Generation of a credible test data set for SORMAS

HZI – 4 November 2020

Stéphane Ghozzi
Hemholtz Centre for Infection Research (HZI)
stephane.ghozzi@helmholtz-hzi.de

Approach

COVID-19 case counts from RKI for Braunschweig, Salzgitter, Wolfsburg:

- corona dashboard: per county, reporting day, onset day, age group, sex
- SurvStat: per county, reporting week, case definition

Generate individual **cases** that reproduce these counts

Add **case features**

- onset date (Gumbel distribution fitted to data)
- infection date (~ exp, scale = 3 days)
- country of residence (Germany/France, 5%)
- first and family names ("German + ~Turkish + ~Polish" / "French")
- address ~ county of reporting LHA (85%)
- hospitalization (10%)
- death (3%)
- symptoms ~ case definition + list (rough probas, incl. asymptomatic)

Create chains of **infections**

- cases of generation 0, 1, 2, 3, 4 (e.g. 0 = 20%)
- added sequentially
- p(infection) ~ degree, exp(diff infection date)
- location ~ close to index (exp), gen 0: random in address

Create **contacts**

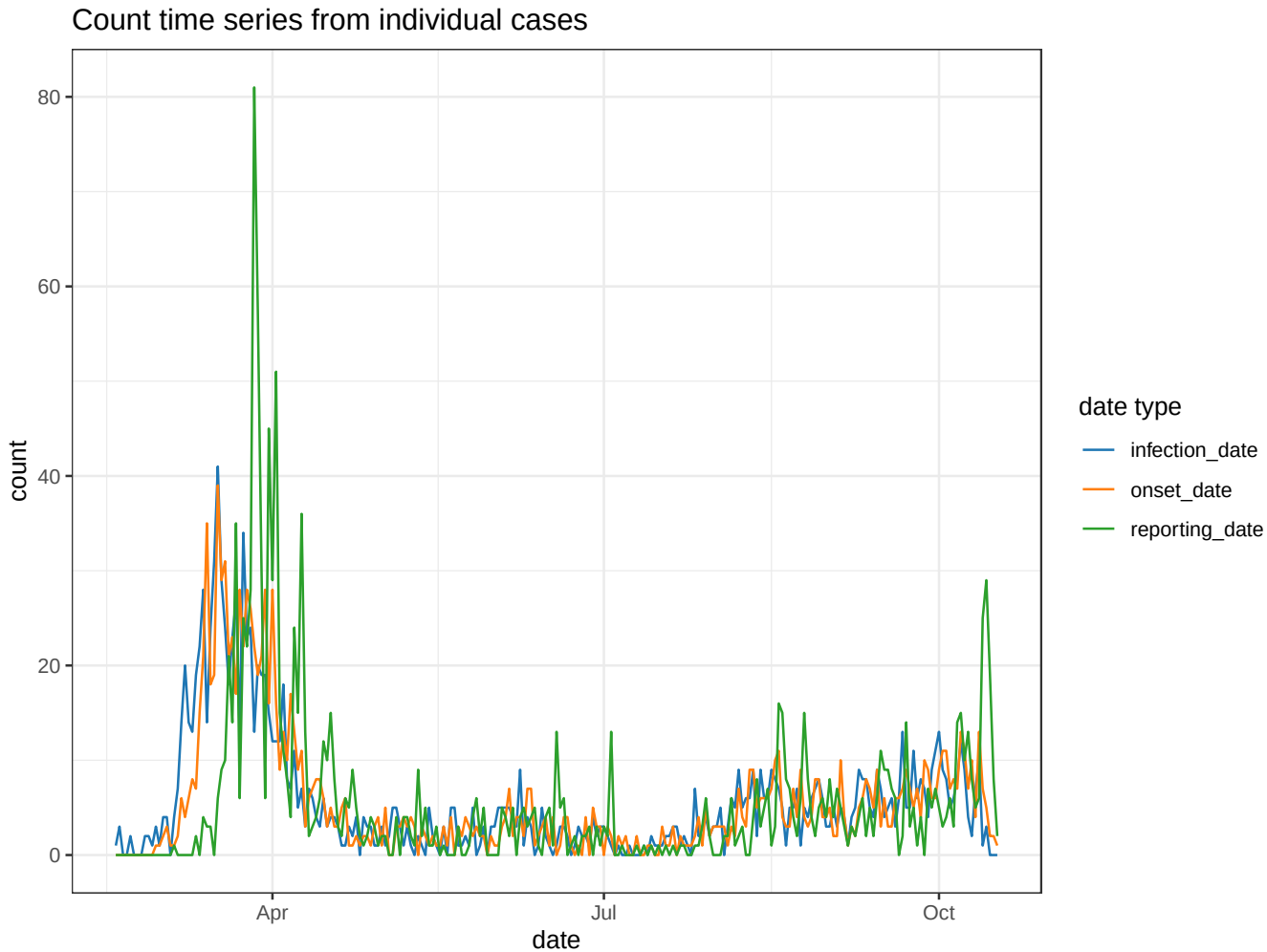
- number of contacts ~ social contact matrix / age (POLYMOD)
- location ~ close to index (exp)
- address ~ address index (67%)
- contacted (50%)
- quarantine (75%)
- tested | contacted (75%)

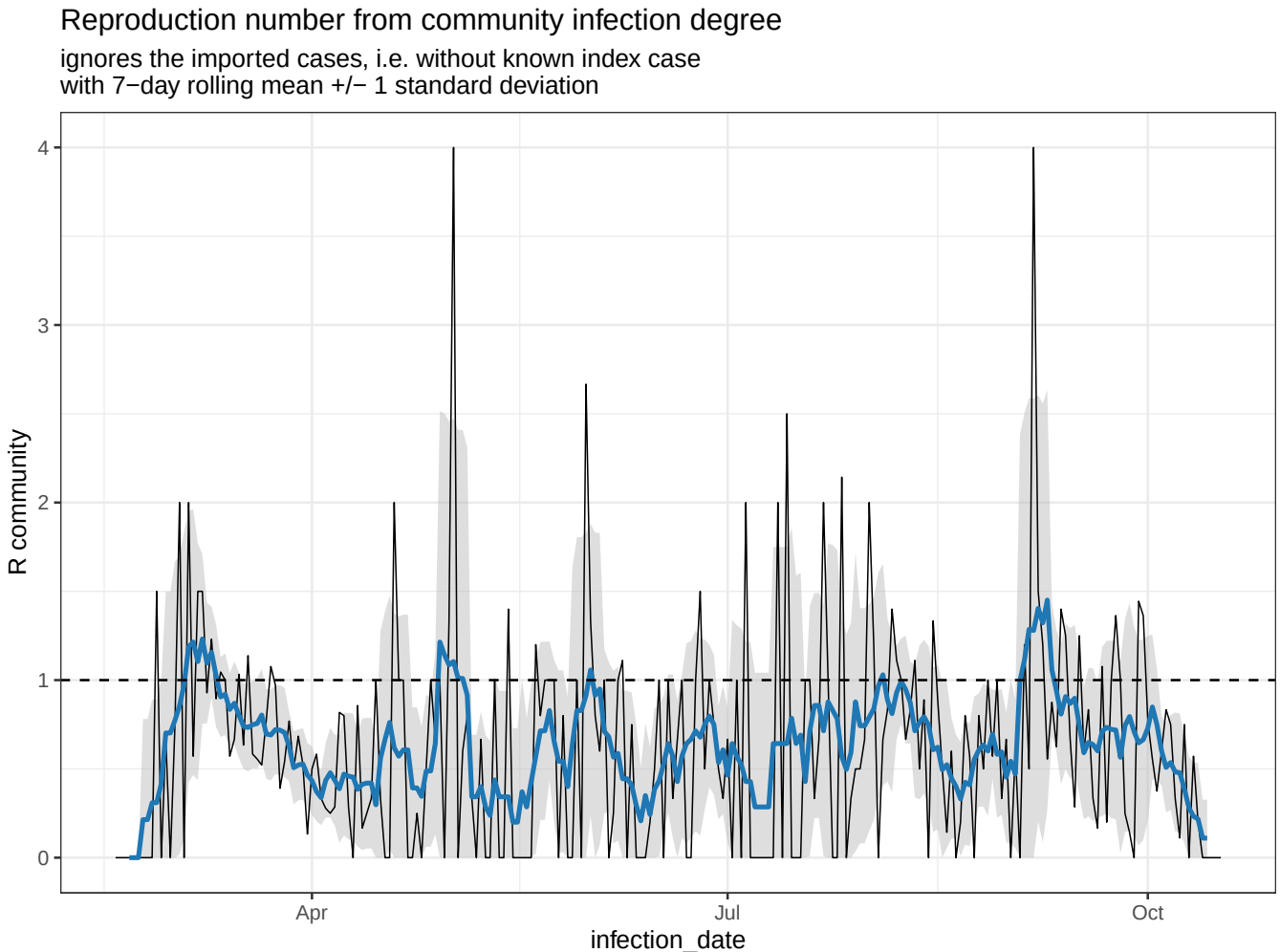
contacts are not shared between cases

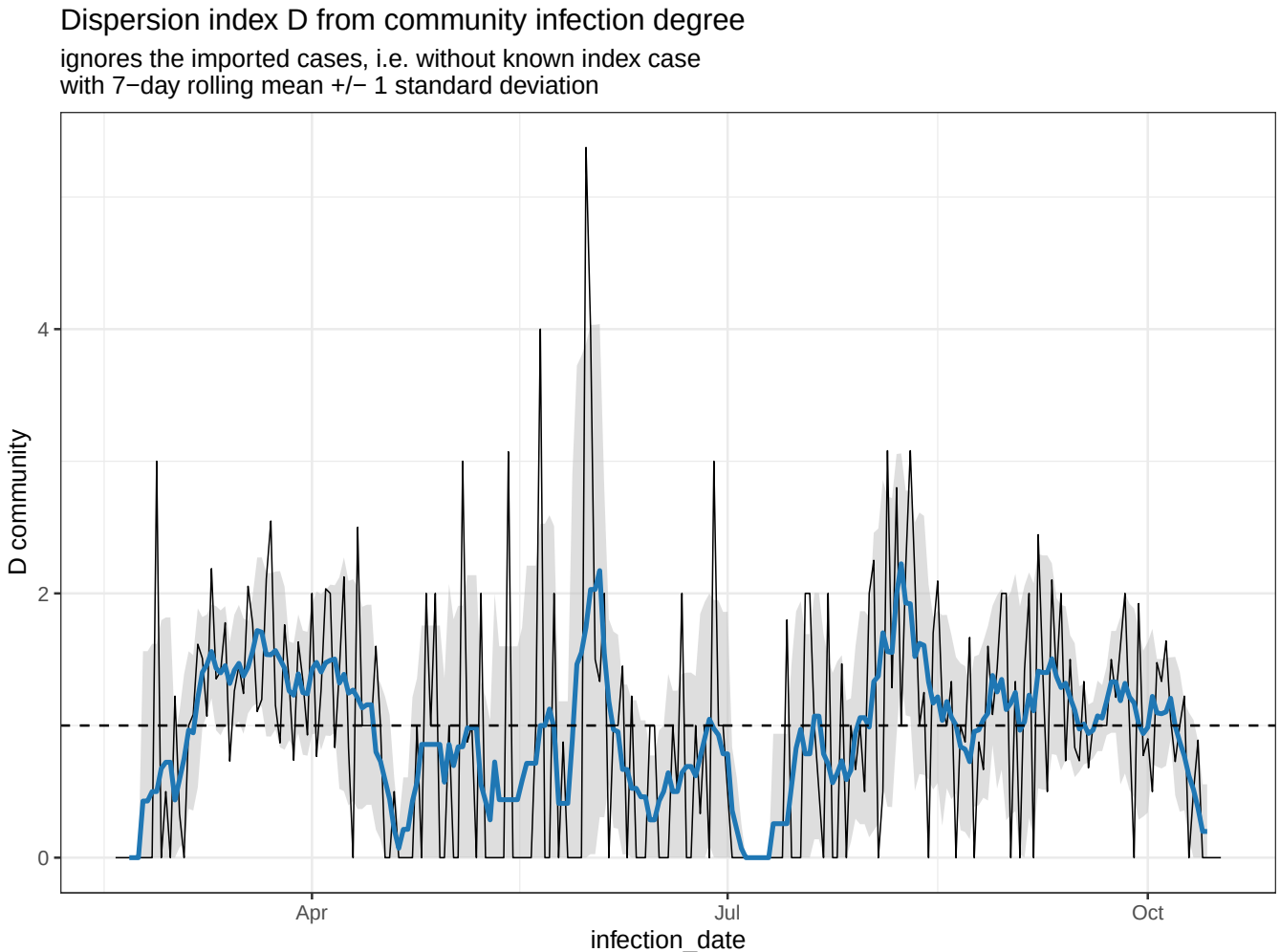
Create **events** and **participants**

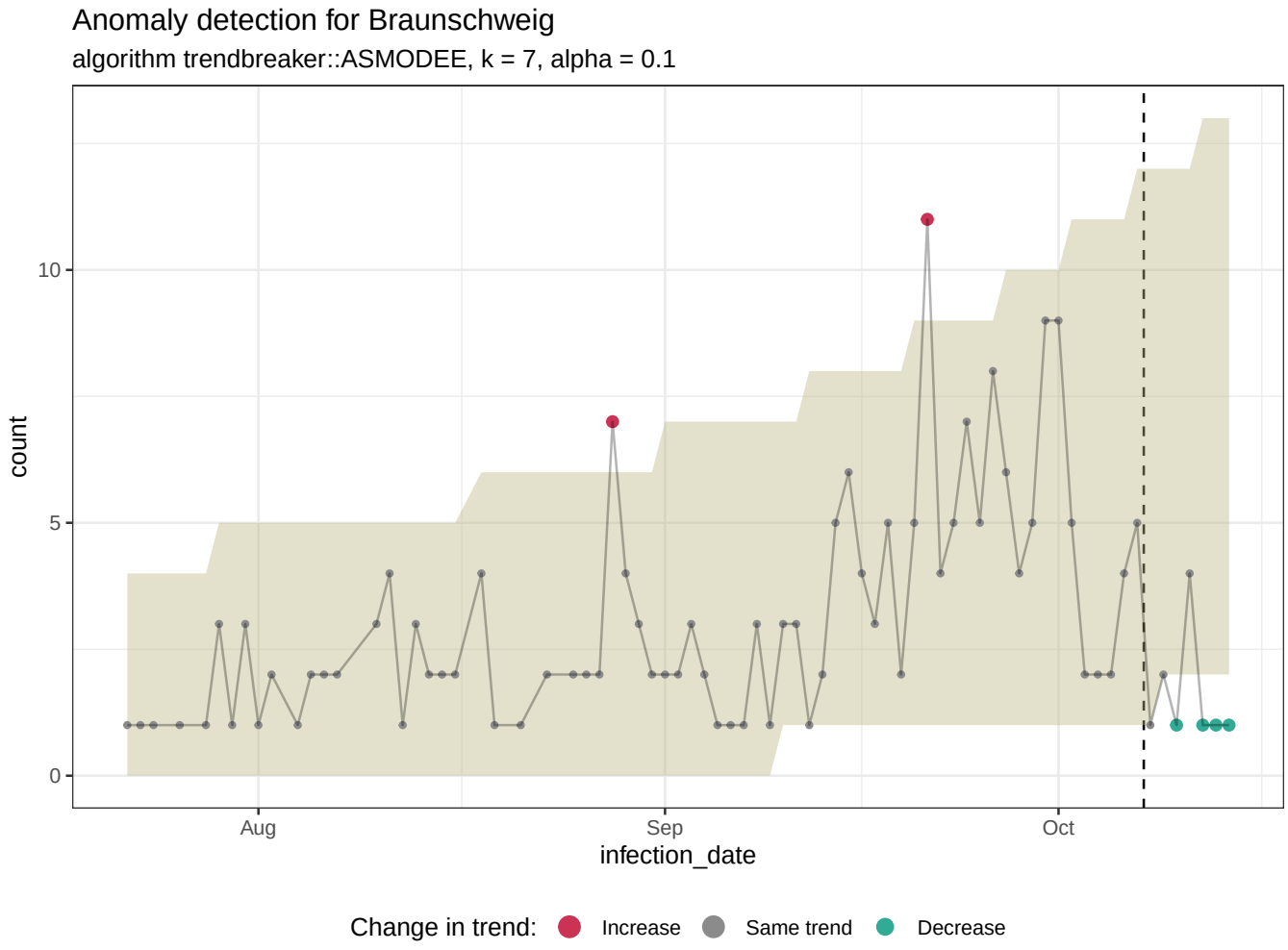
- n participants ~ N(10, 10), at least 5
- at least 1 case
- event date ~ w/in 2 weeks after infection date of case
- 1% chance that a participant is a case or a contact
- participants: random locations
- event: location = center of mass

Times series




Reproduction number from community infection degree
ignores the imported cases, i.e. without known index case
with 7−day rolling mean +/− 1 standard deviation
R community
4
3
2
1
0
Apr
Jul
Oct
infection_date


Dispersion index D from community infection degree
ignores the imported cases, i.e. without known index case
with 7−day rolling mean +/− 1 standard deviation
D community
4
2
0
Apr
Jul
Oct
infection_date


Anomaly detection for Braunschweig
algorithm trendbreaker::ASMODEE, k = 7, alpha = 0.1
count
10
5
0
Aug
Sep
Oct
infection_date
Change in trend:
Increase
Same trend
Decrease

Nowcasting and $R_e(t)$

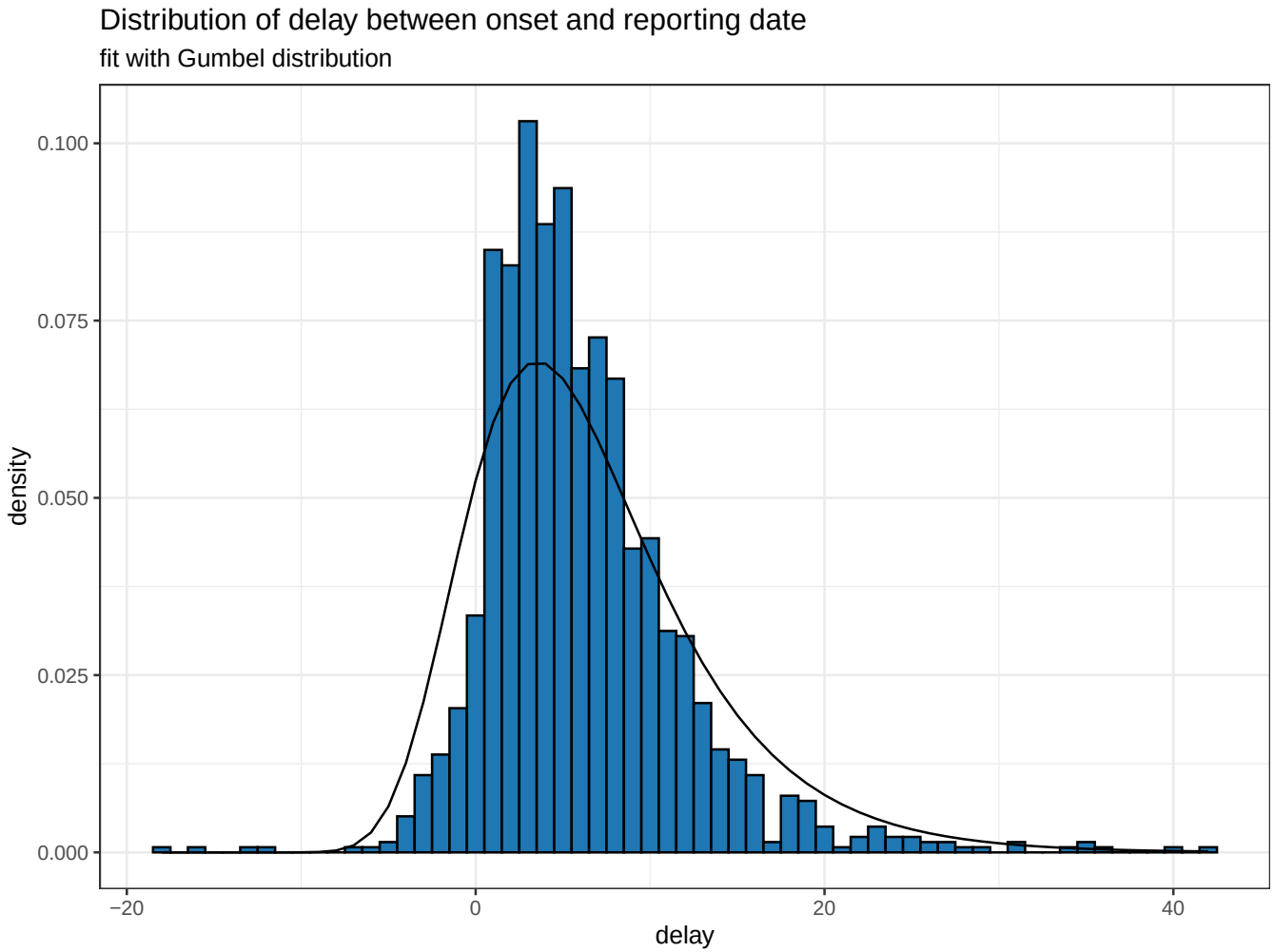
from

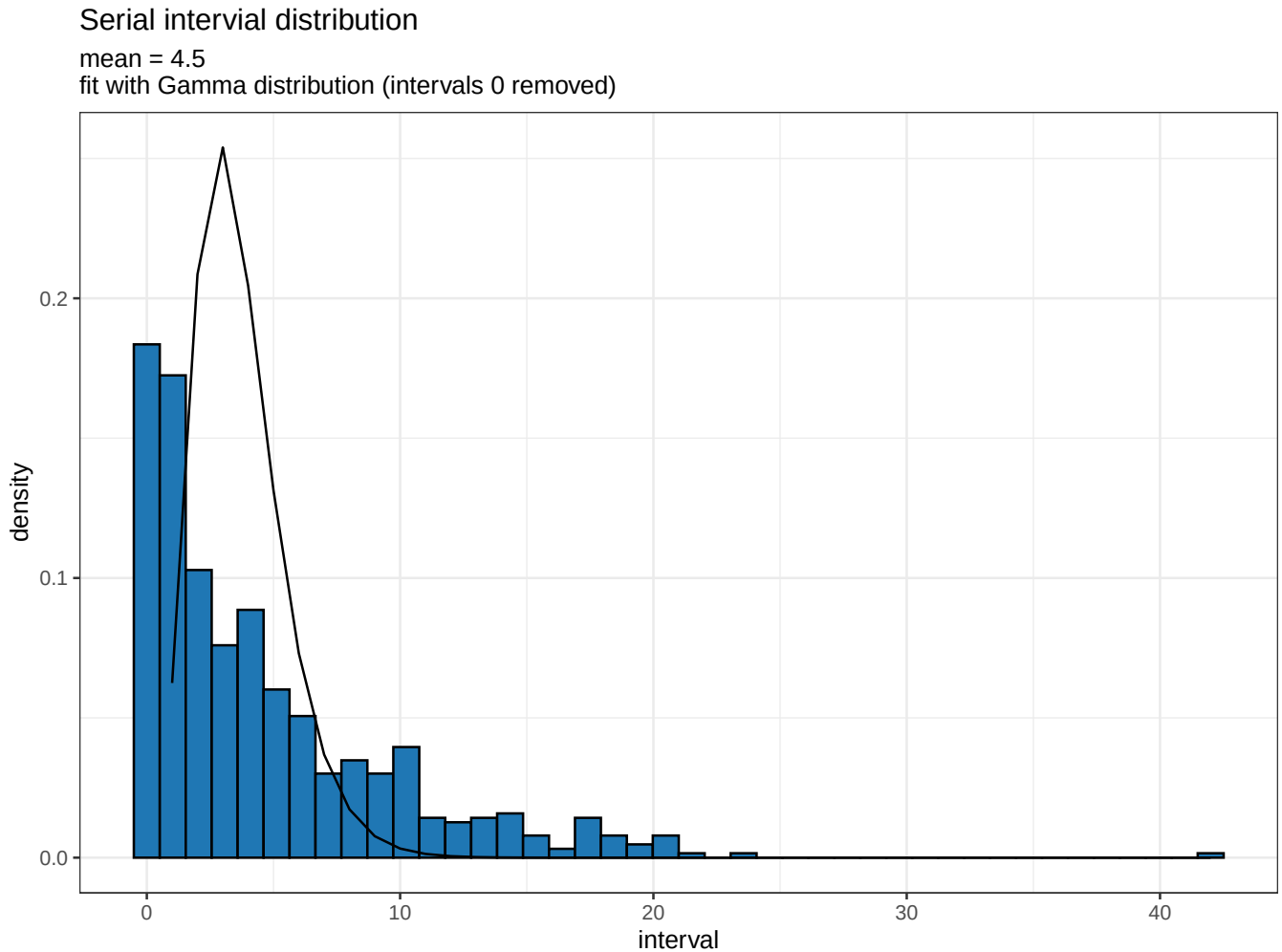
Felix Günther, Andreas Bender, Katharina Katz, Helmut Küchenhoff, Michael Höhle
Nowcasting the COVID-19 pandemic in Bavaria
https://www.medrxiv.org/content/10.1101/2020.06.26.20140210v2
to be published in Biometrical Journal

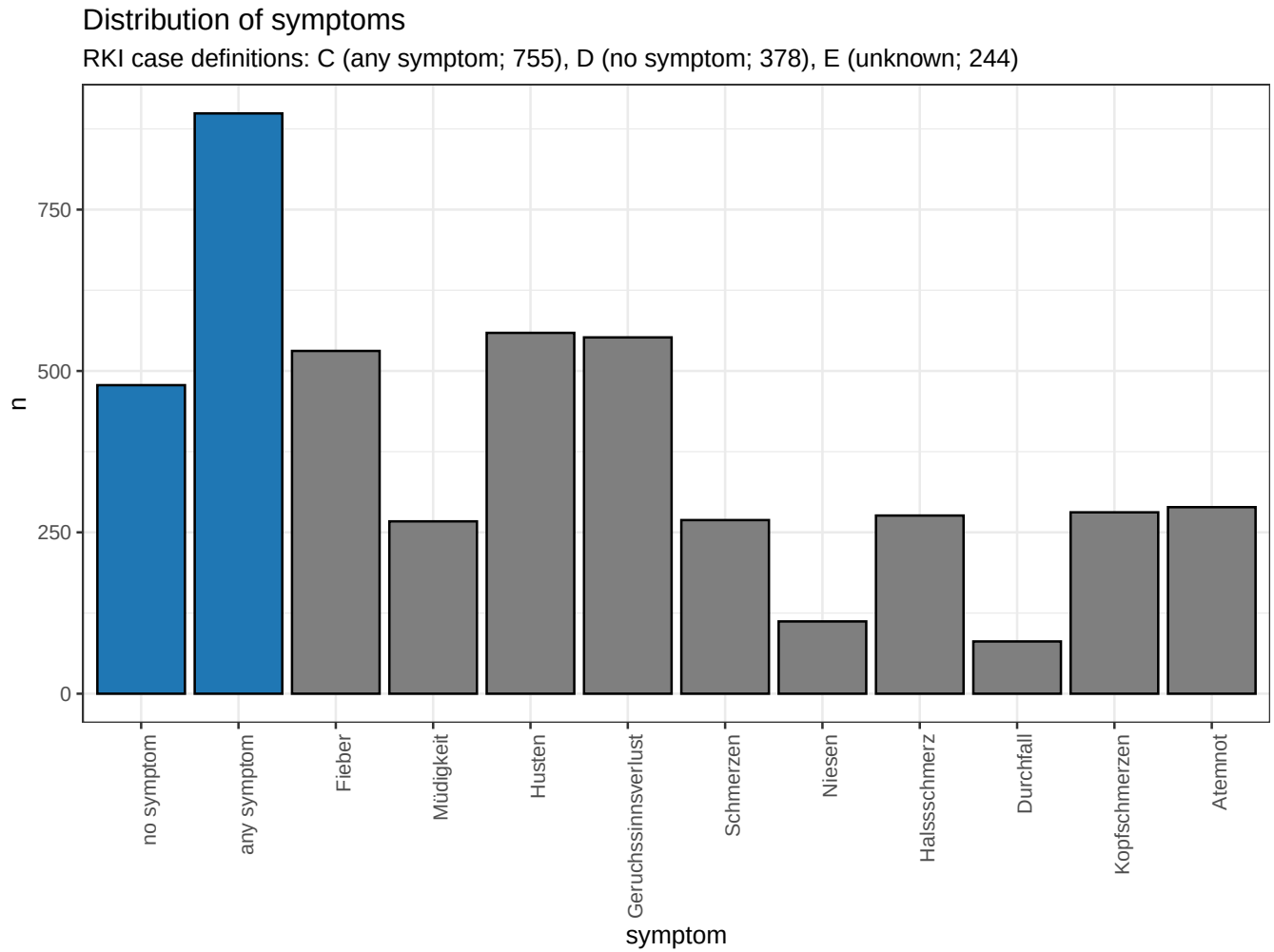
https://github.com/FelixGuenther/nc_covid19_bavaria

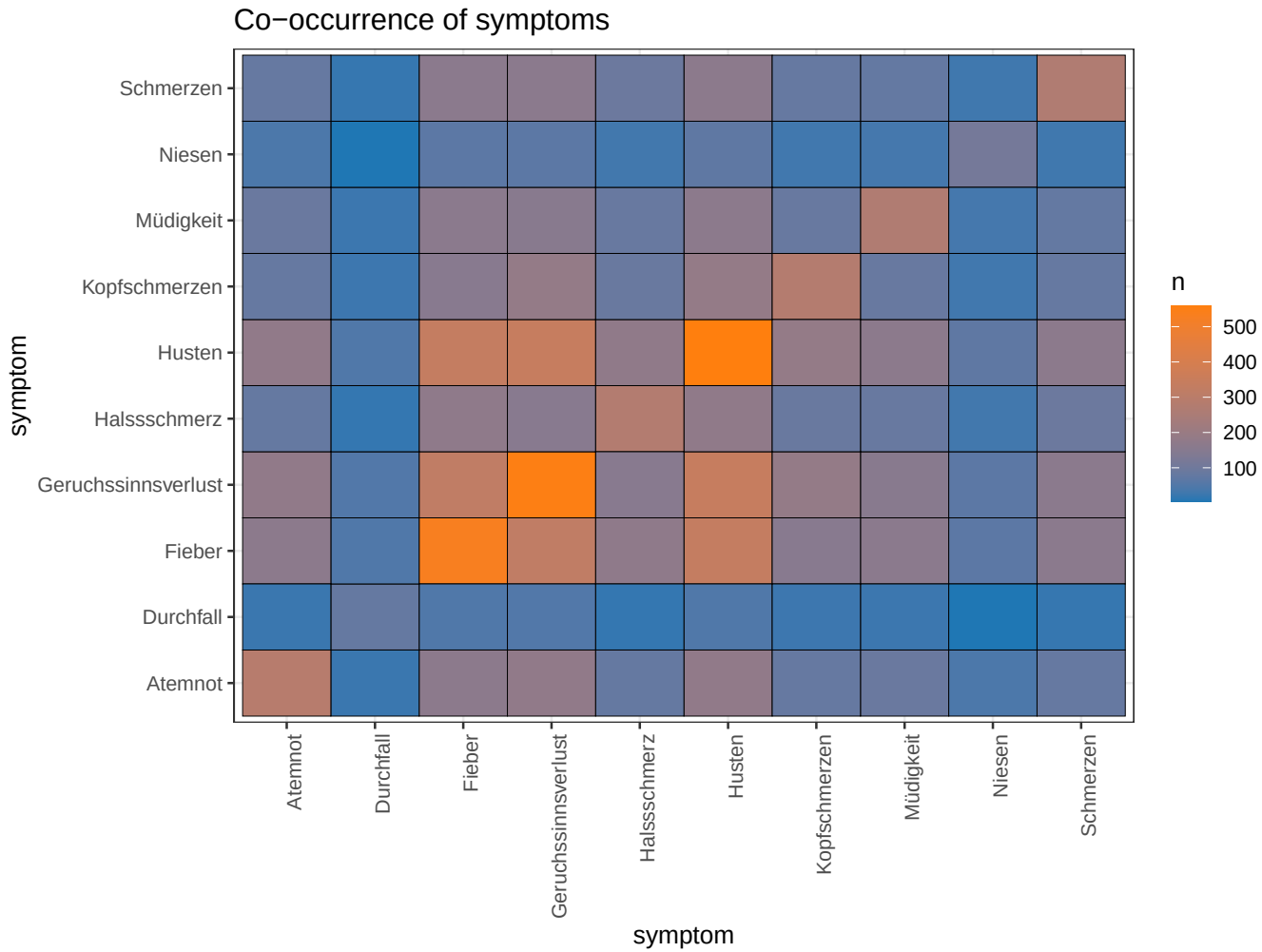
coming soon!

Distributions



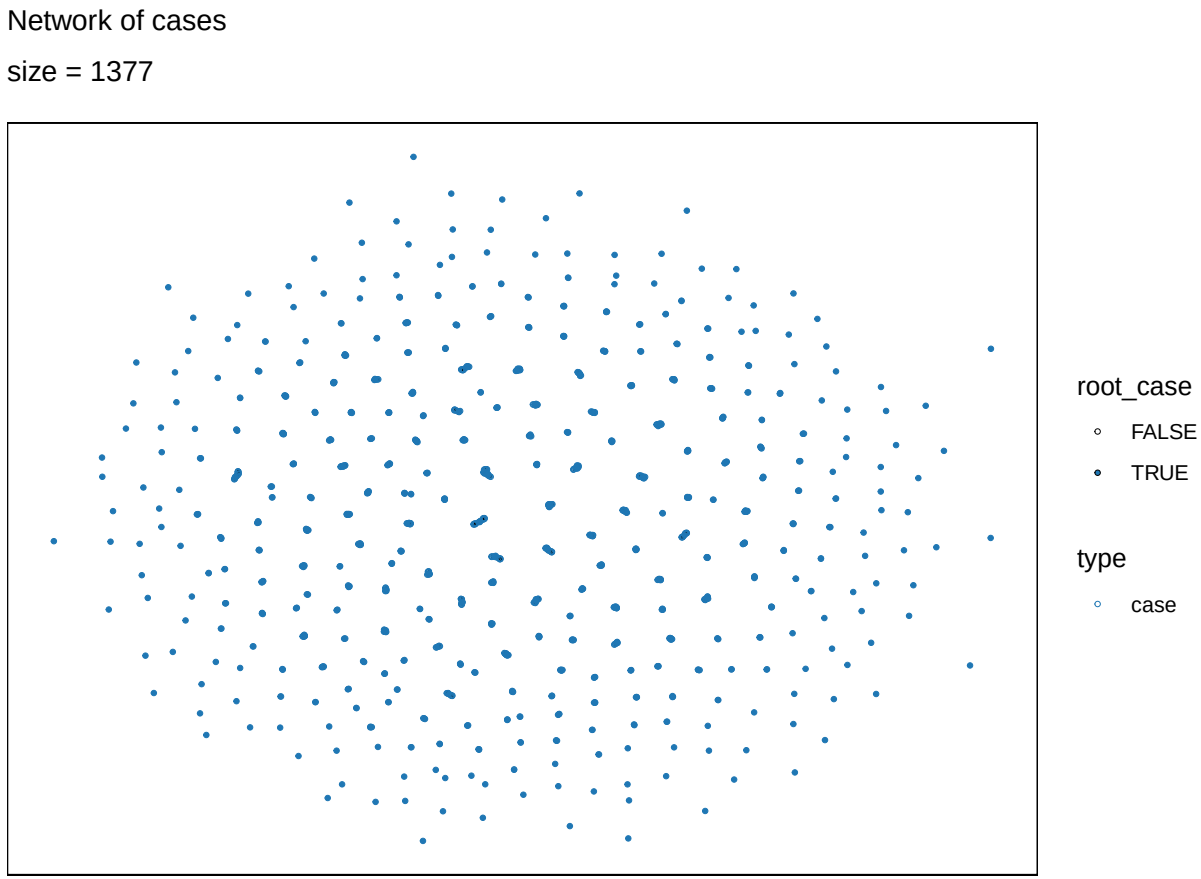

Serial intervial distribution
mean = 4.5
fit with Gamma distribution (intervals 0 removed)
density
0.2
0.1
0.0
0
10
20
30
40
interval


Distribution of symptoms
RKI case definitions: C (any symptom; 755), D (no symptom; 378), E (unknown; 244)
n
750
500
250
0
no symptom
any symptom
Fieber
Müdigkeit
Husten
Geruchssinnsverlust
Schmerzen
Niesen
Halssschmerz
Durchfall
Kopfschmerzen
Atemnot
symptom


Co−occurrence of symptoms
symptom
Schmerzen
Niesen
Müdigkeit
Kopfschmerzen
Husten
Halssschmerz
Geruchssinnsverlust
Fieber
Durchfall
Atemnot
Atemnot
Durchfall
Fieber
Geruchssinnsverlust
Halssschmerz
Husten
Kopfschmerzen
Müdigkeit
Niesen
Schmerzen
symptom
n
500
400
300
200
100

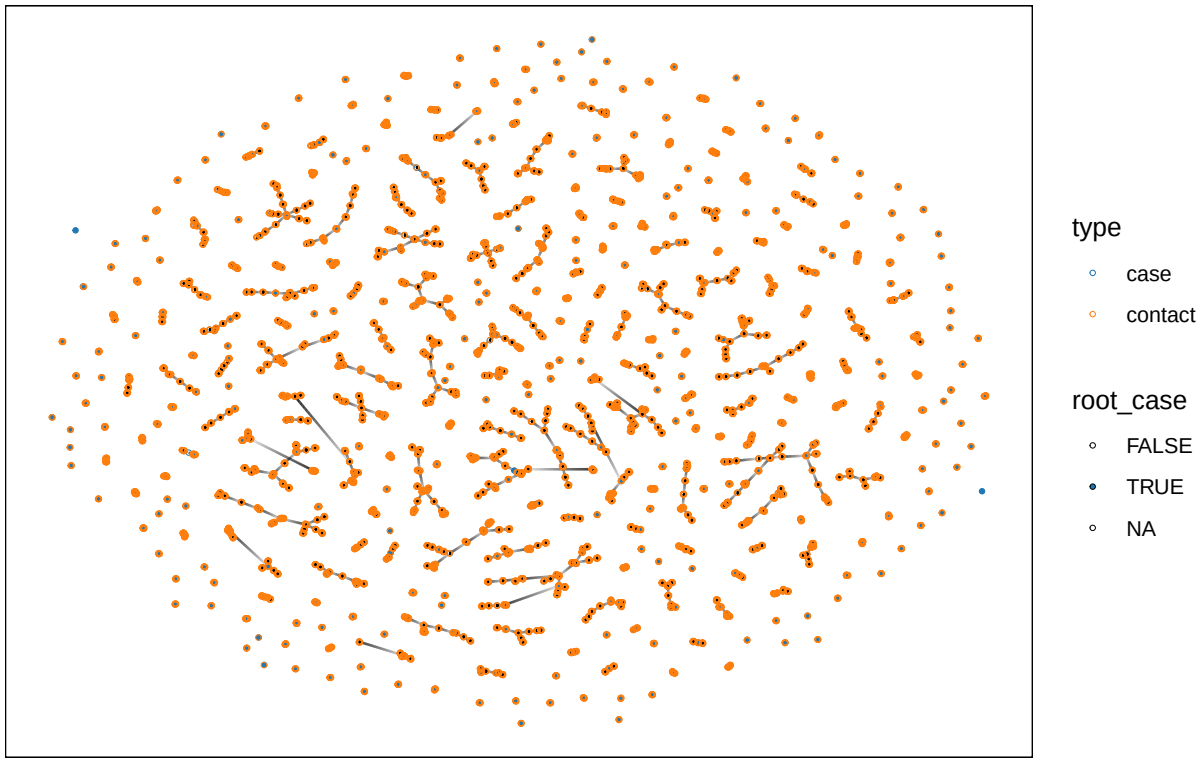
Graph

building the complete graph ~ 8 s



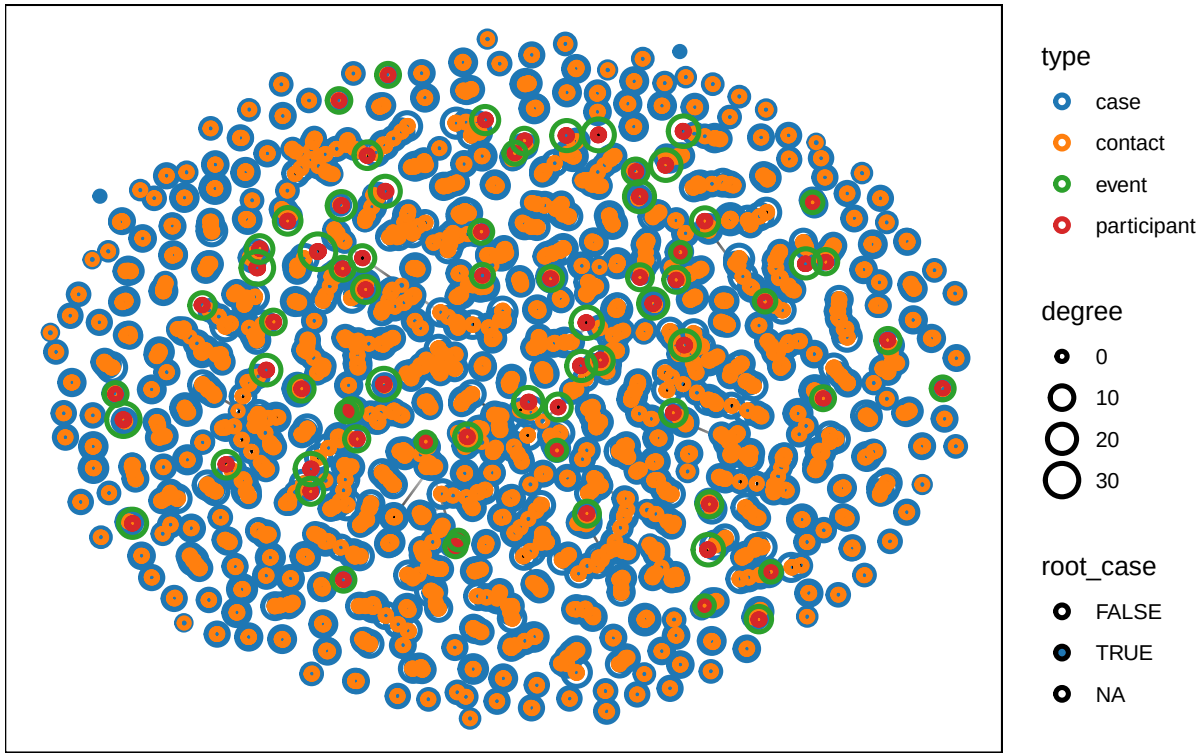
Network of cases and their contacts

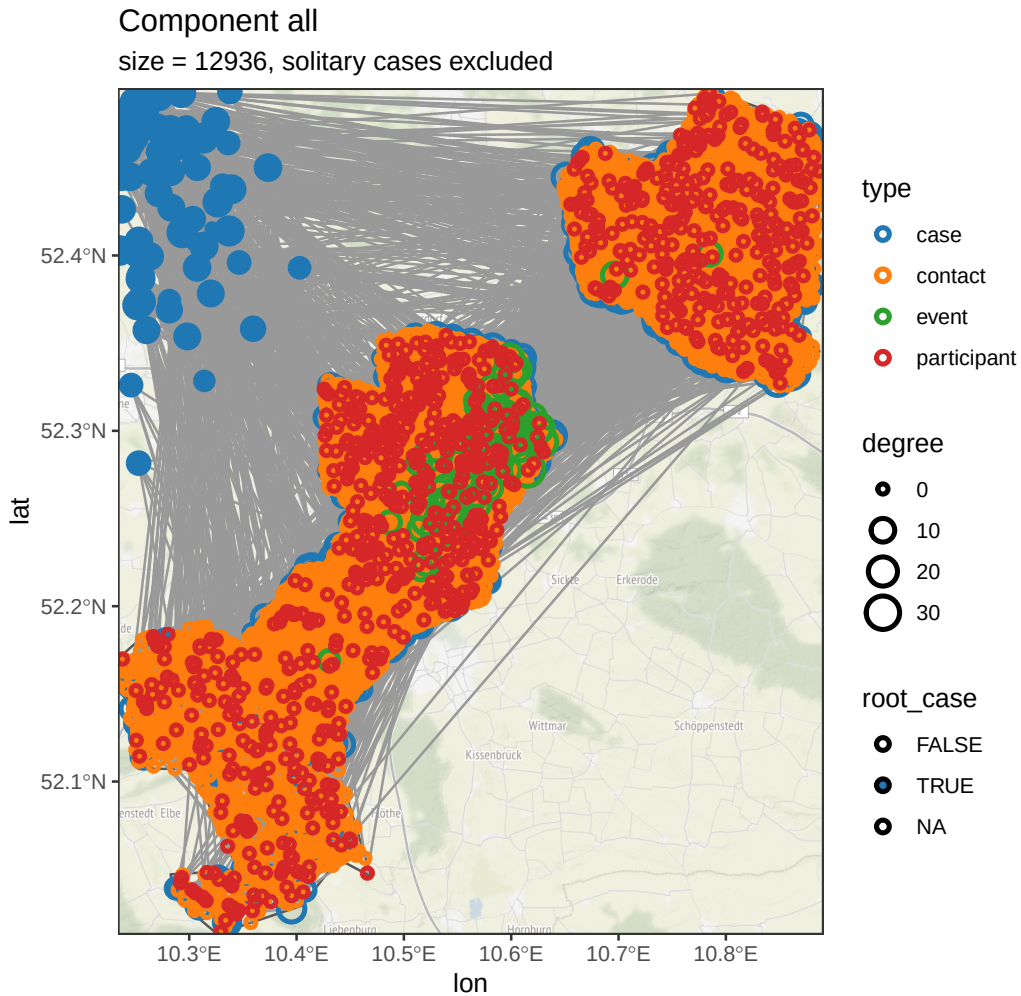
size = 11948

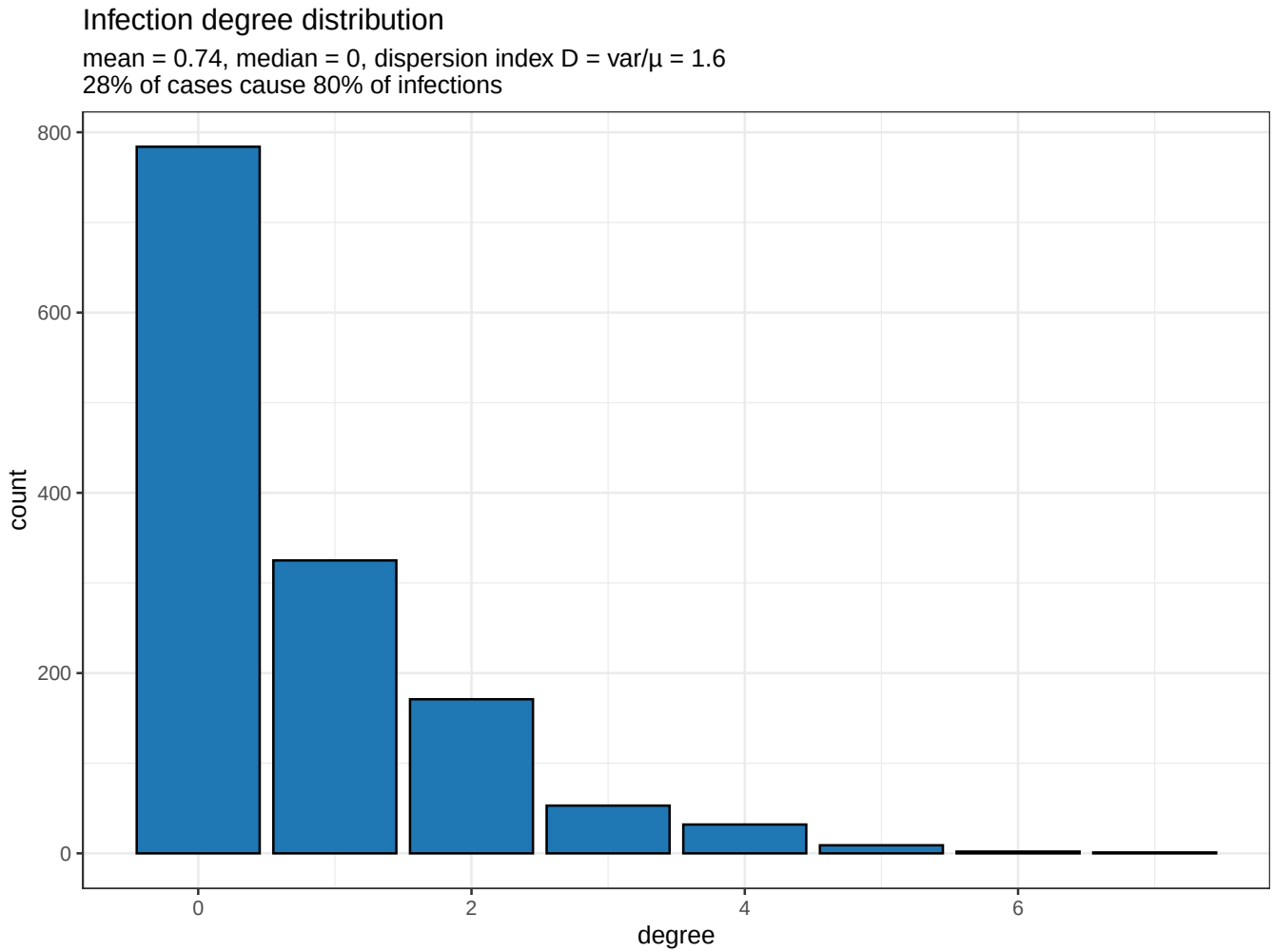


Component all

size = 12936, solitary cases excluded

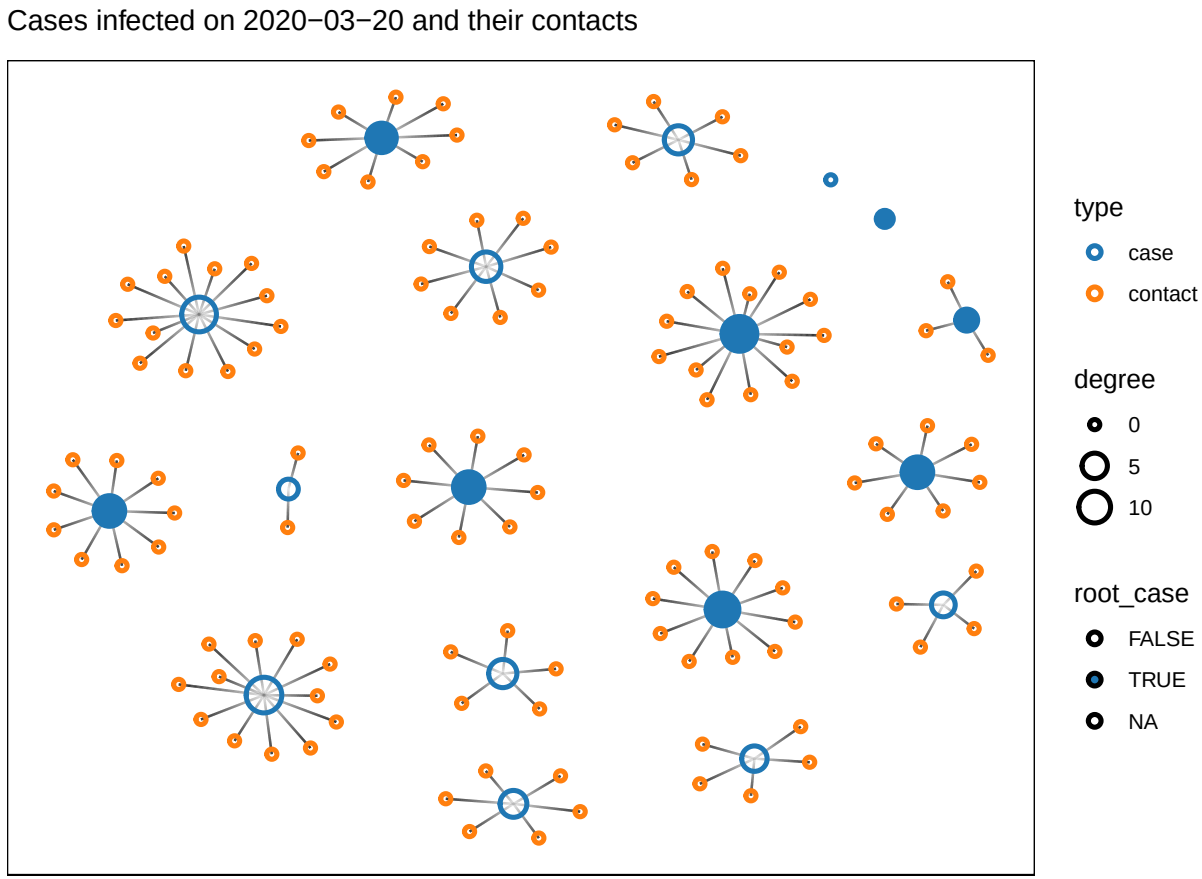



Component all
size = 12936, solitary cases excluded
lat
52.4°N
52.3°N
52.2°N
52.1°N
10.3°E
10.4°E
10.5°E
10.6°E
10.7°E
10.8°E
lon
type
case
contact
event
participant
degree
0
10
20
30
root_case
FALSE
TRUE
NA


Infection degree distribution
mean = 0.74, median = 0, dispersion index D = var/μ = 1.6
28% of cases cause 80% of infections
800
600
400
200
0
count
0
2
4
6
degree

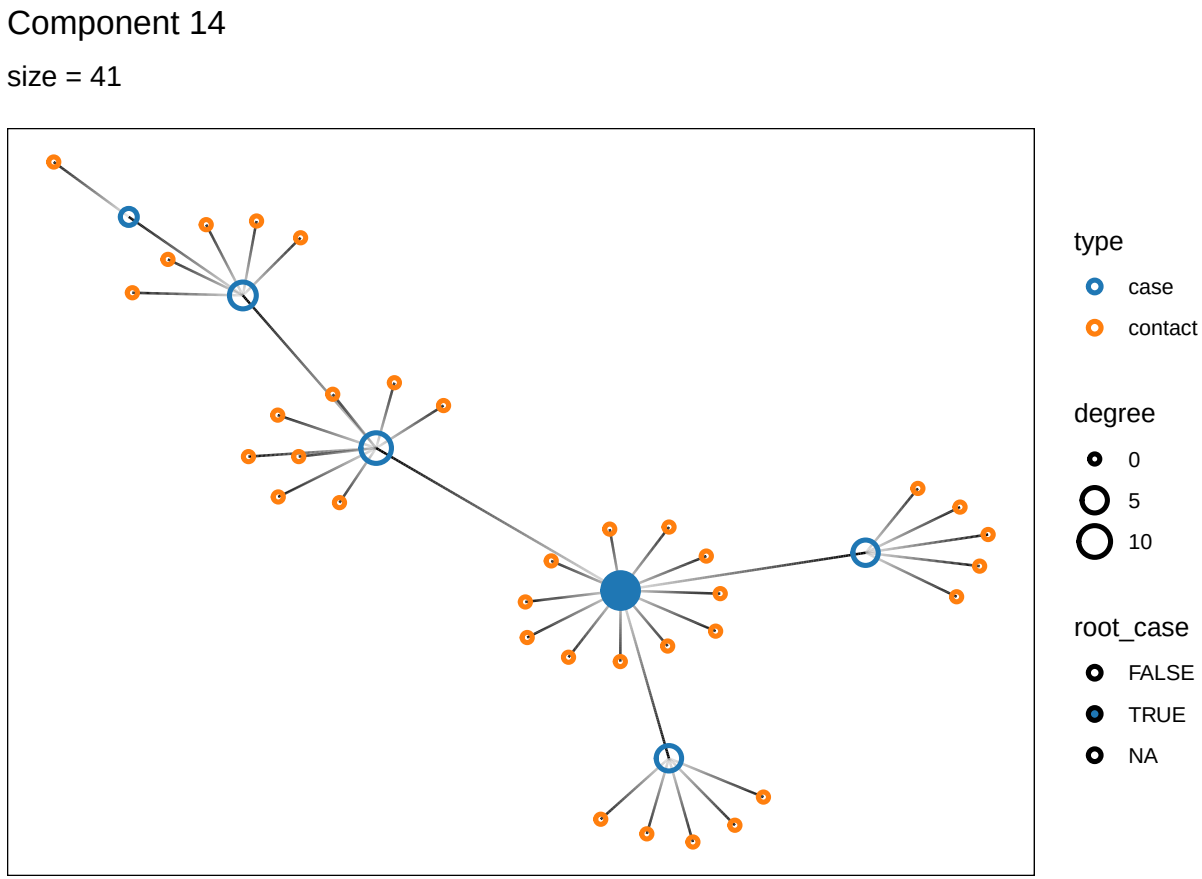
Filter graph by infection date

as an example... *reporting* date might be more relevant

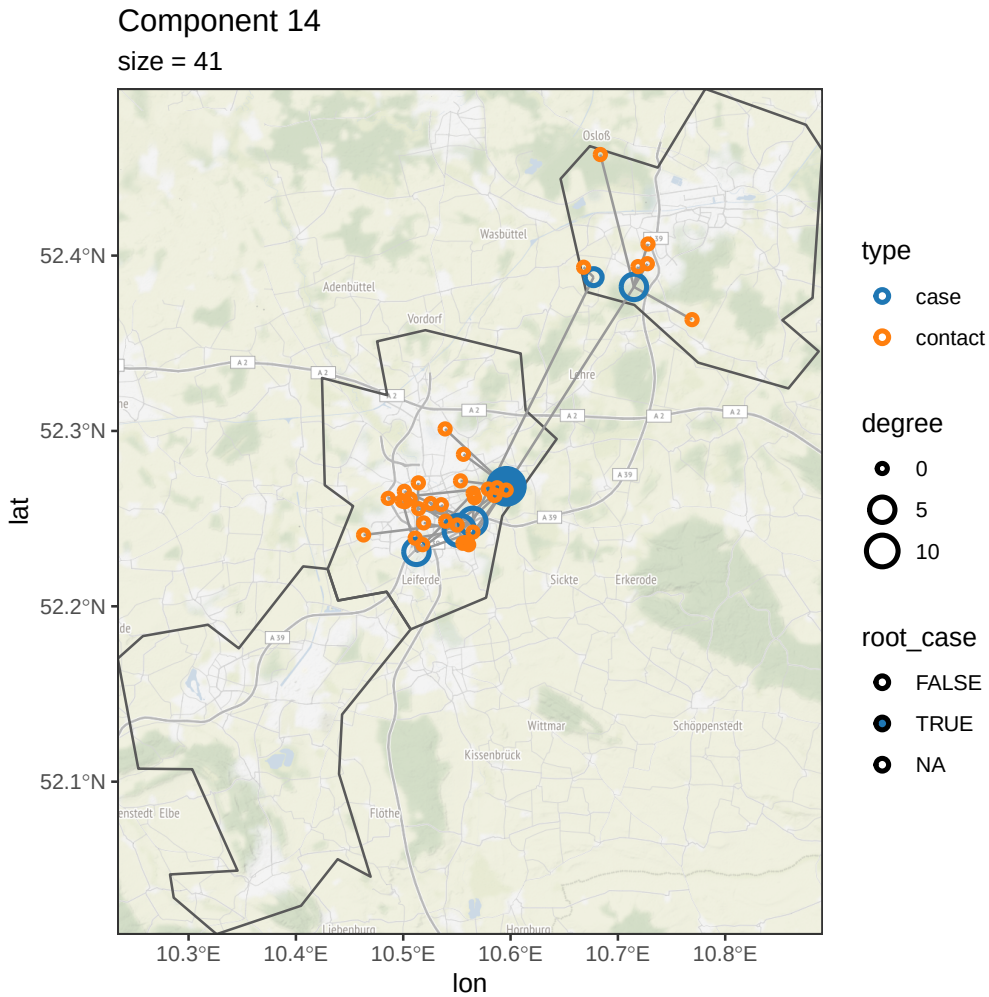


Example 1: component with infection chain

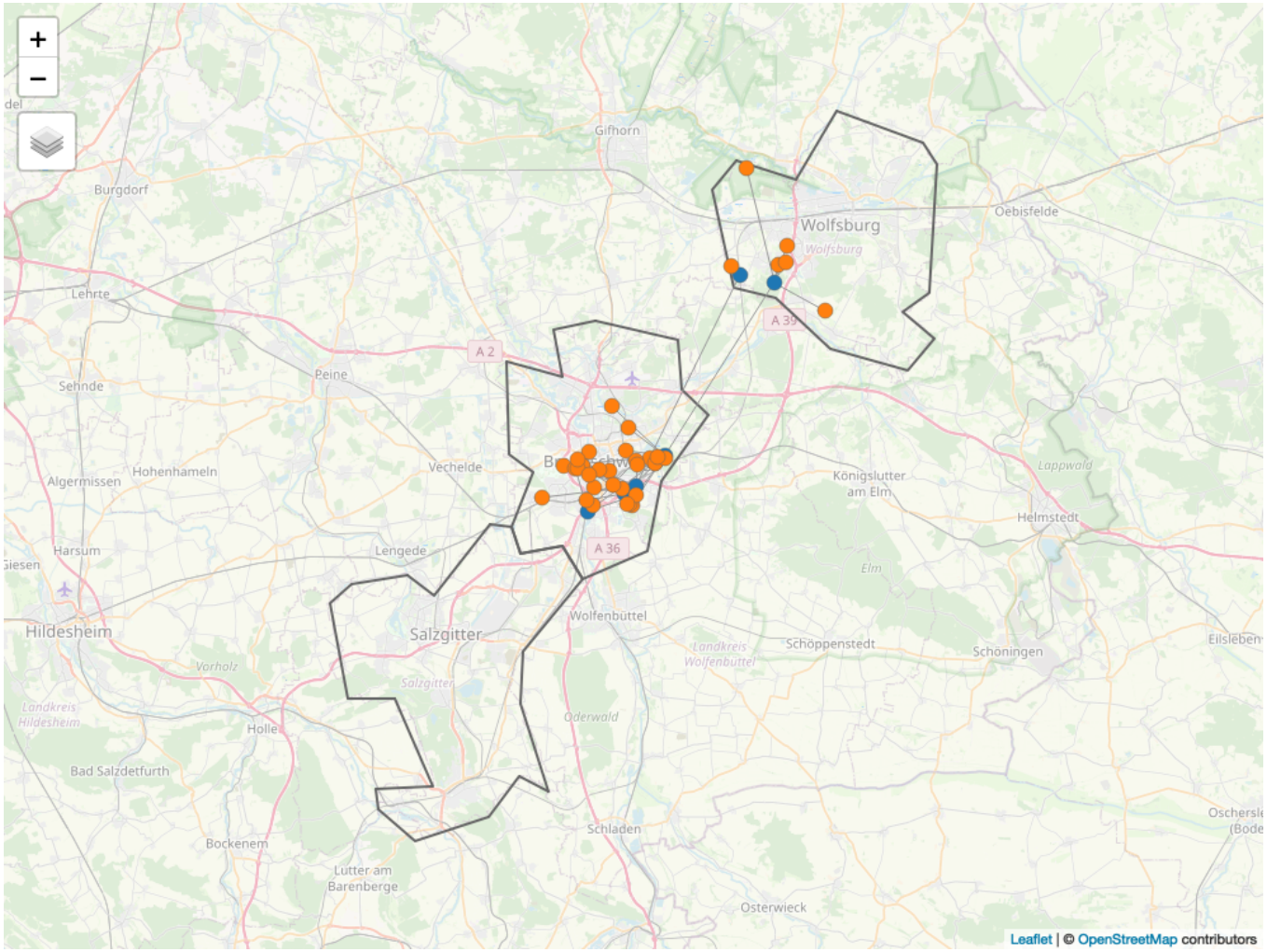
abstract visualization



static visualization in space

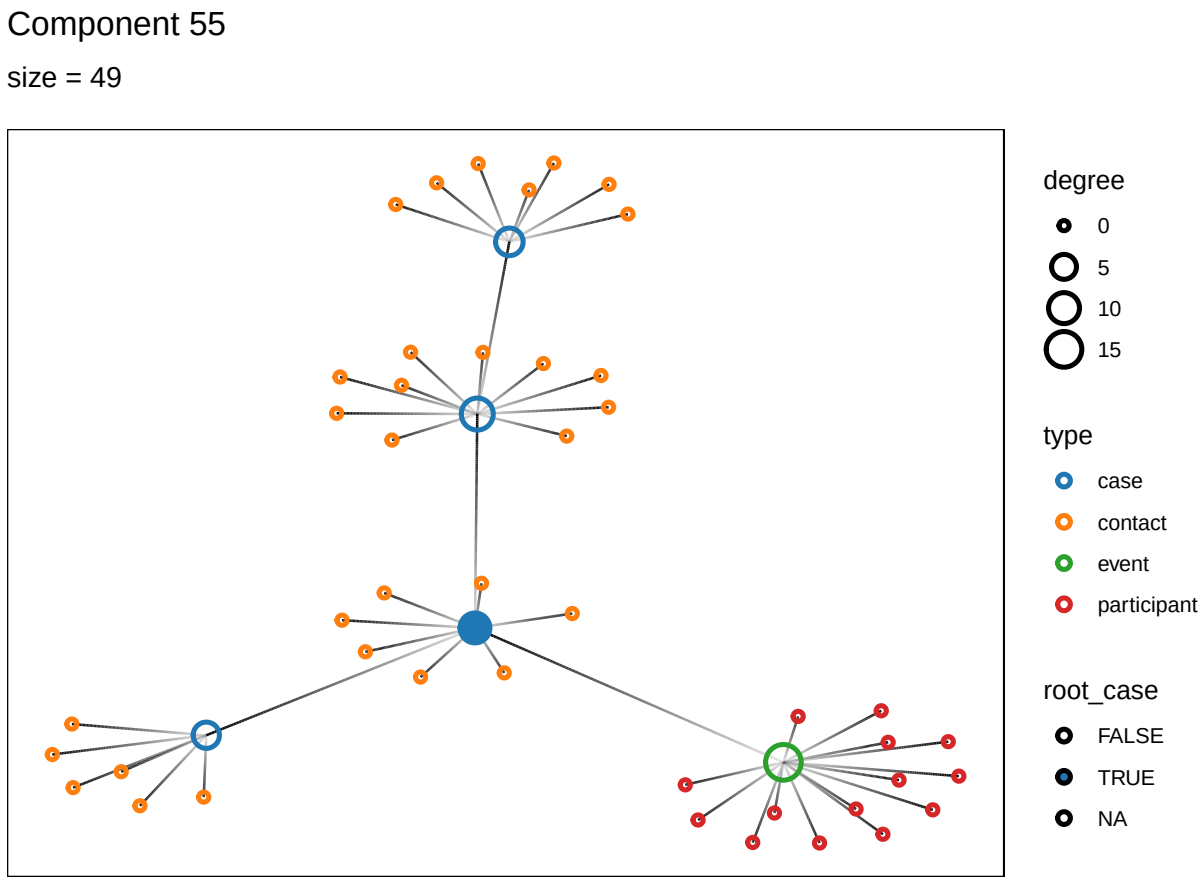


interactive visualization in space

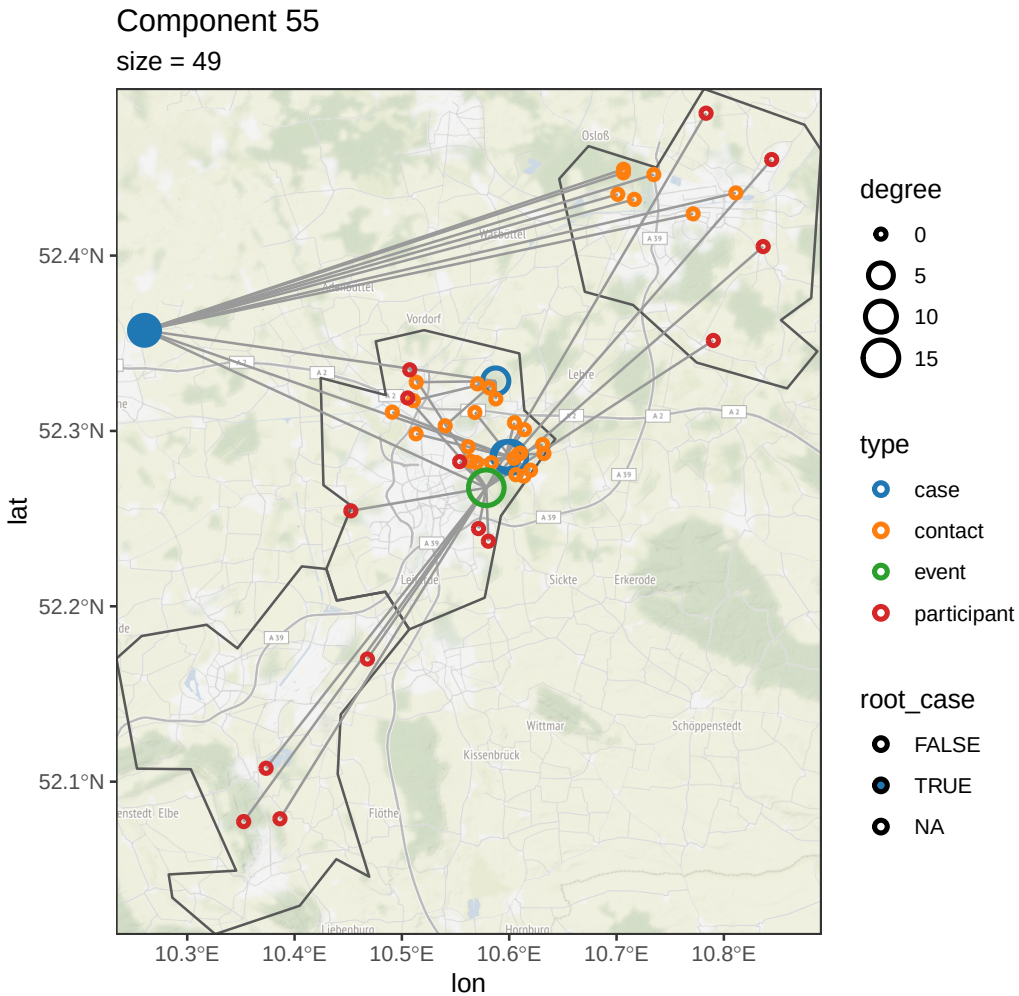


Example 2: component with event

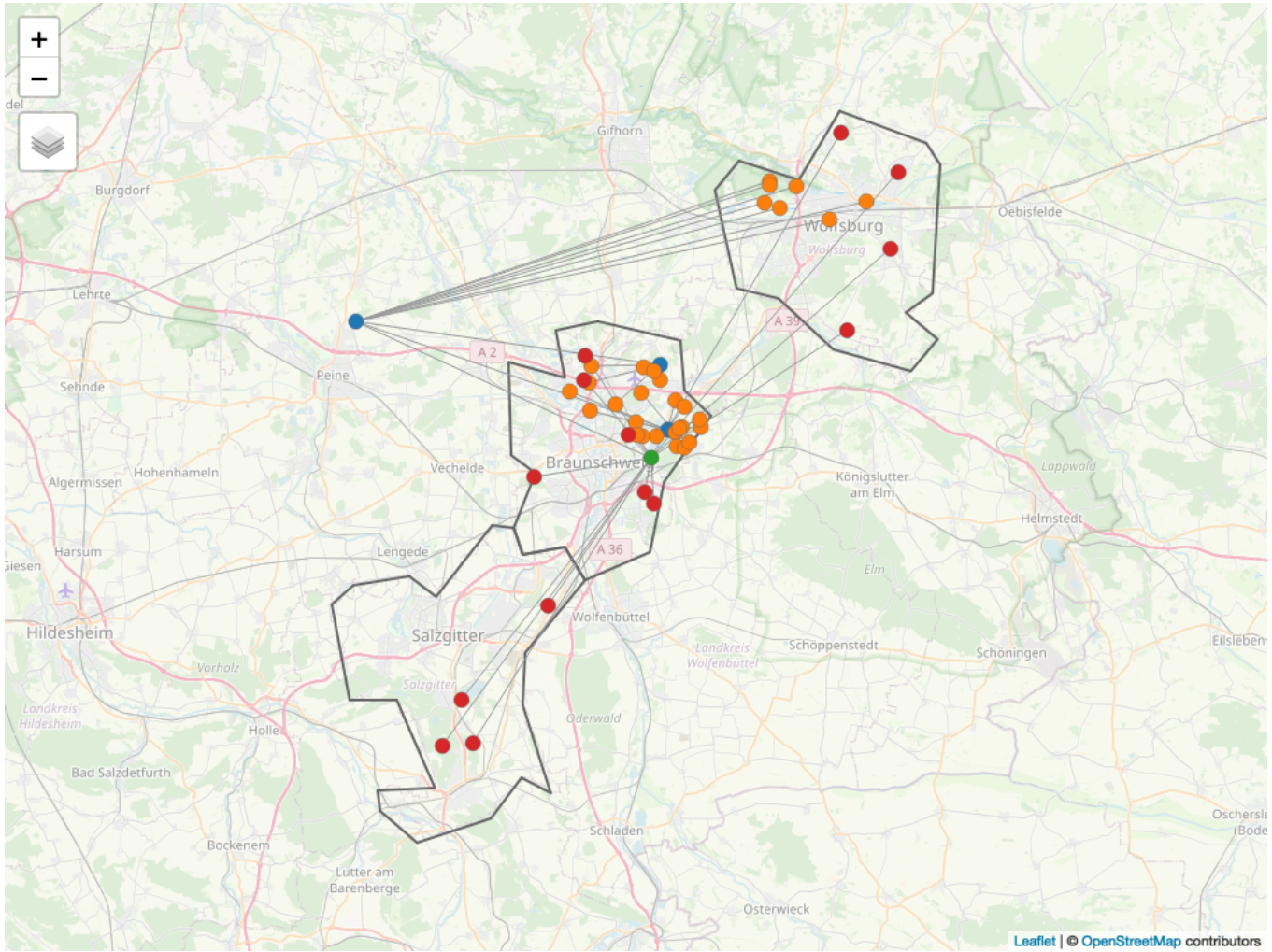
abstract visualization



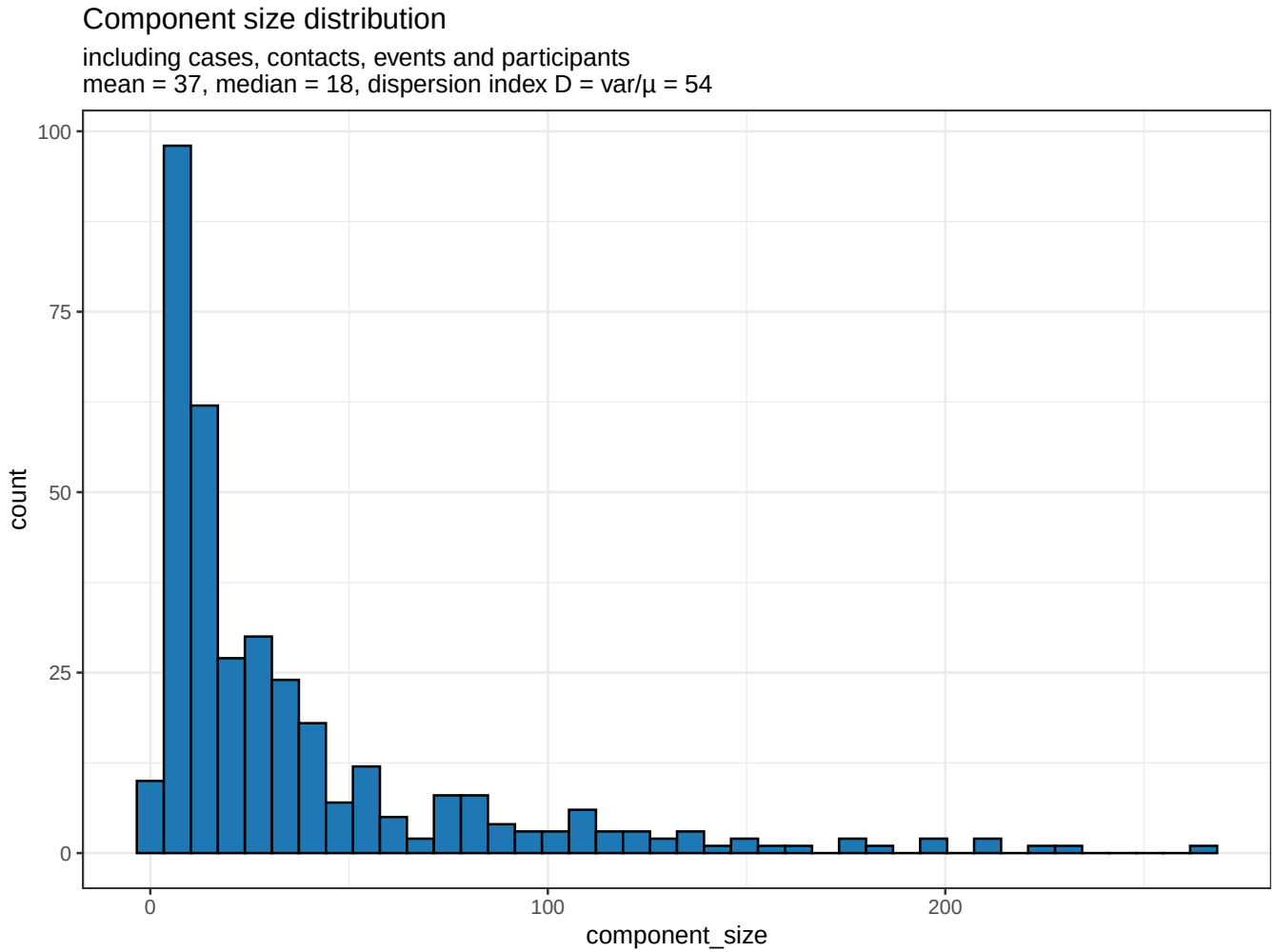
static visualization in space



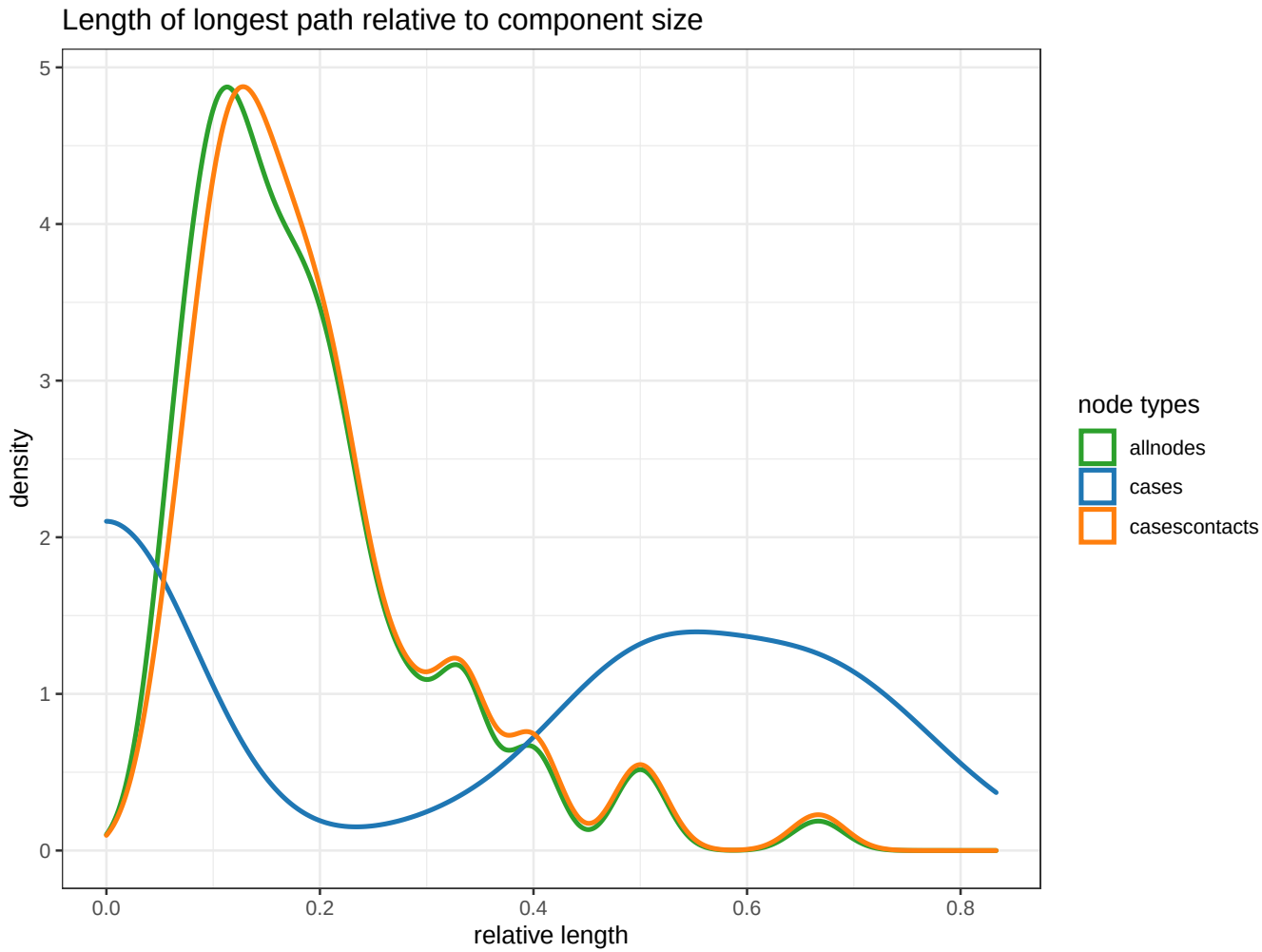
interactive visualization in space



Component features

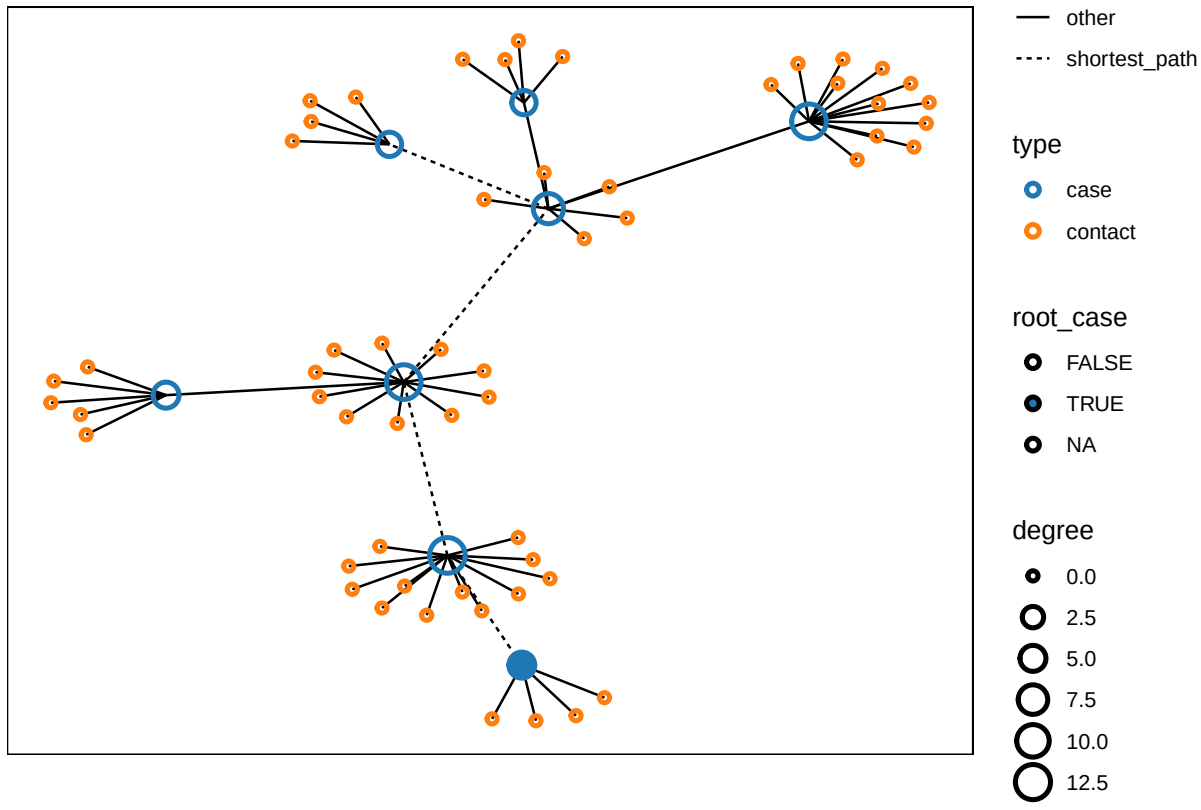


overall: 353 components


Length of longest path relative to component size
density
relative length
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
node types
allnodes
cases
casescontacts

Shortest path from case 824 to case 1306

component 5



Inter-regional graphs

Nodes in county, size = # persons

Edges between counties, width = # contacts or participations

coming soon!

Outlook

Interface to SORMAS and SORMAS-Stats formats:

- export test data to SORMAS
- import SORMAS data to run code (vis + analysis)
- export results to SORMAS-Stats

in progress: Jonas + export from SORMAS test server

Improve the **generation process:**

- reduce **data quality** (add missing, incorrect values)
- statistics-based and age-dependent prob. hospitalization, prob. death
- age-dependent infectiosity
- age distribution in age group
- prob. test ~ testing rate
- non-tree structure (already loops possible through events)
- events are local

Add **features:**

- proper address ~ location / phone number / birth date / sex “divers”
- time and result of test
- isolation/quarantine dates
- hospitalization and death dates
- ICU
- **time dynamics of LHA workflow,** e.g., date of contact, of quarantine
- professional occupation of case
- **setting/context of infection, of event:** type, name

Visualizations and **analyses:**

- nowcast and $R_e(t)$
- aggregation over regions
- **prioritization** (backward tracing), cf. *betweenness*
- aberration detection vs. components
- sizes/number of components in time
- cumulative number of cases, contacts, participants
- dispersion K
- modeling SEIR, under-reporting

Data structure

persons_df

```
tibble [12,867 × 29] (S3: tbl_df/tbl/data.frame)
 $ local_health_authority: chr [1:12867] "Braunschweig" "Braunschweig" "Braunschweig" "Braunschweig" ...
 $ sex                   : chr [1:12867] "w" "w" "w" "w" ...
 $ age_group             : chr [1:12867] "A00-A04" "A00-A04" "A00-A04" "A00-A04" ...
 $ reporting_date        : Date[1:12867], format: "2020-03-22" "2020-03-31" "2020-09-02" "2020-10-06" ...
 $ reporting_week        : chr [1:12867] "2020-W12" "2020-W14" "2020-W36" "2020-W41" ...
 $ onset_date            : Date[1:12867], format: "2020-03-17" "2020-03-30" "2020-09-06" "2020-09-28" ...
 $ id                    : num [1:12867] 1 2 3 4 5 6 7 8 9 10 ...
 $ is_case               : logi [1:12867] TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ age                   : int [1:12867] 4 3 1 0 0 0 11 14 12 13 ...
 $ country_of_residence  : chr [1:12867] "Deutschland" "Deutschland" "Deutschland" "Deutschland" ...
 $ address               : chr [1:12867] "Braunschweig" "Braunschweig" "Wolfsburg" "Braunschweig" ...
 $ infection_date        : Date[1:12867], format: "2020-03-17" "2020-03-28" "2020-09-01" "2020-09-28" ...
 $ case_def_id           : chr [1:12867] "D" "C" "E" "E" ...
 $ case_def              : chr [1:12867] "labordiagnostisch bei nicht erfüllter Klinik" "klinisch-labordiagnostisch" "labordiag
 $ has_symptoms          : logi [1:12867] FALSE TRUE TRUE TRUE TRUE TRUE ...
 $ hospitalized          : logi [1:12867] FALSE FALSE TRUE TRUE TRUE TRUE ...
 $ died                  : logi [1:12867] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ generation            : chr [1:12867] "4" "4" "3" "0" ...
 $ infected_by           : int [1:12867] 117 342 289 NA 1098 NA NA 1106 NA NA ...
 $ degree                : num [1:12867] 0 0 0 0 0 1 2 0 0 1 ...
 $ has_contact           : logi [1:12867] TRUE TRUE TRUE FALSE TRUE FALSE ...
 $ contact_contacted     : logi [1:12867] NA NA NA NA NA NA ...
 $ contact_quarantine    : logi [1:12867] NA NA NA NA NA NA ...
 $ contact_tested        : logi [1:12867] NA NA NA NA NA NA ...
 $ first_name            : chr [1:12867] "Andrea" "Andrea" "Sabine" "Monika" ...
 $ family_name           : chr [1:12867] "Nowak" "Braun" "Koch" "Yılmaz" ...
 $ longitude             : num [1:12867] 10.6 10.6 10.7 10.5 10.6 ...
 $ latitude              : num [1:12867] 52.3 52.3 52.4 52.3 52.2 ...
 $ was_in_event          : logi [1:12867] FALSE FALSE FALSE FALSE FALSE FALSE ...
```

symptoms_cases_df

```
# A tibble: 3,217 x 3
      id symptom             id_person
   <dbl> <chr>                   <int>
 1     1 Fieber                      2
 2     2 Geruchssinnsverlust         2
 3     3 Niesen                      2
 4     4 Kopfschmerzen               2
 5     5 Atemnot                     2
 6     6 Husten                      3
 7     7 Geruchssinnsverlust         3
 8     8 Schmerzen                   3
 9     9 Halssschmerz                3
10    10 Kopfschmerzen               3
# ... with 3,207 more rows
```

contacts_df

```
# A tibble: 11,589 x 4
      id id_index id_contact is_infection
   <dbl>    <int>      <dbl> <lgl>
 1     1      117          1 TRUE
 2     2      342          2 TRUE
 3     3      289          3 TRUE
 4     4     1098          5 TRUE
 5     5     1106          8 TRUE
 6     6      333         15 TRUE
 7     7     1369         16 TRUE
 8     8      897         17 TRUE
 9     9      185         18 TRUE
10    10       22         19 TRUE
# ... with 11,579 more rows
```

events_df

```
# A tibble: 69 x 5
      id date       address      longitude latitude
   <int> <date>     <chr>            <dbl>    <dbl>
 1     1 2020-10-17 Braunschweig      10.6     52.3
 2     2 2020-10-05 Braunschweig      10.6     52.3
 3     3 2020-09-19 Braunschweig      10.5     52.2
 4     4 2020-03-13 Braunschweig      10.6     52.3
 5     5 2020-08-07 Braunschweig      10.5     52.3
 6     6 2020-04-09 Braunschweig      10.6     52.3
 7     7 2020-08-04 Braunschweig      10.6     52.3
 8     8 2020-04-18 Braunschweig      10.6     52.3
 9     9 2020-03-31 Braunschweig      10.6     52.3
10    10 2020-08-06 Braunschweig      10.6     52.3
# ... with 59 more rows
```

event_participants_df

```
# A tibble: 994 x 3
      id id_event id_participant
   <dbl>    <int>          <dbl>
 1     1        1           1243
 2     2        1          11950
 3     3        1          11951
 4     4        1          11952
 5     5        1          11953
 6     6        1          11954
 7     7        1          11955
 8     8        1          11956
 9     9        1          11957
10    10        1          11958
# ... with 984 more rows
```

network_graph

```
# A tbl_graph: 12936 nodes and 12583 edges
#
# A rooted forest with 353 trees
#
# Node Data: 12,936 x 6 (active)
  id    type  root_case date                  geometry degree
  <chr> <chr> <lgl>     <date>                 <POINT>  <dbl>
1 1     case  FALSE     2020-03-17 (10.62963 52.29662)     12
2 2     case  FALSE     2020-03-28  (10.57941 52.2834)      8
3 3     case  FALSE     2020-09-01 (10.70493 52.39894)      7
4 4     case  TRUE      2020-09-28  (10.46876 52.2861)      4
5 5     case  FALSE     2020-10-03  (10.5652 52.24183)      4
6 6     case  TRUE      2020-09-08 (10.48293 52.27645)      6
# ... with 12,930 more rows
#
# Edge Data: 12,583 x 3
   from    to                              geometry
  <int> <int>                            <LINESTRING>
1   117     1 (10.69714 52.37936, 10.62963 52.29662)
2   342     2  (10.61909 52.29242, 10.57941 52.2834)
3   289     3 (10.55414 52.27636, 10.70493 52.39894)
# ... with 12,580 more rows
```