The SARS-CoV-2 effective reproduction rate has a high correlation with a contact index derived from large-scale individual location data using GPS-enabled mobile phones in Germany

Sten Rüdiger*[1,2], Stefan Konigorski[3], Jonathan Edelman[3], Detlef Zernick[1], Christoph Lippert†[3], and Alexander Thieme‡ [4,5]

[1]NET CHECK GmbH
[2]Humboldt-Universität zu Berlin
[3]Digital Health - Machine Learning, Hasso-Plattner-Institut, Universität Potsdam
[4]Department of Radiation Oncology, Charité - Universitätsmedizin Berlin, Germany
[5]Berlin Institute of Health (BIH), Berlin, Germany

Abstract

The novel coronavirus (SARS-CoV-2), which was first discovered in Hubei, China in December 2019, has caused an ongoing pandemic. Due to pauci-symptomatic cases, the virus may spread invisibly in a community. In the absence of vaccination, non-pharmaceutical interventions (NPIs) like interpersonal distancing were implemented in several countries and have been key to effectively reduce viral spreading. In Germany after an exponential growth of case numbers in March 2020, NPIs were able to effectively control the pandemic and sufficiently reduced the daily reported new infections allowing for partial release of NPIs. We developed a novel statistical method to evaluate contacts between individuals, which is essential for virus transmission. We derived the *contact index*, an index for the intensity and heterogeneity of contact behavior from spatial proximity between individuals as proxy for physical interaction based on complex network science. We estimated the contact index from large-scale GPS mobile phone data of 1.15 to 1.4 million users in Germany per day (March to July 2020). A high correlation between the contact index and the effective reproduction number six days later could

*co-correspondence, sten.ruediger@netcheck.de
†co-correspondence, christoph.lippert@hpi.de
‡co-correspondence, alexander-henry.thieme@charite.de

be observed (Pearson correlation $r = 0.96$, P-value < 0.001 for all reported Pearson correlations). This correlation was observed in three different phases of the virus spread in Germany 1) the early phase of the first wave with the highest reproduction rate, 2) phase of strict NPIs (lockdown) with the lowest reproduction, 3) release of NPIs accompanied with an increase of reproduction. The results show that the contact index is able to model and potentially forecast the time evolution of the pandemic in Germany.

1 Introduction

In December 2019 a novel coronavirus, namely the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), first discovered in Wuhan (Hubei province, China), has rapidly caused an ongoing pandemic with more than 20 million confirmed cases and more than 700,000 deaths as of August 2020. SARS-CoV-2 is highly contagious with an estimated basic reproduction number R0 between 1.5-4 [24, 27, 11, 12]. Invisible community spread may occur until a local outbreak becomes evident by a larger number of severe clinical cases. Healthcare systems and their infrastructure have been repeatedly challenged with a rapid onset of critically ill patients requiring hospitalization and intensive care treatment [17, 6]. Since there was no vaccine available, non-pharmaceutical interventions (NPIs) were considered an important instrument in containing viral spread. The first case of SARS-CoV-2 was confirmed in Germany on January 27, 2020 [1]. At first, Germany did not install strict NPIs, which led to an exponential growth of case numbers. On March 8, 2020 public events with more than 1000 visitors were prohibited, following a nationwide lockdown with a general prohibition of public contacts on March 23, 2020. A peak of 6933 daily new cases was reached on March 27, 2020. A continuous decline in new cases per day could be noticed in April, following a plateau with an average number of daily new cases below 1000 from May till July, 2020. Sufficiently low levels of new daily infections allowed for release of NPIs and reintroduction of containment strategies, like contact tracing. Since SARS-CoV-2 could not be entirely removed from the population, interpersonal distancing and rapid detection of local outbreaks remain of pivotal role.

Currently, the gold standard of SARS-CoV-2 detection is a specific real-time reverse polymerase reaction from respiratory specimen, e.g. nasopharyngeal swab [2]. Due to asymptomatic and pauci-symptomatic carriage of SARS-CoV-2 [21, 19, 3], a larger portion of infected individuals remain undetected. Furthermore, it is estimated that 44% of secondary cases are infected through pre-symptomatic transmission events [8]. With an estimated incubation time of 5.2 days [11], a reported detection time for laboratory testing from symptoms to diagnosis of 6.0 days [18] and considering that a larger portion of infected individuals even remain undetected, laboratory testing does not appear to be ideal for rapid outbreak detection given the short infection doubling time of SARS-Cov-2 in the range of 1.4 to 2.5 days [15].

Methods are needed which allow for the detection of local outbreaks at the earliest moment possible. Contact between individuals is essential for virus transmission and represents the first observable event. We hypothesize that individual location history

data assessed by the Global Positing System (GPS) of mobile phones can provide deep insights in the contact behavior of the population and therefore allows for accurate prediction of the time evolution of new SARS-CoV-2 cases.

In our main analyses, we investigated the association of contact numbers with infection rates in Germany provided by the Robert Koch Institute (RKI). For this, we used a contact-related measure known from complex network science, which takes into account the presence of super-contacters that have many contacts during a single day. We call this quantity the contact index C and hypothesized a temporal correlation between it and infection rates.

2 Methods

2.1 Study population

The investigation relies on GPS location history data that is collected via a Software Development Kit (SDK) developed for the primary purposes of assessing the quality of cell phone networks. Cell phone data is collected by the SDK implemented in more than one million cell phones in Germany. Per day data was received from 1.15 to 1.4 million cell phones during March to July 2020. The legal conditions for the processing of the data were described in a report by A. Böken on May 11, 2020. Data records are anonymous. In a first step the number of contacts for each device is determined so that no positional information is retained. Then the data is aggregated by the number of devices that have a certain number of contacts. Only these aggregated numbers are used for further analysis.

2.2 Number of contacts

The approach of contact detection relies on GPS data from embedded software in cell phones originally used to sample the quality of mobile phone networks. It provides precise GPS location and non-aggregated mobility data from a panel of more than one million anonymous individual users which are representative of the German population. GPS data can be used to simultaneously determine location, mobility and contacts between the users with high accuracy. We assume that each cell phone is used always by the same individual. For each cell phone we obtain records with pings from the SDK that contain up to several hundreds messages per day and device. Each message contains, among others, the GPS coordinates of the cell phone. We then project the positions for each ping to a predefined tile of about 8m $\times$ 8m. Using an identification number of the tile we then scan for coincident presence of two different individuals on the same tile with a maximal difference in the time stamp of 2 minutes which we count as one contact for each of the individuals. The number of contacts attributed to each individual is then the number of contacts during one day. The collected data covers the entire period of the pandemic, including weeks before its beginning.

2.3 Sampling of contacts

The nature of our data collection allows estimating the number of real-world contacts for the entire population of Germany. However, a large part of these real contacts is missing from our cell phone sampling for two main reasons: (A) We only cover a fraction of devices. (B) We only cover times when the cell phone is sending a ping. In more detail regarding (A), we cover about 800,000 GPS-enabled devices per day, so that the majority of contacts for an individual goes undetected. As there are about 83 million persons in Germany, we can expect to cover about 1% of persons. Regarding (B), a typical cellphone sends about 200 pings per day. In order to cover the entire day, one ping every two minutes i.e. 720 pings per day are needed. Thus, only about 28% of the time of the day is covered for the average device. Assuming for simplification that the time of pings are independent for different devices, a lower bound estimate of the probability that a contact between two devices was observed is $0.28 \times 0.28 \approx 0.1$. So, according to this rough calculation, we can expect to track at least $0.01^2 \times 0.1 = 0.001\%$ of all contacts between any two persons living in Germany.

2.4 Effective R calculation

The effective reproduction number R values in our analysis have been obtained from the RKI Nowcasting website [1]. For a given day d, R is calculated as the ratio of the sums of infections for days d to $d+6$ and $d-7$ to $d-1$ [22]. This number is then attributed to day d. For regional evaluations, confirmed cases are counted by the district where the individual has their home address.

2.5 Contact index

For a detailed analysis of the data we turn to methods from complex network science. We first define a network or graph of users for every day by assigning one device/individual to a node of the network while each contact defines an edge between nodes. The presence of hosts with many contacts and the ensuing broadness of the degree distributions has strong implications for the occurrence of pandemics. As well, there is a large body of literature in epidemiology that relates a pandemic to an increase of a measure different from the simple number of contacts and which also takes into account the number of large degree nodes. This measure is related to the heterogeneity of the degree distribution (for a review see e.g. Pastor-Satorras et al. [16]) and is the relevant quantity at least for networks which do not exhibit strong degree correlations. This parameter is called contact index C in the following. A mathematical definition of the quantity C can be found in the appendix.

[1] https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting_Zahlen.xlsx?__blob=publicationFile

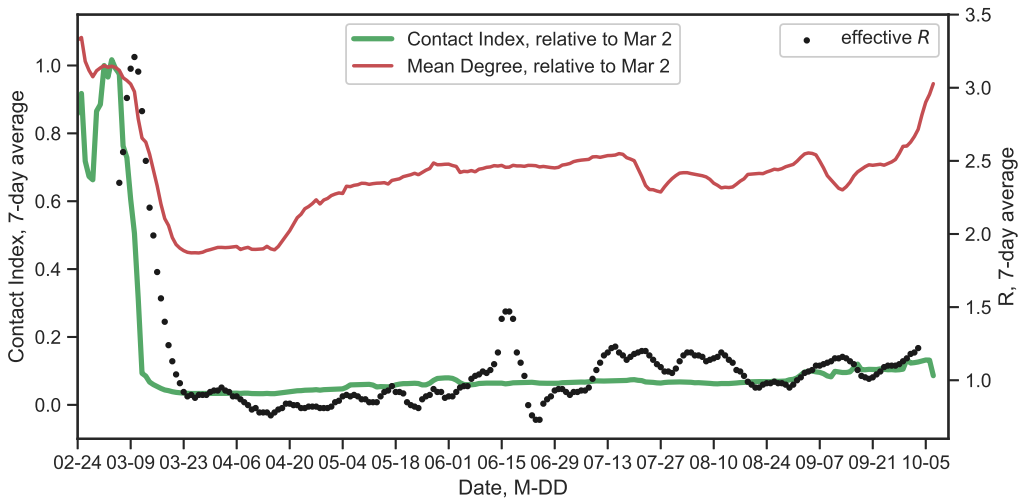

Figure 1: Mean number of contacts from cell phone records (red, left scale), contact index (green, left scale) and effective R (seven day R [22], blue dots, right scale). Each curve has been smoothed by using a seven day sliding window. Mean contacts and contact index are scaled relative to their values on March 2, 2020.

2.6 Statistical analysis

In the analysis, we estimate the contact index from a sample of nodes from the full network of cell phones in Germany. We present descriptive statistics and temporal trends of the number of contacts as well as of the contact index. Finally, we investigate their association with infection rates assessed by R by estimating their Pearson correlation coefficient. The Pearson correlations and their p-values were determined using Python's Scipy package, version 1.3.1.

3 Results

3.1 Association of number of contacts with infections

The average number of cell phones registered during a day was about 800,000. Per day we find between 20,000 and 160,000 devices that had at least one match. The total number of matched pairs varied between 150,000 (before lock-down) and 12,000 (during lock-down) per day.

Figure 1 illustrates that the mean number of contacts of individuals per day (in red) clearly relates to the evolution of R (in black). This holds particularly in the initial phase of the outbreak. The Pearson correlation coefficient between the two curves depends on the time shift and is maximized (over shifts between 0 and 14 days) at 0.79

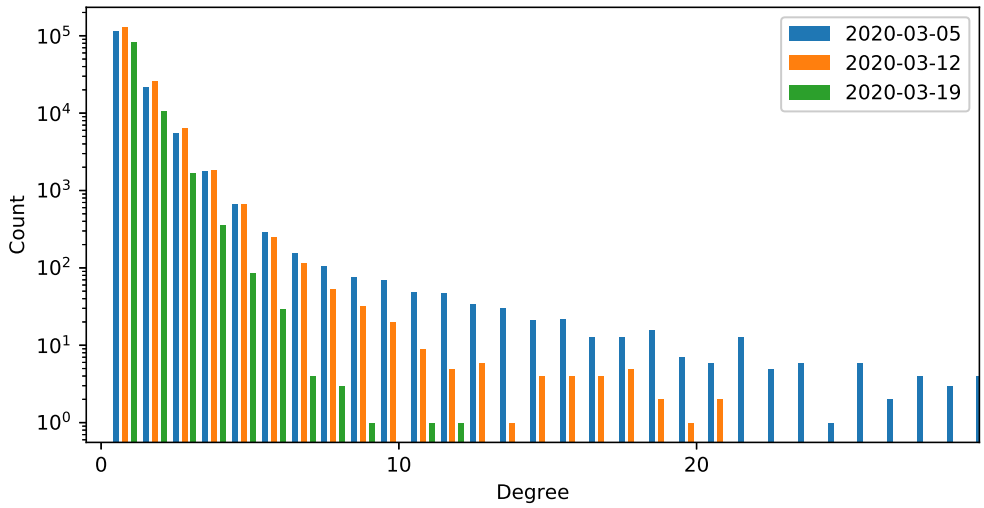

Figure 2: Histogram of number of contacts per person on three different days: March 5, March 12 before official lock-down, March 19 (histogram cut off at 30 contacts) after ban of large gatherings and local regulations closing restaurant and other public venues.

for a shift of seven days, where the evolution of R follows that of the number of contacts in time. The number of contacts is maximized at March 2, 2020 and drops to a fraction of 45% of its maximal value by March 27. However, we note that the R value seems to decrease to levels around 1 before the mean degree hits minimal levels. As well, there is a substantial increase of the number of contacts after the relaxation of the German lock-down in April which is not accompanied by an equally strong increase in R. While the number of contacts reaches 62% of its maximal value by May 1, R stays at values around 1.0. We thus conclude that the number of contacts does not sufficiently model the subsequent evolution of R. For this, more complex methods are needed to investigate further properties of the contact network.

3.2 Association of contact index with infections

Evaluating the network we find that the degree of nodes, i.e., the number of contacts of each person, is broadly distributed with a long tail before the lock-down (Fig. 2) while at later dates during the pandemic, the distribution has a much shorter tail and is highly concentrated around few contacts. Thus initially there are many individuals with large numbers of contacts who would be potential 'super-spreaders', but the lock-down clearly led to a reduction of the number of such individuals at later weeks. In other words, we can ask whether the heterogeneity of the contact network plays a large role in the

infection behavior.

In Fig. 1 the green curve shows the evolution of the contact index in comparison to R. Compared to the number of contacts the contact index stays relatively constant after relaxation of lock-down and thus reflects better the evolution of R as is expected from the theoretical research on epidemics on complex heterogeneous networks.

Estimation of correlation indices suggest a much better predictive power of the contact index for predicting R compared to the simple number of contacts. The maximal (over time shifts between 0 and 14 days) correlation between the contact index and R equals 0.96 (compared to 0.79 for the number of contacts) and is maximal for a time shift of six days. As a six day time shift maximized the correlations with R, we kept using a time shift of six days for further correlation analyses.

3.3 Detailed investigation of evolution after the first wave

Importantly, the contact index also shows a strong association with the outbreak's evolution in the phase after the first wave. This can be seen as correlation in the time series (Fig. 3). The top panel shows a zoom of the time evolution after the first wave, exhibiting a concurrent increase of R and C. The correlation of both quantities can also be assessed from the bottom panel of the figure which plots the R values versus the contact index at a week earlier. The figure also allows to speculate that a contact index of about 50 or more would drive the infection behavior into a sustained super-critical regime with $R > 1$. The data supports this hypothesis and we show that changes in contacts can be a useful signature for outbreaks. The contact index was the best predictor of the subsequent evolution of the epidemic.

Both the correlation between the contact index C and R (for a delay of six days) and as between the mean number of contacts and R estimated to be 0.62 and 0.63, respectively. These similar estimates in the post lock-down phase can be understood from the fact that in this phase the number of super-contacters is much smaller which make the expectation value of $\langle k^2 \rangle$ closer to $\langle k \rangle^2$ so that the contact index becomes similar to $\langle k \rangle$ in the limit of a narrow distribution. This implies that the contact index becomes more important when further measures of the social lockdown are lifted.

4 Discussion

A strong correlation between the effective R and the GPS assessed contact behavior of the German population could be shown. We found the highest correlation between the contact index C, which accounts for a heterogeneous contact behavior of the general population and the importance of superspreading events, and the effective R with a time delay of six or seven days, which is in accordance with the reported SARS-CoV-2 incubation time of 5.2 days (95%CI: 4.1-7.0 days) [11]. C was associated with the effective R in all 3 phases so far observed of evolution of SARS-CoV-2 in Germany: 1) phase of high contact behavior and exponential growth of cases numbers, 2) phase of lowest contact behavior during the lock-down and stark reduction of the viral spreading,

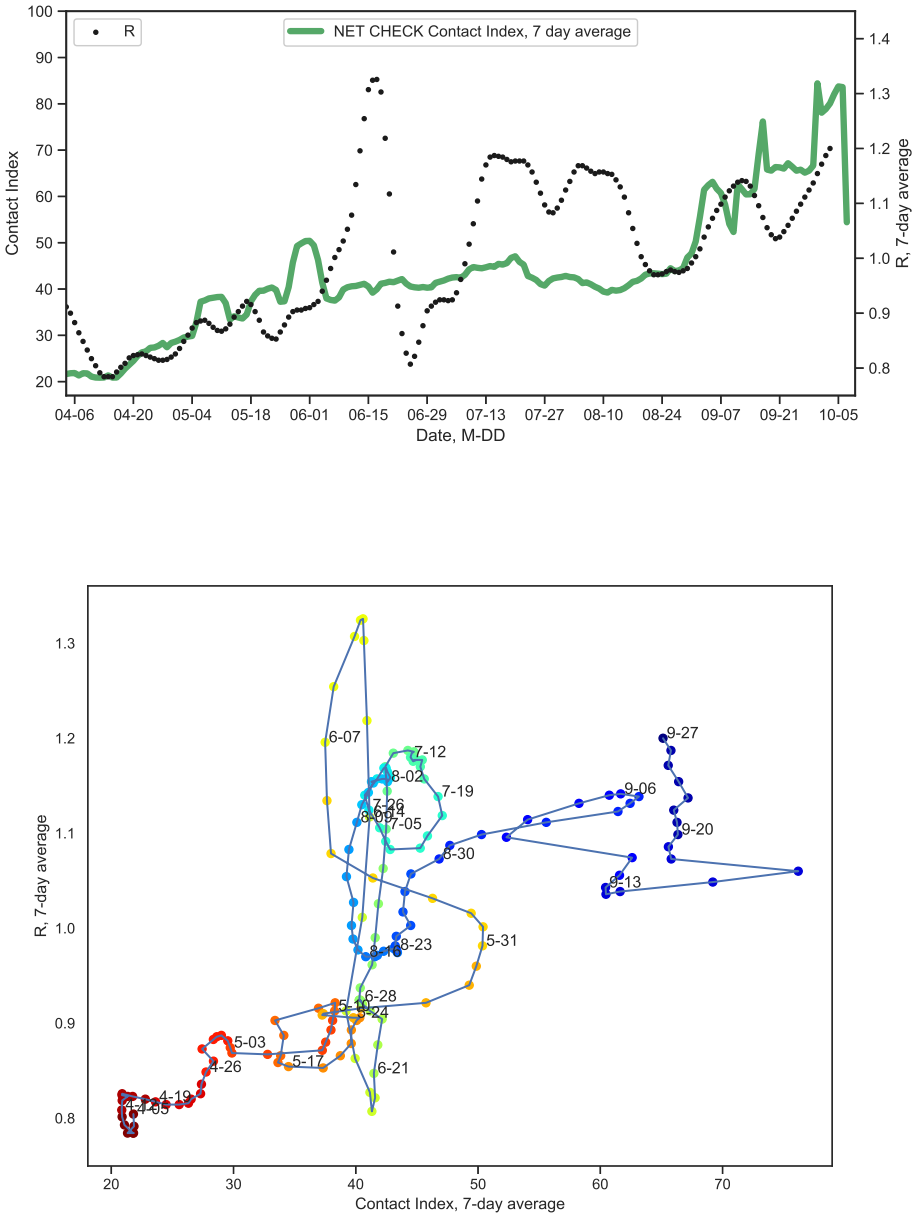

Figure 3: (Top) Evolution post first wave shows a substantial and related increase of R and contact index. (Bottom) R and the contact index clearly exhibit an almost linear relation if plotted with a shift of seven days. Colors denote different weeks as indicated.

3) phase of resume of contacts and slight increases in the effective R. As previously suggested from theoretical work, the average number of contacts had a lower correlation with the effective R than the contact index and would have suggested a higher number of infections especially after release of the lock-down measures. Hence, we propose to use the contact index to assess the effectiveness of social distancing policies and for decision support of social distancing policy-making. Overall, GPS data and various metrics derived from it can provide a means to assess the effect of lock-down measures on contacts and mobility and their association with reductions in infection rates.

There exist other studies which have demonstrated the benefit of rapid deployment of mobile phone applications to receive valuable epidemiological information to combat the pandemic. A model based on mobile phone data on individual mobility accurately predicted the frequency and geographical distribution of SARS-CoV-2 infections in China after the outbreak in Wuhan [9]. Mobility data based on cell tower location is used in many approaches to record mobility of aggregated groups of the general public. For example, Deutsche Telekom has provided daily aggregates of cell tower location data to the RKI [10]. However, mobility data cannot be used to determine social contacts which are a key cause of infection. Several solutions have been developed that determine contacts between users based on Bluetooth low energy (BLE) to inform individuals about contacts to infected individuals. In Germany, a contact tracing BLE app has been released by SAP and Deutsche Telekom on June 15, 2020 [4]. BLE does not provide information about location and mobility and still needs to gather a significant number of users to enable identification and interruption of infection chains [7]. While the technology is promising and the app has already been downloaded 16 million times [13], the results in several countries have so far not met expectations.

Furthermore, there is a growing body of evidence that self-assessed symptoms allows for identification for SARS-CoV-2 hotspots several days before the outbreak [5, 14, 20, 23, 25]. Contact between individuals and development of first symptoms are the very first events which can be observed during the course of the disease. The main advantage the method of GPS based contact tracing presented in this work is that only a relatively small number of users is required to obtain meaningful representative insights when compared to other solutions.

The project's further aim is to develop an early warning system based on the contact behavior and symptom burden of the population, which will be offered as a complementary app. Since the method allows for generation of maps, users do not need to be active users, like with contact tracing apps, to receive valuable information regarding the regional infection risk.

5 Conclusion

A contact index derived from mobile phone location history data was developed which showed a high correlation to the effective R as observed six days later. This novel method provides new insights in the time evolution of the SARS-CoV-2 pandemic in Germany and could be used as a component of an early warning system for country-wide and local

outbreak prediction.

6 Appendix

In this Appendix, we describe our specially designed algorithm that identifies "contacts" from the traces based on the following rationale: If GPS pings arrive from two distinct cell phones that are close in space and time, then we denote this event as a "contact" and use it as a proxy for a human physical contact.

6.1 Sampling of nodes

We now describe more formally how measures of the sampled network, such as mean contacts and second moment of contacts, relate to the respective measures of the original full contact network for all cell phones. We focus on sampling of devices described in restriction (A) in section 2.3, and ignore restriction (B) for simplification, since it is similar to the sampling of nodes in restriction (A) and would just require a re-scaling of the parameter p in equation (14).

In the following, let G denote the full network or graph of all cell phones and let M denote the maximal degree of a node in G. As a reminder, the degree of a node (i.e. person) equals the number of contacts of this person. Following Zhang et al. [26] we let N denote the vector containing the degree counts of the nodes. N has length $M+1$ and the k-th entry of N contains the number of nodes that have degree k, i.e. the number of devices that have k contacts. Thus N contains the counts of the number of cell phones having k links (contacts) to other cell phones.

In the sampling of phones according to (A), we assume that each phone is sampled from G with the same probability p, resulting in the sampled graph G^*. This situation is also described as *induced network sampling* in network theory [26]. The induced network G^* includes all sampled nodes as well as all links from G that connect the sampled nodes in G^*.

The vector of the expected values of the degree counts of the sampled network, N^*, is $E(N^*) = PN$, Here, P is a matrix of entries $P(k,k')$ that describe the probability that a node of degree k' in G is selected and has degree k in G^*. For induced sampling, P is:

$$P_{\text{ind}}(k,k') = \begin{cases} \binom{k'}{k} p^{k+1}(1-p)^{k'-k} & \text{for } 0 \leq k \leq k' \leq M, \\ 0 & \text{for } 0 \leq k' < k \leq M. \end{cases} \tag{1}$$

Thus the k-th entry $E(N^*(k)) = \sum_{k',k\leq k'} N(k')\binom{k'}{k}p^{k+1}(1-p)^{k'-k}$.

In the following we assume that the particular sampling given by our mobile phone records gives rise to a N^*_{ind}, which can be approximated by $E(N^*)$ for large networks, from which we can calculate the degree moments for the original network.

6.2 Derivation of the contact index C

Let $\langle k \rangle$ denote the mean degree of nodes in G: $\langle k \rangle = \sum_{k=0}^{M} kN(k)/(\sum_{k=0}^{M} N(k))$. We first show that the mean $\langle k \rangle_{\text{ind}}$ of the sampled graph is linearly related to the mean of the original graph:

$$\begin{aligned}
\langle k \rangle_{\text{ind}} &\approx \frac{\sum_{k=0}^{M^*} kE(N^*(k))}{\sum_{k=0}^{M^*} E(N^*(k))} && (2)\\
&= \frac{\sum_{k,k',k\leq k'} kN(k')\binom{k'}{k}p^{k+1}(1-p)^{k'-k}}{\sum_{k,k',k\leq k'} N(k')\binom{k'}{k}p^{k+1}(1-p)^{k'-k}} && (3)\\
&= \frac{p\sum_{k'} N(k')\sum_{k,k\leq k'} k\binom{k'}{k}p^{k}(1-p)^{k'-k}}{p\sum_{k'} N(k')\sum_{k,k\leq k'} \binom{k'}{k}p^{k}(1-p)^{k'-k}} && (4)\\
&= \frac{p^2\sum_{k'} k'N(k')}{p\sum_{k'} N(k')} && (5)\\
&= p\langle k \rangle && (6)
\end{aligned}$$

The equality of (4) and (5) follows since $\sum_{k,k\leq k'} k\binom{k'}{k}p^k(1-p)^{k'-k}$ is the mean value of the binomial distribution $B(k',p)$ which equals $k'p$, and $\sum_{k,k\leq k'} \binom{k'}{k}p^k(1-p)^{k'-k}$ is the sum of all probabilities in $B(k',p)$ which is 1. Similarly, we find for the second moment:

$$\begin{aligned}
\langle k^2 \rangle_{\text{ind}} &\approx \frac{\sum_{k} k^2E(N^*(k))}{\sum_{k} E(N^*(k))} && (7)\\
&= \frac{\sum_{k,k',k\leq k'} k^2N(k')\binom{k'}{k}p^{k+1}(1-p)^{k'-k}}{\sum_{k,k',k\leq k'} N(k')\binom{k'}{k}p^{k+1}(1-p)^{k'-k}} && (8)\\
&= \frac{p\sum_{k'} N(k')\sum_{k,k\leq k'} k^2\binom{k'}{k}p^{k}(1-p)^{k'-k}}{p\sum_{k'} N(k')\sum_{k,k\leq k'} \binom{k'}{k}p^{k}(1-p)^{k'-k}} && (9)\\
&= \frac{p\sum_{k'}(k'(k'-1)p^2+k'p)N(k')}{p\sum_{k'} N(k')} && (10)\\
&= p^2\langle k^2 \rangle - (p^2-p)\langle k \rangle && (11)
\end{aligned}$$

Here, (7) is the definition of the second moment, (10) follows from (9) since the second moment for the binomial distribution $B(k',p)$ is $p^2k'^2+k'(p-p^2)$ and $\sum_{k,k\leq k'} \binom{k'}{k}p^k(1-p)^{k'-k} = 1$, and (11) follows from (10) because of the definitions of the first and second moments of $N(k')$. Finally, we describe how the ratio $\langle k^2 \rangle / \langle k \rangle$ of the original graph can

be obtained from the sampled graph via $\langle k\rangle_{\text{ind}}$ and $\langle k^2\rangle_{\text{ind}}$:

$$\frac{\langle k^2\rangle}{\langle k\rangle} \approx \frac{\frac{1}{p^2}\left(\langle k^2\rangle_{\text{ind}} - (p-p^2)\langle k\rangle\right)}{\langle k\rangle} \tag{12}$$

$$= \frac{\langle k^2\rangle_{\text{ind}}}{p\langle k\rangle_{\text{ind}}} - \left(\frac{1}{p} - 1\right) \tag{13}$$

$$= \frac{1}{p}\left(\frac{\langle k^2\rangle_{\text{ind}}}{\langle k\rangle_{\text{ind}}} - 1\right) + 1. \tag{14}$$

This ratio $\langle k^2\rangle/\langle k\rangle$ is of interest, since it describes the growth rate of an infection phase in an uncorrelated network [16]. Since $\langle k^2\rangle_{\text{ind}}$ is larger or equal to $\langle k\rangle_{\text{ind}}$, (14) is non-negative and since p is small in our sampling, we can ignore the addition of the constant 1. Thus we define the contact index C as

$$C := \frac{1}{N_{\text{obs}}}\left(\frac{\langle k^2\rangle_{\text{ind}}}{\langle k\rangle_{\text{ind}}} - 1\right),$$

where $p = N_{\text{obs}}/N_{\text{tot}}$, where N_{obs} is the number of devices observed during a day and N_{tot} is the total number of devices/consumers in the considered area. Since N_{tot} can be considered constant, we drop this factor from the definition of C.

References and Notes

References

[1] Merle M Böhmer et al. "Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series". In: The Lancet Infectious Diseases (2020).

[2] Jasper Fuk-Woo Chan et al. "Improved molecular diagnosis of COVID-19 by the novel, highly sensitive and specific COVID-19-RdRp/Hel real-time reverse transcription-PCR assay validated in vitro and with clinical specimens". In: Journal of Clinical Microbiology 58.5 (2020).

[3] Robert Cohen et al. "Assessment of spread of SARS-CoV-2 by RT-PCR and concomitant serology in children in a region heavily affected by COVID-19 pandemic". In: medRxiv (2020).

[4] Corona-Warn-App [Internet]. Available from: https://www.bundesregierung.de/breg-de/themen/corona-warn-app/unterstuetzt-uns-im-kampf-gegen-corona-1754756. 2020.

[5] David A Drew et al. "Rapid implementation of mobile technology for real-time epidemiology of COVID-19". In: Science (2020).

[6] Giacomo Grasselli et al. "Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy". In: Jama 323.16 (2020), pp. 1574–1581.

[7] Gary F Hatke et al. "Using Bluetooth Low Energy (BLE) signal strength estimation to facilitate contact tracing for COVID-19". In: arXiv preprint arXiv:2006.15711 (2020).

[8] Xi He et al. "Temporal dynamics in viral shedding and transmissibility of COVID-19". In: Nature medicine 26.5 (2020), pp. 672–675.

[9] Moritz UG Kraemer et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China". In: Science 368.6490 (2020), pp. 493–497.

[10] Astrid Krenz and Holger Strulik. The benefits of remoteness: Digital mobility data, regional road infr Tech. rep. cege Discussion Papers, 2020.

[11] Qun Li et al. "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia". In: New England Journal of Medicine (2020).

[12] Ruiyun Li et al. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)". In: Science 368.6490 (2020), pp. 489–493.

[13] Medicus M. Corona-Warn-App: Downloads knacken 16-Millionen-Marke [Internet]. connect. `https://www.connect.de/news/corona-warn-app-download-zahlen-3200860.html`. Accessed: 2010-09-30.

[14] Cristina Menni et al. "Real-time tracking of self-reported symptoms to predict potential COVID-19". In: Nature medicine (2020), pp. 1–4.

[15] Kamalich Muniz-Rodriguez et al. "Doubling time of the COVID-19 epidemic by province, China". In: Emerging infectious diseases 26.8 (2020), p. 1912.

[16] Romualdo Pastor-Satorras et al. "Epidemic processes in complex networks". In: Reviews of modern physics 87.3 (2015), p. 925.

[17] Simone Piva et al. "Clinical presentation and initial management critically ill patients with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in Brescia, Italy". In: Journal of Critical Care (2020).

[18] Jinjun Ran et al. "Quantifying the improvement in confirmation efficiency of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) during the early phase of outbreak in Hong Kong in 2020". In: International Journal of Infectious Diseases (2020).

[19] Lucy Rivett et al. "Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission". In: Elife 9 (2020), e58728.

[20] Hagai Rossman et al. "A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys". In: Nature Medicine 26.5 (2020), pp. 634–638.

[21] Huan Song et al. "A considerable proportion of individuals with asymptomatic SARS-CoV-2 infection in Tibetan population". In: MedRxiv (2020).

[22] Tabelle mit Nowcasting-Zahlen zur R-Schätzung [Internet]. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting_Zahlen.xlsx.

[23] Thomas Timmers et al. "Using eHealth to Support COVID-19 Education, Self-Assessment, and Symptom Monitoring in the Netherlands: Observational Study". In: JMIR mHealth and uHealth 8.6 (2020), e19822.

[24] Joseph T Wu, Kathy Leung, and Gabriel M Leung. "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study". In: The Lancet 395.10225 (2020), pp. 689–697.

[25] Daisuke Yoneoka et al. "Early SNS-based monitoring system for the COVID-19 outbreak in Japan: a population-level observational study". In: Journal of Epidemiology (2020), JE20200150.

[26] Yaonan Zhang, Eric D Kolaczyk, Bruce D Spencer, et al. "Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks". In: The Annals of Applied Statistics 9.1 (2015), pp. 166–199.

[27] Shi Zhao et al. "Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak". In: International journal of infectious diseases 92 (2020), pp. 214–217.

Funding

This project has been initiated within the initiative Mittelstand-digital of the German Federal Ministry of Economic Affairs (BMWi), in the _Gemeinsam digital, the Mittelstand 4.0 Centre of Excellence Berlin. Dr. Thieme is a fellow of the Digital Clinician Scientist program by the Berlin Insitute of Health (BIH).

Ethics and data safety

The effective reproduction number R values in our analysis have been obtained from the RKI Nowcasting website (https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting_Zahlen.xlsx?__blob=publicationFile). They are based on register routine data and are publicly available for secondary analyses. The contact data is collected via a Software Development Kit (SDK) developed for the primary purposes of assessing the quality of cell phone networks. The general terms and conditions of this data collection (see www.netcheck.de/datenschutz) also cover the use for any secondary data analysis through a broad consent.

Competing interests

None.