# Master thesis title tbd

Simon Westfechtel

December 13, 2022

**Abstract**

Abstract goes here

## 1 Introduction

When news of a novel coronavirus, latter dubbed coronavirus disease 2019, or COVID-19, first emerged from the People's Republic of China in late 2019, there was a lot of uncertainty regarding its **threat level**. Especially in western countries, governments' warned about 'fake news' and downplayed the severity of the virus, e.g. discouraging the purchase of protective face masks **cite**. After the epidemic character of the disease had become apparent, the focus soon shifted to slowing or halting the spread of the virus. The efficient tracing of infected individuals' social contacts plays an integral part in preventing the spreading of a disease, and had been a well-established procedure well before the emergence of COVID-19, where for instance a verified case of tuberculosis requires a notification of the health departments **cite**.

While a known and tested method, the sheer number of cases soon threatened to overwhelm healthcare systems and health departments alike **cite**. In order to prevent this worst-case scenario, it became paramount to identify sources of infection and transmission chains. Where medical research focused primarily on virological factors (e.g. which individual features would predispose someone to infection and/or a severe course of illness, how is the virus transmitted from person to person, etc.), the field of computational social sciences offers a different approach.

Although contact tracing initiatives differ in the level of detail that is being recorded, in general they at least identify the source of infection, i.e. a symptomatic or, once they had become widely available, positively tested patient being reported to the local health department, and said patient's social contacts, i.e. the people he had come into contact with prior to being aware of the infection, and who thus could have been infected themselves, and possibly spread the virus further (i.e. the transmission chain).

With this information alone, it is possible to model transmission chains as social networks, where individuals constitute the set of vertices, and contacts between individuals are represented by edges between the respective nodes

**visualisierung**. Thus, through the usage of methods of social network analysis, it becomes possible to identify possible dynamics in the spread of the virus which might otherwise have been overlooked. Additionally, if further information like age, health status and profession were also collected during the contact tracing procedure, modern machine learning approaches can reliably identify covariates that influence the viral spread (e.g. whether younger people are more likely to get infected and/or infect others compared to other age groups).

## 2 Related work

As one would expect, there is ample scientific work on the topic of COVID-19, and the same goes for case contact networks. Typically, case reports from health authorities are translated into a social network structure, which is then analysed using established methods from computer science. These include, but are not limited to:

- Degree centrality as an indicator for transmission rate
- PageRank score to determine individuals' positions in the network
- Community detection to identify cases who are interconnected
- Triadic effects to investigate the spread pattern of the virus

**cite**

Although useful information can already be extracted from the data using aforementioned methods, they have two main shortcomings. First, the raw data is (ideally) structured in a time-stamp format, i.e. case information includes the date when they were reported to/recorded by the health authorities. Since most previous work on this topic ignores this temporal aspect and case contact data is modelled as static networks where interaction between actors takes the form of relational states as opposed to relational events (e.g. "Actor A is contact of Actor B" vs. "Actor A came into contact with Actor B on date D"), newer methods might be suited better to analyse data of this kind. Relational event models, or REMs, (**cite**) are a promising way to model relational event models. They make it possible to understand and explain interactions and discover a multitude of network effects (**examples?**).

While this is a great improvement over static models, relational event models, or RHEMs, only consider dyadic interaction, i.e. in any interaction, only two actors are present; in reality, however, interactions often involve multiple actors, which is especially true for disease transmission, e.g. a COVID-positive individual coming into contact with customers and staff while grocery shopping. Common ways to alleviate this shortcoming are treating all dyads stemming from a polyadic interaction as independent, or clustering all targets of the interaction into a single multi-actor node **cite, visualisierung**.

Of course, these methods are only approximations and may fail to accurately represent reality, as well as be unfeasible or even intractable to model. Relational

hyperevent models build on relational event models and introduce hyperedges, which may connect any number of nodes. Thus, RHEMs make it possible to incorporate higher-order dependencies present in the data which might be missed by purely dyadic network models like REMs, and therefore produce accurate estimations of network effects **cite**. Previous work by **cite**, where relational hyperevent models have been applied to email data of the Enron company, show RHEMs outperform comparable dyadic methods in terms of model fit and predictive performance. Recently, RHEMs were used to model and analyse a case contact dataset from Bucharest, Romania, showing relational hyperevent models can be utilised to effectively investigate transmission chains in the on-going pandemic **cite**.

# 3 Research question & goal

The objective of this work is to apply different models to case contact data in order to answer the following questions:

1. Which network effects and covariates can be identified using various methods?
2. How do static, relational event and relational hyperevent models compare against each other in terms of model fit and predictive performance?
3. How do case contact networks from different regions compare against each other? Are results generalisable or are the regional differences to great to justify drawing conclusions concerning the overall dynamic of the pandemic?

# 4 Data exploration

The basis of this research comprises of four case contact datasets from the People's Republic of China, three of which are smaller, regional sets from the provinces of Yunnan, Hainan and Shanxi, and one larger set from all of China. The regional datasets have already been analysed by **cite** using static networking models, while the larger dataset has, to the knowledge of the author, not been analysed in the context of social networks so far. In the following, an exploratory summarisation of the data will be presented.

**Yunnan dataset** This dataset contains COVID case information from the province of Yunnan in southern China. There are 171 entries, with dates ranging from 2020-01-17 to 2020-02-16, i.e. this dataset is from the very early stages of the pandemic. Covariates deemed to be relevant to this research are listed in table 1. 114 cases have no ties at all, 25 have one tie, eleven have two ties, six have three ties, two have four and five ties, respectively, and 13 have 12 ties. An average degree of $1.3 \pm 3.2$ and a median degree of 0 suggest the corresponding contact network is very sparse. Only three of 18 connected components comprise

| Covariate | Description |
|---|---|
| Date | Date when the case was confirmed/reported |
| Gender | Male/Female |
| Age | In years |
| Relatives | Whether the patient is a relative of previously recorded cases |
| **Strangers** | ? |
| **Arrivedate** | Arrival date in province |
| **Feverdate** | Date of onset of symptoms |

Table 1: Relevant covariates for the Yunnan and Hainan datasets

| Covariate | Description |
|---|---|
| Date | Date when the case was confirmed/reported |
| Gender | Male/Female |
| Age | In years |
| Hukou | Place of residence |
| Relatives | Whether the patient is a relative of previously recorded cases |

Table 2: Relevant covariates for the Shanxi dataset

| Covariate | Description |
|---|---|
| Date | Date when the case was confirmed/reported |
| Gender | Male/Female |
| Age | In years |
| Place of residency | |
| Virus type | Virus variant (e.g Delta) |
| Occupation | |
| Place and event | Activity where infection might have happened |
| Venue | Location where infection might have happened (e.g. School, Workplace) |
| With whom | Who might have been the source of infection |
| Symptom | The patient's symptoms |
| Symptom severity | |

Table 3: Relevant covariates for the Shanxi dataset

of more than three nodes, suggesting transmission chains were effectively broken to prevent further spread **cite**. 58 and 62 cases are female and male, respectively; there is no information on gender for 51 patients. This suggests neither sex is more susceptible to the virus. The average age is $41 \pm 18$ and the median age is 40; with the 75 percentile being 40, patients seem to be mostly younger or middle-aged. 31 out of 60 cases are family members of previously recorded cases; this information is not available for the remaining 111 data records.

**Hainan dataset** This dataset contains COVID case information from the province of Hainan, which is the southernmost province of China. There are 162 entries, with dates ranging from 2020-01-22 to 2020-02-14, i.e. this dataset is also from the early stages of the pandemic. Relevant covariates are the same as for the Yunnan data, listed in table 1. 71 cases have no ties at all, 27 have one tie, 21 have two and three ties, respectively, six have four ties, and eight have five and six ties, each. An average degree of $1.5 \pm 1.8$, a median degree of 1 and only 43% of patients without ties compared to 67% for the Yunnan dataset suggest this network is slightly more dense. Ten out of 27 connected components are comprised of more than three nodes mean that in this network, infections are happening in larger clusters and containment methods might not have been as effective. 84 and 78 cases are female and male, respectively; this supports the hypothesis that neither sex is more susceptible to an infection. An average age of $48 \pm 17$, a 25 percentile of 36, a median age of 51 and a 75 percentile of 62 mean that the patients in this dataset are from an older age group compared to the Yunnan data. 75 out of 162 patients are family members of previously recorded cases.

**Shanxi dataset** This dataset contains COVID case information from the province of Shanxi in northern China. There are 237 entries, with dates ranging from 2020-01-23 to 2020-02-16, so this dataset too is from the early stages. Relevant covariates mostly correspond to those of the previous two discussed sets, with the addition of the place of residence; they are listed in table 2. 108 cases have no tie at all, 68 have one tie, 36 have two ties, 20 have three ties, two have four ties, and three have six, nine or eleven ties, each. An average degree of $.99 \pm 1.3$, a median degree of 1 and a 45% no-tie to tie ratio make this dataset and the Yunnan network comparable in density. 12 out of 40 connected components are comprised of more than three nodes, placing this network between the previously discussed networks in terms of infection spread. With 129 versus 108, there are slightly more men in this dataset, albeit not statistically significant. The average age of patients is $46 \pm 16$, the 25 percentile is 35, the median is 45 and the 75 percentile is 59. Only 87 of the 237 patients are family members of previously recorded cases, suggesting that the majority of infections were transmitted from stranger to stranger. The three most common places of residence are *Xian*, *Ankang* and *Hanzhong*.

**China dataset** This dataset contains COVID case information from all of China. There are 26961 entries (subject to change, depending on possible further preprocessing), with dates ranging from 2020-01-01 to 2022-08-14; therefore, this set stretches over most of the pandemic (as of writing this). Relevant covariates are listed in table 3. In contrast to the former datasets, this one contains quite a few more that might yield interesting insights into infection dynamics. 2012 cases have no tie at all, 6179 have one tie, 12818 have two ties, 3194 have three ties, 1619 have four ties, 966 have five ties, 504 have six ties and 2237 cases have more than six ties. 1333 out of 3573 connected components are comprised of more than three nodes, making this network comparable to the ones discussed previously. The mean age of patients is $42 \pm 18$, the 25 percentile is 30, the median is 41 and the 75 percentile is 54, meaning this dataset contains members of all age groups. There is no information on age for 6915 cases. 9218 women versus 11414 men shows a slight bias towards men; 6329 entries contain no information on sex. The three most common occupations are *student* (970), *worker* (548) and *employee* (326). The three most common places of residence are *Xi'an in Shanxi* (1984), *Wuhan in Hubei* (1715) and *Shijiazhuang in Hebei* (888). Where information was recorded, 462 cases have the Delta and 139 cases the Omicron variant. Among the most common activities suspected to be linked to the infection are *travel to Wuhan* (635), *dinner* (359) and *residence in Wuhan* (337). Infections mostly happened during a family gathering (938), outdoors (791) and in a social setting (503). Sources of infection include confirmed cases, family members and *Wuhan personnel*.

# 5 Methodology

Preliminary methodology goes here