UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics and Data Science

Master's thesis

About social and spatial proximity and leprosy occurrence: a whole network analysis in a Comorian village

Doortje Theunissen

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :
Prof. dr. Andrea TORNERI

SUPERVISOR :
Tine VERDONCK

2024
2025

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

About social and spatial proximity and leprosy occurrence: a whole network analysis in a Comorian village

Doortje Theunissen

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :
Prof. dr. Andrea TORNERI

SUPERVISOR :
Tine VERDONCK

Abstract

Leprosy remains a persistent public health concern in many parts of the world, particularly in endemic areas like the Comoros. Despite the wide-spread occurrence of leprosy, its exact mechanisms of transmission are still not fully understood, with current research indicating that prolonged social contact may play a significant role. In previous research spatial proximity to a leprosy case has been used as a risk factor for a leprosy diagnosis. This study investigates if social proximity can be a better proxy than spatial proximity when it comes to strategies for detecting leprosy cases. Based on survey data on social contacts, four definitions of close contact were created to construct whole-village networks: contacts within the home area, blood ties and partners, links through activities, links through public places. The drivers behind the links in these networks were studied using exponential random graph models. The overlap between social and spatial proximity was assessed, and associations between individual network characteristics and leprosy diagnoses from 2002 to 2023 were analyzed using logistic regression. The results reveal substantial variation in the structure of the networks based on close contact definitions and none of the networks have a complete overlap with the spatial proximity diameters used in previous work. The strongest associations with leprosy occurrence were found in networks based on household and familial ties. The results highlight the importance of integrating social network analysis into disease control strategies, and suggest that targeting both household and close social contacts may improve post-exposure prophylaxis programs.

1 Introduction

Leprosy is a well-known and widespread chronic infectious disease that circulated among human populations for thousands of years. The disease causes skin lesions and affects peripheral nerves. Leprosy is caused by *Mycobacterium leprae*, which was first identified by Hansen in 1873 [8]. Reliable accounts of leprosy can be found in ancient Greek and Roman literature as early as the first century CE. Moreover, probable descriptions of the disease have been identified in Egyptian texts dating back to 1600 BCE and Indian text from around 600 BCE. Though the exact origin of leprosy is unknown, it is believed to have originated in present-day India, with East Africa as a possible alternative. Researchers have identified two major expansion events in the last two millennia. The first started around 250 CE and is linked to the Roman expansion. The second started around 1600 CE, after the arrival of Europeans in the Americas [16].

1.1 Study Area

The Comoros is a country in southeastern Africa that consists of three islands. The study area for this project is a village on the island of Anjouan. Leprosy is known to be highly endemic in the Comoros. In 2023, there were 241 new leprosy cases out of a population of 850 387 [25], [26]. Anjouan has been reported to have an annual incidence of 5-10 cases per 10.000 inhabitants [10]. Anjouan is a volcanic island where people live concentrated in a small number of cities and villages, with the largest cities being the capital Mutsamudu, Tsimbeo and Domoni. A few, mostly coastal, roads connect the cities and villages and leave large uninhabited areas in between [6], [7]. The study village has a population of less than 3000. The name of the study village is kept out of the text to avoid stigmatising its inhabitants. The village consists of 7 neighbourhoods, which are called district A-G in the rest of the text.

1.2 Epidemiology of Leprosy

With 200 000 new cases reported annually and the disease present in over 120 countries, the WHO considers leprosy a neglected tropical disease and aims to eliminate it. Through sustained efforts, the global target of reducing prevalence to less than 1 case per 10 000 people was achieved in 2000 and by 2010, in most individual countries. However, the decline in new cases has been slowed since then. In 2023, 182 815 new cases were reported worldwide. Of these, 107 851 were reported in India, 22 773 in Brazil, and 14 376 in Indonesia [24], [25].

Leprosy is not highly contagious and is thought to be transmitted via respiratory droplets from the nose and mouth of untreated individuals. Transmission requires repeated and prolonged close contact. Once treatment begins, the individual is no longer infectious [24]. Leprosy has a long incubation period, which may extend up to 12 years [18]. *M. leprae* has a doubling time of just under 13 days in host tissue and has undergone reductive evolution, resulting in the accumulation of over 1130 pseudogenes. This loss of functional genes is reflected in its severe metabolic constraints. Due to these constraints and its slow doubling time, *M. leprae* is very difficult to culture in a laboratory setting, making it difficult to study [11].

The WHO classifies leprosy into two types: paucibacillary (PB) and multibacillary (MB). These types are defined as follows [24]:

- PB case: a case of leprosy with 1–5 skin lesions, without demonstrated presence of bacilli in a skin

smear.

- MB case: a case of leprosy with more than five skin lesions; or with nerve involvement (pure neuritis, or any number of skin lesions and neuritis); or with the demonstrated presence of bacilli in a slit-skin smear, irrespective of the number of skin lesions.

The cell-mediated immune response to infection in the affected individual determines the type of leprosy that they develop. Better treatment could be developed to prevent nerve damage if the immune response to leprosy were understood better [18]. The current recommendation to treat leprosy is a multi-drug therapy consisting of dapsone, rifampicin and clofazimine. The WHO recommends 6 months of treatment for PB cases and 12 months for MB cases [24].

Relapse and reinfection can occur in individuals who have been treated for leprosy. Relapse can occur soon after treatment or up to many years later. In the former, it is seen as a failed or insufficient treatment, in the latter it is assumed to be caused by the persistence of "hibernating" bacilli. Dormant bacteria are not as susceptible to treatment and can reactivate when conditions are favorable [3]. To distinguish between relapse and reinfection, whole genome sequencing can be used to test for sequence differences of *M. leprae* between the first and second infection. The sequence information can also be used to keep an eye on reinfection caused by drug-resistant bacteria [22].

1.3 Contact Networks

To facilitate a further decline in leprosy cases, efforts have been made to prevent new cases and to discover already existing cases in order to treat them. Both immunoprophylaxis (e.g. the BCG vaccine) and chemoprophylaxis (medication to prevent infection after exposure) have been proven to help prevent disease in contacts of leprosy cases [20]. To make prophylaxis programs as effective as possible it is important to select the right individuals to treat. Efforts have focused on household members, but clustering of leprosy has also been shown beyond the household [15]. In past undertakings, spatial proximity to infected individuals was used to discover new cases. In the PEOPLE project that was also conducted on the island of Anjouan ([9], [1]), screening individuals within the range of 25 m (household members and neighbours) of a detected case led to detecting 31% of all new cases. Going up to a range of 100 m led to detecting another 47% of all new cases. However, the effort required to detect new cases generally increases significantly with expanding the range around an individual.

To possibly minimize the effort of prophylaxis programs and the detection of new cases, this project investigates if social contacts could be the focus of intervention programs instead of individuals that live within a certain perimeter. To do so, whole-village contact networks relying on different types of social contacts have been reconstructed and analyzed for the population of the study village. This means that information was collected about as many individuals as possible from the same village (seen as ego individuals), but they were also allowed to name contacts outside of the village (seen as alter individuals). Three research questions/goals were described as follows as a guide to the project:

1. **Social proximity networks:** Formulate, summarise, visualise, and compare different versions of the whole-village network, using different definitions of close social contact.
2. **Social versus spatial proximity:** Explore to what extent social proximity (based on social network data) overlaps with spatial proximity (based on geolocation data of houses and public places).

3. **Associations between close social contact and leprosy occurrence:** Assess associations between individuals' network characteristics (exposure variable) and the occurrence of leprosy (outcome variable).

Networks are defined by nodes and edges. In our case, the nodes represent the individuals in the study and the edges represent the connections between them. An edge is always defined by the two nodes it connects. In general, edges can be directional, having a sender and receiver. In our context, however, connections have the same meaning for the two individuals that partake in them. As a result, our whole-village networks only contain undirected edges.

Exponential-family random graph models

Network data differs from most traditional datasets in that its data points are not independent. Edges depend on each other, for example, a friend of a friend is often also a friend. Furthermore, there is also an interdependence between edges and nodes, because a node's characteristics can cause an edge to occur, but the presence of an edge can also influence the characteristics of the node.

One approach to addressing the dependence between nodes in a network is the use of exponential-family random graph models (ERGMs). In these models, each predictor represents a specific configuration of the connections in the network, and the term itself is a function of that specific configuration. At the same time, these terms define both the probability of each individual edge as well as the entire network ([12]).

The general form of an ERGM is:

$$\Pr(Y = y) = \frac{\exp\left[\boldsymbol{\theta}^\top g(y)\right]\ h(y)}{k(\boldsymbol{\theta})}$$

where:

- Y is the random variable for the state of the network and y is a particular realization Y could take,
- $g(y)$ is a vector of model statistics for network y,
- $h(y)$ is a reference measure (which defines the baseline behavior of the model when $\boldsymbol{\theta} = 0$),
- $\boldsymbol{\theta}$ is the vector of coefficients for those statistics, and
- $k(\boldsymbol{\theta})$ is the summation of the numerator's value over the set of all possible networks y, typically taken to be all networks with the same node set as the observed network.

([21])

Network characteristics

Networks can be quantitatively described by several metrics. A **component** is a group of nodes that can all be reached using each other as a starting point and using the connections between them to form a path. When a graph consists of only one component, then every nodes is connected to the others in some way (not necessarily directly!). If a node is not part of a component, it is called an **unconnected node**.

The **degree** of a node is the number of direct connections it has with other nodes. If I have four friends, my node in the network has a degree of 4. The **geodesic** refers to the shortest path between two nodes, i.e. the minimum number of edges you need to follow to get from one node to another. The **diameter** of a network is the largest geodesic in that network. The **betweenness** refers to the degree to which nodes stand between each other, i.e. How many geodesics pass through each node? Clustering refers to a group of nodes that are more closely connected to each other than to other nodes. The **clustering coefficient** is a measure of the degree to which nodes tend to cluster together. **Transitivity** refers to closed triangular relationships (meaning that my friend's friend is also my friend).

Networks based on social contacts are characterized by a few aspects specifically. Firstly, they often show high clustering and transitivity. For a disease such as leprosy, high clustering in a social network could explain why pockets of infected individuals occur and why the disease is endemic to certain regions. High transitivity creates a scenario in which an individual is more likely to be exposed to the disease from more than one source at the same time. Secondly, social networks show assortative mixing, meaning that there are more connections between nodes with similar attributes [13]. The realization that connections in social networks are not random, means we can study the drivers behind the formation of connections. I.e., which attributes are important when it comes to assortative mixing?

2 Methodology

2.1 Data and Software

Data collection

Three different datasets were used during this study. The first provided information on the location and size of the houses in the study village. They were recorded and saved as a .shp file. The second dataset provided information on the leprosy diagnoses between 2002 and 2023 (n = 168). This information was made available by the Institute of Tropical Medicine in Antwerp (ITG) with the permission of the participants. The organizations that played a key role in diagnosing leprosy and collecting the related information were the National Tuberculosis and Leprosy Control Program (Moroni, Comoros) and the Damien Foundation (Brussels, Belgium) in collaboration with ITG and other academic and public health organizations). The dataset contained a list of the data ID's of individuals that had a leprosy diagnoses at some point during 2002 and 2023. The third dataset was newly collected data on the inhabitants of the study village. Data was collected for 2548 individuals and is missing for 197 village inhabitants. These data were collected by means of two surveys, on household and individual level. The inhabitants were surveyed both on their personal information and social contacts. The social contacts that were mentioned during the surveys were allowed to be any other person, independent from whether this second person was also part of the study or not. The survey was conducted by researchers who knew the villagers, and were thus able to identify their connections correctly. They used the Network Canvas program ([2]) to convert answers into data files.

Survey 1: Household level

The household survey was answered by the head of the household and was also the only means of collecting information on children under 10 years old. The household survey gives insights into the composition of the household and to which compound the household belonged. Compounds are groups of household that share some of their living spaces and activities. This means that they could be seen as an extended version of the household. Information on household composition includes true household members, but also other individuals who ate or slept in the house but were not part of the household. Eating and sleeping information was saved as xy-coordinates. The head of household would put all individuals on a drawing of a house with several rooms to indicate who ate and slept in the same room. This was done to avoid questioning directly who sleeps with whom. The survey also gave information on French and Coranic school attendance among children under 10 years old. The social connections of young children were also part of this survey, including who their caretakers outside the household are.

Survey 2: Individual level

The individual survey was answered by individuals of at least 10 years old. The survey logic identified three groups for which the questions differed slightly: individuals younger than 15 years, men aged 15 years and older, women aged 15 years and older. In this survey, participants answered questions about themselves and their types of connections. When mentioning the connections, they also provided the age and sex of the connection. The questions addressed subjects such as marriage, past caretakers, playmates, study mates, educational level, school locations (Coranic and French), and occupation.

Data processing

The initial data processing was done by ITG, resulting in seven Excel tables: household information, household composition, individual information, individual connections, age, sex, and consent. The data were received in Excel. This file type is known to contain hidden characters such as invisible spaces, which can prevent proper table joins. As a result, the first step in cleaning was removing unwanted characters in columns such as individual_id.

There were duplicates present in the data. For example, the same group of houses had two different compound_id's so only one was kept. False duplicates, such as individuals appearing twice in the same household but with differing information were not removed from the dataset. These were instead handled in later data processing steps. Minor data quality issues, such as having the string 'Unknown' instead of NA were addressed as well.

Software

The data processing and analyses were ran in RStudio Version 2024.12.1+563, using R version 4.4.3 (2025-02-28) – "Trophy Case". The following R libraries were loaded: base 4.4.3, dplyr 1.1.4, ergm 4.8.1, ggplot2 3.5.2, graphics 4.4.3, grDevices 4.4.3, magrittr 2.0.3, methods 4.4.3, network 1.19.0, openxlsx 4.2.8, png 0.1-8, sf 1.0-2.0, sna 2.8, statnet.common 4.11.0, stats 4.4.3, stringr 1.5.1, tidyr 1.3.1, utils 4.4.3.

2.2 Social Proximity Networks

Four definitions of social contact were described as a starting point for whole-village networks. These definitions are based on human behaviour and habits in daily life. They can be described as follows:

- **Definition 1. Contacts within the home area:** This definition is built upon the space of the home, with individuals moving throughout this space.
- **Definition 2. Blood ties & partners:** This definition represents social bonds that are often stable and long-lasting.
- **Definition 3. Links through activities:** These are social contacts that result from shared interests or activities outside of the home.
- **Definition 4. Links through public places:** This definition describes a social connection through attendance at the same public place, such as a school or religious institute.

Each of these definitions needed to be represented by actual data in order to convert them to social network objects. Table 1 shows which contact types we were able to extract from the surveys to embody each definition. A more detailed explanation of the contact types can be found in the Appendix. An important note to make is that when asked about the contact types in Definition 3, participants were asked to name individuals they had not named before. This means that the individuals they mentioned as a contact in the home area or as a blood tie or partner could not be named again. This does not affect Definition 4, as it is based on mentioned on mentioned schools and religious institutes, not on named contacts.

Table 1: Close contact definitions.

Definition 1	**Definition 2**
living in the same compound	having a child-parent connection
eating together	being co-parents
sharing a bedroom	being siblings
having a child-caretaker connection	being marriage partners
Definition 3	**Definition 4**
perceived close contacts	going to the same Coranic school
being mentioned colleagues	going to the same French school
being mentioned playmates	going to the same mosque
being mentioned study mates	going to the same madrass

For a connection to exist between two individuals according to one of the definitions of close social contact, they need to meet at least one of the conditions that describe that particular definition. For each definition, a whole-village network was created. These networks contain all the individuals as a node and the connections between them as edges.

The networks are described visually and numerically to be able to see how they differ and are alike and discuss possible consequences when conclusions are drawn from them. The visual description is in the form of plots. Even though the plots are an objective representation of the network, they are meant to give a more subjective impression and a 'feel' for the results of the four definitions. Numerical summary metrics were computed to have a purely objective description of the networks and their structures. The summary metrics are the following: number of components, number of unconnected nodes, degree distribution, geodesic distribution, diameter, betweenness, clustering coefficient and transitivity.

Drivers behind social connections

Looking at the four network definitions, the definition for *Links through activities* differs from the others in one important aspect. It puts the question of connection entirely with the interviewee - "Name your close contacts". It is the definitions that comes closest to the idea of friendship. For this reason we have chosen to study the driving attributes behind these connections further. It could be argued that being marriage partners or co-parents are similar in nature to friendships, since romantic connections are assumed to be a personal choice. The difference between the two is that partners often have a higher influence on each others attributes. They become part of the same household and society, their economy becomes very similar. As a result, the drivers behind the formation of the relationship may no longer exist and are, therefore, impossible to study.

An ERGM was fitted to examine how different attributes influenced the probability of the formation of a connection in the whole-village network that created to represent *links through activities*. To define a model, specific terms can be used. The terms that were used in this project are:

- **nodecov**: Main effect of a covariate
- **nodefactor**: Main effect of a factor attribute
- **nodematch**: Uniform homophily and differential homophily
- **absdiff**: Absolute difference
- **edgecov:** Edge covariate

The model was fitted using backward selection and contained the following set as a starting point: **edges**, **age** (nodecov and absdiff), **sex** (nodefactor and nodematch), **alter_connections** (nodecov), **survey_group** (nodefactor), **quartier** (nodefactor and nodematch), **house_size** (nodecov and absdiff), **individual_distance** (edgecov), **gwdegree** (with decay parameter set to 0.8). Four Markov chains were run to fit the model.

2.3 Social Versus Spatial Proximity

Social proximity

Social proximity is seen as a binary data point for each pair of individuals, and according to each definition. If an edge exists between two individuals according to the definition in question, then this data point is coded as a 1, otherwise as zero. This way of working results in four data points for social proximity for each pair of individuals. For example, if two individuals are connected according to Definition 1, the data point for this pair will be 1 for Definition 1.

Spatial proximity

To obtain the distance between all individuals, the latitude and longitude coordinates of the houses in the study village were used. First, they were transformed into a distance matrix between the houses (household level). Since it was known which individual belonged to which households, the distance matrix of households could be extrapolated into a distance matrix on the individual level. However, some individuals were members of multiple households. In that case, the shortest distance between any of their households and all others was used. See Figure 1 for an example.

After obtaining the distances between the individuals, binary data points were also created for spatial proximity. The data points are based on two perimeters, one of 25 m and one of 100 m, meaning that 1 represents that a set of individuals live within the chosen perimeter from each other. The choice to work with perimeters of 25 m and 100 m is based on the methodology of the PEOPLE project that was mentioned earlier ([9], [1]).

Overlap between social and spatial proximity

For each social proximity definition, the overlap is calculated with the datasets from both spatial proximity definitions. The overlap is stated as the number of pairs that have a 1 for both the social and spatial proximity definitions in question. The results eight datasets are plotted. Age and district are included in the plots to give extra insight into the population.

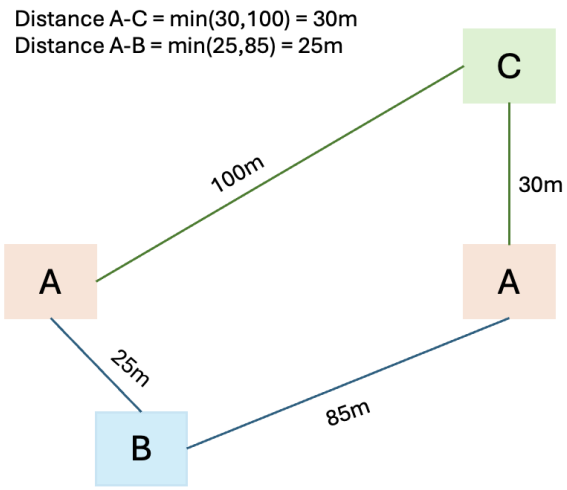


Figure 1: An example of the distance calculation for A-B and A-C.

2.4 Associations Between Close Social Contact and Leprosy Occurrence

To assess associations between individuals' network characteristics and the occurrence of leprosy, logistic regression models were fitted. The included network covariates were the number of contacts according to the different definitions, and the number of diagnosed contacts based on the same definitions. Other covariates were sex, age_group, house_size and quartier.

The number of diagnosed contacts showed a rapid exponential decrease (See Figure 2). To address this, the values were grouped into bins rather than using the raw number of contacts. For the first model, there were three bins per definition. For Definitions 1, 2 and 3, the number of diagnosed contacts was split into: 0, 1, or more than 1 diagnosed contact. For Definition 4 (shared public spaces), they were split into: 0, fewer than 10, and 10 or more diagnosed contacts. For the second model, the number of diagnosed contacts was transformed into a binary covariate (TRUE/FALSE) indicating the presence of diagnosed contacts. The models were fitted using backward selection.

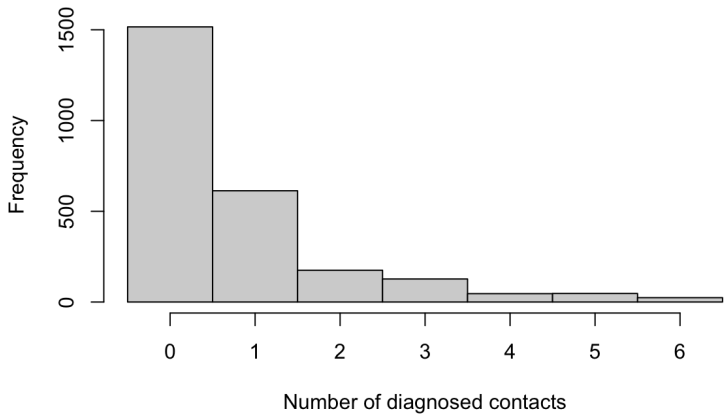


Figure 2: The number of diagnosed contacts that each individual has according to Definition 1

.

3 Results

3.1 Dataset

Base metrics

The household survey produced data on 522 households. The individual survey was answered by 1727 individuals aged 10 and up. Since the individual information on children comes from the household survey, there is a total of 2548 surveyed individuals. These surveyed individuals mentioned 2301 other, non-surveyed individuals as their contacts, of which 197 also live in the same village and 2104 live outside the village.

Age and sex

Age and sex are known for all surveyed individuals. Age is known for almost all non-surveyed individuals, only 11 data points are missing. Sex is known for most non-surveyed individuals, 283 data points are missing. Table 2 and 3 show the age and sex distribution for all individuals. The bins for age distribution align with the survey types. The tables show a difference in age distribution between the surveyed and non-surveyed individuals. The proportion of adults is clearly larger in the non-surveyed individuals, while the number of children is clearly smaller. There also seems to be a difference between males and females in the non-surveyed group.

Table 2: The Age and Sex distribution of surveyed individuals.

	Female	Male	Total
Child (< 10)	384	437	821
Teen ($10 \leq . < 15$)	181	180	361
Adult (≥ 15)	710	656	1366
Total	1275	1273	2548

Table 3: The Age and Sex of non-surveyed individuals as provided by surveyed individuals.

	Female	Male	NA	Total
Child (< 10)	66	87	11	164
Teen ($10 \leq . < 15$)	101	86	5	192
Adult (≥ 15)	657	1020	257	1934
NA		1	10	11
Total	824	1194	283	2301

3.2 Social Proximity Networks

Edge counts

Table 4 shows the number of edges for each connection type and the number of unique edges for each network. The edges represent the connections between the surveyed individuals. The table also shows the number of nodes (individuals) that have at least one connection with another surveyed individual according to the definition used for each network. The networks themselves always include all surveyed individuals (n = 2548) whether or not they are connected to someone else.

When considering the number of edges, we see that Network 4 contains a much higher number of edges than the other networks, especially the contact type *going to the same French school*. The edge count of Network 3 is limited by design. Individuals could only mention a limited number of contacts. Network 2 probably contains missing data as a result from the used definitions (This will be discussed further in the Discussion part of the report).

When considering the number of connected nodes, we see that this number goes down from Network 1 to Network 4. In Network 1, almost all the nodes are connected. In Network 4 around 60% of the nodes are connected.

Table 4: The number of edges and nodes for the different types of contacts and the resulting close contact networks. The number of nodes is kept constant at 2548.

	Edges	Connected nodes
Eating together	1782	1445
Shared bedroom	1833	2196
Child-caretaker link	641	736
Same compound	12409	2541
Network 1 Total	12932	2543
Marriage partners	47	212
Child-parent link	2761	2263
Co-parents	323	582
Siblings	2805	1651
Network 2 Total	5897	2265
Mentioned contacts	1275	1248
Colleagues	715	947
Play together	1708	1469
Study mates	894	745
Network 3 Total	4200	1845
Same Coran school	22684	379
Same French school	137960	919
Same Mosque	71984	722
Same Madrass	2671	142
Network 4 Total	227255	1592

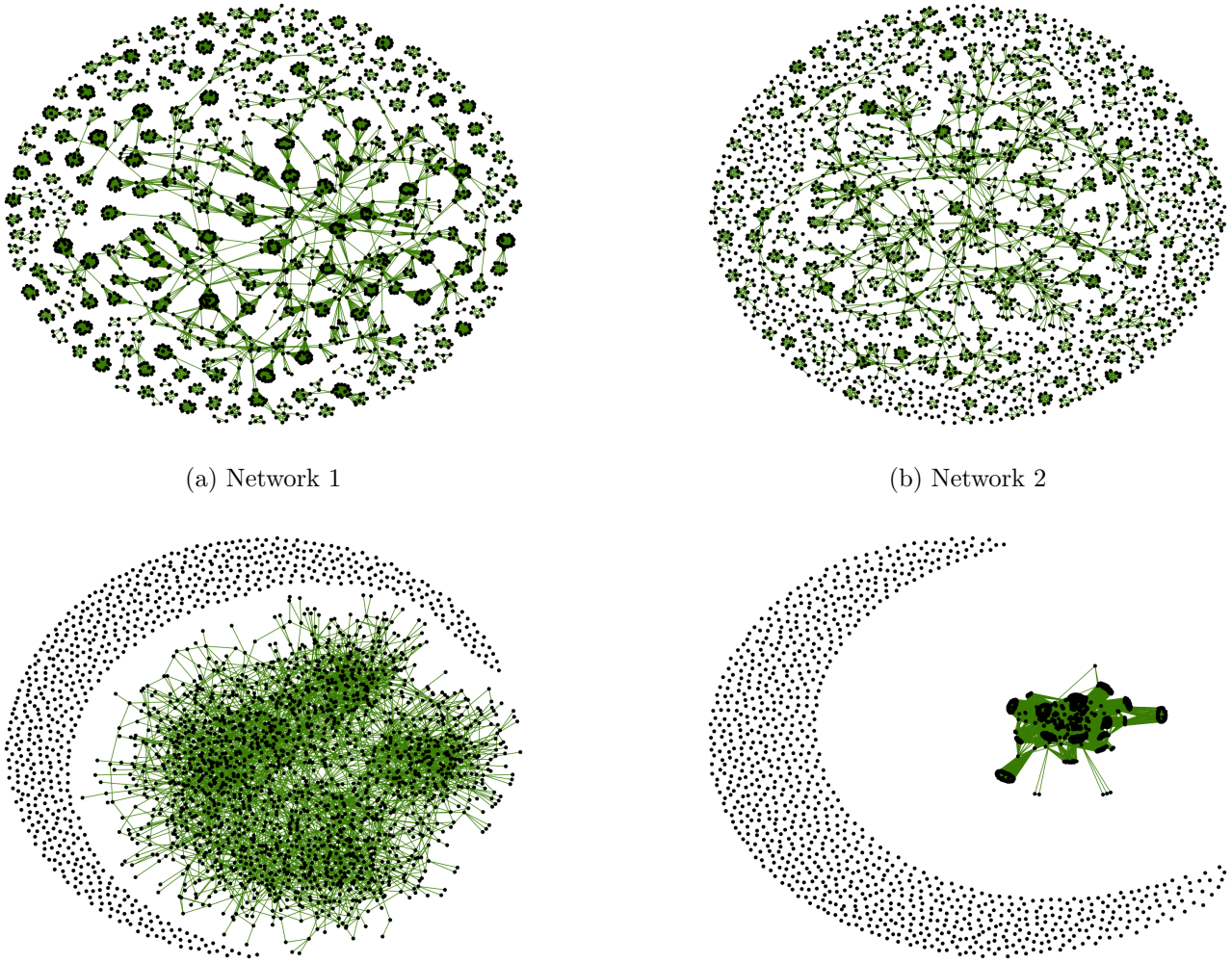
Network visualisations

Visualisations of the four networks can be seen in Figure 3. One of the first aspects about the networks that can be noticed is the clear inclusion of cliques in Networks 1 and 4. This was expected with the definitions that were chosen, since individuals that are a member of the same compound are all connected to each other, and individuals that visit the same public place are also completely connected. Network 2 visually gives the impression of some small cliques as well, but not all of these node groups are. Most households are completely connected when taking partnerships and blood ties into account, but it is not the case per definition.

There is also a significant number of unconnected nodes in Network 3 and Network 4. In Network 3, this is partly the result of including young children in the population. They do not necessarily have friends outside of the compound/family sphere yet, which were not allowed to be mentioned as a playmate. In Network 4 an additional factor could be the data quality of certain information fields. Causing them to not be recognized as being the same.

The connection structure in Network 3 has subjectively a very different look to it. This could be a result

of not displaying clear cliques. It could also be a result of the difference in betweenness distributions (see table 5) for the different networks. Network 3 contains a higher number of individuals that serve as bridges between others.



(a) Network 1

(b) Network 2

(c) Network 3

(d) Network 4

Figure 3: A visualisation of the four whole-village networks using the Fruchterman and Reingold's force-directed placement algorithm. Every black dot represents one individual. Every green line represents a connection between two individuals according to the different definitions for each network. Each visualisation contains the same number of nodes (n = 2548).

Descriptive metrics

Table 5 shows descriptive metrics for each network, providing insight into their structure and key characteristics.

The first metric examined was the number of components. A component is a group of nodes in a network where all nodes are reachable from one another. None of the networks consists of only a single component. In Network 4, all connected nodes can be found in the same component, but it also contains 956 unconnected nodes. Even though the number of components exceeds 100 for all networks, in Networks 1,

3, and 4 over half of the individuals are part of the same component.

Next, we examined the degree distribution. The degree refers to the number of direct connections a node has to other nodes in the network. In all four networks, it can be observed that the number of direct connections differs between individuals, some individuals have more contacts than others. Network 4 is the most extreme in terms of degree distribution. A large part of the population does not have any connections at all, while simultaneously, a large portion of the population has hundreds of connections.

We also examined the geodesic distances and network diameter. The geodesic refers to the shortest path between two nodes, measured by the number of intermediate nodes. This characteristic is only calculated for nodes that are part of a component of more than 1 node, only one component meets this criterion in Network 4. The diameter is the maximum geodesic of a network, showing how far individuals can be separated while still remaining connected within the network. The networks look quite different in this aspect. Individuals in Network 4 are only 2.1 steps removed from each other. Network 2 shows the furthest distance between individuals, with the furthest being 48 steps removed from each other.

Betweenness centrality quantifies how often a node lies on the shortest path between other nodes. It counts how many indirect connections are routed through them. It is a way of showing how central an individual is in the network, how many connections pass through them. In Network 3, more individuals play an intermediary role compared to the other networks. In Networks 1, 2 and 4, more than 75% of the individuals have a betweenness of 0, meaning they do not serve as bridges between other individuals.

The clustering coefficient reflects the cliques that are also visible in the visualizations of the networks for Definition 1, 2 and 4. In Network 1 and 2, the clustering coefficient already reaches a value of 1 at the 1st quantile mark and Network 4 reaches 1 before the median. Network 3 shows low clustering, with a value of 0.2 at the 3rd quantile mark.

Lastly, we looked at transitivity. Transitivity is a measure of the probability that two nodes connected to the same node are also connected to each other. The results show a low transitivity for Network 3 and higher transitivity values in the other three networks.

Drivers behind social proximity

The summary of the Monte Carlo maximum likelihood results of the ERGM for Network 3 can be seen in Table 6. Convergence was achieved when running the model and the chains showed adequate mixing. The only non-significant variable was the nodefactor for sex and was therefore excluded from the model. The effect size for the edges covariate tells us that the baseline tendency for edge formation in the network is extremely low — about 1 in 1,230 dyads are expected to form a tie by default.

According to the output, large drivers behind connections are shared sex and residence in the same quartier. In a quantitative sense, the odds of a connection occurring between same-sex individuals are approximately 15 times higher than the odds of a connection between opposite-sex individuals, controlling for other variables. Individuals living in district G appear to have more contacts than individuals living in other quartiers, they have about 3 times the odds of forming a connections compared to individuals that live in the other districts.

Table 5: Descriptive network metrics for the whole-village networks.

	Network 1	Network 2	Network 3	Network 4
Number of components	137	437	711	957
Nodes in largest component	1479	860	1830	1592
Unconnected nodes	7	283	703	956
Mean degrees	20.3	9.3	6.6	356.8
1st quantile degrees	12	6	0	0
Median degrees	16	10	6	170
3rd quantile degrees	26	14	10	688
Mean geodesic	11	16.7	6	2.1
1st quantile geodesic	8	9	5	2
Median geodesic	11	15	6	2
3rd quantile geodesic	14	23	7	2
Diameter	28	48	14	5
Mean betweenness	4327	2481.6	3301.8	542.3
1st quantile betweenness	0	0	0	0
Median betweenness	0	0	1040.4	0
3rd quantile betweenness	0	0	4472.1	0
Mean clustering coefficient	0.9	0.9	0.2	0.9
1st quantile clustering coefficient	1	1	0	0.8
Median betweenness	1	1	0.1	1
3rd quantile betweenness	1	1	0.2	1
Transitivity	0.9	0.8	0.1	0.8

Table 6: Summary statistics for the ERGM of Network 3

	Estimate	Std. Error	Pr($>\|z\|$)
edges	-7.1200292	0.1029417	<1e-04
nodecov.age	0.0175608	0.0006307	< 1e-04
absdiff.age	-0.1167529	0.0023403	< 1e-04
nodematch.sex	2.7055806	0.0628212	< 1e-04
nodecov.alter_connection	-0.0461907	0.0205691	0.0247
nodefactor.survey_group.child	-1.0405665	0.0705308	<1e-04
nodefactor.survey_group.teen	0.4458862	0.0256299	<1e-04
nodefactor.quartier.districtB	0.0833227	0.0414895	0.0446
nodefactor.quartier.districtC	0.0372337	0.0279914	0.1835
nodefactor.quartier.districtD	0.0404698	0.0303719	0.1827
nodefactor.quartier.districtE	0.6891886	0.0384915	< 1e-04
nodefactor.quartier.districtF	0.0197066	0.0286749	0.4919
nodefactor.quartier.districtG	1.0855464	0.0754795	< 1e-04
nodematch.quartier	0.7786907	0.0440697	<1e-04
nodecov.house_size	0.0022502	0.0002662	< 1e-04
absdiff.house_size	-0.0049313	0.0004935	< 1e-04
edgecov.individual_distance	-0.0061977	0.0002164	< 1e-04
gwdeg.fixed.0.8	-0.9062357	0.0716867	< 1e-04

3.3 Social Versus Spatial Proximity

Figures 4 and 5 show plots on the proportion of overlap between contacts according to the chosen spatial proximity perimeters and the social proximity definitions.

Figure 4 shows the number of contacts on the x-axis and the proportion of these contacts living within the specified range on the y-axis. The lined pattern observed in subfigures (a) to (f) is a result of mathematical properties. For example, if an individual has two contacts the proportion of their contacts that live within a specific perimeter can only be 0, 0.5 or 1. For Network 1, most contacts live within 25 m, which aligns with the definition *contacts within the home area.* There is not much visual difference between the plots for 25 m and 100 m, meaning that the contacts that do not live within 25 m, live further away than 100 m. The plots of Network 2 do not show a large difference between the 25 m and 100 m perimeters either. The plots also show a large number of individuals that have most of their connections living close by. However, in contrast to Network 1, a significant portion of individuals has little to no connections in close proximity. Network 3 shows a larger difference between the 25 m and 100 m ranges. There is a smaller proportion of contacts within 25 m compared to the previous two networks, but a higher number of contacts fall within the 25–100 m range. Network 4 shows a different pattern from the other three, due to its much higher number of edges. The vertical lines represent the large cliques that were already visible in the network plots. It is expected that this network would show a large difference between the 25 m and 100 m ranges. When more than 100 individuals attend the same village mosque, it is highly unlikely they all live within 25 m of each other.

Figure 5 shows histograms of the proportion of contacts within the specified ranges for different age groups. For Definition 1, there are only minor differences. For Definition 2, the number of individuals with all their contacts within the specified range goes down with age. For Definitions 3 and 4, there are clear differences between age groups as well as between the two specified ranges.

3.4 Associations Between Close Social Contact and Leprosy Occurrence

Leprosy metrics

Out of the 2548 surveyed individuals, 168 individuals were diagnosed with leprosy. Figure 6 presents the age distribution of the population. Subfigure (a) shows all the individuals in the population, while subfigure (b) only shows diagnosed individuals. The most noticeable difference occurs among individuals under the age of 20. The proportion of younger individuals is higher in the entire population compared to those diagnosed with leprosy. It is important to note that both plots of the age distribution represents the age of the individuals in 2023. The date of diagnosis was not available, so the age at diagnosis could differ from the age at the time of the survey by up to 21 years.

Fitted models

The results of the logistic regression model can be seen in Table 7. The following variables showed no significance and, as a result, were dropped during backward selection: house size, sex, quartier, number of contacts in Networks 1, 3 and 4.

The age group variable was significant in both models, with the highest effect being in individuals aged 15-35 (compared to reference group aged 0-7). The number of contacts according to Definition 2 *blood*

(a) Network 1 - 25m

(b) Network 1 - 100m

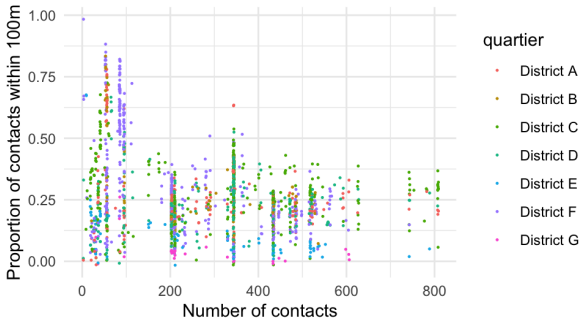
(c) Network 2 - 25m

(d) Network 2 - 100m

(e) Network 3 - 25m

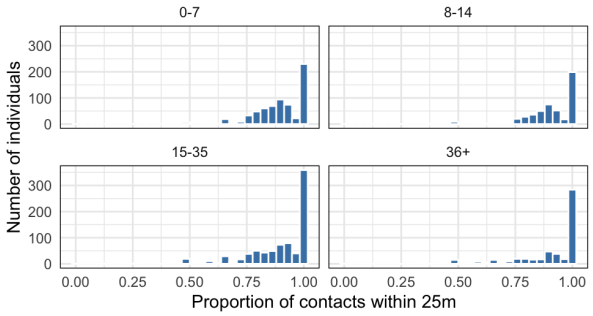
(f) Network 3 - 100m

(g) Network 4 - 25m

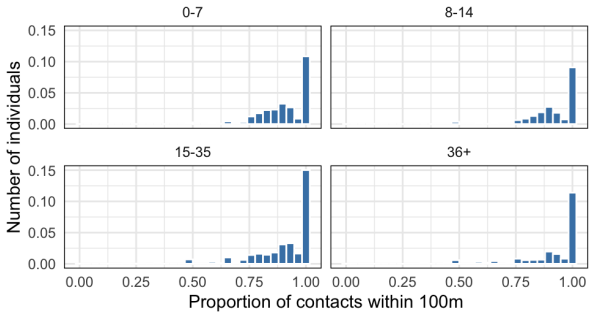


(h) Network 4 - 100m

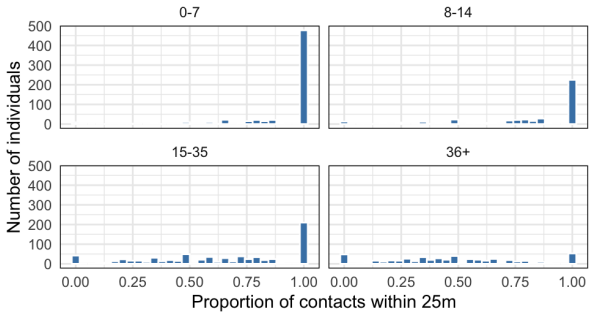
Figure 4: Visualisations of the proportion of contacts within a spatial range of 25m or 100m for the four whole-village networks, contrasted with the quartier individuals reside in. A small amount of jitter was added to the points in the plots to give a better view of the number of points.



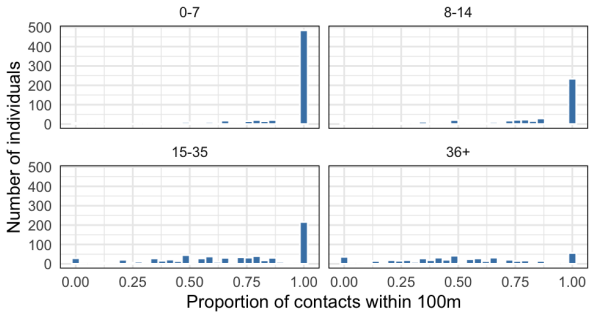
(a) Network 1 - 25m



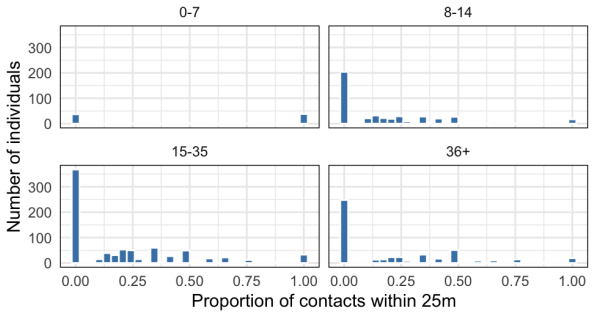
(b) Network 1 - 100m



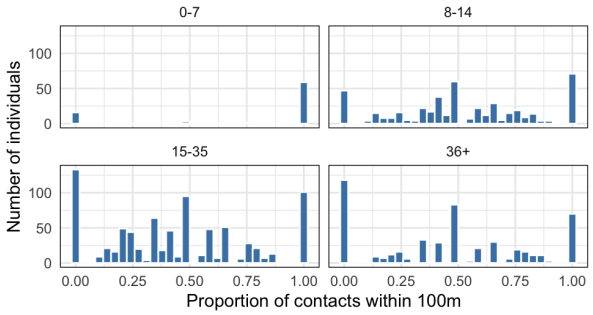
(c) Network 2 - 25m



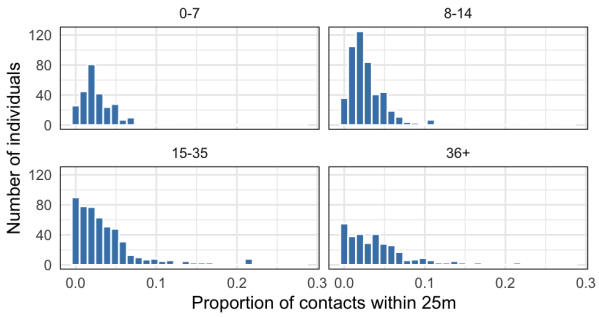
(d) Network 2 - 100m



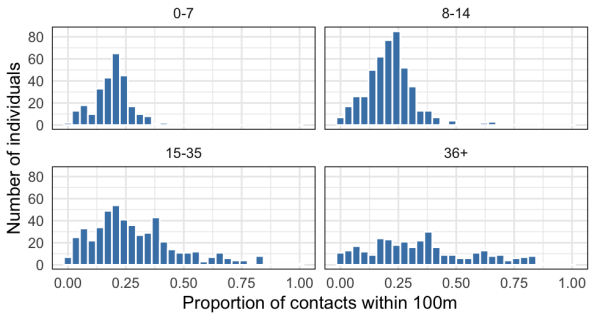
(e) Network 3 - 25m



(f) Network 3 - 100m

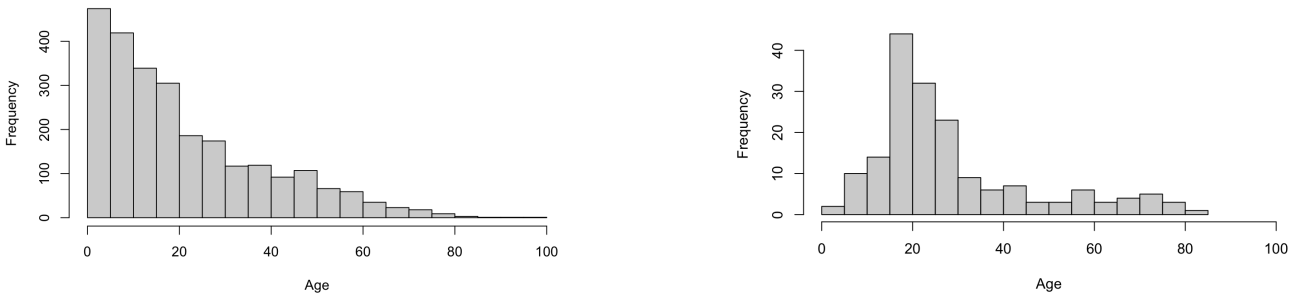


(g) Network 4 - 25m



(h) Network 4 - 100m

Figure 5: Visualisations of the proportion of contacts within a spatial range of 25m or 100m for the whole-village different networks, displayed per age group.



(a) Surveyed individuals

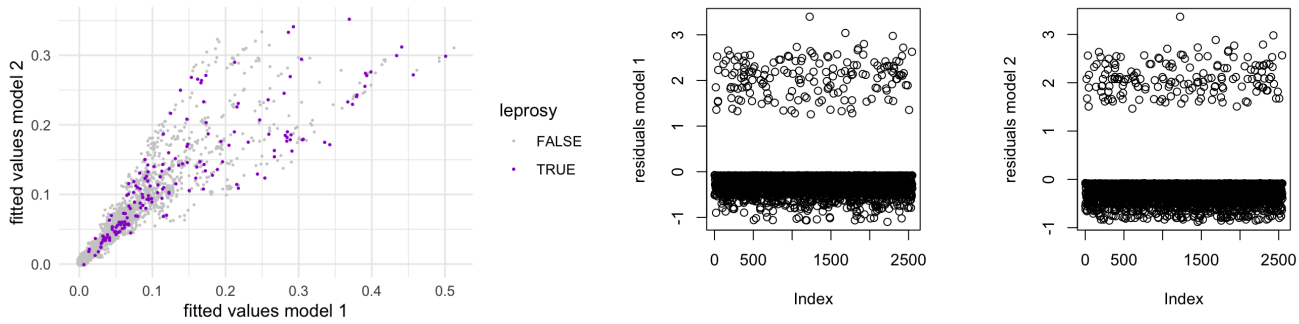
(b) Surveyed individuals diagnosed with leprosy

Figure 6: The age distributions of individuals in the population at the time of the survey.

ties & partners was also significant in both models. In Model 1, there was a significant effect of having multiple diagnosed Definition 1 contacts, whereas having only one did not show a significant effect. In Model 2, Definition 1 had a significant effect. All forms of diagnosed contacts under Definitions 2 and 3 showed significant effects.

Diagnostics

A complete separation between the residuals of diagnosed and undiagnosed individuals can be seen in subfigures (b) and (c) of Figure 7. This is likely due to the low incidence of leprosy in the population and is a limitation of using a basic logistic regression model. Subfigure (a) shows the fitted values for both models. Diagnosed individuals are represented by purple dots, undiagnosed individuals are represented by smaller, light gray dots. Diagnosed and undiagnosed individuals cannot be separated, but diagnosed individuals seem to have higher fitted values overall.



(a) Fitted values

(b) Residuals Model 1

(c) Residuals Model 2

Figure 7: Basic diagnostics plots for logistic regression models.

Table 7: Summary measures for the logistic regression models.

	Model 1: Bins			Model 2: TRUE/FALSE		
	Estimate	Std. Error	Pr(> |z|)	Estimate	Std. Error	Pr(> |z|)
(Intercept)	-4.8199	0.5521	< 2e-16	-5.1898	0.5434	< 2e-16
age_group8-14	1.6175	0.5661	0.00427	1.7076	0.5625	0.002401
age_group15-35	2.8632	0.5225	4.26e-08	2.8858	0.5218	3.18e-08
age_group36+	2.4620	0.5404	5.22e-06	2.4645	0.5399	5.00e-06
n1_leprosy_contact_Single	0.1395	0.2266	0.53817			
n1_leprosy_contact_Several	0.6092	0.2332	0.00901			
n1_leprosy_contact_True				0.3830	0.1924	0.046516
n2_leprosy_contact_Single	0.6514	0.2428	0.00729			
n2_leprosy_contact_Several	1.4443	0.2919	7.49e-07			
n2_leprosy_contact_True				0.9433	0.2150	1.15e-05
n3_leprosy_contact_Single	0.5176	0.1872	0.00568			
n3_leprosy_contact_Several	0.5170	0.2751	0.06015			
n3_leprosy_contact_True				0.5673	0.1715	0.000942
n4_leprosy_contact_Under10	-0.3138	0.2116	0.13809			
n4_leprosy_contact_10plus	-0.2785	0.2135	0.19202			
n4_leprosy_contact_True				0.1373	0.1839	0.455349
log_n2_nr_contacts	-0.3176	0.1199	0.00807	-0.2929	0.1194	0.014137
AIC	1077.2			1087.4		

Exploring next steps

An additional plot and table were generated to provide further insights. Table 8 presents the number of individuals tested and diagnosed under each definition, considering both spatial and social proximity. Column 3 shows the proportion of diagnosed individuals among those tested, while Column 4 reports the proportion for all diagnosed individuals that would have been identified using each definition. This allows us to discuss strategies for detecting additional cases. Network 1-3 each yielded a higher proportion of diagnosed individuals among those tested. In contrast, the spatial proximity definitions tested more individuals and managed to identify all diagnosed individuals. When combining network 1, 2 and 3, the proportion of diagnosed individuals among those tested is equal to that of the 25 m spatial perimeter.

Table 8: The proportions of diagnosed among tested individuals and proportions of identified diagnosed individuals

	# Tested	# +Diagnosis	Proportion D+/T	Proportion identified
25 m radius	2112	168	0.08	1.00
100 m radius	2532	168	0.07	1.00
network 1	1108	101	0.09	0.60
network 2	795	77	0.10	0.46
network 3	726	83	0.11	0.49
network 4	1577	114	0.07	0.68
network 1-3	1786	137	0.08	0.82

Figure 8 represents a first step into considering how individuals' network characteristics could relate to their leprosy status. The figure shows boxplots of betweenness centrality for diagnosed and non-diagnosed individuals. The low betweenness values for network 1, 2 and 4 are reflected by the figure. In network 3, there seems to be a difference between diagnosed and non-diagnosed individuals in both the median and 1st quantile values.

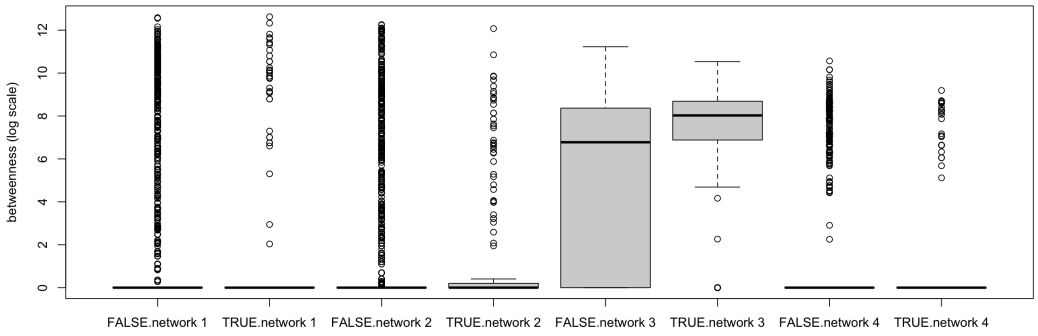


Figure 8: Boxplots to compare betweenness scores of diagnosed (TRUE) and non-diagnosed (FALSE) individuals in the four whole-village networks.

4 Discussion

The first stage of this study focused on developing clear definitions for close social contact. The initial approach was to begin with the most intensive form of contact and then expand the definition step by step, ensuring that each new definition included connections from the previous step. Previous research has shown that a leprosy diagnosis is associated with more intensive social contacts (e.g. [5]) and it is therefore interesting to determine the threshold of whether a contact is intensive enough or not. However, an issue in the construction of constructing such definitions is in deciding which contact is more intensive than the others. For instance, is someone that you mention as sometimes eating in the same room as you a more intensive contact as your sister that lives on the next street, looking at it from a leprosy perspective? It would require more information to be able to actually rank them. In the absence of these rankings, the most practical approach was to define different types of contact. These included contacts within the home space, blood ties & partners, friendships and meeting places are all categories that make sense and that are easy to imagine. Each of these categories presented both advantages and limitations in the context of this study.

Household contacts are known to have a higher incidence of leprosy [23]. Therefore, our initial definition was based around contacts within the home area. The difficulty with this definition is the presence of compounds in the study village. A compound refers to a cluster of households that have some form of shared space. Therefore, we consider them as an extended household, even though they are not necessarily as close as a household. This distinction could explain the difference in results between having one diagnosed contact (non-significant effect) or two or more (significant effect), since not all members of a compound might be equivalently close to each other.

Definition 2 (blood ties & partners) captures long-term social bonds. This is relevant for a disease such as leprosy, which is thought to spread through repeated contact over time. There is also a genetic component to an individual's susceptibility to leprosy and to the clinical form it takes (e.g., paucibacillary versus multibacillary) [4]. This genetic component makes it important to take blood ties into account. While many of the connections in Definition 1 are between blood relatives, they also include non-kin. Therefore, we made Definition 2 to isolate blood-ties, including those outside of the household. The biggest issue with this definition is that it contains no information about contact between individuals. It does not matter what your genetic susceptibility is to a specific disease if you never come in contact with it. A second issue is that (as previously mentioned) many individuals that are connected through blood ties also do live in the same household. The results for Definition 2 could therefore be confounded by contacts within the home space. This is exacerbated by the method that was used to construct this dataset. The survey did not directly ask about blood ties, instead, a question on individuals' children was used to reconstruct as many blood-ties as possible. As a result, missing data for this definition is unlikely to be missing at random, missingness is probably higher for blood ties outside of the household. This is reflected by Figures 4 and 5, which show that contacts, according to Definition 2, mostly live within 25 m of each other, especially in the younger age groups. The results might therefore not completely represent the true effect of blood ties & partners, but overestimate it.

Intensive social contacts outside of the home area are also associated with a leprosy diagnosis [5]. These are represented by Definition 3 (links through activities) and Definition 4 (links through public places). The survey questions used to construct the dataset for Definition 3 specifically asked participants to name up to three close contacts that had not been mentioned yet during questions around the household, partners, eating, sleeping, etc. This restriction is reflected by the relatively low proportion of contacts living within 25 m. As shown in Figure 5, most children do not have any contacts according to Definition 3. The most reasonable explanation is that their close contacts were already mentioned in earlier sections and thus mostly exist in close spatial proximity. Previous research has shown that most diagnosed children

have a relative with a prior leprosy diagnosis [19]. The main limitation of Definition 3 is the restrictive phrasing of the question on close contacts, which both excluded contacts and limited their number. This likely reduced the network's completeness, especially when it comes to existing transitivity. Transitivity can be important in the context of leprosy, as it can cause multiple sources of exposure. Such constraints could also explain the non-significant "several diagnosed contacts" term in the logistic regression model. Definition 4 was the only contact definition that was not associated to a leprosy diagnosis according to the models. These types of contact also, subjectively, feel much less like a true close contact.

Age at diagnosis is a key indicator when it comes to the elimination of leprosy. While most diagnosed individuals are adults, detection also occurs in children. Our dataset reflects this situation. It is important to note that the recorded age in our dataset refers to the time of the survey, not the time of diagnosis. Because diagnoses in this dataset date back as far as 2002, individuals may appear older at diagnosis than they actually were. Even after accounting for this limitation, there are still more diagnoses at adult age within the study population. In general, diagnoses in young children exists, but they are more common in children older than 10, which can be explained by the long incubation time [14]. The presence of cases in children indicate relatively recent transmission of the disease and highlights the need for continued monitoring in our study population.

In the literature, male sex is often reported as a risk factor for leprosy [17], often attributed to men having a higher number of close contacts [5]. In this study, however, sex was not a significant factor in the logistic regression model. According to the ERGM for the drivers of contacts under Definition 3, there was no evidence of men reporting more close contacts than women. However, since the survey limited the number of contacts that could be named, it is not possible to confirm whether men in the study truly have an equal amount of close contacts compared to women. The known number of contacts for all four definitions separately was included as a covariate in the logistic regression models and was only significant for Definition 2. It is possible that social connectedness is not fully captured by contact counts alone. Therefore, Figure 8 was added to the additional section of the results. Nodes with high betweenness centrality could be considered more influential within the network, and it is reasonable to question whether this has an effect on their own health status or of those around them. Figure 8 suggests a potential difference in betweenness between diagnosed and non-diagnosed individuals under Definition 3 (links through activities). However, this pattern could be driven by age instead of disease status. Further investigating into the relationship between network characteristics (such as betweenness) and leprosy diagnoses could be an interesting next step and provide additional insights. However, incorporation network characteristics into prevention strategies might be challenging, since they depend on the entire network rather than an individual in isolation.

Table 8 compares the proportion of diagnosed contacts identified through contact definition networks and spatial proximity perimeters. For networks 1-3 a higher proportion of individuals were diagnosed with leprosy. Spatial perimeters a lower proportion, but were able to identify all diagnosed individuals in the study population. This reflects a trade-off between efficiency and completeness. Close contact definitions result in more detections per person tested, but do not capture all cases in the population. It is important to note that disease information was unavailable for non-surveyed individuals, meaning that diagnosed individuals live within a specific perimeter by definition. As shown in Figures 4 and 5, a considerable number of close contacts do not live within the village area. Incorporating contact definitions as an addition to small spatial perimeters help identify new cases living farther away, and help discover new clusters of infected individuals.

An important consideration is the nature of dataset used in this study. Even though some contacts from earlier years were reported in the survey (mostly ex-partners and caretakers), it is in essence a cross-sectional dataset. Since leprosy is a disease with a long incubation time, it is possible that individuals who were close contacts and became infected in the past, were no longer considered a close contact at the

time of the survey.

4.1 Conclusion

The results of this study indicate that specific types of social contact are associated with leprosy occurrence. Associations were found for contacts within the home area, blood ties and partners, and links through activities, whereas links through public spaces showed no such association. These findings suggest that relying solely on spatial screening risks missing individuals who are socially close but geographically distant, and who may therefore still be at elevated risk.

For future interventions, this highlights the potential benefit of integrating social network information with spatial proximity data. Combining these approaches could improve the efficiency of post-exposure prophylaxis and active case-finding strategies by enabling the identification of both household members and socially connected individuals at higher risk. Such targeted strategies may be particularly valuable in endemic settings with limited resources, where balancing completeness and efficiency is critical.

Overall, the results support the incorporation of social network analysis into leprosy control frameworks as a routine component of intervention planning. This approach could help reveal otherwise hidden chains of transmission, support more precise targeting of preventive measures, and contribute to progress towards the global goal of leprosy elimination.

Ethics, Society and Stakeholders

The protocol of the original study that generated the data used in this project was reviewed and approved by ethics committees in both Belgium and in the Comoros. This dual approval process underscores a commitment to conducting research in a manner that is ethically sound and respectful of local regulations, cultural norms, and participant rights.

For this project, there was direct collaboration with The Institute of Tropical Medicine in Antwerp, who stated as their mission to deliver science for worldwide health. Studying effects of socials networks can help support targeted interventions. Better targeted interventions are useful from a global health perspective for diseased individuals to be discovered as early as possible, as well as from an economic perspective, making interventions more cost effective.

The datasets that were made available for this project were the result of great efforts of a large group of people. Key contributors include the study participants, who devoted their time and shared their personal information; The local study team and fieldworkers who carried out the research in situ; The National Tuberculosis and Leprosy Control Program and The Damian foundation who have both made significant progress in reducing the prevalence of leprosy.

Beyond the local context, a broader network of stakeholders should also be mentioned. These include global actors such as the World Health Organization and international partners engaged in leprosy control efforts, as well as people affected by leprosy around the world. The insights gained from leprosy research can inform global strategies, support more effective interventions, and ultimately contribute to the worldwide goal of eliminating leprosy as a public health problem.

Acknowledgments

I would like to thank my internal supervisor, Andrea, for his guidance throughout this thesis. His feedback and advice during our regular meetings were very valuable, particularly regarding the statistical aspects of my work. I also thank Tine and Claudia for making the topic available and for providing background knowledge on leprosy. Their feedback and input were an important contribution to this project.

Software Code

R code can be found in the following repository: https://github.com/uh-2025/thesis_code

References

[1] Sofie Braet. *Using novel molecular approaches to understand transmission of Mycobacterium leprae in the Comoros.* Phd thesis, Universiteit Antwerpen, Antwerp, 2023. Available at `https://repository.uantwerpen.be/docman/irua/02e798motoM7e\#page=135`.

[2] Complex Data Collective. Network canvas, 2023.

[3] Ana Cláudia Mendes do Nascimento, Diogo Fernandes dos Santos, Douglas Eulálio Antunes, Maria Aparecida Gonçalves, Marcela Araujo de Oliveira Santana, Bruno de Carvalho Dornelas, Luiz Ricardo Goulart, and Isabela Maria Bernardes Goulart. Leprosy relapse: A retrospective study on epidemiologic, clinical, and therapeutic aspects at a brazilian referral center. *International Journal of Infectious Diseases*, 118:44–51, 2022.

[4] Vinicius M. Fava, Monica Dallmann-Sauer, and Erwin Schurr. Genetics of leprosy: today and beyond. *Human Genetics*, 139(6):835–846, Jun 2020.

[5] S. G. Feenstra, Q. Nahar, D. Pahan, L. Oskam, and J. H. Richardus. Social contact patterns and leprosy disease: a case-control study in bangladesh. *Epidemiology and Infection*, 141(3):573–581, 2013.

[6] Geodatos. Población de comoras, 2025. `https://www.geodatos.net/poblacion/comoras` [Accessed: 2025-07-31].

[7] Google. Satellite view of anjouan, 2025. `https://maps.google.com` [Accessed: 2025-07-31].

[8] G. H. A. Hansen. (supplement). *Norsk Mag. Laegervidenskaben*, 1, 1874.

[9] Epco Hasker, Younoussa Assoumani, Andriamira Randrianantoandro, and Bouke Catharina de Jong. Post-exposure prophylaxis in leprosy (people): a cluster randomised trial. *The Lancet Global Health*, 12:2214–109X, 2024.

[10] Epco Hasker, Abdallah Baco, Assoumani Younoussa, Aboubacar Mzembaba, Saverio Grillone, Tine Demeulenaere, Guido Groenen, Philip Suffys, and Bouke C DE Jong. Leprosy on anjouan (comoros): persistent hyper-endemicity despite decades of solid control efforts. *Leprosy review*, 88(3):334–342, 2017.

[11] Marc Monot, Nadine Honoré, Thierry Garnier, Romulo Araoz, Jean-Yves Coppée, Céline Lacroix, Samba Sow, John S Spencer, Richard W Truman, Diana L Williams, et al. On the origin of leprosy. *Science*, 2005.

[12] Martina Morris, Mark S. Handcock, and David R. Hunter. Specification of exponential-family random graph models: terms and computational aspects. *Journal of statistical software*, 24:1–24, 2008.

[13] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, Sep 2003.

[14] Marcela Bahia Barretto de Oliveira and Lucia Martins Diniz. Leprosy among children under 15 years of age: literature review. *Anais Brasileiros de Dermatologia*, 91(2):196–203, Mar 2016.

[15] Nimer Ortuno-Gutierrez, Abdallah Baco, Sofie Braet, Assoumani Younoussa, Aboubacar Mzembaba, Zahara Salim, Mohamed Amidy, Saverio Grillone, Bouke C. de Jong, Jan Hendrik Richardus, and Epco Hasker. Clustering of leprosy beyond the household level in a highly endemic setting on the comoros, an observational study. *BMC Infectious Diseases*, 19(1):501, Jun 2019.

[16] Saskia Pfrengle, Judith Neukamm, Meriam Guellil, Marcel Keller, Martyna Molak, Charlotte Avanzi, Alena Kushniarevich, Núria Montes, Gunnar U. Neumann, Ella Reiter, Rezeda I. Tukhbatova, Nataliya Y. Berezina, Alexandra P. Buzhilova, Dmitry S. Korobov, Stian Suppersberger Hamre, Vitor M. J. Matos, Maria T. Ferreira, Laura González-Garrido, Sofia N. Wasterlain, Célia Lopes, Ana Luisa Santos, Nathalie Antunes-Ferreira, Vitória Duarte, Ana Maria Silva, Linda Melo, Natasa Sarkic, Lehti Saag, Kristiina Tambets, Philippe Busso, Stewart T. Cole, Alexei Avlasovich, Charlotte A. Roberts, Alison Sheridan, Craig Cessford, John Robb, Johannes Krause, Christiana L. Scheib, Sarah A. Inskip, and Verena J. Schuenemann. Mycobacterium leprae diversity and population dynamics in medieval europe from novel ancient genomes. *BMC Biology*, 19(1):220, Oct 2021.

[17] Victoria Grace Price. Factors preventing early case detection for women affected by leprosy: a review of the literature. *Global Health Action*, 10(sup2):1360550, 2017.

[18] Laura C Rodrigues and Diana NJ Lockwood. Leprosy now: epidemiology, progress, challenges, and research gap. *The Lancet Infectious Diseases*, 11(6):464–470, 2011.

[19] Raisa Rumbaut Castillo, Laura C. Hurtado Gascón, Jenny Laura Ruiz-Fuentes, Fernanda M. Pastrana Fundora, César R. Ramírez Albajés, Andres F. Henao-Martínez, Carlos Franco-Paredes, and Ángel Arturo Escobedo. Leprosy in children in cuba: Epidemiological and clinical description of 50 cases from 2012–2019. *PLOS Neglected Tropical Diseases*, 15(10):1–13, 10 2021.

[20] Anne Schoenmakers, Liesbeth Mieras, Teky Budiawan, and Wim H van Brakel and. The state of affairs in post-exposure leprosy prevention: A descriptive meta-analysis on immuno- and chemoprophylaxis. *Research and Reports in Tropical Medicine*, 11:97–117, 2020.

[21] Statnet Development Team. Exponential random graph models (ergms) using statnet, 2024. `https://statnet.org/workshop-ergm/ergm_tutorial.html` [Accessed: 2025-05-29].

[22] Mariane M. A. Stefani, Charlotte Avanzi, Samira Bührer-Sékula, Andrej Benjak, Chloé Loiseau, Pushpendra Singh, Maria A. A. Pontes, Heitor S. Gonçalves, Emerith M. Hungria, Philippe Busso, Jérémie Piton, Maria I. S. Silveira, Rossilene Cruz, Antônio Schetinni, Maurício B. Costa, Marcos C. L. Virmond, Suzana M. Diorio, Ida M. F. Dias-Baptista, Patricia S. Rosa, Masanori Matsuoka, Maria L. F. Penna, Stewart T. Cole, and Gerson O. Penna. Whole genome sequencing distinguishes between relapse and reinfection in recurrent leprosy cases. *PLOS Neglected Tropical Diseases*, 11(6):1–13, 06 2017.

[23] Camila Silveira Silva Teixeira, Júlia Moreira Pescarini, Flávia Jôse Oliveira Alves, Joilda Silva Nery, Mauro Niskier Sanchez, Carlos Teles, Maria Yury Travassos Ichihara, Anna Ramond, Liam Smeeth, Maria Lucia Fernandes Penna, Laura Cunha Rodrigues, Elizabeth B. Brickley, Gerson Oliveira Penna, Maurício Lima Barreto, and Rita de Cássia Ribeiro Silva. Incidence of and factors associated with leprosy among household contacts of patients with leprosy in brazil. *JAMA Dermatology*, 156(6):640–648, 06 2020.

[24] WHO. Leprosy, 2025. `https://www.who.int/news-room/fact-sheets/detail/leprosy` [Accessed: 2025-06-08].

[25] World Health Organization. Weekly epidemiological record. *Weekly Epidemiological Record*, 99(37), September 2024. Accessed 2025-06-08.

[26] Worldometers.info. Population of comoros, 2025. `https://www.worldometers.info/world-population/comoros-population/` [Accessed: 2025-06-08].

Appendix

The mentioned types of social contacts are defined as follows:

- **Living in the same household:** Individuals that are considered to be members of the same household by the head of the household. They have the same household ID in the household survey. Individuals can belong to multiple households.
- **Eating together:** The head of the household was given an image of a house with several rooms. They placed individuals that eat together in the same room on the image. This also includes individuals that are not a member of the household.
- **Shared bedroom:** This definition includes two forms. (1) The head of the household was given an image of a house with several rooms. They placed individuals that share a bedroom in the same room on the image. This also includes individuals that are not a member of the household. (2) Being mentioned as someone's sleeping friend.
- **Having a child-caretaker connection:** Individuals that have been said to be a caretaker of children in the household by the head of the household, have a child-caretaker connection with those children. This is a boolean column variable in the dataset.
- **Living in the same compound:** Members of households that are connected to the same compound ID are seen as living in the same compound.
- **Having a child-parent connection:** All surveyed individuals listed their children, but they did not name their parents. This means that we only have a child-parent connection when the parent lives in the village of Vouani. Where the child lives does not have an impact.
- **Being co-parents:** Individuals were seen as co-parents if they had both listed the same child. This means that both individuals had to live in Vouani to be seen as co-parents.
- **Being siblings:** Individuals were seen as siblings if they were mentioned as a child by the same parent. This means that we only have sibling information about the people that have a parent living in Vouani.
- **Being marriage partners:** Surveyed individuals listed their marriage partners. Only marriage partners that do not live in the same household count in this definition. If only one of the couple listed the marriage, it already counted for both.
- **Being a mentioned close contact:** Individuals were asked to list a maximum of three close contacts that were not already mentioned in the previous categories.
- **Being a mentioned colleague:** Individuals were asked to list a maximum of three colleagues they felt close to.
- **Being a mentioned playmate:** Individuals were asked to list a maximum of three playmates, for children under 10 years old, the head of household was asked for this information.
- **Being a mentioned study mate:** Individuals were asked to list a maximum of three study mates, for children under 10 years old, the head of household was asked for this information.
- **Going to the same Coranic school:** Individuals indicated to which Coranic school they went. Multiple selections were permitted. For individuals under 10 years old, this question was answered by the head of the household.

- **Going to the same French school:** Individuals indicated to which French school they went. Multiple selections were permitted. For individuals under 10 years old, this question was answered by the head of the household.
- **Going to the same mosque:** Individuals indicated to which mosque they went. Multiple selections were permitted.
- **Going to the same madrass:** Individuals indicated to which madrass they went. Multiple selections were permitted.