^{\mathbf{1}}} - \hat{P_i^{\mathbf{2}}}}{\hat{P}_i(1-\hat{P}_i)(\frac{1}{1000} + \frac{1}{1000})}$$

where $\hat{P}_i$ is the pooled proportion under the null hypothesis ($P_i^{\mathbf{1}} = P_i^{\mathbf{2}}$). This is a well-known test statistic which follows a standard normal distribution when the true proportions are equal. In order to test for equality of the vectors $P^{\mathbf{1}}$ and $P^{\mathbf{2}}$, we must adjust the significance level of each individual test to correct for multiple comparisons. Power and Type I error

rate are two important factors which influence our choice of adjustment. Power is defined as the probability of falsely accepting the null hypothesis, whereas the Type I error rate is the probability of falsely rejecting the null hypothesis. Multiple comparison adjustments aim to reduce the family-wise Type I error rate by controlling the significance level of each individual comparison. We will discuss several common adjustment methods and their use cases:

Bonferroni Correction: The Bonferroni correction, proposed by Dunn (1961), is an adjustment which controls the family-wise Type I error rate. For a given significance level α, the Bonferroni correction guarantees a family-wise Type I error rate which is $\leq \alpha$. It does this by setting the test-wise significance level to $\frac{\alpha}{N}$, where N is the number of tests. The Bonferroni correction is ideal for this experiment because it makes no assumptions about independence between the individual tests. Note that the Bonferroni correction is conservative, meaning it lacks power for rejecting the null hypothesis.

Holm's Method: Holm's method, proposed by Holm (1979), is a powerful alternative to the Bonferroni correction. Holm's method tests the hypotheses iteratively, updating the p-value at each step. First, we sort the list of p-values in increasing order, and we begin the sequential significance tests from the lowest p-value. The algorithm starts with a significance level of $\frac{\alpha}{N}$. If the first result is non-significant, we test the second result with a significance level of $\frac{\alpha}{N-1}$. In general, the i'th test statistic is tested with a significance level of $\frac{\alpha}{N-i+1}$. Holm's method also guarantees a family-wise Type I error rate of $\leq \alpha$, however it offers greater power than the Bonferroni correction.

Hochberg Procedure: The Hochberg procedure is similar to Holm's method, however it assumes non-negative correlation between tests. We begin by testing the largest p-value, adjusted for a single comparison. If the p-value is insignificant, we test the next largest p-value sequentially. In general, the i'th largest p-value is tested by adjusting for i comparisons. By doing this, we guarantee a greater power than Holm's method and the Bonferroni Correction.

Pairwise comparisons were carried out with Fisher's Exact Test and p-values were simulated due to computational constraints. The Z-test was rejected due to the presence of proportions

Table 1: Number of rejections for each method

Method	Number of Rejections	Proportion of Rejections
Bonferroni	24	0.03196
Holm	24	0.03196
Hochberg	24	0.03196

close to 0, which violated the $np > 5$ assumption. Multiple comparisons were controlled for by the three methods. The results are shown in Table 1.

The null hypothesis was rejected 24 times for all multiple comparison adjustments, which suggests that the constant transmission probability assumption is invalid. It is worth noting that the Bonferroni Correction is extremely conservative for 751 comparisons, therefore this result provides strong evidence of a difference between the algorithms. Alger (2020) promotes the use of multiple experiments to increase the reliability of a conclusion. Accordingly, we use a goodness of fit test to support the hypothesis that applying the constant transmission probability assumption produces different results.

Goodness-of-fit test

In this section, we test the constant transmission probability assumption by a likelihood ratio test. First, we sum I_i over all runs to get the total times infected, n_i. By doing this for every node, we obtain a contingency table of the form:

Node	1	2	3	4	...
Algorithm 1 (Constant Transmission)	n_{11}	n_{12}	n_{13}	n_{14}	...
Algorithm 2 (Varying Transmission)	n_{21}	n_{22}	n_{23}	n_{24}	...

Where n_{ij} is the number of times node j is infected for algorithm i. Denote by n_{i+} the i'th row sum, and similarly by n_{+j} the j'th column sum. Under the assumption of independence between the two variables, algorithm and node, the expected value of the cell count, E_{ij}, is given by:

$$E_{ij} = \frac{n_{i+}n_{+j}}{n}$$

where $n = \sum_i n_{i+}$. The likelihood ratio test statistic

$$G = 2 \sum_{ij} n_{ij} ln \left(\frac{n_{ij}}{E_{ij}} \right) \tag{14}$$

asymptotically converges to a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom, where I and J are the number of rows and columns respectively in the contingency table.

Table 2: Likelihood Ratio Test Results

statistic	p.value	parameter	method
2330.044	0	750	Log likelihood ratio (G-test) test of independence without correction

The likelihood ratio test (Table 2) corroborates the conclusion that the two algorithms produce different outcomes, with $p - value < 1 \times 10^{-20}$.

References

Admiraal, Ryan, Jules Millen, Ankit Patel, and Tim Chambers. 2022. "A Case Study of Bluetooth Technology as a Supplemental Tool in Contact Tracing." *Journal of Healthcare Informatics Research* 6 (2): 208–27.

Alger, Bradley E. 2020. "Scientific Hypothesis-Testing Strengthens Neuroscience Research." *Eneuro*. Society for Neuroscience.

Bavelas, Alex. 1950. "Communication Patterns in Task-Oriented Groups." *Journal of the Acoustical Society of America*.

Dunn, Olive Jean. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56 (293): 52–64.

Erdos, Paul, Alfréd Rényi, et al. 1960. "On the Evolution of Random Graphs." *Publ. Math. Inst. Hung. Acad. Sci* 5 (1): 17–60.

Fetzer, Thiemo, and Thomas Graeber. 2021. "Measuring the Scientific Effectiveness of Contact Tracing: Evidence from a Natural Experiment." *Proceedings of the National*

Academy of Sciences 118 (33): e2100814118.

Freeman, Linton C. 1977. "A Set of Measures of Centrality Based on Betweenness." *Sociometry*, 35–41.

Grindrod, Peter, Mark C Parsons, Desmond J Higham, and Ernesto Estrada. 2011. "Communicability Across Evolving Networks." *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 83 (4): 046120.

Hanke, Moritz, and Ronja Foraita. 2017. "Clone Temporal Centrality Measures for Incomplete Sequences of Graph Snapshots." *BMC Bioinformatics* 18: 1–18.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 65–70.

Holme, Petter. 2018. "Objective Measures for Sentinel Surveillance in Network Epidemiology." *Physical Review E* 98 (2): 022313.

———. 2021. "Fast and Principled Simulations of the SIR Model on Temporal Networks." *Plos One* 16 (2): e0246961.

Katz, Leo. 1953. "A New Status Index Derived from Sociometric Analysis." *Psychometrika* 18 (1): 39–43.

Kempe, David, Jon Kleinberg, and Amit Kumar. 2000. "Connectivity and Inference Problems for Temporal Networks." In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, 504–13.

Kermack, William Ogilvy, and Anderson G McKendrick. 1927. "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115 (772): 700–721.

Kim, Hyoungshick, and Ross Anderson. 2012. "Temporal Node Centrality in Complex Networks." *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 85 (2): 026107.

Kiss, István Z, Joel C Miller, Péter L Simon, et al. 2017. "Mathematics of Epidemics on Networks." *Cham: Springer* 598 (2017): 31.

Klovdahl, Alden S. 1985. "Social Networks and the Spread of Infectious Diseases: The AIDS Example." *Social Science & Medicine* 21 (11): 1203–16.

Lin, Nan. 1976. "Foundations of Social Research." *(No Title)*.

Macdonald, George. 1952. "The Analysis of Equilibrium in Malaria." *Tropical Diseases Bulletin* 49 (9): 813–29.

Piraveenan, Mahendra, Mikhail Prokopenko, and Liaquat Hossain. 2013. "Percolation Centrality: Quantifying Graph-Theoretic Impact of Nodes During Percolation in Networks." *PloS One* 8 (1): e53095.

Rocha, Luis EC, and Naoki Masuda. 2014. "Random Walk Centrality for Temporal Networks." *New Journal of Physics* 16 (6): 063023.

Seidel, Scott Y, and Theodore S Rappaport. 1992. "914 MHz Path Loss Prediction Models for Indoor Wireless Communications in Multifloored Buildings." *IEEE Transactions on Antennas and Propagation* 40 (2): 207–17.