2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 29 Nov. – 2 Dec. 2021, Lloret de Mar, Spain

Contact Tracing with the Exposure Notification Framework in the German Corona-Warn-App

Steffen Meyer*, Thomas Windisch*, Adrian Perl*, Daniel Dzibela*, Robert Marzilger*, Nicolas Witt*, Justus Benzler†, Göran Kirchner†, Tobias Feigl*, and Christopher Mutschler*

*Fraunhofer Institute for Integrated Circuits IIS
Precise Positioning and Analytics Department
{firstname.lastname}@iis.fraunhofer.de

†Robert-Koch-Institut (RKI)
Berlin, Germany
{BenzlerJ | KirchnerG}@rki.de

***Abstract*—Digital Contact Tracing (CT) protocols based on Bluetooth are best implemented at the system level to save resources and preserve security aspects. Combined with a government-monitored software platform, these CT-protocols can then be used to support controlling pandemics such as COVID-19. However, it is unclear how these protocols have to be parameterized to ensure the most accurate and reliable CT.**

This paper describes how we derived optimal parameters for a decentralized CT from extensive measurement campaigns that we carried out together with Deutsche Telekom (DT) and SAP under the supervision of the Robert Koch Institut (RKI). We examined the Google/Apple Exposure Notification Framework (ENF), which in combination with the front-end, i.e., the German Corona-Warn-App (CWA), enables digital CT in Germany. With centimeter accurate optical reference systems we show that optimal parameters are application-specific. However, they cause impractical high resource costs. In contrast, optimized general parameters offer an everyday compromise between energy costs, applicability, accuracy, and reliability of the ENF.

***Index Terms*—Digital Decentralized Contact Tracing, COVID-19, SARS-CoV-2, German Corona-Warn-App, CWA, ENF.**

I. Introduction

Contact tracing (CT) followed by quarantining identified exposed people in infected index cases is one of the most important activities to hold back the spread of communicable diseases that are transmitted from person to person, such as COVID-19. While the manual tracing of contacts is important and inevitable it is extremely resource-intensive and time-consuming [1]–[3]. Especially with high incidence values, manual CT becomes impractical when used in isolation [4].

The probability of infection usually increases with the pathogen dose, which in turn increases with the excretion rate of the contagious individual and the duration of the interaction, while it decreases with distance. Thus, the duration of a contact with a contagious person, and the distance between those involved, are obviously measures of infection risk [5]. The automatic exposure notification (EN) of conventional CT improves notification, privacy, scalability [6], [7], and handling of potential chains of infection [8], [9]. The main challenge is the error-free detection and tracking of the contacts of an infected person during their (unnoticed, but contagious) initial phase of the infection in everyday life [8], [10].

Epidemiological models indicate an advantage of such smartphone-based automatic EN technologies to detect and warn COVID-19 infected people. Google and Apple joined forces and provided a consistent privacy-preserving API with which applications notify exposure (Google/Apple Exposure Notification - GAEN) based on the attenuation of a Bluetooth signal emitted by an infected person's device, the duration of its reception, and its timing in relation to the occurrence of symptoms and the date of laboratory detection of SARS-CoV-2. Software applications based on the GAEN protocol, such as Immuni (Italy) [11] and SwissCovid (Switzerland) [9] use two main decision criteria for notifying users after SARS-CoV-2 exposure and recommend appropriate need for action: (cumulative) contact duration of 10 min within a certain radius such as 1 to 2 m [12]. This mirrors criteria and rules as they are also applied to conventional CT, based on epidemiological evidence and assumptions [13], [14].

However, the challenges are two-fold. First, the attenuation of Bluetooth signals is not a reliable measure of the distance without a well-tuned parameterization [15]. While fluctuations in the power level of BLE devices can be calibrated out, objects between transmitters and receivers (such as furniture, walls, clothing, etc.), antenna patterns, and 2.4 GHz radio interference attenuate and influence the signals non-deterministically [16], [17]. Second, the risk of infection is not a binary function of distance, duration or timing [18]. Some studies [19] even suggest that the GAEN protocol does not reliably identify and trace contacts. Hence, the main goal of the GAEN protocols is to reduce overestimated distances (i.e., false negatives) and thus minimize undetected contacts [20].

This paper describes how we determined the parameters of the German Corona-Warn-App (CWA) with respect to the GAEN protocol. In a cooperation with Deutsche Telekom, SAP, Google/Apple, and under the supervision of the Robert-Koch-Institut (RKI) and the federal government of Germany, we proposed optimized parameters and investigated their effects on CT accuracy in typical everyday situations.

In our experiments we simulated a range of everyday situations, including (semi-)controlled public transport, e.g., bus, airplane, and subway, as well as typical everyday scenarios with (n=283) individual (synthetic) test persons to cover differ-

ent motion dynamics and propagation environments and to derive parameters and their effects on the expected CT accuracy in selected scenarios. Our analysis shows that optimal parameter sets are application- and situation-specific. However, a prior classification of the respective situation and environment for the optimal selection of certain parameter sets is impractical even for modern phones and would limit their usability. Hence, a general parameter set is proposed which provides almost identical accuracy, is resource-saving, and can be transferred to typical everyday situations. In contrast to similar studies [21], our optimized GAEN parameterization identifies contacts even in an unfavorable scenario (F1-score=53%, F2-score=58%) and yields almost few false negatives.

The rest of the paper is structured as follows. Sec. II discusses related work. Sec. III presents the CWA system and architecture, and our method. Sec. IV describes experimental setups and Sec. V evaluates the results. Sec. VI concludes.

II. Related Work

A total of 49 nations use CT: 2 with unknown software; 8 out of 47 use sensitive absolute positions, e.g., GPS; 18 of 47 use GAEN; the rest uses proprietary developments; 30 only use BLE; 17 combine Bluetooth with other information, mostly GPS or QR code [22]. Thus, the EN System (ENS) developed by Google/Apple has emerged as a *de facto* standard for digital CT. The keys to the success of the ENS are integrated privacy and expected performance. So far, only a few publicly available studies have examined these aspects and shown that ENS, like any other BLE-based range estimation system, performs unpredictably due to wireless propagation effects [20], [23], [24]. They have also shown that all ENS configurations tend to miss real contacts. In response, many countries are continuously monitoring their configurations.

Leith et al. [19], [20], [23] report results of a COVID-19 CT measurement study with a Google Pixel 2 and GAEN (<v1.5) carried out on a light-rail tram (7 participants) [19] and a bus (60 pairs of phone locations from 5 participants) [23]. In both studies they found only little correlation between the RSS and the phone distances. A follow-up study [19], [23] used the same phones for the bus [23] and tram [19] scenarios to reduce the variability in the data and examined the effects of the environment and the way people hold their phones. While it is generally known that variations between different chipsets, antennas and housings cause a high variability in signal attenuation, they identified the following key factors [20] that influence the measurements: (i) differences between Bluetooth chipsets; (ii) (small) changes in the relative orientation of Bluetooth sensors; (iii) radio environments with an obstructed or NLoS connection; (iv) signal reflection from walls, floors and objects; and (v) channel hopping. However, from their studies it remains unclear how parameter calibration mitigates (i), and it is unclear how (ii) - (v) can be compensated at all. Thus, we pre-examined general calibration parameters to compensate for differences between phone models.

Simula et al. [24] conducted (semi-)controlled experiments on a selection of representative phones to test the performance of two (custom Smittestopp and GAEN) CT systems. Their GAEN-based (v1.5) CT used the same parameters (i.e., similar levels of attenuation and definition of close contact) as the applications in Germany and Great Britain [25]. They found that GAEN almost always recognizes contacts on iOS (max. error: 1 out of 25), while it fails more often on Android. Instead, Smittestopp always yields undiscovered contacts both under iOS (max. error: 6 out of 30) and Android (max. error: 30 out of 30). In the semi-controlled experiments they simulated 8 everyday situations (fitness center; two nearby restaurants; restaurant; bar; shopping center; bus; café; office) and found that GAEN is more accurate (85% high risk detection) than Smittestopp (80% high risk detection) and yields a higher accuracy (recall: 84.6, precision: 78.6, and accuracy: 84.8, true positive (TP) = 84.6% and false negative (FN) = 15%) than Smittestopp (recall: 84.6, precision: 73.3, and accuracy: 81.8). However, their study design does not show the impact of each scenario on the final results. In addition, they do not report whether the phones are calibrated and how they have been placed and how accurate their reference metrics or systems are. In contrast, we designed a similar study concept, but report important facts and results.

We agree with previous studies and think that the GAEN with optimized and calibrated parameters for everyday situations presents a valuable solution for digital CT. We suggest to use the number of undetected critical contacts (FN) as a common metric for the assessment of CT systems. In contrast to previous studies of Leith et al. [19], [20], [23] our experiments prove that sensitive, well-thought-out parameters enable to detect almost all important infectious contacts at a moderate oversensitivity. Although available studies suggest that (Bluetooth) RSS-based CT is possible, only few practical GAEN-based studies are available today and (to the best of our knowledge) non of them uses an exact reference system which leaves the functional limits of CT applications based on RSS still unknown.

III. System Architecture

A. The German Corona-Warn-App (CWA)

Fig. 1 shows the (simplified) processing pipeline of the German GAEN-based Corona-Warn-App. Applications based on GAEN regularly search for exposure information from users who have been positively tested for COVID-19. The current implementation of GAEN receives beacons (advertisement messages) from neighboring devices every 120 to 300 s and announces beacons every 250 ms. A received beacon can be interpreted as an indication of proximity and its attenuation level indicates the likelihood that it is within a certain distance to the device sending the beacon. From these messages a receiver estimates the duration and distance of a contact.

BLE devices can be configured to transmit beacons at regular intervals. To distinguish between beacons each device running GAEN generates a random Temporary Exposure Key (TEK) once a day. Based on the TEK, Rolling Proximity Identifiers (RPIs) are generated and updated approximately every 10 min (hence, around $24 \cdot 60/10$=144 RPIs

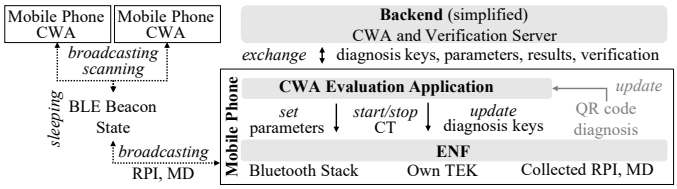


Fig. 1. Pipeline of the German Corona-Warn-App (CWA).

are generated per day) to improve privacy. The transmitted beacons carry the current RPI as well as encrypted metadata (MD) that contains the wireless transmit power level.

In case of an infection the personal TEKs are uploaded to a central server. Other CWA applications on other devices can then download these TEKs and use them to identify relevant RPIs that were received via beacons and are stored on the device. If there is a match, the values reported by GAEN can be used to determine the attenuation time to estimate the risk of infection and to trigger an EN if this risk is sufficiently high. A typical requirement of an epidemiological risk model (regulated by the RKI) is that a person is, e.g., from 10 min within a radius of 1.5 m to an infected person. The assignment of the GAEN attenuation duration to the EN is therefore largely based on the attenuation level as an indicator of the proximity between persons.

The CWA is open source[1] and is based on the GAEN system. CWA reviews exposures from the latest 14 days several times a day. This captures the vast majority of contacts with positive cases.[2] This recommendation is based on current knowledge about the contagious period of COVID-19 and the incubation period, i.e., the time window from infection to symptom onset [26].

B. GAEN-based Exposure Estimation

GAEN enables the CWA to request the characteristics of exposure to COVID-19 positive people. The CWA uses a series of thresholds a_* that divide the range of attenuation values within a 30 min time window into 4 buckets (`near`: B_n: $[0, a_n)$; `mid`: B_m: $[a_n, a_m)$; `far`: B_f: $[a_m, a_f)$; `very far`: B_{vf}: $[a_f, \infty)$). The Exposure Score (ES) is an estimate of the duration of exposure in the immediate vicinity, i.e., the weighted sum of the duration per attenuation bucket:

$$\mathrm{ES} = w_n \cdot t_n + w_m \cdot t_m + w_f \cdot t_f + w_{vf} \cdot t_{vf}, \qquad (1)$$

where t_n, t_m, t_f and t_{fv} are the exposure durations in the attenuation areas B_n, B_m, B_f, and B_{vf} and w_n, w_m, w_f, and w_{vf} are their individual weights. To estimate the overall risk score, the ES is also influenced by the Transmission Risk Value (TRV) of the contact. In the following, the TRV is assumed to be 1.0 and is therefore neglected. In the end all ESs per day are summed up. If the total duration exceeds a threshold, a risk is reported by the CWA. This threshold-duration is based on the RKI's epidemiological model, which defines a risk encounter as $\mathrm{ES}_{EPI} \geq 10$ min for a single calendar day. This is calculated for each day for the current and the last 14 days.

[1] `https://github.com/corona-warn-app`

[2] Assuming the positive contact also uses the CWA and reports its infection.

C. Parameter Optimization

The performance of the CWA is mainly influenced by choosing the best set of thresholds a_n, a_m, a_f, a_{vf} and corresponding weights w_n, w_m, w_f, w_{vf}. The key idea is to observe reference contact times, their duration, and their proximity from phones that best approximate the real proximity of two people in everyday situations.

To find optimal parameters for CWA, we conducted a large-scale parameter search with approximately 0.5 million combinations based on data from all experiments, see Sec. IV. Based on the epidemiological model we derive condition positive (P) as a risk encounter with $ES \geq 10$ min measured by a reference system, and the opposite with $ES < 10$ min as condition negative (N). Accordingly, we define TP (if P is confirmed by CWA), TN (if N is confirmed by CWA), FP (if CWA failed to detect N) and FN (if CWA failed to detect P).

The search for optimal parameters was carried out with the following objectives: minimize FN (undetected critical contacts) and maximize TP (detected critical contacts). At the same time FP (falsely reported noncritical contacts) and TN (correctly detected noncritical contacts) are less critical. Thus, we optimize the attenuation thresholds a_*, their weights w_*, and ES_{CWA} above which a contact is classified as positive, w.r.t. the smallest FN values and order according to the highest F_2 (β=2) score. Note that the F_β-score measures the accuracy of a test using precision and recall. Precision is the ratio of true positives (TP) to all predicted (true and false) positives (TP+FP). Recall is the ratio of true positives to all actual positives (TP+FN):

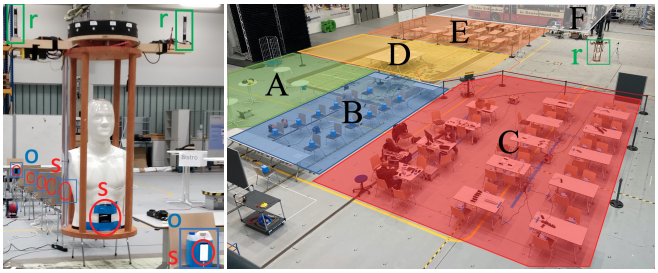
$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \qquad (2)$$

$$= \frac{(1+\beta^2) \cdot TP}{(1+\beta^2) \cdot TP + \beta^2 \cdot FN + FP}. \qquad (3)$$

With an increasing β (> 1), the recall is weighted higher than the precision and a high F-Score ensures lower FN values. RKI recommends β=2 to weight FN more heavily and to highlight possible undetected infected cases on the one side but not to exhaust the test capacities on the other side. We report the FN values because a single undetected positive case can have major consequences and because there is no β to weight the F-Score so that a single case has an effect on the final F-Score.

IV. Experiments

We designed our experiments so that we can (to the best of our knowledge) derive a number of optimal parameters for everyday situations in Germany. First, Sec. IV-A describes the hardware and software infrastructure that we use for our experiments. Then, Sec. IV-B describes our three studies: the first evaluates the influence of different materials on the attenuation and accuracy of the distance estimation, and the second and third examine (semi-)controlled everyday situations with a robot or human participants.



(a) Reference system. (b) Overview of our controlled experiment.

Fig. 2. **Left:** Robot reference system r, obstacles o and phones s. Phones s are attached to the obstacles o, which form our synthetic human participants (PSP), see bottom right corner. Windshield washer fluid in a canister has attenuation characteristics (about 15 to 20 dB) that are comparable to those of the real human body; **Right:** Overview of our large scale experiments with reference robot r. A: bar, B: queue, C: dining restaurant, B+C: fast food restaurant, D: crowd, E: school, A-D: large office, A-E: fairground and shopping center, and F: bus.

A. Hardware and Software Infrastructure

Measurement Systems. Our controlled experiments use 40 different phones with Google's Android OS (8×Pixel 4, 20×Pixel 4a, 1×Huawei Honor 20 and 11 different Samsung phones: 2×A40 and 1×: S10+, A20e, A40, A50, S8, S9+, Note 8, Note 9, Note 10+). We resort to 20 Google Pixel 4a (Android version 10) phones for our semi-controlled experiments to reduce variability. The phones are calibrated according to the Google guidelines [27]. We deploy a modified CWA with GAEN-API version 1.7 to start and stop CT using REST and the Android Debug Bridge (adb).

CWA Modification. We use a modified CWA on all devices and register them in the Google GAEN whitelist so that they can utilize the API. The modified app serves as an intermediary between a Python application and the GAEN-API. It allows to control the API using REST, e.g., to enable and disable the contact tracing, start and stop the transmission of BLE beacons, report the devices as infected and upload their TEKs to a custom server, download shared TEKs, and to query generated ExposureWindows. A fresh install resets the TEK (the automatic reset only happens once a day).

Reference Systems. To collect reference data on the actual temporal and spatial relationships of the individual phones, we use two different time-synchronized optical systems: For our controlled large-scale studies on an area of 44 × 33 m, we use a NIKON iGPS system (CEP_{95} ≤0.9 cm on average) in the Fraunhofer IIS L.I.N.K test center [28], see Fig. 2 (a). For our semi-controlled real-world studies, we use a (mobile) Qualysis system with a sufficient number of markers and cameras to cover the motions in an area of interest as best as possible (CEP_{95} ≤1 cm on average), see Fig. 3. In a post-processing step, we linearly interpolate, re-sample, and synchronize the reference data with the GAEN measurements. We analyze the data visually and statistically and drop invalid measurements.

B. Experiment Design

Effects of Materials. We evaluate the effects of different materials and surfaces. We disturbed the line of sight (LoS) of two calibrated phones (i.e., combinations (Pixel 4; Pixel 4), (Pixel 4; S10+), (Pixel 4; A50), (A50; S10+)), by (1) glass, (2) acrylic glass, (3) wall, (4) ceiling, (5) handbag, (6) trouser pocket, and (7) winter clothing (wet and dry), each with different orientations of the devices. We also investigated the effect of different orientations such that the devices were placed either with the front (display side) or with the back in two different orientations to one another. A reference recording with direct LoS between two phones (same combinations) was recorded in advance to calculate the attenuation of the LoS without interference. We varied the distance between two phones (0.5, 1.0, 1.5, 2.0, 2.5 in [m]) with a fixed obstacle between them ((1) to (4): 0.5 m and (5) to (7): 0 m).

Controlled Synthetic Experiments. To derive optimal parameters that generalize well to typical everyday contact situations, we designed a large-scale experiment with 9 different synthetic scenarios. A robotic reference system r (see Fig. 2 (a)) autonomously follows a pre-defined path, and records both GAEN contacts with an attached phone and synchronized reference data. Each scenario is designed to mimic the typical propagation environment of its real-world counterpart. Fig. 2 (b) shows an overview of the measurement area with the different scenarios. The static smartphones are attached to synthetic human bodies (5-liter bucket with windshield washer fluid) at a location that is most likely for each scenario. The NIKON iGPS system provides reference measurements.

Semi-controlled Real-world Experiments. To derive optimal parameters for typical everyday contact situations and human movement dynamics, we designed a large-scale, semi-controlled experiment with 3 different real-world scenarios with human participants. All of them were informed about the aim and procedure of the study and gave written informed consent for their participation. They wore a FFP2 mask during testing and underwent a COVID-19 rapid test beforehand. Each scenario is designed to mimic the typical propagation environment of its real counterpart. A supervisor instructs each participant on the individual schedule (exemplary shown in Fig. 3). Markers of the reference system are attached to helmets or caps worn by the participants who carry Google Pixel 4a phones. Using the modified CWA and the Python application, all measurements can be started and stopped synchronously. See Tbl. I for a detailed overview of our controlled synthetic and semi-controlled real world experiments.

V. Results

We discuss the effects of materials on the Bluetooth attenuation (Sec. V-A) before we present the parameters that we opti-

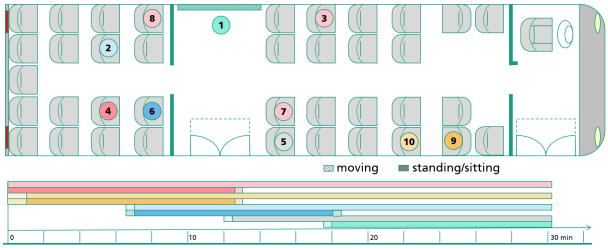


Fig. 3. Exemplary time course of our bus study. Colors indicate the individual seating positions, (de)boarding- and travel times.

TABLE I
DESCRIPTION AND STATISTICS OF EXPERIMENTS, SEE FIG. 4 FOR THEIR ILLUSTRATION. (TOTAL MEASUREMENT DURATION, TDUR). ONLY IN THE SHOPPING CENTER AND ALL SEMI-CONTROLLED EXPERIMENTS ARE RANDOM PEOPLE INFECTIOUS, OTHERWISE IT IS ALWAYS r.

Type	Scenario	Description	Runs [#]	Phones [#]	TDUR [min]
Controlled, synthetic scenarios	Bar Fig. 2 A	The robot (r) simulates a waiter moving to one or more passive synthetic participant (PSP) tables, talks to them for a while, and leaves the table to switch to another.	29	13	1247
	Queue Fig. 2 B	Typical queuing situation, e.g., a cash register with 2 queues with 6 PSP each. The robot (r) simulates a shopper, moves through the queue at different speeds and stops at different locations for different periods of time.	38	13	684
	Dining Rest. Fig. 2 C	40 PSP are seated at 14 tables and a waiter (r) takes orders, serves and picks up the customers' dishes from their tables. r moves randomly between the tables and waits for different lengths of time at each table.	7	40	784
	Fairground Fig. 2 A-E	A visitor (r) moves slowly through an exhibition and visits several exhibits in a random order. r waits for a few min at each exhibit. 40 PSP were placed randomly (almost evenly distributed) to mimic other fair visitors.	12	40	840
	Fast-food Restaurant Fig. 2 B+C	A guest (r) waits (for different periods of time per test run) at a service desk, eats at a randomly selected table (for approx. the same time per test run) and then returns the tray. 28 PSP were evenly distributed on 14 tables in the measuring area to mimic other people.	16	28	720
	Large Office Fig. 2 A-D	An employee (r) works in an office, moves to the printer room, visits a colleague, moves to the meeting room and stops in the kitchen. r waits for a different length of time in each room. 26 PSP are evenly distributed over the area to mimic other people.	6	26	420
	School Fig. 2 E	A teacher (r) moves in front of a class and writes on the board (waits). The teacher walks through the class rows and stops at each table for different periods of time. 26 PSP are nearby and are evenly distributed.	15	26	705
	Crowd Fig. 2 D	A person (r) moves randomly either in the middle or around a crowd, waiting in random places and for any length of time. 37 PSP with phones in trouser pockets are evenly distributed on chairs to imitate other people.	37	37	1147
	Shopping center Fig. 2 A-E	A person (r) moves naturally within a crowd and waits in random places for various periods of time. 17 PSP are evenly distributed across the area to mimic other people. In contrast to the other scenarios, we attach 3 to-be-exposed mobile phones to r instead of one.	8	17	944
Semi controlled	Bus Fig. 4(a)	10 participants stood or sat and had the phone in their hands or in their pockets during a bus ride. Participants were asked to take specific seats and to get on or off the bus according to the supervisor's instructions. Not all passengers adhered to the rules of social distancing, i.e., there are distances <1.5 m between participants.	7	10	201
	Subway Fig. 4(b)	This scenario covered 2 different conditions: static - 10 participants sit or stand still after boarding the subway; dynamic - they get on and off randomly and move within the train.	15	11	344
	Airplane Fig. 4(c)	For each test run we varied: the positions of the phones of 20 participants and 4 PSP, e.g., table, bag, hand, bag, and headrest, and the a priori defined seats. Random participants go to the toilet or take luggage from the rack. We also included 2 crew services with drinks and food and random requests to the flight attendants.	3	22	277

TABLE II
RESULTS OF OUR EXPERIMENTS ON THE EFFECTS OF DIFFERENT MATERIALS AND SURFACES ON THE SIGNAL ATTENUATION

Material	Acryl (0.05 m)				Ceiling		Coat				Coat (wet)				Glass (0.015 m)				Wall (0.3 m)			
Pose	B2B		F2F		B2B	F2F	B2B		F2F		B2B		F2F		B2B		F2F		B2B		F2F	
	45	90	45	90	90	90	45	90	45	90	45	90	45	90	45	90	45	90	45	90	45	90
$\mu(max(e))$	3.6	**3.8**	3.4	3.6	**43.4**	22.4	3.8	4.2	4.7	**4.9**	6.7	5.7	7.0	**7.7**	9.0	12.4	**20.8**	14.5	15.1	14.8	**23.6**	18.0
$\sigma(\mu(max(e)))$	1.7	3.7	1.9	1.6	2.6	1.9	1.5	2.6	1.5	3.2	2.0	1.1	2.3	5.5	1.3	1.6	4.1	2.2	2.2	2.5	3.8	2.4
$\mu(e)$	1.0	1.4	1.4	1.4	10.8	5.6	1.0	1.6	2.8	1.6	3.1	2.5	4.2	2.9	5.4	7.8	14.6	10.4	10.4	11.2	18.9	13.8
$\sigma(\mu(e))$	0.5	1.2	1.1	0.7	0.7	0.5	0.5	1.5	0.7	1.0	1.1	0.8	2.1	1.8	0.9	0.6	2.6	2.1	0.5	0.4	2.8	2.5

Signal attenuation measured between devices with Back-to-back (B2B), Front-to-Front (F2F), and 45° or 90° rotation about the pitch-axis, and all errors e in [dB] w.r.t. to device combinations (Pixel/Pixel, Pixel/S10+, Pixel/A50, A50/S10+) and distances (0.5, 1.0, 1.5, 2.0, 2.5 in [m]). Worst cases per material in **bold**. F2B and B2F combinations are not listed as we saw that they are almost exactly between the respective F2F and B2B errors. We analyzed the data and removed outliers (negative or extreme values).

mized on both (semi-)controlled experiments (Sec. V-C). Next, we discuss the specific results of the controlled (Sec. V-D) and semi-controlled (Sec. V-E) experiments.

A. Effects of Materials

We found that the attenuation of various materials weakens the signal strength of the transmitter by up to 43.4 dB. Tbl. II shows the mean and the standard deviations of the attenuation of the various materials averaged over approx. n=1,000 measurements each. F2F attenuations are often worse than B2B ones. We think this is due to the device specific antenna pattern. The signal strength is also influenced by the orientation of the phone. Orientations of 90° (antennas aligned parallel to each other) almost always lead to far lower attenuations than at 45°. The attenuation of glass depends on both the material thickness and the type. The influence of brick walls and concrete ceilings (0.6 m) is highest, as these materials cause a high level of attenuation. (Wet) coats also influence the signal strength. This suggests adjusting parameters depending on the season to compensate for the materials worn in cold periods.

TABLE III
RESULTS OF OUR (SEMI-)CONTROLLED EXPERIMENTS. ∅ - MEAN AND σ - STANDARD DEVIATION OF F_1, F_2, TP, FP, FN AND TN. THE CAPITAL LETTERS IN EACH SCENARIO INDICATE THE SCENARIO AREA IN FIGURE 2.

	Controlled									Semi-controlled		
	Bar (A)	Queue (B)	Dining rest. (B+C)	Fairground (A-E)	Fast-food rest. (B+C)	Large office (A-D)	School (E)	Crowd (D)	Shopping center (A-E)	Bus	Airplane	Subway
$\varnothing F_1$	0.83	0.5	0.88	0.53	0.59	0.55	0.46	0.74	0.78	0.81	0.75	0.48
$\varnothing F_2$	0.88	0.66	0.88	0.58	0.61	0.66	0.68	0.82	0.9	0.86	0.88	0.69
TP	239	104	173	59	95	23	44	670	744	361	275	195
FP	70	181	21	68	72	31	103	377	408	126	181	420
FN	25	23	25	36	59	7	0	88	0	40	0	2
TN	14	148	15	227	152	89	129	161	0	67	0	84
∅TP	8.24	2.74	28.83	5.9	6.79	3.83	3.67	18.61	1.94	5.01	11.46	2.6
∅FP	2.41	4.76	3.50	6.8	5.14	5.17	8.58	10.47	1.06	1.75	7.54	5.6
∅FN	0.86	0.61	4.17	3.6	4.21	1.17	0	2.44	0	0.56	0	0.03
∅TN	0.48	3.89	2.50	22.7	10.86	14.83	10.75	4.47	0	0.93	0	1.12
σ_{TP}	3.78	0.44	2.03	1.45	1.52	0.37	0.94	10.73	1.43	2.55	3.54	2.39
σ_{FP}	2.5	1.48	0.96	2.44	0.74	1.21	2.18	9.0	1.43	1.92	5.58	2.9
σ_{FN}	0.86	0.84	2.03	1.2	1.52	0.37	0	2.96	0	1.56	0	0.16
σ_{TN}	0.5	1.89	0.96	2.79	0.74	1.21	2.74	5.42	0	1.5	0	2.6

B. Optimized Parameters

In our material experiments, we found that the first direct LoS path is influenced more by objects in the immediate vicinity along the path than by walls at a great distance. Based on our findings of the effects of materials, signal propagation environment, and human movements, we thus propose to derive general parameters based on (semi-)controlled experiments that take into account all possible (static and dynamic) types of obstacles along a transceiver line.

Based on the collected data from all of our experiments and using Google's attenuation calibration tables [27], the following parameters yield the highest score based on our epidemiological model and the chosen error metric: a_n=63 dB, a_m=73 dB, a_f=79 dB and w_n=0.8, w_m=1.0, w_f=0.1, w_{vf}=0.0 at ES_{CWA}=9.0 min. Attenuation values in the narrow range B_n correspond to narrow exposures which are weighted with w_n=0.8. Likewise, mid-range attenuations B_m correspond to narrow exposures and are therefore weighted with w_m=1.0. w_f=0.1 reduces the influence in B_f and w_{vf}=0 discards highly attenuated values in B_{vf}, as they correspond to negligibly large distances between devices. An EN is triggered if ES_{CWA} $\geq$9 min within a single day (in contrast to $\mathrm{ES}_{EPI} \geq 10$ min).

C. General Insights

We measured a total of 6,431 contacts for the controlled synthetic (4,680) and semi-controlled real-world (1,751) scenarios and found differences in the results between and within the two scenario types. This indicates that a scenario-specific parameterization would yield the best results. We first discuss important findings of the individual scenarios of both experiments (or groups of scenarios if they have a comparable study concept, similar dynamics, and propagation environment or show comparable effects). Although we have derived a general parameter from all scenarios, we cannot compare them directly due to their individual study design. Tbl. III shows the results.

Bars, **Dining rest.**, **Crowd**, **Shopping center**, **Bus**, **Airplane**, and the **Subway** scenarios show high positive values ($\mathrm{TP_{rel}}$+$\mathrm{FP_{rel}}$>80% with $\mathrm{TP_{rel}}$ = TP/total contacts) that demand high test capacity. Instead, **Fairground**, **Large Office**, and **Fast-food** yield high negative values ($\mathrm{TN_{rel}}$+$\mathrm{FN_{rel}}$>55%). The FPs for the experiments **Queue**, **Large Office**, **School** and dynamic **Subway** (result not shown separately) are almost twice as high as the corresponding TPs. These are side effects of our general parameters and the F2-score, as low FNs come at the expense of higher FPs. Since the main goal is to reduce the FN rate, this is acceptable with sufficient test capacity.

The high TPs of the **Bar**, **Dining rest.**, **Shopping center**, **Bus**, and **Airplane** scenarios are caused by a longer exposure time at a bar- or at a dining table (>1h), many infected people shopping ($\geq$3), and the particular propagation environment (metallic tube) in buses and airplanes, which promotes signal propagation. The highest $\mathrm{FN_{rel}}$ (>9%) were obtained in **Fast-food**, **Dining rest.**, and **Fairground**. Some contacts were not detected as the distances between people were often large and contact times were short, as the scenarios are more static with test persons sitting at the table or standing in front of exhibits for a long time. In contrast, **School**, **Shopping center**, **Airplane** and **Subway** yield the lowest FNs ($\approx$0). Almost every contact is correctly detected in the **School** and **Airplane** scenarios. In the **Airplane** and **Subway** scenarios the good propagation conditions cause almost all contacts to be detected as positive. Instead, the **Shopping center** scenario is very dynamic with moving people, which leads to fewer contacts with shorter durations. However, all contacts are correctly detected when customers stay in front of the shop windows for a longer time. The separate evaluation of the **Subway** scenario's static and dynamic sub-tests shows that more motion (dynamics) leads to shorter contact times and possibly fewer TP and more TN contacts. In contrast, the static variant leads to lower FN and high TP values, which is in line with the **Airplane** scenario. The results therefore suggest that there are differences between static and dynamic scenarios. Thus, it would make sense to separate the parameters according to static and dynamic scenarios, e.g., using motion classifiers.

D. Results of the Controlled Synthetic Experiments

We first address general findings, before we discuss individual clusters. The results show that most of the critical contacts are correctly classified as critical cases (recall of 89%). In contrast, the number of all FNs (=263) is relatively low (TN=935). The highest ∅FNs are reported in the **Fast-food** (4.21±1.52), **Dining rest.** (4.17±2.03), **Fairground** (3.6±1.2), and **Crowd** (2.44±2.96) scenarios but with high variances. In contrast, the other scenarios resulted in remarkably lower ∅FN values (<1.17) and variances (<0.86). The higher FNs in these scenarios correlate with the density and amount of people absorbing signals. Standing/sitting near infectious subjects whose signal is shadowed can also lead to increased FNs (compare the high FNs of the **Crowd** scenario). From this we can follow that simple social rules, such as placing

(a) Bus. (b) Subway. (c) Airplane (Airbus A321).

Fig. 4. Scenarios of our semi-controlled experiments. d_c - close distance [<1.5 m], d_m - middle distance [<2.5 m], d_f - far distance [<7.5 m].

the phone on the table (e.g., in **Bar**, **Dining rest.**, **School** and **Office**) or around the neck (e.g., in **Queue**, **Shopping center** or **Crowd**) enables better LoS conditions and thus helps to reduce FNs. Although we weight the FNs more heavily (thus the FPs also increase), most scenarios still yield higher TP and TN values. These findings are also reflected in the F2-scores, that are below 70% in scenarios with high FPs and FNs.

The **Bar** and the **Dining rest.** experiments show that staying at tables in a bar or restaurant yields high TP+FP values and puts an additional load on the test capacity and results in higher FN than TN values. We think that this may be caused by the large distances between phones in these scenarios. The **School** and the **Queue** scenarios show a low number of infections and missed detections. While the detection is more accurate for standing at the beginning or the end of a line, standing in the middle of the line showed higher FN rates.

Even if the **Dining rest.** (∅FN=4.17) and **Fast-food** (∅FN=4.21) scores show similar FNs, their FPs differ: FPs in **Fast-food** (∅FP=5.14) are clearly higher than in **Dining** (∅FP=3.5). The reason may be that in a dining restaurant typically less dynamic human bodies affect the radio propagation environment than in a typical fast-food restaurant. Thus, the dining scenario is more controllable and lowers the FPs.

The **Fairground** and **Large office** experiments show low TPs and high TNs, this indicates a low risk of infection. We think this is related to the rather high distance between phones in the large measurement area in these scenarios. Yet, their FP rate is higher than the TP rate which implies that our general parameters are not best suited for these types of scenarios.

The results of the **Fast-food** experiment are worst: the highest FN_{rel}=15.6±5.6% and a high unnecessary load on the test capacity (FP_{rel}=19.0±2.8%) across all experiments. We think that this represents a Fast-food restaurant with more dynamic human bodies affecting the radio propagation environment. Thus, the scenario is less controllable and raises FP.

The results of our **Crowd** experiment show that standing or moving around a crowd of people yields a low FN=0, while moving in the middle causes a high FN=88 and a high infection rate (TP=670). We think that moving in the middle of the crowd of infected people is a very high risk, but it gets even worse as we cannot detect all contacts.

The individual study design of the **Shopping center** experiment allows no comparison to other scenarios (we placed three to-be-exposed phones on the robot r and random people were infected). Here, the highest number of true (TP_{rel}=64.6±47.8%) and false infected contacts (FP_{rel}=35.4±47.8%) are generated. We think that the higher FPs in a mall are caused by reflections from nearby objects along the propagation path. These effects indicate again, that a specific parameter set may be a better choice for such a scenario. However, the low number of FNs and TNs compensates for this. We think both the movement dynamics of people that walk along or stand to look at the shop windows enable this.

E. Results of the Semi-controlled Real-world Experiments

In the semi-controlled scenarios the average recall (95%) is higher (6%) than in the controlled scenarios. This is because the in-between diversity of the propagation environments of the semi-controlled scenarios is much lower than that of the controlled and thus our general parameters perform better. Across all semi-controlled experiments ∅FN is very low (<0.6). Fig. 4(a-c) show examples of our scenarios for the measured distances of the CWA and reference system.

The higher FN rate in the **Bus** scenario (FN_{rel}=6.4±17.6%) causes the lowest recall (90%). In comparison, the low FN rates (≈0.0) in the **Subway** and **Airplane** scenarios result in a high recall (≈100%). In contrast, the FP_{rel} is significant among the scenarios ($p<0.01$): this time **Bus** (FP_{rel}=20.8±22.8%) shows the lowest rate, and **Airplane** (FP_{rel}=36.7±24.5%) as well as **Subway** (FP_{rel}=59.1±29.6%) show much higher ones. Even if we split the **Subway** scenario into static and dynamic sub-tests (with passengers getting on and off and movements in the train), we see that both deliver highest FP values across all three experiments. However, the three scenarios highly increase the load of the test capacity (TP+FP). **Bus** (∅FN=0.56±1.56) shows the highest FN value on average, but also the highest variances since we measured both on the bus and at the bus-stop. In contrast, the other scenarios resulted in lower ∅FN values (around 0.0) and variances (<0.25).

Our test design could be responsible for the higher FN scores in the **Bus** scenario because users were allowed to put their phones in their pockets, some sat far away, and others were not exposed long enough as they only traveled for a short time. This is in contrast to the **Subway** and **Airplane** scenarios, where the users are inherently encouraged to place their phones in a (static) way that enables better LoS conditions and thus, helps to sustainably lower the FN (=0) rate. Interestingly, for **Airplane** there are no TNs. We think the propagation environment of an aircraft contributes to the signal spread and hence all contacts were detected

to be positive, as they all underestimated the real distance. Although we weight the FN values more heavily and thus the FP values also increase, all scenarios (except **Subway**) nevertheless yield higher TP and TN values. These findings are also reflected in the F2-scores, which barely fall below 70% in the semi-controlled scenarios even for an excessive number of FP as the FN values are very low. Interestingly, the FP values for the **Subway** experiments are almost twice as high as the corresponding TP values. When we split the **Subway** scenario into static and dynamic variants we see that dynamics in a subway reduce the FP rate gently at the cost of slightly lower TP and higher TN rates. However, since the main goal is to reduce the FN rate, even a static situation in a **Subway** is not a problem if there is enough test capacity.

VI. Conclusion

We presented an extensive measurement campaign that was carried out together with DT and SAP under the supervision of the RKI to derive optimal parameters for a decentralized CT (CWA from Germany) based on GAEN. We show that while optimal parameters are application- and environment-specific (and also depend on motion dynamics), our optimized general parameters are a practical compromise between energy costs, applicability, accuracy and reliability of the ENF. We show that our parameters only lead to few false negatives.

However, care must be taken in scenarios where people stay together closely, hence attenuate or reflect the signals such that contacts are not detected. We suggest that scientific studies should report a uniform metric based on calibrated measurements, namely false negative results, i.e., an infected contact was not detected, to enable better comparability of the methods and parameters in the future.

Parameters have to be revised when epidemiological models change, e.g. to cope with new variants of COVID-19 (Delta variant, etc.). Furthermore, other respiratory diseases can be tracked via CT protocols.

Acknowledgments

We thank Airbus and the Verkehrs-Aktiengesellschaft Nürnberg (VAG) for providing their measurement environment, and the participants that took part in our tests.

References

[1] A. J. Kucharski, P. Klepac, A. J. Conlan, S. M. Kissler, M. L. Tang, H. Fry, J. R. Gog, W. J. Edmunds, J. C. Emery, G. Medley, *et al.*, "Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 10, pp. 1151–1160, 2020.

[2] A. Aleta, D. Martin-Corral, A. Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, *et al.*, "Modeling the Impact of Social Distancing, Testing, Contact Tracing and Household Quarantine on Second-Wave Scenarios of the COVID-19 Pandemic.(2020)," *Publisher Full Text*, 2020.

[3] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, *et al.*, "Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study," *The Lancet Infectious Diseases*, vol. 20, no. 8, pp. 911–919, 2020.

[4] B. Armbruster and M. L. Brandeau, "Contact tracing to control infectious disease: when enough is enough," *Health care management science*, vol. 10, no. 4, pp. 341–355, 2007.

[5] C. N. Haas, J. B. Rose, and C. P. Gerba, *Quantitative microbial risk assessment*. John Wiley & Sons, 2014.

[6] C. Fraser, L. Abeler-Dörner, L. Ferretti, M. Parker, M. Kendall, and D. Bonsall, "Digital contact tracing: comparing the capabilities of centralised and decentralised data architectures to effectively suppress the COVID-19 epidemic whilst maximising freedom of movement and maintaining privacy," *University of Oxford*, 2020.

[7] S. Von Arx, I. Becker-Mayer, D. Blank, J. Colligan, R. Fenwick, M. Hittle, M. Ingle, O. Nash, V. Nguyen, *et al.*, "Slowing the spread of infectious diseases using crowdsourced data," *Covid Watch*, 2020.

[8] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser, "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 368, no. 6491, 2020.

[9] M. Salathé, C. L. Althaus, N. Anderegg, D. Antonioli, T. Ballouz, E. Bugnion, S. Capkun, D. Jackson, S.-I. Kim, J. Larus, *et al.*, "Early evidence of effectiveness of digital contact tracing for SARS-CoV-2 in Switzerland," *medRxiv*, 2020.

[10] M. E. Kretzschmar, G. Rozhnova, M. C. Bootsma, M. van Boven, J. H. van de Wijgert, and M. J. Bonten, "Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study," *The Lancet Public Health*, vol. 5, no. 8, pp. e452–e459, 2020.

[11] G. Avitabile, V. Botta, V. Iovino, and I. Visconti, "Towards Defeating Mass Surveillance and SARS-CoV-2: The Pronto-C2 Fully Decentralized Automatic Contact Tracing System.," *IACR Cryptol. ePrint Arch.*, vol. 2020, p. 493, 2020.

[12] N. Ahmed, R. A. Michelin, W. Xue, S. Ruj, R. Malaney, S. S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, and S. K. Jha, "A Survey of COVID-19 Contact Tracing Apps," *IEEE Access*, vol. 8, pp. 577–601, 2020.

[13] Robert Koch Institut, "Kontaktpersonen-Nachverfolgung bei SARS-CoV-2-Infektionen." Accessed May 31, 2021. [Online].

[14] European Centre for Disease Prevention and Control, "Contact tracing: public health management of persons, including healthcare workers, who have had contact with COVID-19 cases in the European Union – third update, 18 November," tech. rep., ECDC, 2020.

[15] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communic. Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[16] G. Li, E. Geng, Z. Ye, Y. Xu, J. Lin, and Y. Pang, "Indoor positioning algorithm based on the improved RSSI distance model," *Sensors*, vol. 18, no. 9, p. 2820, 2018.

[17] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Computing Surveys*, vol. 46, no. 2, 2013.

[18] A. Trivedi and D. Vasisht, "Digital contact tracing: Technologies, shortcomings, and the path forward," 2020.

[19] D. J. Leith and S. Farrell, "Measurement-based evaluation of Google/ Apple Exposure Notification API for proximity detection in a light-rail tram," *PLoS ONE*, vol. 15, no. 9 September, pp. 1–16, 2020.

[20] D. J. Leith and S. Farrell, "Coronavirus Contact Tracing: Evaluating The Potential Of Using Bluetooth Received Signal Strength For Proximity Detection," *arXiv*, pp. 1–11, 2020.

[21] J. Lelieveld, F. Helleis, S. Borrmann, Y. Cheng, F. Drewnick, G. Haug, T. Klimach, J. Sciare, H. Su, and U. Pöschl, "Model Calculations of Aerosol Transmission and Infection Risk of COVID-19 in Indoor Environments," *Intl. J. of Environmental Research and Public Health (JERPH)*, vol. 17, no. 21, 2020.

[22] P. H. O'Neill, T. Ryan-Mosley, and B. Johnson, "A flood of coronavirus apps are tracking us. Now it's time to keep track of them," May 2020.

[23] D. J. Leith and S. Farrell, "Measurement-based evaluation of Google/Apple Exposure Notification API for proximity detection in a commuter bus," *arXiv*, pp. 1–8, 2020.

[24] Simula Research Laboratory, "Sammenligning av alternative løsninger for digital smittesporing," 2020.

[25] S. Meyer, T. Windisch, N. Witt, and D. Dzibela, "Fraunhofer IIS - Google Exposure Notification Api Testing," June 2020.

[26] C. McAloon, Á. Collins, K. Hunt, A. Barber, A. W. Byrne, F. Butler, M. Casey, J. Griffin, E. Lane, D. McEvoy, P. Wall, M. Green, L. O'Grady, and S. J. More, "Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research," *BMJ Open*, vol. 10, no. 8, 2020.

[27] G. Inc., "Exposure Notifications BLE attenuations - Calibration," 2021.

[28] M. Stahlke, S. Kram, C. Mutschler, and T. Mahr, "NLOS detection using UWB channel impulse responses and convolutional neural networks," in *Intl. Conf. Localization and GNSS*, (Tampere, Finland), 2020.