Attack vectors to re-identify individuals from the anonymised Smittestopp dataset

Hagen Echzell

Independent Researcher
Oslo, Norway
hagen.pub@echzell.org

Abstract. The first version of "Smittestopp", the Norwegian Institute of Public Health's (NIPH) contact tracing application, centrally stored data about the population's contact patterns with reference to a static personal identifier, a decision that has been widely discussed and criticised. After the Norwegian Data Protection Authority had temporarily forbidden further data collection and processing in June 2020, NIPH announced to discontinue the app and stated that all data related to the application would be deleted. Nevertheless, in October 2021, researchers from an institution involved in the development of the app published a paper called "Nationwide rollout reveals efficacy of epidemic control through digital contact tracing" [3]. In their paper, they analysed a derived dataset based on the Smittestopp data that was announced to be deleted. The authors claim that this derived dataset was anonymised and therefore does not include any personal data. We challenge this assumption by explaining how different external sets of personal data can be matched with the dataset, which potentially leads to a re-identification of persons and a disclosure of their private contacts. We conceptually show how some of these methods can be applied on an example case using publicly available information on Erna Solberg, Norway's former prime minister. We conclude that it appears reasonably likely that individuals can be re-identified and that the dataset should not be considered anonymised.

Keywords: Anonymisation · Contact Tracing · GDPR · Privacy · Smittestopp

1 Introduction

After the outbreak of COVID-19, in 2020 the Norwegian Institute of Public Health (NIPH) published a contact tracing app called "Smittestopp", like many other countries did. The privacy-relevant decision to store data about the population's contact patterns in a central facility and with reference to static personal identifiers – phone numbers, in this case – has since been widely discussed and

criticised [10,14,13].
The Norwegian Data Protection Authority declared the way the app works as a disproportionate intrusion into the users' rights. The discourse seems to have come to an end after NIPH announced to discontinue the app and promised a deletion of all obtained data. [6]

In October 2021, however, researchers from Simula Research Laboratory and Simula Metropolitan, institutions that were involved in the development of the app, published a paper called "Nationwide rollout reveals efficacy of epidemic control through digital contact tracing" [3]. In their paper, the authors evaluate the effectiveness of the contract tracing app, based on contact data from the Smittestopp app that they claim to be anonymised. Whether the data really is anonymised plays an important role: If it was possible to re-identify persons from the dataset, this would contradict NIPH's statement that all data has been deleted. Furthermore, there is potential for legal consequences. Should it turn out that the data includes personal information, the General Data Protection Regulation (GDPR) would apply. But most importantly, the continued existence of the dataset may pose a serious risk to former users of the Smittestopp app: Since the app registered real world contacts between persons, a successful re-anonymisation would reveal intimate details about people's private lives, could expose romantic relations, journalistic sources and compromise trust in the institutions that vouched for the trustworthiness and safety of the system.

In this paper, we will first introduce the legal concepts of anonymisation and pseudonymisation given by the GDPR. We will further have a look at the ethical and legal considerations that lead Simula to argue that the data can be considered anonymised and legally be used for research. In the main part of our paper, we will analyse what kind of data the researchers have access to. We will propose attack vectors to match the dataset with external, person-specific data from various sources. Finally, on the example of Norway's former prime minister Erna Solberg and publicly available data about her behaviour, we will demonstrate the use of these attack vectors and conclude whether Solberg could likely be re-identified, if a motivated intruder obtained access to the dataset. We will present another case with Ola Normann, a fictional person with a more common lifestyle.

2 Pseudonymisation vs. anonymisation

The demarcation between pseudonymisation and anonymisation is crucial for deciding whether data protection acts like the GDPR apply.

The GDPR defines *pseudonymisation* as

> the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use

> of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (Art. 4(5) GDPR)

Naturally, the question arises, whether pseudonymised data – i.e. data that can only be related to a natural person using additional information – fulfills the definition of *personal data*, which the GDPR defines as

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (Art. 4(1) GDPR)

If the specific mention of a by reference to an identifier indirectly identifiable natural person does not convince the reader to conclude that question yet, the recitals to the GDPR explicitly clarify the intentions behind the law:

> Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. (Recital 26(2) GDPR)

> The explicit introduction of 'pseudonymisation' in this Regulation is not intended to preclude any other measures of data protection. (Recital 28(2) GDPR)

In other words, even if pseudonymisation might reduce the risk for the affected natural persons, pseudonymised data should be considered as and treated like personal data, according to the GDPR.

2.1 Assessment of the Smittestopp dataset

Elmokashfi et al. claim that the Smittestopp dataset was not just pseudonymised, but in fact anonymised [3, Supplementary Note 2]. This does not only mean that the data holder is not in possession of additional information that could lead to an identification of natural persons, but that natural persons are not identifiable at all; a much stronger statement. Consequently, the law firm Wiersholm confirmed that under this stronger condition, the dataset can be legally used:

> [...] the dataset can be legally used for research purposes if there does not exist additional information that would make it possible to re-identify persons under the assumption that all reasonable means for re-identification is used. ([3, Supplementary Note 10])

This assessment is in line with the recitals to the GDPR:

> To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out [...] (Recital 26(3) GDPR)

For deciding whether a means is "reasonably likely", the recitals to the GDPR recommend to take "all objective factors" into account. These include:

- the costs of an identification
- the amount of time required for identification
- the available technology at the time of the processing,

as well as "technological developments", (Recital 26(4) GDPR) which even requires an anticipatory or dynamic consideration.

With respect to a given dataset, concluding that no natural persons are identifiable is a non-trivial task. To review "all the means reasonably likely to be used" would be time-consuming, if possible at all, particularly as seemingly unrelated information might be combined to re-identify an individual. As the Norwegian Centre for Research Data (NSD) puts it: "It is also possible that a combination of data can be linked to a person. For example, if exact age, place of residence and field of study is collected, and there is only one person who is 57 years of age from Geilo studying theatre science, then this is personal data." [11]

However, should it be possible to find means of re-identifying even a single natural person, it would be much easier to formulate an argument that the dataset in question should be regarded as pseudonymised and not as anonymised – and would therefore be subject to the GDPR or other data protection acts. In the following, we will describe the data in question, as it has been presented by Elmokashfi et al. in [3]. Subsequently, we will provide some potential attack vectors to re-identify individuals from the Smittestopp dataset, some of which are based on existing external datasets that the authors partially obtained.

3 The Data

Let us first introduce the dataset as it is described in [3, Supplementary Note 2]. Originally, the dataset contained information about interpersonal real-life contacts based on Bluetooth low energy (BLE) measurements conducted by

the Smittestopp-app. These BLE measurements detect phones in a radius of 10 meters, and additionally provide an estimated distance based on the signal strength. The authors only possess technically pseudonymised, daily aggregates of these data points. Specifically, this pseudonymised, daily aggregated dataset consists of records $(p_a, p_b, \Delta t, RSSI_{max}, RSSI_{avg}, N, T_a, T_b)$ each specific to a pair of devices, where the variables have the following meaning:

- p_a, p_b: identifiers of the involved devices, where each original device identifier has been replaced by a random, static pseudonymous number
- Δt: the duration between the first and last contact of the given day, i.e. a value between 0h and 24h
- $RSSI_{max}, RSSI_{avg}$: the strongest / average signal strength of measurements between the two devices
- N: the number of registered contacts on the given day
- T_a, T_b: the device type (iOS / Android) of both devices

In their paper, the authors included several graphs that show e.g. the ratio of active users or the average number of contacts for different days (Figures 1c, 1d, 2a, 3a) [3]. Therefore, we assume that these daily aggregated tuples can also be assigned to specific dates. Even though a *date field* is not explicitly mentioned in the description of the records, creating these graphs would not have been possible without being able to assign specific dates to each individual record. When the authors used data that was aggregated while the app was still operational (e.g. in Figure 1b in the main text), they explicitly pointed this out [3, Supplementary Note 1]. The authors' ability to disregard "all devices that were not present on seven different days" [3, Supplementary Note 2] confirms our assumption.

The dataset contains data from 17th of April to the 4th of June 2020. The Smittestopp-app was downloaded at least 1.5 million times, and had up to around 800k daily active users. The authors' research was conducted based on a partial dataset that includes about 26.8 millions contacts between about 545k phones. The authors report a rate of daily active users between 50% and 70%.

Considering the legal status of the dataset, the authors consulted NSD and the Norwegian law firm Wiersholm. Wiersholm concluded "that the dataset can be legally used for research purposes if there does not exist additional information that would make it possible to re-identify persons […]" [3, Supplementary Note 10]. With regard to the question whether such a re-identification might be possible, "Simula leans on the received assessment from NSD that re-identification of individuals from the dataset is hard to imagine." [3].

After we requested the mentioned assessment from NSD (as per the Norwegian Freedom of Information Act (FIA)), we learned that the text that was likely cited here[1] stems from an informal email that NSD sent in extension of a phone

[1] Original paragraph in Norwegian: "Basert på [det som fremgår av dokumentet dere sendte], er det rimelig å anta at den første tilnærmingen vil innebære at det er

conversation with Simula. In this email, NSD outlines two different approaches to formally assess the question whether the data contains personal information. However, upon further request, they could not find any further correspondence regarding such a formal assessment. It should be noted that NSD assumed that all data about who used the app was deleted – an assumption we will challenge in the following section.

We sent a similar FIA request to NIPH and asked whether they assessed the risks of re-identifying individuals before sharing the dataset or allowing its future use. NIPH denied this request, "as [NIPH] do not have any written case documents related to this topic."

We requested a copy of the dataset from NIPH as per §9 FIA. However, NIPH denied our request as they claimed that "All personal data related to the first version of Smittestopp has been delated (sic) [...]". Upon pointing out that the requested dataset provably continues to exist and that Smittestopp's privacy policy explicitly defines NIPH as the responsible data controller, even for research on anonymised data [5], we were directed to their data processor Simula: "Simula's usage of anonymous data for their own purposes, does not fall under the data controller-responsibility of NIPH". Even though Simula offers to provide a copy of the dataset to research institutions [3] – subject to the approval of Simula's own management – we decided to prioritise our research based on the available description of the dataset, at this point.

4 Attack vectors

Now that we have an impression of what kind of data still is available after the claimed deletion, we will explore possible ways to extract patterns and information from these data, among others by combining with external datasets. Finally, we will demonstrate how these attack vectors may be combined to potentially re-identify individuals from the dataset. Listing 1.1 shows an example data point we created for illustrative purposes. We refer to the unidentified data subjects represented by the pseudonymous numbers p_a or p_b as *entities*. *Re-identification* in this context means mapping an entity from the dataset to a directly or indirectly identifiable person. If we fall short of a full re-identification, we might still be able to specify a subset of *candidate entities* that we expect to include

snakk om personopplysninger. [...] Dersom man velger den andre og mer risikobaserte tilnærmingen, er det mer sannsynlig at man kan argumentere for at datasettet er anonymt. Sannsynligheten for å identifisere noen, fremstår som lav. Dataene som er registrert inngår sannsynligvis ikke i andre registre, og er heller ikke offentlig tilgjengelige. Det er dermed vanskelig å se for seg hvordan en motivert inntrenger skulle gå frem for å klare å identifisere noen. Det kan imidlertid være nyttig å presisere her at det er relevant å ta med i betraktningen teknologisk utvikling, det vil si om det i fremtiden kan være mulig å identifisere noen med ny teknologi."

```
{
        "p_a": 484263,
        "p_b": 116099,
        "delta_t": 85500,
        "rssi_max": 97,
        "rssi_avg": 50,
        "N": 34,
        "T_a": Android,
        "T_b": Android,
        "date": 2020-05-17
}
```

Listing 1.1: An example record we created for illustrative purposes. We specified Δt, the difference between the first and last contact on a given day, in seconds and RSSI on a scale from 1 to 100. Since Δt, corresponds to 23¾h, we can assume that the two individuals associated with p_a and p_b spent two consecutive nights close to each other.

the entity corresponding to a person. Hence, re-identification is equivalent to reducing the candidate entity set to a size of 1.

4.1 Working conditions and schedule

As Elmokashfi et al. explain, it is possible to classify what they call "known contacts", which might be family members or co-workers. As "known contacts", they selected "device pairs that met on at least seven different days". They state that "Close contacts were longer and can last a whole day (i.e., household contacts) or several hours with a peak around 7 hours (i.e., work contacts)". The authors even suggest that it is possible to extract information about some entities' working conditions: "a non-trivial fraction of highly connected users have an occupation that exposes them to excessive close contacts, e.g., health personnel, shop keepers, rail conductors" [3].

We go a step further and claim that some of these individuals' living and working conditions create highly individual patterns that will be visible in the dataset. Since we can classify work-related contacts, we can also estimate how many co-workers an entity has, whether they work in changing teams, etc. Even further, we can figure out on which days and how long a certain entity worked over the time-period of the dataset. People who have a lot of sick days or who work under flexible conditions – e.g. those that work on weekends, have highly individual schedules or multiple jobs – are under risk to be identified. Such a re-identification may happen by comparing entities' working characteristics with personal data from electronic time tracking systems.
Specifically the before-mentioned groups of health personnel, shop keepers and

rail conductors are at risk with this approach; not only because they were not able to work from home, like many other employees in Norway could, but also since they might work night shifts that leave an even more characteristic trace in the dataset: An entity that worked 8 hours on Monday, 3:25 hours on Tuesday, 4:35 hours on Wednesday, and 8 hours on Thursday likely showed up for night shift on Tuesday, precisely at 20:35.
The same logic holds for students. Periods of presence-teaching in schools and universities are publicly available knowledge, which helps classifying entities as students. With access to students' individual class schedules and presence registers, we could gain highly individual profiles that can be compared to entities' contact patterns.

4.2 Household conditions, events and travel

Parallel to the working conditions, the data also holds potentially revealing information about living conditions. We already figured out that household contacts can be distinguished from other ones. Hence, the information from the dataset allow us to estimate how many flatmates an entity has. We even expect finer-grained living conditions to be visible in the data: Assuming that all household-members actively used the app, a couple with a child, where two people sleep in the same room, with a third person out of BLE reach, is expected to generate substantially different contact patterns than a student collective with three inhabitants. We could identify family or couple travels, where entities from a common household continue their mutual contact pattern, but all external contacts, like work contacts, stay out for a period of time. A lot of this information can be compared to data from the Norwegian National Population Register [18], records of booked plane and train tickets and again to electronic time tracking records.

4.3 Phone type and app download date

Initially, the Smittestopp-app was exclusively downloadable in Apple's App Store and Google's Play Store [7]. Therefore, all users of the Smittestopp app had to be signed up for an account with either Google or Apple. Downloading the app from their app stores left digital traces: Making use of our GDPR-given right for a copy of our personal data, we requested such a copy from both Google's [9] and Apple's [2] data export portals and found a detailed history of our own app downloads (see Listing 1.2). Note that these records can be matched with the email-addresses (which are considered personal data) that were used to create the respective accounts. Of course, the Smittestopp app only started generating contacts after the moment it was installed on the device. Hence, when we know an individual downloaded the app on a certain day, we expect the first contact of the corresponding entity to appear shortly after, but never before that time.

```
"libraryDoc": {
  "doc": {
    "documentType":
    "Android Apps",
    "title": "Ruter"
  },
  "acquisitionTime":
    2018-03-21T17:34:44.
        581Z
}
```

(a) Google Play Store

```
{
  "Apple ID Number": ****,
  "Item Purchased Date":
    2022-05-10T07:45:57.082Z,
  "Item Description":
  "Ruter - Mobility in Oslo/Viken",
  "Seller": "Ruter AS",
  "Device Identifier": ****,
  "Device IP Address": ****
}
```

(b) Apple App Store,
some data edited for privacy reasons

Listing 1.2: Both Google and Apple precisely record when a particular user downloaded an app. These data were requested using Google's and Apple's data export portals [9,2]. The examples show when the requesting user downloaded the public transport app "Ruter". In another file, Apple even recorded the app update history.

Furthermore, we know from the dataset whether entities' phones are of the type "Android" or "iOS", which corresponds to whether the app was downloaded from the Google Play Store, or Apple's App Store. Given that there were particularly few Android users in the last days spanned by the dataset [3, Supplementary Figure 1], the number of candidate entities in the dataset is reduced to about 55 000 (estimated by the referenced Figure), if an Android user of the app is expected to have had a registered contact on e.g. 4th of June.

A special role comes to an update of the app's iOS version rolled out in early May [3, Supplementary Note 1.1] that according to the authors improved on the detection rate of iOS devices. The included Figure [3, Supplementary Figure 2] indeed suggests an approximately doubled daily number of discovered devices from about 2 to 4 . One of the files we received from Apple's data export portal actually provided us with a detailed log of when a specific update of an app had been installed on an iOS device. This information could be matched with the characteristic jump in the detection rate we expect to see for each entity of the type iOS in the dataset, given they installed this update. Again, people exposed to excessive contacts are at particular risk, since such a jump will be determined most precisely for entities that are represented with lot of data points.

4.4 Isolation periods

The most visible pattern in an entity's contact data will likely be the time after a confirmed infection with the SARS-CoV-2 virus. In Norway, after a positive test,

people were ordered to isolate for 14 days – later reduced to 10 days [12]. Again, we looked for external datasets that could be used in this context and made a find after logging in to the the patient portal of "Fürst Medisinisk Laboratorium", a medical laboratory, as well as by requesting a record of personal data from "Dr. Dropin AS", a company providing physician services. Both of them provided a list of the requesting person's SARS-CoV-2 test results going back until 2020, some of them including whether the patient reported symptoms of the virus before the test, had a close contact, or the location of the test centre, such as "Arrival Oslo Airport".

If people followed the ordered isolation period, we expect to have clearly identifiable periods of little to no contacts in the dataset. If we have access to the date a person tested positive, we will expect such a period of at least 10 or 14 days – starting at latest on the day of the positive test and ending exactly 10 or 14 days after the test – for the corresponding entity.

According to NIPH [8], on 17th of May 2020 – Norwegian Constitution Day –, as few as 4 people tested positive in total. Of course, there might be a higher number of entities in the dataset whose period of very little contact ends exactly 10 days after that. This could be due to travel abroad, or the temporary abandonment of the Smittestopp app. The question of how many of these characteristic periods exist in the dataset can only be answered with access to the data.

5 Combining attack vectors

We next give some examples of how these attack vectors can be combined. To demonstrate this, we collected publicly available data about Erna Solberg, Norway's former prime minister. In an example case, we will show how to use this data to systematically reduce the set of candidate entities that could represent her phone in the dataset, based on the attack vectors described in the previous section. On the example of the fictional construction worker Ola Normann, we will show that even people with a more common lifestyle are at risk to be re-identified, in our case by knowledge of sick days, installation date and phone type.

5.1 Case: Erna Solberg

We will start with the phone type and the app download date. In the night between 16th and 17th of April 2020, Solberg uploaded a picture to her verified Facebook profile, showing her and minister Bent Høie [15]. Each of them present what looks like an iPhone with the active Smittestopp app. The phones are held within an approximate distance of 30cm. The photo was most likely taken on a press conference on 16th of April, the first day of Smittestopp's availability.

Thereby, we know that Solberg uses an iPhone and – assuming that the app worked correctly – the app registered Høie's phone (and vice versa). Her phone therefore generated Solberg's first contact on exactly the 16th of April. We can therefore reduce the set of candidate entities to about 360k, the number of active iPhone users on the first day of the app's availability [3, Supplementary Figure 1]. It would be particularly interesting to know whether the dataset holds any contacts before the first date of official availability. If such contacts exist, but were omitted in the report, we would assume to have a very small cohort of people having access to a development version of the app. We would expect Solberg to be one of these. However, without access to the dataset, this possibility remains speculative.
Further than only using Solberg's device type and download date, we can also use Høie's phone type and some characteristics of that contact. By knowing that there must be a contact between two devices of the type "iOS", we can remove all iOS entities from the candidate set that did not have contact to another iOS entity on 16th of April. This contact will likely also have one of the highest $RSSI_{max}$ values observable in the dataset: Two devices, held 30cm apart with no visible obstruction, while the app is active, will yield an almost perfect signal strength. In times of social distancing, such a close device contact between people from different households would be unusual. We can further look for regularities: Since Solberg and Høie had regular common public appearances in the following weeks, the device pair belonging to the two of them needs to have at least one contact on each of these days.
Note that this process can be repeated with any public figure that is known to use the Smittestopp app – for example by showing it to the media – and has been seen together with Solberg at least once, in the given period.

We are not aware of Solberg having spent any time in isolation. Otherwise, as explained, we could have used this information to further narrow down the set of candidates – Solberg would certainly have respected the rules and, compared to her otherwise busy life, would have heavily cut down on her physical contacts during the ordered isolation period.
However, there are other characteristic events that should have left traces in the dataset: On the occasion of Nurses' Day, 12th of May 2020, Solberg visited an Oslo hospital – videos of her dancing in closer contact to health workers are available on TikTok [17]. Elmokashfi et al. noted that they can classify "highly connected users" that include "health personnel" [3]. Hence, on this specific day, the entity representing Solberg's phone will have registered contacts to a cohort of such highly connected users. If she is not a regular visitor of that specific hospital, these registered contacts between Solberg and entities from that specific cohort will remain a one-off encounter.
In the same way, Solberg attending the reopening of a school [1] will have left a similarly characteristic pattern in the dataset.

Without access to the dataset, we do not know how many of the about 360k first-day adopters of Smittestopp (iOS) created a profile in the dataset that could

look similar to Solberg's footprint we just described. However, we believe it is likely that this footprint is unique. If in doubt, there are plenty of additional resources we can use to further narrow down the set of candidate entities that might be Erna Solberg. Based on media reports, we would know if and when Solberg went travelling to other cities or even abroad. We should be able to find out how many people live in Solberg's household, and what type of phone they use. For close contacts of hers, we can employ a similar investigation.

If by such an approach, we could narrow down e.g. both Bent Høie and Erna Solberg to sets B and E of each 100 candidate entities (both a subset of the 360k first-day iOS users, denoted F), and assume 10 average daily contacts per person (which is an overestimation, as [3, Supplementary Figure 2] shows), we would likely be able to re-identify both of them: On 16th of April, there are $100^2 = 10\,000$ possible unique contacts between members from B and E. For the given day and a random pair of entities from F, the probability of a contact between them is $10/(360\,000 - 1) \approx 1/36\,000$. Thus, we expect $10k/36k \approx 0.28 < 1$ pairs of entities from B and E to have met by coincidence, on 16th of April. Hence, the actual Høie and Solberg, who are known to have met on that day, would likely be re-identified because of their registered contact.

Of course, this simplified calculation is not entirely realistic, as it assumes uniform, independent distributions of contacts. However, what we illustrate here is that by combining information about people we know to have a certain relation, the number of potential candidates can be dramatically reduced. We can iterate this technique with more people.

5.2 Case: Ola Nordmann

Due to her very particular, publicly documented behaviour, Erna Solberg serves as a well-suited example to demonstrate many vectors for re-identification at once. However, it is natural to ask whether people with a more ordinary lifestyle are at risk of being re-identified, as well. We will show this with Ola Normann, a fictional construction worker. Ola works on regular work days and in a team, with the same colleagues every day. On the first day of Smittestopp's availability, the team collectively decides to install the app. Ola's employer tries to find out whether his workers joined a union gathering. If he obtained access to the dataset and managed to re-identify Ola, he could check whether Ola had unusually many contacts on the day of the gathering, matching the expected duration of such a meeting. His boss knows that Ola uses an Android phone and installed the app on 16th of April, just like his colleagues. Thereby he can reduce the number of candidate entities to 112k, the number of first-day Android users (as estimated by [3, Supplementary Figure 1]). Going through the time tracking system, his boss also finds that in the 7 weeks spanned by the dataset, Ola had 3 isolated sick days, due to occasional migraine. As we discussed earlier, it is visible in the dataset whether an entity showed up at work or not, on a given day. Will

Ola's pattern of work contacts be unique in the dataset? As we do not have data about the frequency of isolated sick days, we will use the rate of self-certified sickness absence in males in the 2nd quarter of 2020 as an upper bound, which is 0.5% [16] – assuming that single sick days would usually be self-certified. An estimate upper bound for the number of remaining candidate entities that reveal the same 3 days of absence would be $112\,000 * (0.5/100)^3 = 0.014 < 1$. It is therefore likely that Ola can be re-identified by his employer.

As in the previous case, these calculations only serve illustration purposes. For example, there might be other reasons than sickness to not show up at work. On the other hand, however, we can further reduce the number of candidate entities by removing the fraction of people who worked from home, by adding more linkable information, like Ola's household structure, or just by increasing the number of sick-days in our example.

6 Concluding the claims on anonymisation

We introduced several attack vectors to match data from the Smittestopp dataset with data available from external sources. Among others, these external datasets include electronic time tracking systems, COVID-19 test data, the Norwegian National Population Register, news articles and installation logs from Google's and Apple's app stores. For each of the datasets we mentioned, the data holders can easily reference identifiable persons. In many cases, household members, employers or government entities will be able to access these external datasets.

We therefore do not share NSD's assessment "that re-identification of individuals from the dataset is hard to imagine" [3]. Among others, this assessment was made on the false assumption that all data about who used the app was deleted. We furthermore question whether such an informal and subjective statement should be a legal base to lean on, given the considerable damage a re-identification of individuals would lead to.

The second consideration mentioned by Elmokashfi et al. was one given by the law Firm Wiersholm [3, Supplementary Note 10]:

> [...] the dataset can be legally used for research purposes if there does not exist additional information that would make it possible to re-identify persons under the assumption that all reasonable means for re-identification is used.

Based on the article [3] as well as our correspondence with NIPH and NSD, we could not find that either NIPH or Simula have thoroughly and formally investigated the question whether such additional information and reasonable means exist. Our paper can therefore be understood as a supplement to Wiersholm's statement. Of course, without access to the dataset, we can not prove

that any entity can be re-identified. However, we demonstrated a case where only by using publicly available data about Erna Solberg, the amount of candidate entities could be dramatically reduced from the original size of 1.5 million users. Additionally, we demonstrated that the risk for re-identification spreads among social contacts. Using Høie's and Solberg's common public appearances, we demonstrated how their respective sets of candidate entities can be further reduced by orders of magnitude; a process that can be iterated with as many other contacts as we know about. On the case of a fictional construction worker with 3 sick days we made an even stronger point: even rather common behaviour can likely generate unique patterns. It therefore appears reasonably likely that entities from the anonymised Smittestopp dataset can be re-identified and that the dataset should not be considered anonymised. As a consequence, we argue that the dataset contains personal data and the GDPR applies. It is however up to the Norwegian Data Protection Authority to conclude on our findings from a legal perspective.

None of the approaches we described is particularly time-consuming, expensive or technically sophisticated. Should the Smittestopp dataset be disclosed – that could be through a technical fault, a data request by false researchers or perhaps even a FIA request – the described attacks could be executed by people with a bit of technical knowledge. It is therefore easy to construct a capable, motivated attacker with access to some of the external datasets: That could be a jealous partner, a nosy employer or someone who happens to take control over someone's Google or Apple account.

6.1 Further research

Our collection of attack vectors is by no means complete. According to a privacy auditing platform, the app binaries included third-party tracking tools like "Microsoft Visual Studio App Center Analytics" [4]. It would be interesting to analyse whether Microsoft has collected data that could be used for re-identification. If we obtained access to the dataset, we could further check whether the order of records has been shuffled after the claimed anonymisation or whether their order reveals information about the original device identifiers or temporal relations. Lastly, but more sophisticated, it would be very interesting to try to exploit the graph structure of the data. This could happen by comparing the contact graph derived from the dataset to for example the social graph of Facebook.

Acknowledgements

A great deal of acknowledgement goes to Hans Heum, whose paper [10] first sparked my interest in the topic and who helpfully answered many of my ques-

tions. I would also like to thank the organizers of the MNSES9100 course at UiO, in the context of which a first draft of this paper was written.

References

1. Statsminister Erna Solberg besøker Apalløkka skole, https://tv.vg.no/video/196813/se-statsminister-erna-solberg-besoeker-apalloekka-skole, accessed: 2022-08-31
2. Apple Inc.: Data and privacy, https://privacy.apple.com/account, accessed: 2022-06-22
3. Elmokashfi, A., Sundnes, J., Kvalbein, A., Naumova, V., Reinemo, S.A., Florvaag, P.M., Stensland, H.K., Lysne, O.: Nationwide rollout reveals efficacy of epidemic control through digital contact tracing **12**(1), 5918 (2021). https://doi.org/10.1038/s41467-021-26144-8, https://www.nature.com/articles/s41467-021-26144-8
4. Exodus Privacy: Smittestopp, https://reports.exodus-privacy.eu.org/en/reports/191581/, accessed: 2022-06-21
5. Folkehelseinstituttet (FHI): Bruk av smittestopp og personvern. Internet Archive: https://web.archive.org/web/20200602134246/https://www.fhi.no/sv/smittsomme-sykdommer/corona/bruk-av-smittestopp/, https://www.fhi.no/sv/smittsomme-sykdommer/corona/bruk-av-smittestopp/, accessed by the Internet Archive: 2022-06-02
6. Folkehelseinstituttet (FHI): FHI stopper all innsamling av data i smittestopp (arkivert), https://www.fhi.no/historisk-arkiv/covid-19/nyheter-2020/jun/fhi-stopper-all-innsamling-av-data-i-smittestopp/, accessed: 2022-08-31
7. Folkehelseinstituttet (FHI): Sammen kan vi knekke korona – last ned smittestopp. Internet Archive: https://web.archive.org/web/20200419161647/https://www.helsenorge.no/smittestopp, https://www.helsenorge.no/smittestopp, accessed by the Internet Archive: 2020-04-19
8. Folkehelseinstituttet (FHI): Statistics about coronavirus and COVID-19, https://www.fhi.no/en/id/infectious-diseases/coronavirus/daily-reports/daily-reports-COVID19/, accessed: 2022-05-30
9. Google: Google takeout, https://takeout.google.com/, accessed: 2022-06-22
10. Heum, H.: Stupid, evil, or both? Understanding the smittestopp conflict. NISK Norsk informasjonssikkerhetskonferanse (3) (2021), https://ojs.bibsys.no/index.php/NIK/article/view/964
11. NSD: Notification form for personal data, https://nsd.no/en/data-protection-services/notification-form-for-personal-data, accessed: 2022-06-13
12. NTB nyheter: FHI: Et halvt års karantenefritak om du har vært smittet, https://www.dagsavisen.no/nyheter/innenriks/2020/05/08/fhi-et-halvt-ars-karantenefritak-om-du-har-vaert-smittet/, accessed: 2022-08-31
13. O'Caroll, T., Egenæs, J.P.: Concerns regarding the government of norway's smittestopp app (Jun 2020), https://www.digi.no/filer/Amnestys_brev_til_regjeringen_om_Smittestopp.pdf
14. Sandvik, K.B.: Smittestopp: If you want your freedom back, download now. Big Data & Society **7**(2) (7 2020). https://doi.org/10.1177/2053951720939985

15. Solberg, E.: Timeline photos, https://www.facebook.com/ernasolberg/photos/a.394689651831/10158047528326832/, accessed: 2022-10-31
16. Statistics Norway: 12439: Sickness absence for employees (per cent), by type of sickness absence, contents, quarter and sex, https://www.ssb.no/en/statbank/sq/10074236, accessed: 2022-10-31
17. Støre, M.: Erna Solberg danser med sykepleiere på tiktok, https://www.vg.no/i/1nLpbe, accessed: 2022-08-31
18. The Norwegian Tax Administration: This is the national registry, https://www.skatteetaten.no/en/person/national-registry/about/this-is-the-national-registry/, accessed: 2022-08-31