An efficient algorithm to estimate Covid-19 infectiousness risk from BLE-RSSI measurements

Jean-Marie S Gorce, Malcolm Egan, Rémi Gribonval

An efficient algorithm to estimate Covid-19 infectiousness risk from BLE-RSSI measurements

Jean-Marie GORCE , Malcolm EGAN, Rémi GRIBONVAL

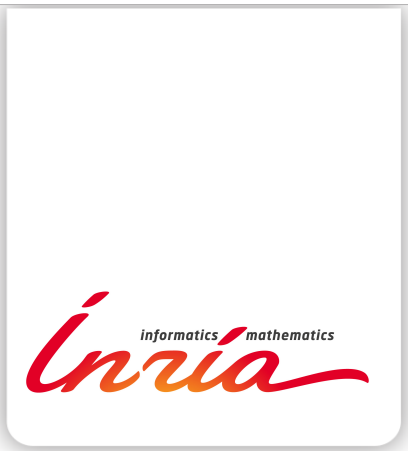
RESEARCH REPORT

N° 9345

May 2020

Project-Teams MARACAS and DANTE

ISSN 0249-6399 ISRN INRIA/RR--9345--FR+ENG



An efficient algorithm to estimate Covid-19 infectiousness risk from BLE-RSSI measurements

Jean-Marie Gorce* †, Malcolm Egan*†, Rémi Gribonval*‡

Project-Teams MARACAS and DANTE

Research Report n° 9345 — May 2020 — 36 pages

Abstract: This report describes an algorithm to predict the Covid-19 infectiousness risk from physical contacts through BLE-RSSI measurements. In order to derive a robust risk estimator, the proposed algorithm relies on known physical wireless propagation effects and on technical properties of current BLE interfaces. The proposed algorithm has been tested on the data acquired by the German teams from the Fraunhofer institute in the PEPP-PT European project. We thank them for sharing their data, for their helpful comments and answers.
Our algorithm is currently under evaluation on the French data acquired from May 18 to 20, 2020.

Key-words: Covid-19, StopCovid, Bluetooth, BLE, distance prediction, infectiousness propagation prediction

* Mail: {firstname.lastname}@inria.fr
† Univ Lyon, Inria, INSA-Lyon, CITI lab (EA3720), F-69621 Villeurbanne, France
‡ Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Un algorithme efficace d'estimation du risque de contagion Covid-19 à partir de mesures BLE-RSSI

Résumé : Ce rapport documente un algorithme de prédiction du risque de contamination par le Covid-19 lors de situations de proximité physique en utilisant des mesures Bluetooth (BLE-RSSI). Pour aboutir à un estimateur robuste du risque, l'algorithme proposé s'appuie sur des propriétés connues de la physique de la propagation sans fil et sur les propriétés techniques des interfaces BLE actuelles. L'algorithme proposé a été testé sur les données acquises par les équipes allemandes de l'Institut Fraunhofer dans le cadre du projet européen PEPP-PT, que nous remercions pour le partage des données ainsi que pour leur aide apportée via des commentaires et réponses à nos questions.

Notre algorithme est actuellement en cours d'évaluation sur les données françaises acquises entre le 18 et le 20 mai 2020.

Mots-clés : Covid-19, StopCovid, Bluetooth, BLE, prédiction de distance, prédiction de propagation d'infection

1 Introduction

Tracking how Covid-19 spreads over a population is a critical aspect that may aid relaxing lockdown conditions. The European project PEPP-PT, and the French protocol ROBERT built on it, both aim at using Bluetooth Low Energy (BLE) measurements obtained from HELLO messages to estimate the risk of infection spreading from diagnosed individuals.

In this document, we develop an algorithm to improve the accuracy of risk estimation within the structure and parameters of the ROBERT protocol. Underlying the algorithm is mathematical modeling of the physical wireless communication link and experimental data based on Bluetooth Received Signal Strength Indication (BLE RSSI) traces. In particular, the algorithm is evaluated on experimental data obtained by the Bundeswehr experiments, which provide a large number of device-to-device BLE RSSI traces in realistic scenarios. Descriptions of these scenarios have been provided by the working group *Bluetooth Measurements and Proximity* within the PEPP-PT project [1, 2, 3, 4]. These documents include a general description of the project which was delivered on April, 9, 2020 [5], and intermediate documents including reports on measurements and descriptions of experimental data [1, 2, 3, 4].

In the solution proposed by the PEPP-PT project, a core element is a logistic regression matrix used to predict a risk score from a small number of parameters acquired from a 15-minute long RSSI vector; namely, the mean RSSI, max RSSI and the number of received packets. In our approach—inspired by the data obtained in the PEPP-PT project—we propose a risk score that relies directly on a physical interpretation of the experimental RSSI data by incorporating pathloss and fading. We also adopt a Bayesian formulation of the risk estimation problem.

The document is organised as follows. First, Section 2 develops the proposed algorithmic solution, to give a general overview of the solution. All parameters involved in the model will be properly introduced and defined in the subsequent sections. Section 3 describes standard knowledge on propagation facts. In Section 4 we describe a Bayesian model for distance estimation from RSSI measurement. Exploiting PEPP-PT data from the Bundeswehr experiments, we also tune the model parameters and evaluate the accuracy of the RSSI-distance relation. Section 5 derives a model of the probability of contamination, and shows how the RSSI measures can be aggregated to compute a risk score.

2 Description of the proposed algorithm

In this section, we develop an algorithmic solution to compute an estimate of the infectiousness risk, which has been developed upon the design of ROBERT algorithm used in the Stopcovid application.

The proposed algorithm is first detailed and its integration at the different layers in the Robert protocol is presented to provide a general and practical overview of the algorithm. Then, the following sections detail the theoretical foundation and the experimental approach used to develop this algorithm.

2.1 Assumptions and interactions with higher layers

The proposed algorithms involve three layers :

- The calibration function is hosted in the BLE *physical layer*.
- The risk scoring function is hosted in the *application layer* (on the mobile phone). Note that according to Robert protocol, Risk scores are kept on the phone, unless the user is tested positive.

- The scores associated to the temporary received Ids are sent to the server by the mobile, when the owner is declared infected.
- A user who wants to evaluate his risk, may decide to send the list of his former temporary Ids to the server.
- The server then compares this list to the list of infected temporary Ids returned by infected persons, and compute a global infectiousness risk. This risk is a multiple-contacts risk measure and not a one-to-one risk evaluation.

This structure has been imposed for privacy reasons by the ROBERT protocol.

The objective of this document is to detail the process of converting a set of BLE RSSI measurements into a final aggregated risk. The code of our application is spread over three layers: the BLE-PHY calibration layer, the application layer and the server.

The BLE-PHY calibration layer interacts with the application layer, through the following functions as illustrated in Fig.1:

- Start/stop function.
- Send data.
- Update (or initialize) parameters.
- Update software.

The main block in this layer is the `Format_contact` which gathers the packets received with the same temporary Ids, and compensate the raw RSSI values by correction factors. This bloc creates a *contact* which contains a temporary Id, a set of RSSI values and the corresponding timestamps.

At the second layer, the application layer exploits these contacts by applying first the `Filter_contact` function which suppresses the non significant contacts (duration of less than $5mn$). It then computes score vectors and a cumulative score associated to each contact. The final filter function `Filter_cum_score` allows to select only the most significant contacts associated to the time period.

2.2 System parameters

The following parameters are critical for the efficiency of the scoring function.

- *Rate* $R_b(i)$: Each device which hosts the StopCovid application sends i HELLO packets at a rate $R_b(i)$. The nominal value of this rate is $R_b(i) = 0.1Hz$ (i.e. one packet each 10 seconds, but the exact value may varies with the devices). As discussed in section 3.3, a rate of at least 3 packets per minute is recommended. Note that a regular sampling is in fact not necessary. The rate in average, should not be lower that 4 packets per minute ($0.066Hz$). A higher rate, e.g. $R_b(i) = 0.2Hz$ could provide a significant improvement on the results, but at the price of an energy consumption increase and more packet collisions in dense environments.
- *Slots and synchronization*: in order to preserve the privacy, the devices change their identifier (Id) every 15 minutes. The synchronization of the slots is granted by the application. It is supposed here that the application is able to synchronize the devices and ensure that the time shift between devices is below $1min$. The slots are thus guaranteed to be at least of $13min$.

  We denote T_{sl} the nominal time duration of slots.

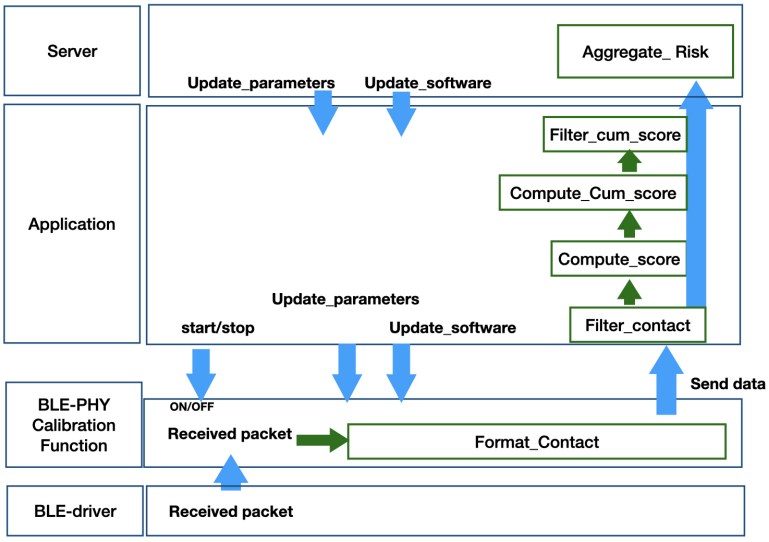


Figure 1: Representation of the proposed algorithm with the main processing blocs : `Filter_contact`, `Compute_score`, `Compute_cum_score`, `Filter_cum_score` and `Aggregate_risk`.

- *Calibration :* G_{tx}, G_{rx}: It has been widely observed that RSSI values in BLE suffer from a lack of calibration. The RSSI values may differ from more than $20dB$ (we shall refer the reader to the work of the PEPP-PT group [2] or to [6, 7]. Therefore, a raw RSSI measure should be understood as a non calibrated value that cannot be directly interpreted to measure neither the distance, nor the risk. A calibration is necessary, with two compensation gains G_{tx} and G_{rx}, to compensate respectively for the transmission gain/power and the reception gain. Additional compensation can be applied.

  - *Static calibration:* $G_{0,tx}, G_{0,rx}$ Each device i possesses its own (static) calibration values $G_{0,tx}(i)$ and $G_{0,rx}(i)$. In StopCovid, these values will be issued from the work of the GSMA group, and are provided to the device when the application is installed. These values can be later updated by the server, through the application. *To avoid confusion, the values of* $G_{0,tx}(i)$ *and* $G_{0,rx}(i)$ *in this document, are related to the GSMA coefficient as follows*

$$G_{0,tx}(i) = -\texttt{tx\_RSS\_correction\_factor} - \texttt{ref\_loss} \quad (1)$$
$$G_{0,rx}(i) = -\texttt{rx\_RSS\_correction\_factor}, \quad (2)$$

    where `tx_RSS_correction_factor` and `rx_RSS_correction_factor` are the GSMA correction factors and `ref_loss` is the reference pathloss measured at one meter with the reference phones in the GSMA procedure.

    When a device i sends a HELLO packet with its temporary `Id`, it also sends a compressed dynamic gain $G_{tx}(i)$ that possibly combines the device specific static gain $G_{0,tx}(i)$ with other dynamic calibration factors resulting from the environment.

    In the algorithm V1, the received RSSI will be compensated only with these static gains : $G_{tx}(i) = G_{0,tx}(i)$ and $G_{rx}(i) = G_{0,rx}(i)$. Therefore, the receiver j will apply

the two compensation values : $G_{tx}(i)$ (sent by the transmitting device along with its temporary Id Id(i,t), but without direct information on its actual identity i) and $G_{rx}(j)$ (its own compensation value). The use of these values is described in the algorithm below.

- *Posture calibration*: Besides device-specific static calibration gains, the RSSI values are impacted by the current position of the device [7]: if the device (typically a smartphone) is in the pocket rather than in the user hand, near the ear, we can expect few dB less in the measured RSSI in both directions (emit and receive). Therefore the sender and the receiver should modify their overall calibration gains according to their postures. These corrections will be integrated if significant impact is confirmed during our experimentation.

  This requires to assume that the device is able to evaluate the posture (hand or pocket) based on different sensors : screen on/off, light sensor, motion, ... This is under investigation either from external information, or (partially) from dynamics of RSSI time series.

  Therefore, if a device is identified to be in a pocket, a gain of XdB can be applied to the gains. The posture gain at the transmission and the reception is the same, therefore an additional penalty $G_{post}(i) \in \{0, X\}$ could be updated every minute from the posture evaluation routine.

  The local Tx and Rx gains are updated accordingly :

$$G_{tx}(i) = G_{0,tx}(i) + G_{post}(i),$$

  and

$$G_{rx}(i) = G_{0,rx}(i) + G_{post}(i).$$

  Note the the TX gain applied to the received RSSI is the gain received in the HELLO packets.

- *Environment calibration*: This part is also optional. Based on environmental measurements, mostly based on the density of HELLO packets, an environment density parameter is estimated. A parameter, called crowded state: $S_{cr} \in [0, 1]$, is evaluated. Then, an additional compensation gain $G_{env}(i)$ can be introduced into $G_{tx}(i)$ and $G_{rx}(i)$ for each device :

$$G_{tx}(i) = G_{0,tx}(i) + G_{post}(i) + G_{env}(i),$$

  and

$$G_{rx}(i) = G_{0,rx}(i) + G_{post}(i) + G_{env}(i).$$

Overall, when both posture calibration and environment calibration are used, the emitter first computes $G_{tx}(i)$ as above and sends this value, along with its temporary Id. The receiver uses this value together with its own value $G_{rx}(j)$ to obtain a calibrated RSSI value as further described in the algorithm below.

- *Risk scoring:* the conversion of the calibrated RSSI values into a risk is detailed in section 5. It is parameterized by a value r_0, which relates the cumulative score coefficient to a probability of infection. Considering HELLO packets received on device j (which represents the "current device/receiver" on which the BLE and Application are run) from a transmitting device with temporary Id `Id`, the objective is to estimate the conditional probability $\mathbb{P}(\texttt{Id} \text{ is infected } \mid j \text{ is infected})$. This is an estimate of the probability that the

owner of the device with temporary Id `Id` becomes infected if we (later, through a positive test) become aware that the owner of device j was actually infected during the considered 15min period of contact.

This probability is represented by a score $\pi_{j,\texttt{Id}}$ which is kept locally on device j. If the owner of device j is tested positive and declares this event to the application layer by providing a code obtained from a physician, and *only in this case*, then this triggers the transmission to the server of the temporary Ids corresponding to scores above a certain threshold, together with the corresponding scores. The threshold value (or no threshold) could be updated by the server layer to adjust false positive / false negative tradeoffs. Moreover, the server can adjust the value r_0 which serves to translate scores to actual estimates of probabilities of infection.

We propose the following options, where $\hat{\mathbf{Q}}_{j,\texttt{Id}}$ is a vector of calibrated RSSI values:

- A unique score: This score is the output of the `Compute_cum_score` function for each 15-minute period and each device Id `Id` met during this period, the cumulative score $\pi_{j,\texttt{Id}} = R(\hat{\mathbf{Q}}_{j,\texttt{Id}})$ given by (45) is returned to the application after being filtered to keep only the most significant contacts.

  If the owner of device j declares itself as infected then this cumulative score is sent to the server, and the probability $\mathbb{P}(\texttt{Id} \text{ is infected} \mid j \text{ is infected})$ can be computed given a parameter r_0 that translates the score into a risk level. This is the version intended to be used in algorithm V1.
- A more detailed metric with the following set of parameters $(\bar{\pi}_{j\texttt{Id}}, D_{j\texttt{Id}}, N_{j\texttt{Id}})$ which are respectively the average score (either with (53) or (55)), the duration of the contact and the number of non null RSSI values (or higher than a threshold).
- Other options upon request are to send either the output of `Compute_score` or `Filter_contact` to explore other risk computation. But the advantage of computing the risk in the application is to account for local information, as mentioned earlier to compensate for posture or environment.

Whatever the final choice the parameters to be sent to the server are named $\Theta_{j\texttt{Id}}$.

Remark. *When receiver j encounters transmitter i over different 15-minute periods, the application layer (in device j) cannot accumulate the corresponding scores since the transmitter's `Id` changes from one period to another. However, if the receiver j is later tested positive and sends to the server the list of Ids it has potentially infected, it will do so for all of the Ids it has met (possibly restricted to the infectious period, and/or to scores above a threshold). Then the final score will aggregate all scores. As discussed in section 6.0.3, this allows compute the probability that the owner of device i is infected, $\mathbb{P}(i \text{ is infected})$, by aggregating scores over possibly different Ids from the same user i, and over all different j that have declared themselves as infected.*

Remark. *It could also be worth adapting the risk scoring to some posture or environmental variables. For example, when a person is walking, the critical distance below which there is a high risk of infection is considered to be of the order of $d_c \approx 2m$, while $d_c \approx 10m$ when the person cycling. Another example is that of healthcare personnel who are wearing high quality protective equipment. Allowing them to declare this as a parameter would potentially decrease the rate of false positives.*

In order to simplify the implementation, we summarize in Table 1 the list of the parameters to be taken into account.

Name	Role	Procedure	Nominal value
R_b	HELLO Tx rate		$1/10Hz$
$G_{0,tx}$	Tx calibration	Compensate_gain()	see table GSM-A
$G_{0,rx}$	Rx calibration	Compensate_gain()	see table GSM-A
G_{post}	Posture calibration	Compensate_gain()	$0dB$; (TBD)
G_{env}	Environment calibration	Compensate_gain()	$0dB$ (TBD)
D_{th}	Minimum significant duration	Filter_contact()	$2min$
T_{win}	Time window for aggregation	Bufferize_RSSI()	$120s$
$T_{overlap}$	Overlap between successive time windows	Bufferize_RSSI()	$60s$
T_{ref}	Reference duration for risk evaluation	Aggregate_risk()	$15min$
r_0	Risk coefficient	Aggregate_risk()	TBD
P^*_{min}	Minimal significant RSSI	Compute_score()	$-66dBm$
Δ_P	Power range	Compute_score()	$\{15, 20, 21, 23, 27, 39\}$

Table 1: List of parameters to be used in the application. The parameters indicated by *not necessary* are used in the theoretical mode but do not need to be used in the implementation. the parameters with the symbol $.^*$ are those for which the nominal values have been estimated from the PEPP-PT datasets and used to compute the other parameters, but they are not used explicitly in the algorithms.

2.3 General algorithm description

The implemented functions follow:

`Compensate_gain()` This function, according to the known compensation gains, should apply the correction on the raw RSSI values:

$$Q_{j\texttt{Id}}(n) = RSSI_{j\texttt{Id}}(n) + G_{rx}(j) + G_{tx}(\texttt{Id}). \tag{3}$$

where the correction gains have been defined above.

`Bufferize_RSSI()` The RSSIs and the corresponding timestamps, associated to a series of HELLO packets sent by the same `Id` are bufferized and represent what we call a contact. These contacts are sent to the application.

`Filter_contact()` This function suppresses contacts made of less than D_{th}. By default, $D_{th} = 2min$.

`Compute_score` This function computes for each contact, a scoring vector. Three sub-functions are used :

- Clipping : RSSI values higher than $RSSI_{th}$ are clipped. Indeed, we observed some HELLO packets with critically high RSSI values. However to keep this information, this RSSI value is replaced by the min of its two adjacent RSSI values.
- Windowing : to compute a score as a function of time, a sliding window of length T_{win} is used with an overlap $T_{overlap}$. Default values are $2min$ and $1min$ respectively.

- Fading reduction : in each window, the mean power is computed according to (8) in section 3.3 where $P_{j\texttt{Id}}[k], k \in (1, N)$ stand for the corrected RSSI values received in the current window. The interest of this averaging in the linear domain relies on the fact that it enhances the impact of high RSSI values, which are more relevant, than low values which can be due to a long distance or to a deep fading. The output is the averaged RSSI $\hat{P}_{j\texttt{Id}}(k)$ produced each minute during a slot of at most 15 minutes.

- Score computation : Taking a $15min$ vector of average RSSI values (in the nominal setup we should have 15 values), the scores is computed according to the risk model described in section 5. To be precise, from the vector of compensated and averaged RSSI values $\hat{\mathbf{Q}}_{j\texttt{Id}} = [\hat{Q}_{j\texttt{Id}}(1), \ldots, \hat{Q}_{j\texttt{Id}}(K)]$, the risk parameters are computed.

`Compute_cum_score` This function computes a unique risk score according to (52).

`Filter_cum_score` This is an optional function which allows to keep at most N_{max} contacts per slot. Additional information such as the number of simultaneous contacts can be used for further improvement of this filter.

`Aggregate_risk` As described in section 6.0.3, the probability of contamination aggregates all cumulative scores according to (56) or (57). This function is hosted at the server.

3 Propagation model and parameters adjustment

In this section we first describe the main aspects of the propagation model which relates the received powers to the emitted power and the distance, using various nuisance parameters. We then discuss averaging techniques to reduce the effect of nuisance parameters such as fading and shadowing, and experiments to assess the performance of these techniques.

3.1 Propagation model

The individual RSSI measures suffer from static errors, shadowing and fading.

The received power at a receiver j from a transmitter i is known to be governed by the following law

$$P_{ji}(t) = P_E + G_E + G_R + L_0 - L(d_{ji}) + S_{ji}(t) + F_{ji}(t), \tag{4}$$

where $P_{ji}(t)$ and the emitted power P_E are in dBm. The distance d_{ji} between transmitter and receiver is in meters. The emitting gain Δ_E, receiving gain Δ_R (which will be respectively compensated by G_{tx} and G_{rx}), the theoretical pathloss function $L(\cdot)$, and the theoretical reference pathloss at one meter $L_0 = L(d = 1m)$ (which should be close to the experimental one, `ref_loss`, cf (1)), are all in dB, as well as the shadowing offset S_{ji} and the fading offset F_{ji}. The time dependence t indicates the quantities that are evolving with time.

Considering a reference distance $d_{\max}$, in the absence of any shadowing or fading, the associated *theoretical uncompensated received power* is

$$P(d_{\max}) := P_E + G_E + G_R + L_0 - L(d_{\max}), \tag{5}$$

and we can rewrite the actually received uncompensated received power through the following form

$$P_{ji}(t) = P(d_{\max}) + L(d_{\max}) - L(d_{ji}) + S_{ji}(t) + F_{ji}(t).$$

The pathloss function $L(d_{ji})$ is modeled with the very usual model [8] as

$$L(d_{ji}) = 10 \cdot n \cdot \log_{10}\left(\frac{d_{ji}}{d_{\max}}\right) + L(d_{\max})$$

where n is the pathloss slope coefficient. Then, a relative pathloss can be defined as:

$$L_r(d_{ji}) = 10 \cdot n \cdot \log_{10}\left(\frac{d_{ji}}{d_{\max}}\right), \tag{6}$$

leading to the reference equation describing the uncompensated received power as:

$$P_{ji}(t) = P(d_{\max}) - L_r(d_{ji}) + S_{ji}(t) + F_{ji}(t). \tag{7}$$

In this document, and for the model, we will use $d_{\max} = 5m$. The values of the pathloss slope n and of the theoretical uncompensated received power $P(d_{\max})$ are critical modeling values that can only be estimated from real data and for a given pair of devices. However, knowing the compensation gains allows to apply the same model to other devices.

In a real environment, the received power suffers from two additional offsets: shadowing S_{ji} and fading F_{ji} effects. Although both effects are related to the complexity of the environment, and to multipath propagation, they characterize two different phenomena. The shadowing, also called *slow fading* represents the randomness introduced by obstacles and surrounding reflections due to human bodies, walls, furniture [9, 10, 11]. It is called slow fading because its effect is rather static and evolves only if people move significantly. Therefore, in the target application, here, for one estimation of several minutes this shadowing variable is constant and introduces a strong bias on the average power received during one observed slot. This shadowing effect usually behaves like a Gaussian random variable of zero mean. Its standard deviation is a critical parameter of the model and represents the shadowing strength. We denote it σ_{Sh}. Typically, the better the calibration process, the smaller the shadowing strength.

The fading F_{ji} refers to so-called *fast fading*. Typically it introduces a random effect due to the random summation of multipaths [12]. It varies very quickly, and at the acquisition rate we are concerned with (about $1Hz$), we can consider that each RSSI measure provides a new realization of this fading parameter. This random variable in the linear domain usually follows an exponential law, or more generally a Nakagami-m law. Due to fading, a one shot RSSI value may suffer from deep fading such as a loss of more than $20dB$! This is why it appears critical to suppress the fading effect by applying an averaging over multiple RSSI measures.

3.2 Fading reduction

The fading $F_{ji}(t)$ is due to multipath propagation and induces fast power variations. Typically, the received power (at a rate below $R_b <= 1Hz$) is corrupted by a strong fading. The analysis of real traces shows that the probability distribution of the fading effect is close to a Rayleigh fading in almost all scenarios. According to [13] and other references, the worst situation comes when the fading is Rayleigh, i.e. the received power follows an exponential distribution. If the shadowing is constant over the observations, the power variations are only due to fading. A good estimator that allows to reduce fading effect under these assumptions is derived by averaging in the linear domain according to:

$$\hat{P}_{ji} = 10 \cdot \log_{10}\left(\frac{1}{m}\sum_{k=1}^{m} 10^{P_{ji}(k)/10}\right), \tag{8}$$

where $P_{ji}(n)$ are the individual log-domain power measures and m is the number of independent measures.

To evaluate the gain of doing this averaging, let us have a look at the distribution of the variable F_{ji}. It is exponentially distributed in the linear domain: as $\mathsf{X} := 10^{\mathsf{F}/10}$ is a power we have

$$\mathsf{X} \sim \exp(-x). \tag{9}$$

Then, the log-fading follows :

$$\mathsf{F} \sim c \cdot \exp(cf - e^{cf}), \tag{10}$$

with $c = \log(10)/10$.

When multiple measures $X_{ji}(k)$, $1 \leq k \leq m$, are obtained with the same shadowing and distance conditions, the linear average estimator is given by :

$$\bar{X}_{ji} = \frac{1}{m} \sum_{k=1}^{m} X_{ji}(k), \tag{11}$$

which follows the distribution

$$\bar{X}_{ji} \sim \frac{m^m}{(m-1)!} \cdot \bar{x}^{m-1} \exp(-m\bar{x}). \tag{12}$$

Then, the linear domain log-fading $\bar{F}_{ji} = 10 \cdot \log_{10}(\bar{X}_{ji})$, follows :

$$\bar{F}_{ji} \sim \frac{cm^m}{(m-1)!} \cdot \exp\left(cm\bar{f} - m \cdot e^{c\bar{f}}\right). \tag{13}$$

The corresponding distributions are shown in Fig.2. The asymmetry of the distribution for small

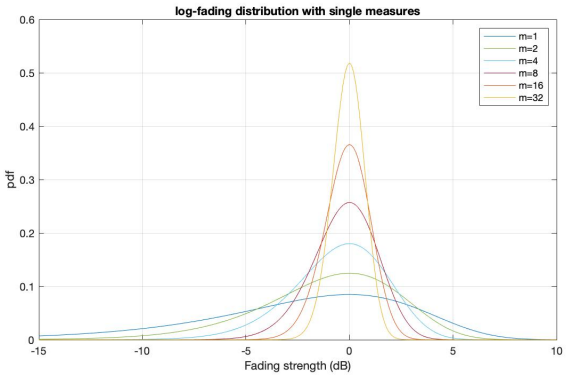


Figure 2: The fading distribution as a function of the number of measures is represented. For $m = 1$ the distribution is strongly asymmetric and presents deep fading which affect the risk estimation probability.

m is clearly visible, as well as the large spreading of the distribution. The standard deviation for different values of m is summarized in Table 2.

The impact of the fading does not exceed roughly $1dB$ in terms of standard deviation when at least 8 values are used for the averaged RSSI, which is made feasible, at a rate $R_b = 0.25Hz$ and an averaging over $2min$ slots. It is worth mentioning that this correction is valid only in a

m	1	2	3	4	5	8	16	32
std σ_{fad} (in dB)	$5,6$	$3,5$	$2,7$	$2,3$	2	1.6	1.1	0.78

Table 2: Standard deviation (std) related to fading after averaging over m values.

relatively static scenario. Luckily, this is precisely the most relevant scenario for the considered application.

Consequently, we consider that if the number of RSSI measurements is less than 5, the confidence in the resulting averaged RSSI (and therefore in the subsequent risk score) will be low, and should be accounted for in the algorithm. For this reason, specific corrections have been introduced in Table 5 for cases where $m < 5$.

These results provide a justification for the parameters used in the procedure `Compute_mean_RSSI()`, described in Section 2. In particular, using windows of $2min$ with an overlap of $1min$, provides a means of obtaining an averaged RSSI, each minute, with a sufficiently low variance.

3.3 Impact of Shadowing

Once this fading compensation has been applied, any known correction gains G_{tx} or G_{rx} (see Section 2) should be applied to replace the averaged RSSI measure $\hat{P}_{ji}$ (see (8)) by a compensated version. Indeed, when no compensation is applied here, then additional uncertainties increase the shadowing effect. For instance, it has been observed that the static gains $G_{0,tx}$ and $G_{0,rx}$ may be as large as $\pm 20dB$. As such, if these gains are not known, they can increase the shadowing offset by more than $10dB$.

If it is possible to detect specific postures of the device (phone in hand *vs* phone in the pocket) or to estimate characteristics of its environment (crowded area, ...), this information may contribute to better predict the shadowing offset as described now, and therefore to eventually improve the estimation of the risk score. These posture / environment dependent compensations take place in the `Gain_compensate()` procedure, which we now describe.

Assume that the fading has been removed (or the remaining effect is included in shadowing), and known static gains have been applied.

Denoting $P_0(d_{\max}) := P_E + L_0 - L(d_{\max})$ the theoretical *ideally compensated* received power at distance $d_{\max}$ (compare with (5)), we then have a compensated and averaged RSSI measure given by (cf (7)):

$$\hat{Q}_{j\mathtt{Id}}(t) = P_0(d_{\max}) - L_r(d_{ji}) + \tilde{S}_{ji}(t), \tag{14}$$

where now $\tilde{S}_{ji}(t) = (G_E + G_{tx}) + (G_R + G_{rx}) + S_{ji}(t) + \tilde{F}_{ji}(t)$ accounts for both the shadowing effect $S_{ji}(t)$ and residual errors in compensating the emission/reception gains and the fading offset. This compensated and averaged RSSI, $\hat{Q}_{j\mathtt{Id}}(t)$, is computed each minute during time slots of $15min$, for the temporary Id `Id` of transmitter i. The remaining shadowing $\tilde{S}_{ji}(t)$ is a random variable which is known to follow a normal distribution $\mathcal{N}(0, \sigma_{Sh})$. Without shadowing we would have a direct relation between the $\hat{Q}_{j\mathtt{Id}}$ and the distance d_{ji}. Unfortunately the shadowing introduces strong perturbations in this relation and has to be carefully considered. The impact of this shadowing is studied in Section 4 and is accounted for in two ways :

1. Compensation : the averaging function used over each slot allows to compensate the shadowing if some variations occur during the $15min$ slots. This is likely to appear, in crowded

environments and when people are moving around the devices involved in a contact. However, in some configuration (long meetings in a room, with fixed situations), the shadowing may remain strong for a long time.

2. Risk scoring : the uncertainty about the power level is included in the model to take into account the shadowing distribution. The risk is weighted according to the confidence we have on the measures, in order to minmize the error probability on the risk at the end.

3.4 Description of the Dataset

To parameterize the model, we used the dataset described in [3]. All individuals involved in the experiment we used for this evaluation hold their phones in their hands. All devices used in these experiments are SAMSUNG SM-A405SFN. Because the final values of the compensation gains are not yet publicly available, we work in the rest of this document, with the raw RSSI values.

An example of a trace is provided in Fig.3, corresponding to Room1.

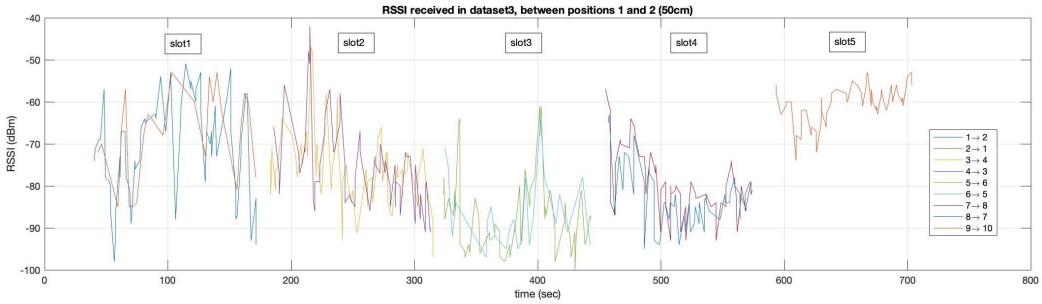


Figure 3: These plots illustrate the variability of the raw RSSI values since they corresponds to successive slots, with two soldiers positioned at the same places, marked on the ground. In this example, the inter-body distance is $50cm$. Fading is characterized by the fast RSSI variations (along a slot) while shadowing is illustrated by the difference in average power between the different slots.

As shown in this Figure, the variability of the RSSI is large. Typically, the average signal strength observed in slot 3 is much lower with no evident justification. A rough analysis of the traces gives an idea of the difficulty to predict a distance or a contamination risk from these measures. But, as suggested in [7], the true question is to evaluate the achievable tradeoff between false positive and true positive decisions.

Possible reasons for significantly lower RSSI in some scenarios include:

- Uneven behavior of the phones. Further analyses by comparing the traces over time may help to see if there is a correlation with the phone Id.
- Unlucky fading situation which creates an attenuation by self-interference. Yet, since fading variations are clearly seen, this is probably not the case.
- A shadowing occurs (the phone is handled in the other hand or is masked by the owner). Note that in the scenario there are no other people between the devices.
- The overall environment changes. It is worth noting that during slots 3 and 4, the environment is more crowded.

In addition, in this dataset, we see that the RSSI trace from transmissions between location 10 and location 9 in Room 1 is absent for some technical reason. For the other traces, we can observe that the symmetry of the links is quite good. It means that using the reverse pathloss link should not bring too much information to stabilize the estimates.

Interestingly, to give an idea of the variability of the RSSI measures with the distance we give in Fig 15 (in the appendix) the raw RSSI data received by receiver 2 from transmitter 1. They both enter the playground at slot 1, they move simultaneously according to the two different paths described in [3]. The distance between them is successively $[50cm, 200cm, 403cm, 180cm, 364cm]$. In this scenario, slot 1 presents a high risk, slots 2 and 4 are moderately critical and slots 3 and 5 are not risky.

A further illustration of the main problem we have to deal with is in room 4, during slot 1. The distance is only $50cm$ but the RSSI is much lower than the RSSI measured during the following sets. An important message highlighted in these traces is the necessity to compensate for fading first and then to track shadowing variations and to compute a risk estimate accounting for uncertainties in the RSSI measurements.

3.5 Pathloss models, mean power evaluation and fading compensation

The objective of this section is to evaluate the model parameters from real data.

The mandatory parameters are the received RSSI ($P(d_{\max} = 5m)$), the attenuation slope (n) and the shadowing strength.

We use the whole dataset, acquired in 5 different rooms (3 indoor, 2 outdoor). We extracted the theoretical ground truth distances for all traces and we compare the curves providing the mean RSSI as a function of distance (see Fig.4).

The estimation of the average distance obtained with three estimators is provided : classical average power (red), average power in the linear domain (magenta, according to Eq.(8)) or by averaging over the 25% highest values (blue).

The optimal one is the one that presents the highest ratio between the pathloss variation with distance and the standard deviation. This occurs for the linear domain average, and justifies the use of Eq.(8) to reduce fading.

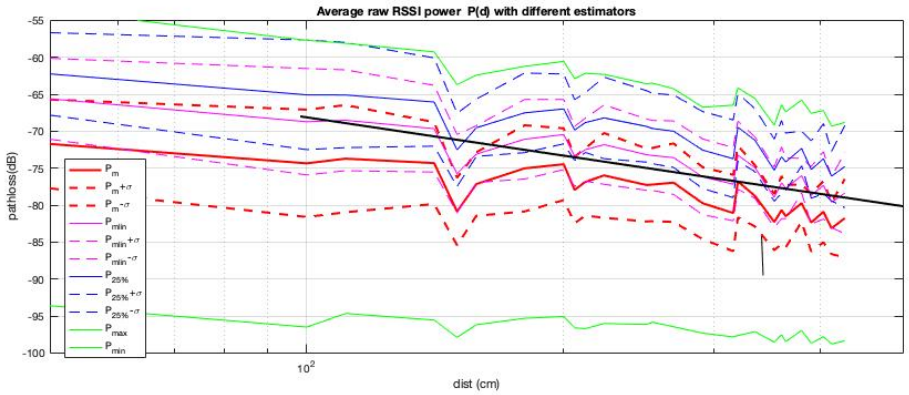


Figure 4: Average received power as a function of the distance for three estimators (mean power, linear domain mean power, and mean of the 25% largest values.

From these data we can determine an average pathloss model given by $P(d_{\max} = 5m) = -80dBm$, $n = 1.7$ and $\sigma_{Sh} = 5dB$.

The slope has been chosen to fit the measurements in the range $[1m, 4m]$ which is the most important part to discriminate the risk. Of course, this model has been built with a specific phone model. For future exploitation, the calibrated value of $P_0(d_{max})$ should be computed by adding the compensation factors. Interestingly, Fig.5 shows the power distribution for three classes

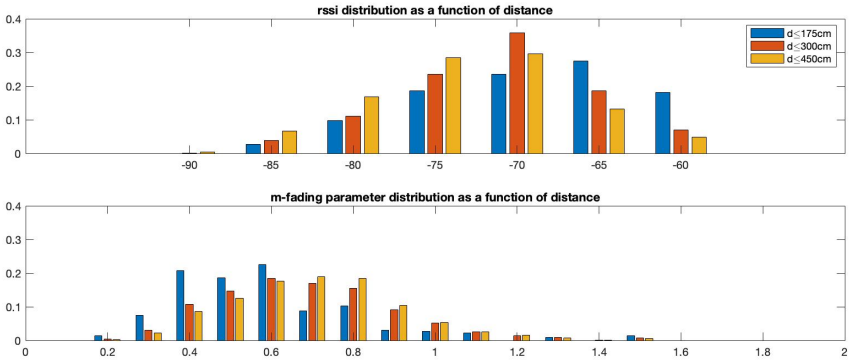


Figure 5: Histogram of the averaged raw RSSI values over dataset1, for three distance classes $(C1 : d < 175cm, C2 : 175cm < d < 300cm, C3 : 300cm < d < 450cm)$

of pairwise devices, based on the ground-truth distance, which is known in the experimental dataset.

Clearly the overlap between the different classes is important. Let us emphasize that in this model the fading was carefully compensated by an averaging process over $10min$ traces. The estimation of the fading strength parameter m (used to define the linear domain averaging (8) for $m = 1$) exhibits an interesting right shift with distance. However, its variability is too high to make it relevant for classification.

The main parameters and notations used in this section and the next one are summarized in Table 3.

Name	Role	Nominal value
d_c	Critical distance for risk evaluation	$2m$
d_{max}	Reference distance for pathloss	$5m$
$P^*(d_{max})$	Non calibrated reference RSSI at d_{max}	$-80dBm$
σ^*_{Sh}	Shadowing strength	$4dB$
$n*$	Pathloss slope coefficient	1.7

Table 3: List of parameters used in the propagation model and the risk model. The parameters with the symbol .* are those for which the nominal values have been estimated from the PEPP-PT dataset. They are used to compute the parameters P^*_{min} and Δ_P used in the algorithm. Note that at the time of this study, the calibration gains were not known and all the parameters have been computed without compensation for static gains. To be reused with other dataset, these parameters should be compensated with the GSMA gain factors. .

4 Distance-RSSI relation in a Bayesian formalism

In light of the model (14), we now develop a Bayesian model of the conditional probability distribution of the distance given the observed (calibrated) RSSI values.

4.1 Mathematical notations

We adopt the following notations: x, $\boldsymbol{x}$, $\boldsymbol{X}$ respectively denote deterministic scalar, vector and matrix quantities and X, $\mathbf{x}$, $\mathbf{X}$ respectively denote random scalar, vector and matrix variables. Moreover, we denote by $(\cdot)^*$, $(\cdot)^\mathsf{T}$ and $(\cdot)^\mathsf{H}$, the conjugate, transpose and conjugate transpose operations respectively.

4.2 Bayesian formalism

We denote d the distance between a pair of devices. In this section the objective is to establish a Bayesian model characterizing the relation between d and the Bluetooth Low Energy (BLE) RSSI measurements gathered from HELLO packets.

Consider the scenario where mobile i acts at the transmitter (Tx) and j as the receiver (Rx). Of course, the process is in practice bi-directional, but we consider estimating the risk from the point of view of a receiver, without any feedback from j to i. While the symmetry of the links could be used to strengthen the estimation but such approach would need additional information in the HELLO packets.

As mentioned in the former section, the raw RSSI measurements evaluated from one receiver are subject to various effects, including fading, shadowing, and receiver noise. Receiver noise and fading are short term effects and can be significantly reduced with local averaging, as already presented in section 3.3. Therefore, in the rest of this section, we assume that an estimation of the average RSSI has been done, after the application of all known compensations. The remaining RSSI vector is still affected by a shadowing effect. The received RSSI is given by (14) and (6) hence, given the two devices i and j communicating at a given distance d, the pathloss (PL) between both devices, assumed to be symmetric, is a random variable given by:

$$\hat{\mathsf{Q}} = P_0(d_{\max}) - a\log(d/d_{\max}) + \mathsf{S}, \tag{15}$$

where $P_0(d_{\max}), d_{\max}, a$ are the model parameters (given in the former section), and S is a random variable which stands for shadowing and other effects related to the environment. Parameter $P_0(d_{\max})$ is the theoretical compensated average power observed when $d = d_{\max}$, and parameter a is related to the pathloss slope parameter n in (6) by $a := \frac{10}{\log(10)}n$. As we have seen in section 3, the measurements done by the Bundeswehr [2, 3] in the PEPP-PT project give credit to such a model.

Consider now the distance-RSSI problem reformulated for the scenario of interest. From a given signal duration acquired within one 15-minute time slot, a compensated average RSSI is provided according to the algorithm described above and leads to $\hat{\mathsf{Q}}$. It is further assumed that the measures are made for a distance in the range $d \in [0; d_{max}]$, at most. For the sake of simplicity and without loss of generality, this maximum distance is set to the reference distance $d_{\max}$ in the model (15). In the experimental part, we will use $d_{\max} = 5m$.

Let be define the gap Γ between the theoretical compensated average received power $P_0(d_{\max})$ and the measured value as

$$\mathsf{\Gamma} = \hat{\mathsf{Q}} - P_0(d_{\max}). \tag{16}$$

Clearly if the ground truth distance is $d = d_{\max}$, $\mathsf{\Gamma} = -\mathsf{S}$. Otherwise,

$$\mathsf{\Gamma} = -a\log(d/d_{\max}) - \mathsf{S}. \tag{17}$$

Now, the problem can be clearly identified as an estimation problem in a noisy channel.

In order to estimate a contamination risk, which is assumed to be directly related to the distance d and the duration of the contact, we propose to exploit a Bayesian formulation of the problem. For the moment, we consider a static distance to be computed from Γ.

The Bayesian formulation leads to:

$$p_{\mathsf{D}|\Gamma}(d|\gamma) = \frac{p_{\Gamma|\mathsf{D}}(\gamma|d) \cdot p_{\mathsf{D}}(d)}{p_{\Gamma}(\gamma)}; \tag{18}$$

where $P_{\mathsf{D}}(d)$ is the prior knowledge on d, $p_{\Gamma}(\gamma)$ is the marginal of Γ and $p_{\Gamma|\mathsf{D}}(\gamma|d)$ is its likelihood given d. In the StopCovid problem we are interested in the posterior distribution $p_{\mathsf{D}|\Gamma}(d|\gamma)$ since it provides a basis to construct risk scores. For example, given an observation γ, the conditional probability density function (pdf) of the distance can be used to evaluate the CDF $F(d_c|\gamma) = P(d \leq d_c|\Gamma)$. Here, d_c is the maximal distance at which the virus can be transmitted. We note that epidemiologist have evaluated a typical maximal distance to be $d_c = 2m$, but this can be extended to $d_c = 10m$ when the persons holding the devices are cycling.

We will discuss the transformation of this posterior probability into a risk score in next section. We however anticipate that such risk measure should answer the question: *Given an observed average RSSI, what is the probability that $d \leq d_c$?*

4.3 Prior information

The Prior probability represents the prior knowledge about the distance d. We will essentially work with a probabilistic model assumes that the devices are located uniformly at random over the area. That is, for a pair of devices i, j, fix the position of device i and then consider that device j is located uniformly at random in the disc of radius $d_{\max}$, the maximal distance we are concerned with. Therefore, if device j is randomly positioned in this disc, the pdf of the distance is :

$$p_{\mathsf{D};d_{\max}}(d) = 2 \cdot \frac{d}{d_{\max}^2}. \tag{19}$$

As we will see later, it may be more convenient to work with a logarithmic distance that we define as

$$\mathsf{Z} := -\log(\mathsf{D}/d_{\max}). \tag{20}$$

The pdf of this inverse log distance is thus given by:

$$p_{\mathsf{Z};d_{\max}}(z) = 2 \cdot e^{-2z}, \tag{21}$$

where $z \in [0; \infty)$. Note the correspondence $d = d_{\max} \to z = 0$ and $d = 0 \to z = \infty$.

Remark. *This basic model will be used as the reference model, but it does not take into account any social effect and considers a pure uniform random distribution of the device's holders. This may overestimate the prior probability of small distances. Social models could suggest to introduce repulsive effects with a distance law such as :*

$$p_{\mathsf{D}}(d) = 1 - e^{-\alpha d^2}, \tag{22}$$

where α is a parameter of the model. This is not pursued further in this document

4.4 Likelihood function

Consider the likelihood of the gap $\mathsf{\Gamma}$, cf (17), conditional to the "log-distance" Z, cf (20), namely $p_{\mathsf{\Gamma}|\mathsf{Z}}(\gamma|z)$. From the model given in (17), we have:

$$\mathsf{\Gamma} = a \cdot \mathsf{Z} - \mathsf{S}. \tag{23}$$

From this equation it follows that the likelihood $p_{\mathsf{\Gamma}|\mathsf{Z}}(\gamma|z)$ is determined by the pdf of S.

It is commonly admitted that this shadowing variable follows a normal distribution in many radio environments, and this seems to be partially confirmed by the experimental data provided by the BLE-RSSI in PEPP-PT [1, 2, 3].

Therefore, the likelihood follows a normal law:

$$p_{\mathsf{\Gamma}|\mathsf{Z}}(\gamma|z) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(\gamma - a\cdot z)^2}{2\sigma_{Sh}^2}}. \tag{24}$$

where a is the pathloss coefficient already defined, and σ_{Sh}^2 is the shadowing variance. The value of these coefficients can be tuned to fit with experimental data. In Section 3, we found $n = 1.7$ and $\sigma_{Sh} = 5dB$ and we recall that $a = 10n/\log(10)$.

4.5 Posterior distribution

The derivation of the posterior probability follows (18), but now expressed as a function of Z:

$$p_{\mathsf{Z}|\mathsf{\Gamma}}(z|\gamma) = \frac{p_{\mathsf{\Gamma}|\mathsf{Z}}(\gamma|z) \cdot p_{\mathsf{Z}}(z)}{p_{\mathsf{\Gamma}}(\gamma)}; \tag{25}$$

Injecting the derivations above, one obtains:

$$p_{\mathsf{Z}|\mathsf{\Gamma}}(z|\gamma) = K_0(\gamma) \cdot \exp\left[-\frac{\left(z - \frac{\gamma}{a} + \frac{2\sigma_{Sh}^2}{a^2}\right)^2}{2(\sigma_{Sh}/a)^2} \right] \tag{26}$$

where $K_0(\gamma) = \left[\int_0^\infty \exp\left(-\frac{(z-\gamma/a+2(\sigma_{Sh}/a)^2)^2}{2(\sigma_{Sh}/a)^2}\right) dz\right]^{-1}$ is a normalization constant. Note that even if this posterior distribution looks to be normally distributed, this is not exactly the case since in the model z is restricted to the half-line $[0; +\infty)$. This is therefore a truncated normal distribution, where $K_0(\gamma)$ ensures the normalization of the probability density function. The normalization constant can be either computed numerically or from the marginal, likelihood and prior pdfs. However, in some cases, we may not need to compute it explicitly.

4.6 Estimation metrics

From this posterior distribution, different estimators of the log-distance can be derived:

1. The maximum likelihood estimator is the value z which maximizes the likelihood:

$$\hat{z}_{\mathtt{ML}} = \frac{\gamma}{a}.$$

This estimator does not take into account the prior information.

2. The maximum a posteriori (MAP) maximises the posterior distribution, leading to:

$$\hat{z}_{\texttt{MAP}} = \begin{cases} \frac{\gamma}{a} - \frac{2\sigma_{Sh}^2}{a^2} & \text{if} \quad \gamma \geq \frac{2\sigma_{Sh}^2}{a} \\ 0 & \text{otherwise} \end{cases}.$$

   Compared to the ML estimator, this estimator takes into account the prior distribution. When γ is smaller than the threshold $2\sigma_{Sh}^2/a^2$ the estimate of the log-distance corresponds to a distance $\hat{d} = d_{\max}$.

3. The MMSE (minimum mean square error) is given by

$$\hat{z}_{\texttt{MMSE}} = \mathbb{E}_{\texttt{Z}|\Gamma}(z),$$

   i.e. the average of the posterior distribution. If the truncation of the normal distribution is neglected, then this estimator is equal to the MAP, otherwise, the average value needs to be computed.

For the purpose of computing the infectiousness risk, the posterior distribution of the risk is more powerful and may help to do classification more efficiently. Typically, an hypothesis test can be conducted (from a Bayesian or a Neyman-Pearson test), allowing to adapt efficiently the balance between false alarm (FA) and misdetection (MD) rates.

4.7 Simulation results

Using the parameters obtained in Sec.3, i.e. with $P(d_{\max}) = -80dBm$ and $n = 1.7$, the Bayesian model is used, with a maximal coverage range equal to $5m$. Figure 6 below plots the cdf of the distance probability as a function of a given raw RSSI measure for three different shadowing values

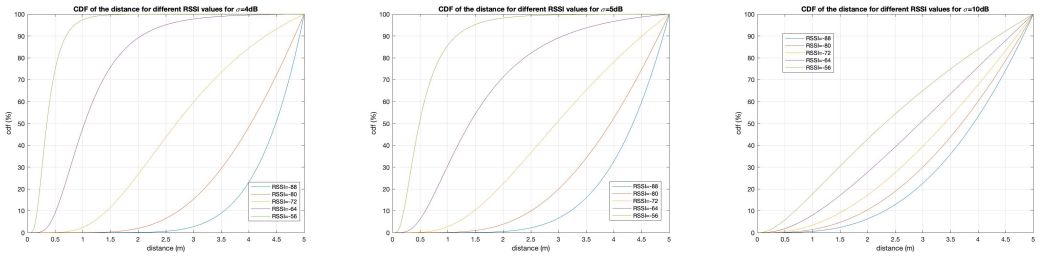


Figure 6: CDF $p_{D|RSSI}(d|rssi)$ for different raw RSSI values (not compensated by the static gains) and for three values of σ_{Sh}, $(4dB, 5dB, 10dB)$.

On Fig 6, the probability that the ground truth distance is below some value for a given averaged raw RSSI, is estimated. We see that even a small variation in σ_{Sh} has a strong impact on the accuracy of the risk prediction. On the left figure with $\sigma_{Sh} = 4dB$, a raw $RSSI = -64dBm$ allows to ensure that $Pr(d < 2m|RSSI) = 0.9$. But on the center, with $\sigma_{Sh} = 5dB$, the probability reduces to $Pr(d < 2m|RSSI) = 0.7$. On the right, with $\sigma_{Sh} = 10dB$, this probability reduces to 45%.

This is why a good calibration is necessary. Therefore, many efforts should be done to reduce this variability with contextual and posture compensations.

Of course, the value of the threshold relies on the model, and especially on the standard deviation of the shadowing, as illustrated in Fig.7.

Clearly, when the standard deviation of the shadowing is given by $\sigma_{Sh} = 4dB$, the accuracy of the distance's measure increases very significantly while a value of $\sigma_{Sh} = 10dB$ contains only very few information (at $10dB$, it seems not reasonable to compute a proximity measure, and this is the level of shadowing we can expect if the prediction is based on single RSSI values). The flatness of the curve at $\sigma_{Sh} = 10dB$ shows the poor selectivity of the RSSI information.

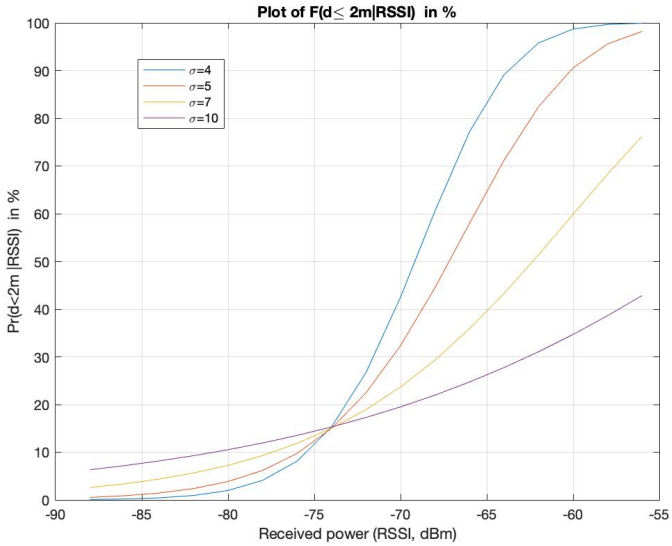


Figure 7: Cumulated distributed function (cdf) of the conditional probability $F(d \leq 2m|rssi)$ when the rssi is the compensated and averaged RSSI, as a function of the shadowing strength.

5 Infectiousness model: risk evaluation

The infectiousness model aims at computing an infection risk from a contact, defined by an exposition duration and a distance. To the best of our knowledge, such a model does not exist. From the state-of-the-art literature, a high risk can be assumed when someone is exposed to an infected person at a distance of 2 meters or less for a duration of 15 minutes or more. Of course this risk depends on many factors : are the people face to face ? are they following safety rules (mask, etc...)? As such information is not available we consider a risk averaged on the (unknown) distribution of these hidden latent factors.

Let start by some general considerations and reminders on the main properties of the ROBERT protocol. At a given time, each user X has a random infection state $\mathsf{IS}_X \in \{I, \bar{I}\}$, where I means infected, and $\bar{I}$ means not infected. When two users A and B get sufficiently close for their mobiles to establish a BLE connection, the system uses BLE RRSI measurements to register a (pseudonymized) contact between them, and an objective is to compute a risk of contamination. After few days, if patient B is declared infected, the system has to determine the probability that patient A is infected given the assumption that B was infected when the contact occurred:

$$\mathbb{P}\left(\mathsf{IS}_A = I | \mathsf{IS}_B = I; c_{AB}\right) \tag{27}$$

where c_{AB} gathers characteristics of the contact, such as its duration and aggregated information

about the distance between A and B during this contact. We investigate a model of how the risk probability scales with time and distance.

5.1 Additivity of the risk with respect to time

Fortunately, the time dependence of the risk can be determined by considering simple probability rules. Denote $c = c_{AB}$ for short and let q be the conditional safety probability for patient A:

$$q = \mathbb{P}\left(\mathsf{IS}_A = \bar{I} | \mathsf{IS}_B = I; c\right). \tag{28}$$

Then, assume that patients A and B have been in contact two times, according to contact characteristics c_1 and c_2. By a natural consideration, we assume that A is safe only if he/she is kept safe after each contact. Assuming statistical independence, this yields

$$\mathbb{P}\left(\mathsf{IS}_A = \bar{I} | \mathsf{IS}_B = I; c_1, c_2\right) = q_1 \cdot q_2. \tag{29}$$

Clearly the overall safety is the product of by-contact safety variables. Given that q is a probability, define a risk level as:

$$r = -\log(q), \tag{30}$$

where r is variable defined on $\mathbb{R}^+$. By definition, the probability of being infected is related to the risk $r(c)$ by:

$$\mathbb{P}\left(\mathsf{IS}_A = I | \mathsf{IS}_B = I; c\right) = 1 - e^{-r(c)}. \tag{31}$$

Then, we arrive to the first important property of successive exposure

Proposition 1 (Risk additivity)**.** *When a patient A is exposed to patient B at the occasion of two contacts c_1 and c_2, the risk function r is additive*

$$r(c_1, c_2) = r(c_1) + r(c_2).$$

Proof. The proof follows from the definition of the risk function and the fact that the safety probability is the product of by-contact safety probabilities (assuming statistical independence). □

By extension, for a contact trajectory $\boldsymbol{c} = \{c_i; i \in [1, N]\}$ defined by a series of contacts we have

$$r(\boldsymbol{c}) = \sum_{i=1}^{N} r(c_i) \tag{32}$$

The probability of being infected during a contact is obviously a function of the distance d between users A and B and of the duration t of their contact. If $\boldsymbol{c}$ is now defined by a continuous trajectory of the distance $d(t)$ as a function of time, $t \in [0, T]$, then the total risk becomes:

$$r(\boldsymbol{c}) = \int_{t=0}^{T} R(d(t))dt. \tag{33}$$

where R is a monotonic decreasing function of distance, to be modeled.

If this trajectory $d(t)$ is a piece-wise constant function, then $\boldsymbol{c}$ turns back to a discrete vector, and (32) becomes:

$$r(\boldsymbol{c}) = \sum_{i=1}^{N} R(d_i) \cdot \Delta t_i. \tag{34}$$

A "ground truth" summary of the contacts between users A and B if therefore $c_{AB} = \{(\Delta t_i, d_i)\}_{i=1}^{N}$ and allow to compute the conditional probability of infection once we will have the function $R(.)$.

To summarize, while the dependency of the risk with respect to time is linear by property of the safety probability, its dependency with respect to the distance d remains unknown.

5.2 Possible models of the instantaneous risk as a function of distance

As we know that the distance measure from RSSI vectors is very uncertain, the simplest model is almost surely the best one to use in the StopCovid context.

Possible functions as suggested in [4] to model the risk as a function of distance are:

$$\begin{array}{llcll} 1. & \text{Heaviside} & R_H(d) & = & \begin{array}{ll} r_0; & d \le d_c \\ 0 & \text{otherwise} \end{array} \\ 2. & \text{Piecewise linear} & R_L(d) & = & \begin{array}{ll} r_0; & d \le d_c \\ r_0 \frac{d_0 - x}{d_0 - d_c} & d_c < d < d_0 \\ 0 & \text{otherwise} \end{array} \\ 3. & \text{Exponential} & R_E(d) & = & \begin{array}{ll} r_0; & d \le d_c \\ r_0 e^{d_c - d} & \text{otherwise} \end{array} \end{array} \tag{35}$$

where d_c is a critical distance and r_0 is parameter characterizing the overall risk level, to be determined by experts in epidemiology. To the best of our knowledge, a reasonable choice could be $d_c = 2m$ in a static or walking situation, while $d_c = 10m$ could be more appropriate for people cycling. Therefore for a given ground truth trajectory $d(t)$, $t \in [0, T]$ one obtains the following risk measures, where $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ should be interpreted as the integrals corresponding to the "probability" and "expectation" over the values taken in the considered time interval (with a uniform distribution):

1. Heaviside function : with $R_H(\cdot)$, the risk associated to a trajectory $d(t)$ is

$$r_H(\boldsymbol{c}) = r_0 \cdot T \cdot \mathbb{P}(d \le d_c). \tag{36}$$

2. Piecewise linear function : with $R_L(\cdot)$, the risk is now

$$r_L(\boldsymbol{c}) = r_0 \cdot T \cdot \left(\mathbb{P}(d \le d_c) + \mathbb{E}\left[\mathbb{1}_{\{d_c < d < d_0\}} \cdot (d_0 - d)\right] / (d_0 - d_c)\right) \tag{37}$$

3. Exponential function : with $R_E(\cdot)$, the risk is now

$$r_E(\boldsymbol{c}) = r_0 \cdot T \cdot \left(\mathbb{P}(d \le d_c) + \mathbb{E}\left[\mathbb{1}_{\{d \ge d_c\}} \cdot e^{d_c - d}\right]\right) \tag{38}$$

Note that for a constant ground truth distance $d(t) = d$, we get :

$$\begin{aligned} r_H(\boldsymbol{c}) &= r_0 \cdot T \cdot \mathbb{1}_{\{d \le d_c\}} \\ r_L(\boldsymbol{c}) &= r_0 \cdot T \cdot \min\left(1, \frac{d_0 - d}{d_0 - d_c}\right) \cdot \mathbb{1}_{\{d < d_0\}} \\ r_E(\boldsymbol{c}) &= r_0 \cdot T \cdot \min\left(1, e^{d_c - d}\right) \end{aligned} \tag{39}$$

5.3 Expectation of the risk – notion of score

In practice, the ground truth distance trajectory is not directly accessible but only estimated from RSSI measurements, hence the corresponding risk is not directly accessible either. It can at best be estimated from concrete contact information, denoted $\tilde{c}_{AB}$, that contains (averaged and compensated) RSSI-based measures, e.g. the vector of corrected RSSI values $\hat{\mathbf{Q}}_{j\mathtt{Id}}$ defined in the former sections (with j the index of the mobile of patient B and $\mathtt{Id}$ the temporary Id of that of patient A) or sequences γ_i of gaps (see (16)) between such measures and a reference RSSI. .

The *score* is a quantity useful to assess the risk level once multiplied by r_0: it is defined as

$$\pi(\gamma) := \mathbb{P}\left(\mathsf{D} \leq d_c \mid \gamma\right). \tag{40}$$

From now on, we rely on the Heaviside risk model $r_H(\cdot)$: given the intrinsic uncertainty in estimating γ itself, and the averaging over this uncertainty provided by the following approach, it seems superfluous to use a more sophisticated parametric risk model. With this model, the expected risk as well as its variance are simple functions of the score:

$$\begin{aligned} \mathbb{E}\left[R_H(\mathsf{D}) \mid \gamma\right] &= r_0\pi(\gamma) \\ \mathtt{Var}\left[R_H(\mathsf{D}) \mid \gamma\right] &= r_0^2\pi(\gamma)(1-\pi(\gamma)) \end{aligned}$$

As illustrated on Figure 8, this allows to quantify uncertainties on the actual value of the risk given an observed value of γ. In the future, it could be envisioned to exploit these to handle false-positive /false-negative tradeoffs.

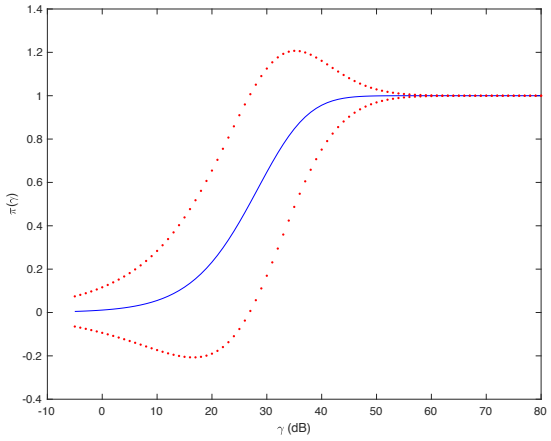


Figure 8: The plot represents $\pi(\gamma) \pm \sqrt{\pi(\gamma)(1-\pi(\gamma))}$ for $a = 5$, $d_{\max} = 10m$, $d_c = 1.8m$ and $\sigma = 7dB$, using the particular expression of $\pi(\gamma)$ from (9).

Given a local contact information $\tilde{c}_{AB} = (\gamma_i, \Delta t_i)_{i=1}^N$ and defining the cumulative score

$$\pi(\tilde{c}_{AB}) := \sum_{i=1}^{N} \pi(\gamma_i)\Delta t_i \tag{41}$$

we obtain estimates of the risk and its variance

$$\mathbb{E}\left[r_H(\boldsymbol{c}) \mid \tilde{c}_{AB}\right] = r_0\pi(\tilde{c}_{AB}) \tag{42}$$

$$\mathtt{Var}\left(r_H(\boldsymbol{c}) \mid \tilde{c}\right) = r_0^2 \sum_{i=1}^{N} \pi(\gamma_i)(1-\pi(\gamma_i))(\Delta t_i)^2. \tag{43}$$

Remark. *Importantly, this shows that we only need to be able to compute a score $\pi(\gamma)$ (or more generally $\pi(\tilde{c})$ on the mobile, as the absolute risk level r_0 can be tuned at the server layer as needed for epidemiology purposes, or to control tradeoffs between false positives and false negatives.*

Proposition 2. *With the distance model previously developed we get an analytic expression of the score. Denoting* $\texttt{erfc}(x) = \sqrt{2/\pi}\int_x^\infty e^{-t^2}dt$ *we have*

$$\pi(\gamma) = \frac{\texttt{erfc}\left[2^{-1/2}\left(\frac{\log(d_{\max}/d_c)}{\sigma_{Sh}/a} - \frac{\gamma}{\sigma_{Sh}} + \frac{2\sigma_{Sh}}{a}\right)\right]}{\texttt{erfc}\left[2^{-1/2}\left(-\frac{\gamma}{\sigma_{Sh}} + \frac{2\sigma_{Sh}}{a}\right)\right]}. \tag{44}$$

In terms of cumulative score we have

$$\pi(\tilde{c}_{AB}) = \sum_{i=1}^{N} \Delta t_i \frac{\texttt{erfc}\left[2^{-1/2}\left(\frac{\log(d_{\max}/d_c)}{\sigma_{Sh}/a} - \frac{\gamma_i}{\sigma_{Sh}} + \frac{2\sigma_{Sh}}{a}\right)\right]}{\texttt{erfc}\left[2^{-1/2}\left(-\frac{\gamma_i}{\sigma_{Sh}} + \frac{2\sigma_{Sh}}{a}\right)\right]} \tag{45}$$

Proof.

$$\begin{aligned}\pi(\gamma) = \mathbb{P}\left(\mathsf{D} \le d_c \mid \gamma\right) &= \mathbb{P}\left(d_{\max}e^{-\mathsf{Z}} \le d_c \mid \gamma\right) = \mathbb{P}\left(\mathsf{Z} \ge \log(d_{\max}/d_c) \mid \gamma\right) \\ &= K_0(\gamma)\int_{\log(d_{\max}/d_c)}^{\infty} \exp\left(-\frac{\left(z - \frac{\gamma}{a} + \frac{2\sigma_{Sh}^2}{a^2}\right)^2}{2(\sigma_{Sh}/a)^2}\right) dz\end{aligned}$$

Since $\int_u^\infty \exp(-(z-\mu)^2/2\sigma^2)dz = \int_{(u-\mu)/(\sigma\sqrt{2})}^\infty \exp(-t^2)dt\sigma\sqrt{2} = \sigma\sqrt{2}\sqrt{\pi/2}\,\texttt{erfc}((u-\mu)/(\sigma\sqrt{2}))$, with any u and $\mu := \gamma/a - 2\sigma_{Sh}^2/a^2$, $\sigma = \sigma_{Sh}/a$ we obtain

$$\pi(\gamma) = \frac{\sigma\sqrt{\pi}\,\texttt{erfc}((\log(d_{\max}/d_c) - \mu)/(\sigma\sqrt{2}))}{\sigma\sqrt{\pi}\,\texttt{erfc}(-\mu)/(\sigma\sqrt{2}))} = \frac{\texttt{erfc}((\log(d_{\max}/d_c) - \mu)/\sqrt{2}\sigma)}{\texttt{erfc}(-\mu)/\sqrt{2}\sigma)}$$

□

5.4 Improved estimates using side information

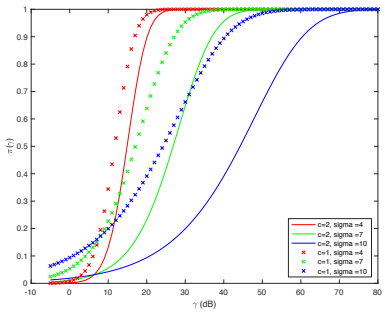


Figure 9: The plots represent $\pi(\gamma)$ for $a = 5$, $d_{\max} = 10m$, $d_c = 1.8m$ and $\sigma_{Sh} \in \{4, 7, 10\}dB$.

If side information is available, it can improve the estimate of the risk (or more precisely of the score), either by providing a better prior for the distance model, or a more accurate estimate of the shadowing variance σ_{Sh}, or a better calibration of the pathloss model. Some examples include:

- if the user is cycling, set $d_c = 10m$, otherwise $d_c = 1.8m$;
- if the phone is in the pocket (or using some other side information), shadowing is more likely hence set $\sigma_{Sh} = 10dB$
- if the number of mobiles around is high, the probability of a short distance d is increased so we may want to change the prior on Z to increase the probability of a large "log-distance". A heuristic version would be $p(z) = ce^{-cz}$ with $c < 2$ (to increase the heaviness of the tails compared to the model with $c = 2$ considered so far), leading to a modified posterior for Z compared to (26):

$$p_{\mathsf{Z}|\Gamma}(z|\gamma, c) = K_c(\gamma) \cdot \exp\left[-\frac{\left(z - \frac{\gamma}{a} + \frac{c\sigma_{Sh}^2}{a^2}\right)^2}{2(\sigma_{Sh}/a)^2}\right] \tag{46}$$

resulting in

$$\pi(\gamma|c) = \frac{\texttt{erfc}\left[2^{-1/2}\left(\frac{\log(d_{\max}/d_c)}{\sigma_{Sh}/a} - \frac{\gamma}{\sigma_{Sh}} + \frac{c\sigma_{Sh}}{a}\right)\right]}{\texttt{erfc}\left[2^{-1/2}\left(-\frac{\gamma}{\sigma_{Sh}} + \frac{c\sigma_{Sh}}{a}\right)\right]}. \tag{47}$$

This is illustrated on Figure 9 where $\pi(\gamma)$ is displayed for an average value of a, $d_{\max} = 10m$, $d_c = 1.8m$ and various values of σ_{Sh}, $c \in \{1, 2\}$. The value $c = 1$ is arbitrary for illustration purposes.

6 A practical risk assessment

In order to build a computationally efficient model, we propose to exploit the former model to calculate a risk associated to a 15-minute time slot. Considering the reference critical contact $c^* = \{\Delta t = 15min, d = 2m\}$, an expert should answer two related questions

1. If someone is exposed for $15min$ at $2m$ from an infected person, what is his/her probability of being infected ?
2. What probability of being infected should be reported as *high risk* to a user of the application ?

Answers to these questions can help determine a value ϵ^* such that:

$$\mathbb{P}\left(\mathsf{IS}_A = I | \mathsf{IS}_B = I; c_{AB} = c^*\right) = 1 - \epsilon^*. \tag{48}$$

Combined with (31), (34), and (35), this should enable tuning r_0, which gives direct relations between the expected risk and computable scores π.

Remark. *We insist again on the fact that r_0 is not needed at the application layer to compute the score $\pi_{j\texttt{Id}}$. The parameters ϵ^* or r_0 are only mandatory if we want to relate the score to a concrete probability of infection using* (42) *and* (31).

At the moment, let us suggest the following numbers, for illustration purposes. We consider that one exposition of $15min$ to a face to face contact at less than $2m$ induces a probability of contamination of 10%. We also assume that a global probability of being infected above 20% should be reported to the device's owner.

According to these assumptions, the following parameters are proposed:

parameter	Critical infec. prob.	Critical safety prob.	Ref. exposure time	Ref. exposure distance	risk coeff.
variable	Pr_I^*	ϵ^*	Δ t	d_c	r_0
value	10%	0.9	$15min$	$2m$	$0.007min^{-1}$

Table 4: Values of the main parameters used to translate scores into concrete risks and probabilities of contamination.

We now turn to the estimation of the probability of contamination in three steps : computing a score $\pi_{j\mathtt{Id}}$ associated to one contact measure ($\hat{Q}_{j\mathtt{Id}}(n)$), then computing an cumulative score associated to a time slot, and finally aggregating the probability of contamination over the day for one receiver.

6.0.1 Score for one contact measure

Let us start with the score $\pi_{j\mathtt{Id}}$. Due to the complexity of the model, we propose to use the risk model based on the Heaviside function $r_H(\cdot)$ given in (35). Now, exploiting the score expressed in (44), we obtain

$$\pi_{j\mathtt{Id}}(n) = \frac{\mathtt{erfc}\left[2^{-1/2}\left(\frac{\log(d_{\max}/d_c)}{\sigma_{Sh+f}(n)/a} - \frac{\gamma(n)}{\sigma_{Sh+fad}(n)} + \frac{2\sigma_{Sh+fad}(n)}{a}\right)\right]}{\mathtt{erfc}\left[2^{-1/2}\left(-\frac{\gamma(n)}{\sigma_{Sh+fad}(n)} + \frac{2\sigma_{Sh+fad}(n)}{a}\right)\right]}. \tag{49}$$

The value of the parameters to be used in this expression have been obtained from the PEPP-PT dataset in section 3 and are:

- $d_{\max} = 5m$. Maximal distance used in the model
- $a = 10n/\log(10) = 7.38$ (according to the pathloss model found in experimental data with $n = 1.7$).
- $\gamma(n) = \hat{Q}_{j\mathtt{Id}}(n) - P^*(d_{max}) = -80dBm$, estimated from the non calibrated experimental data.
- $\sigma_{Sh+fad}(n)$ is the standard deviation due to shadowing and remaining fading, with $\sigma_{Sh} = 4dB$, according to the experimental results and the additional fading term is given in Table 2.

Note that the parameters P_L, a and $\sigma_{Sh} = 5dB$ are the unique parameters to be determined. Tuning these parameters as a function of the environment (crowded/not crowded, indoor/outdoor,...) could improve any algorithm, but this needs to be evaluated from extensive measurement campaigns.

The function (49) is quite complex to evaluate analytically on a mobile device due to the evaluation of the `erfc` function. Yet, as illustrated on Figure 9, its shape is relatively basic, hence we propose a piecewise linear approximation of this function according to:

$$\hat{\pi}(\gamma) := \max(0, \min(1, \alpha(\gamma - \gamma_{\min}))) \tag{50}$$

where the slope α and the reference value $\gamma_{\min}$ can be adapted to the parameters appearing in (49).

In practice, these piecewise linear approximations are determined by fitting the value and slope at the point γ where $\pi(\gamma) = 0.5$.

For the sake of computational efficiency, we can develop piecewise approximations, where a linear piecewise function is proposed for each value of $\sigma_{Sh+fad} = \sigma_{Sh} + \sigma_{fad}$ (see Table 2) for values of σ_{fad}. The corresponding functions are represented in Fig.10 for the different standard deviations. The dashed curve represents the piecewise linear models.

This results concretely in scores expressed as

$$\pi_{j\mathtt{Id}}(n) = \max\left(0, \min\left(1, \frac{\hat{Q}_{j\mathtt{Id}}(n) - P^*_{min}}{\Delta_P}\right)\right) \tag{51}$$

where P^*_{min} and Δ_P are obtained from the curves and thus rely on experimental measurements on which the model is computed.

Note that introducing the calibration gains would simply shift these curves on the x-axis.

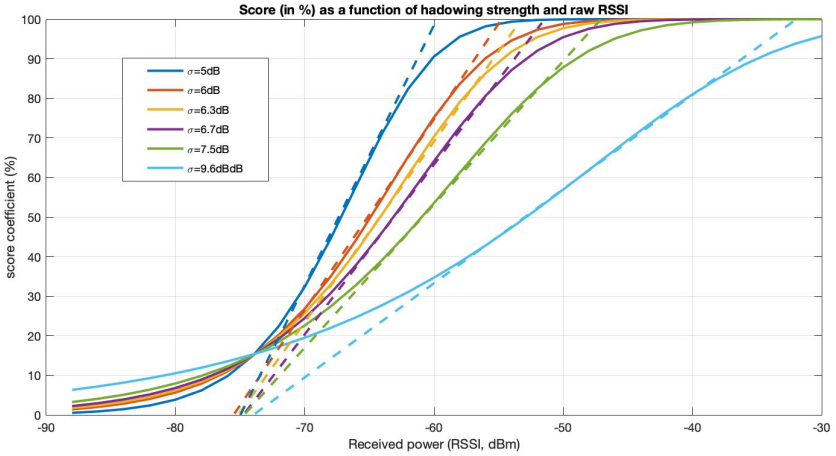


Figure 10: The curves represent the score π as a function of the non calibrated averaged RSSI for the different fading plus shadowing strength values given in Table 5. The dashed lines represent the simplified piecewise model that fits the curve at $\pi = 0.5$ and allow to determine the values of $P_{\min}$ and Δ_P in section 2.

We see that in these simulations, that the piecewise linear model is defined by two extreme values $P_{\min}$ and $P_{\max}$. We can note that the minimal significant power $P_{\min}$ is almost constant for all curves, around $-75dBm$. Clearly, this means that an average power lower than $-75dBm$ can be interpreted as a non-exposure situation with a high probability. On the opposite the other limit $P_{\max}$ is very sensitive to the shadowing strength. The power range is defined as the difference between these two extreme values

$$\Delta_P = P_{\max} - P_{\min}.$$

Table 5 summarizes the parameters of the model.

6.0.2 Cumulative score for one $15min$ time slot

Now, the next question is to compute a score associated to a $15min$ slot.

number of measures $N_p(n)$	1	2	3	4	5	> 5
additional standard deviation $S(n)$ (in dB)	$5,6$	$3,5$	$2,7$	$2,3$	2	1
total standard deviation $\sigma_{Sh+f}(n)$ (in dB)	$9,6$	$7,5$	$6,7$	$6,3$	6	5
Minimal power P_{min} (in dBm)	-75	-75	-75	-75	-75	-75
Max Limit power P_L (in dBm)	-36	-48	-52	-54	-55	-60
Diff power Δ_P (in dB)	39	27	23	21	20	15

Table 5: Additional standard deviations due to fading, as a function of the number of measures per $2min$. This standard deviation adds to the shadowing standard deviation, estimated at $4dB$.

Basic approach According to (45), we can compute the summation of the scores :

$$\pi_{j\mathtt{Id}} = \sum_{n=1}^{N} \pi_{j\mathtt{Id}}(n). \tag{52}$$

This cumulative score will give a number between 0 and 15. Typically, we suggest that a cumulative score less than 3 could be discarded, and a cumulative score higher than 7 would correspond to a strong risk. This metric has been proposed for the first version of the application.

Alternative approach However, to keep track of the exposure duration and intensity, we propose for future versions to compute instead the *average* score :

$$\bar{\pi}_{j\mathtt{Id}} = \frac{1}{N} \sum_{n=1}^{N} \pi_{j\mathtt{Id}}(n). \tag{53}$$

and to build a vector of contact parameters, given by

$$\Theta_{j\mathtt{Id}} = \left\{ \bar{\pi}_{j\mathtt{Id}}, D_{j\mathtt{Id}}, N_{j\mathtt{Id}} \right\}, \tag{54}$$

where $D_{j\mathtt{Id}}$ is the duration of the contact (difference between the indices of the first and the last non null coefficients) and $N_{j\mathtt{Id}}$ is the number of non null scores. Since $N = 15$ is known this is a strictly richer information as we can recover $N\bar{\pi}_{j\mathtt{Id}} = \pi_{j\mathtt{Id}}$.

A last optimization relies on the average computation. Considering channel properties, we believe that strong values of risks are more important, because they reveal almost surely an existing contact. Therefore, we suggest to use a softmax based average according to

$$\bar{\pi}^{sm}_{j\mathtt{Id}} = b \cdot \log \left(\frac{1}{N} \sum_{n=1}^{N} e^{\pi_{j\mathtt{Id}}(n)/b} \right), \tag{55}$$

where b is a parameter to be tuned. A value $b = 0.1$ gives good results and allows to reasonably differentiate high and low risk coefficient values.

6.0.3 Computing the probability of contamination

Now let us turn to the computation of a probability of contamination of a given user i, computed at the server side. The general principle is as follows

- each receiver device i whose owner has been tested positive is storing a list of temporary Ids which have been identified to be potentially infected, as their cumulative scores $\pi_{i\mathtt{Id}}$ exceed the threshold prescribed by the server. Such Ids, are then sent by such devices j to the server.

- on a regular basis, user j queries the server to evaluate its contamination risk by combining all computed risks. The server can retrieve a vector of the corresponding scores $\Pi(j) := (\pi_\ell)_{\ell=1}^L$ (or possibly an array of enriched scores $\Theta(j) = \{(\bar{\pi}_\ell, D_\ell, N_\ell)\}_{\ell=1}^L$). Each (enriched) score comes as a plain value, with no information on the device j which computed it.

Basic approach Let first consider the case where the cumulative score has been computed according to (45). Then according to (31) and using the same arguments leading to the additivity of the risk, the total probability of contamination of user j is

$$\mathbb{P}\left(\mathsf{IS}_j = I | \Pi(j)\right) = 1 - \exp\left(-r_0 \cdot \sum_{\ell=1}^{L} \pi_\ell\right). \tag{56}$$

Note that since $\Delta_T = 1min$ it does not appear in the above expression.

Let us illustrate with few examples :

1. Case 1 : j has been subject to $L = 4$ contacts with respective scores π_ℓ of 5, 7, 10, 10. Typically a score of 10 may represents a strong contact lasting $10min$ or a medium contact lasting $15min$. Then, the probability of being contaminated is

   $$\mathbb{P}\left(\mathsf{IS}_j = I | \Pi(j)\right) = 0.13,$$

   which represents a moderate risk.

2. Case 2 : i has been subject to $L = 4$ contacts, all with respective scores of 10. Then, the infectiousness probability is

   $$\mathbb{P}\left(\mathsf{IS}_j = I | \Pi(j)\right) = 0.244,$$

   which represents a strong risk.

3. Case 3 : j has been subject to $L = 20$ contacts, all with a moderate score of 4. Then, the infectiousness probability is

   $$\mathbb{P}\left(\mathsf{IS}_j = I | \Pi(j)\right) = 0.43$$

   which represents a strong risk. This risk is not due to a long exposition to a unique person, but to the multiplicity of the individual risks.

Alternative approach Now let us consider the case where the average score has been returned with duration and number of contacts. Then the infectiousness probability can be computed according to

$$\mathbb{P}\left(\mathsf{IS}_j = I | \Theta(j)\right) = 1 - \exp\left(-r_0 \cdot \sum_{\ell=1}^{L} \bar{\pi}_\ell \cdot D_\ell\right). \tag{57}$$

Alternatively, D_ℓ can be replaced by N_ℓ which better represents the strength of the contact.

Although the result is equivalent to the direct computation of the cumulative score, having the three parameters $\bar{\pi}_\ell, D_\ell$ and N_ℓ may allow to balance in the future the impact of the three parameters and to develop a more advanced combination function.

7 Experimental results of risk assessment on PEPP-PT data

7.1 Experimental setup

Based on the model derived in sections 4 and 5, and calibrated as described in section 3, we propose to evaluate the model on the data from the German group.

Note that the model has been calibrated from the whole set of data from the German group, taken as a whole set. There is no specific calibration for each room where the data have been acquired. 3 rooms are indoor and 2 are outdoor.

The unique parameters required in our model, are the pathloss slope coefficient ($n^* = 1.7$) and the RSSI at $5m$, i.e. $P^*(d_{\max}) = -80dBm$.

According to the fading model described above, and using the Bayes optimal detector, we computed the parameters of the piecewise linearized risk scoring function, as indicated in table 5 i.e $P_{min} = -75dBm$ and $\Delta_p = 15dB$ (all trace have been recorded at a fast sampling with much more than 8 values per window).

In this section, we exploit only the traces obtained when the soldiers were standing in each position for $10min$ (long slots), in each of the 5 rooms used in this experiment.

Each dataset includes the traces along 6 slots. Soldiers 1 and 2 entered first the playground, then 3 and 4, and so on. The soldiers walk around two paths for odd and even numbers.

7.2 Reference risk

In order to evaluate the capability of our algorithm to estimate the infectiousness risk, we first need to determine a reference risk. For that we use the reference distances between pairs of soldiers, according to the scenario described in [3].

The reference risk classes are defined between 0 and 5, where 5 represents the class with the maximal risk. The correspondence between distances and theoretical risk is reported in Table 6. Not that these numbers are here for the purpose of classification, but they do not represent a

Tx-Rx Distance $d(m)$	$\leq 1m$.	$\leq 1.5m$ $> 1m$	$\leq 2m$ $> 1.5m$	$\leq 2.5m$ $> 2m$	$\leq 3m$ $> 2.5m$	. $> 3m$
Reference risk (arbitrary)	5	4	3	2	1	0

Table 6: risk classes associated to the different signals.

quantification of the risk. Further, we will not use this classification for the purpose of classifying all observed signals in these categories. But we use these classes to observe how they are classified from the received signals. The primary objective is to identify class 4 and class 5 with the highest probability, with a low false positive rate for the class 5 at least.

In addition, when in the scenarios two soldiers are separated by a third soldier positioned between the two others, it can be anticipated that the average RSSI will be lowered due to non line of sight (nLOS) conditions. However, it is also worth noting that the risk associated to these two soldiers is also reduced. Therefore, the RSSI reduction due to the presence of someone in the middle is correlated to a risk reduction.

To track this fact, and to avoid too much non detection due to this masking effect, we explored the data from the German group and identified all pairwise situations where a third soldier was masking the line of sight. Then the risk class of the corresponding radio link was reduced by 2. For instance, if two soldiers have a risk class 3 when considering their mutual distance, the presence of a soldier between them reduces this link to class 1.

7.3 Risk evaluation

We describe herein the procedure for the risk evaluation applied to these data. In the StopCovid application, a contact $C(A \to B)$ is made of a list of HELLO packets received from node A to node B during a slot, with a maximal duration of $15min$. Indeed, each $15min$, the terminal Ids are randomly changed and the receiver cannot track the current Tx.

In the data we are considering here, the duration of the slots was $10min$. Therefore, the contact duration we are looking for is aligned on the experimental slot duration, i.e. $10min$. From each raw trace, the signal corresponding to one slot is extracted and processed as if it was corresponding to a StopCovid trace of $10min$.

The RSSI trace is windowed by a sliding rectangular window of $120sec$, positioned each $60sec$. For instance, at time t_0, the first operation is to get the RSSI values received from t_0 to $t_0 + 120$. From this set, the linear domain mean power is computed according to (8). For each slot of $10min$, we get between 9 and 11 averaged RSSI measurements. The RSSI traces, over 4 consecutive slots of $10mn$, as well as the distance and risk estimations are provided in Fig. 11 and Fig. 12

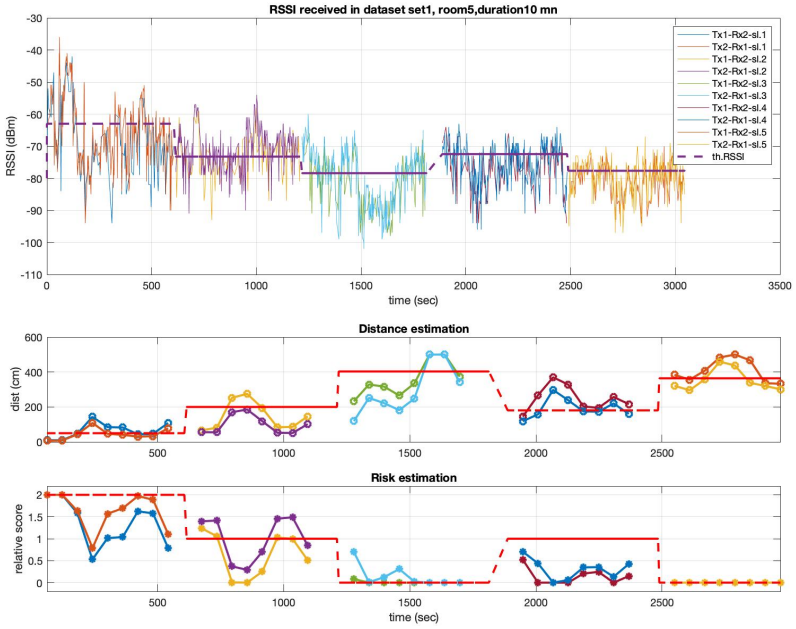


Figure 11: Traces between devices 1 and 2, room 5 from PEPP-PT data. Red lines are ground-truth values. Note that the scores plotted in (c) are on a scale $[0; 2]$. The ground-truth score is 2 for a distance lower than $1m$ and 1 for a distance lower than $2.5m$. The estimated score is also scaled by a factor 2 for consistency.

For each averaged RSSI measurement, the risk is computed with (51), and provides a number between 0 and 1. According to (52), the score associated to the contact is the average of individual risks.

Considering one dataset (first data set on April, 9 and for room 5), the results presented in Fig. 13 have been obtained. On the left, the figure presents the cumulative distribution (CDF) of the estimated score for a given class. For the class $d > 4m$, the CDF is represented in the classical orientation (from left to right). For the other classes, the CCDF is presented, from right

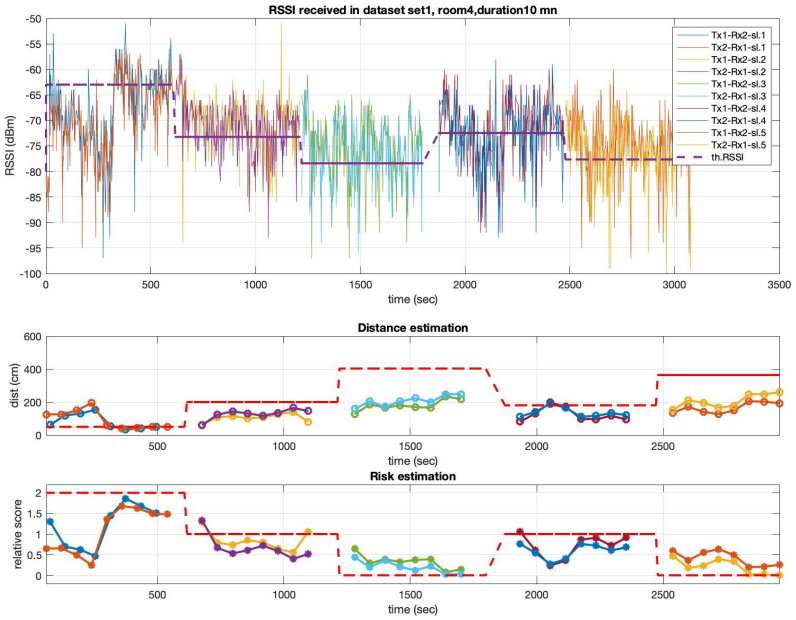


Figure 12: Traces between devices 1 and 2, room 4 from PEPP-PT data. Note that the scores plotted in (c) are on a scale $[0; 2]$. The ground-truth score is 2 for a distance lower than $1m$ and 1 for a distance lower than $2.5m$. The estimated score is also scaled by a factor 2 for consistency.

to left.

According to the proposed algorithm, three scores are defined : high score for $\bar{\pi} > 7/15$, medium score for $3/15 > \bar{\pi} > 3/15$ and negligible score if $\bar{\pi} < 3/15$. These limits are represented in the figure by the vertical black lines. With these values, we obtain the conditional classification in Table 7. The thresholds can be shifted to balance false positive and true positive rates.

ground truth risk class	0	$1-2$	$3-5$	$4-5$	5
distance	$d > 3m$	$2m \leq d \leq 3m$	$d < 2m$	$d < 1.3m$	$d < 0.8m$
no risk	80%	30%	30%	21%	12%
medium risk	18%	40%	20%	12%	14%
high risk	2%	30%	50%	67%	74%

Table 7: Class detection probabilities under the risk scores identified in Fig. 13.

Note that a unique threshold could be used : risk/no risk, with the objective to discriminate the class C_0 against a subset of the other classes. For instance, a threshold value at 0.35, would provide a false alarm rate of 5%, while the non detection probability for the class C_5 would be only of 12%. If the critical distance to be considered is $2m$, then the non detection probability would approach 35% with this threshold. If the $2m$ distance is the most critical, we may move the threshold to 0.1, providing a false alarm rate of 30% and a non detection probability of 20% for the distance class of $2m$.

The take away message is that it is possible to detect 80% of contacts at less than $2m$ and 90% of contacts at less than $1m$ at the price of false alarms at a rate of 30% (with a threshold at 1m).

The choice of the threshold may depend on the ratio between the packet received from class c_0 compared to the positive class. Let consider the case of a critical distance at $2m$. If the number of packets is equivalent in both classes, then one packet classified in the class *risk* has a probability of 3/4 to be a true positive and 1/4 to be a false positive. Then if one packet is classified in the class *no risk*, it has also a probability of 3/4 to be a true negative and 1/4 to be a false negative.

But if the number of packets in class c_0 is twice the number of packet in the positive class, then, one packet classified in the class *risk* has a probability of 0.57 to be a true positive and 0.43 to be a false positive. Then if one packet is classified in the class *no risk*, it has a probability of 87% to be a true negative and only 13% to be a false negative.

These examples are just here to remind that the knowledge of a prior information about the distribution of packets may be useful to tune the thresholds to be applied in the risk function. If it is anticipated that the prior probability of class 0 packets is higher, then the thresholds should be shifted to the right.

This is why it seems more constructive to track the score (the number between 0 *and* 1*) and to work on these numbers at the server side. We indeed believe that such application may be efficient to determine an average risk over days than granting the detection of one event.*

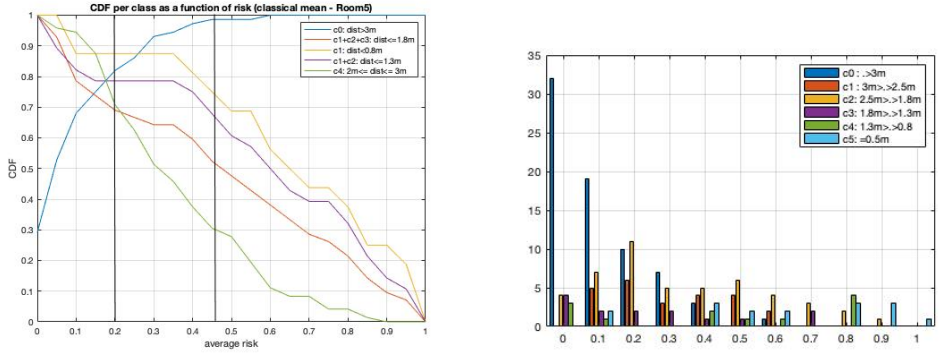


Figure 13: Distribution of risk estimation as a function of the ground truth risk class (based on distance).

7.3.1 Impact of the average computation method

In the former result, we applied a classical mean on the scores computed for each minute. However, it appears in the data that some risk scores are more significant than others. Especially, when a high score is measured, then the probability of a contact is much more likely, while when a low score is obtained, the probability of a false negative is higher. Therefore we propose the use of the weighted softmax score according to (55).

The improvement induced by this estimator is illustrated in Fig.14. While the class C_0 curve remains comparable, the other curves shift to the right characterizing an improvement in the false positive-misdetection tradeoff. Remarkably, we can see that at a threshold $\pi = 0.45$, the false positive rate remains equivalent while the detection probability grows up to 90% for class C_1.

We therefore applied the same model, with the same parameters to the 4 other rooms in the dataset. The results are at as good for rooms $1, 3, 4$ but are less convincing in room 2. In this

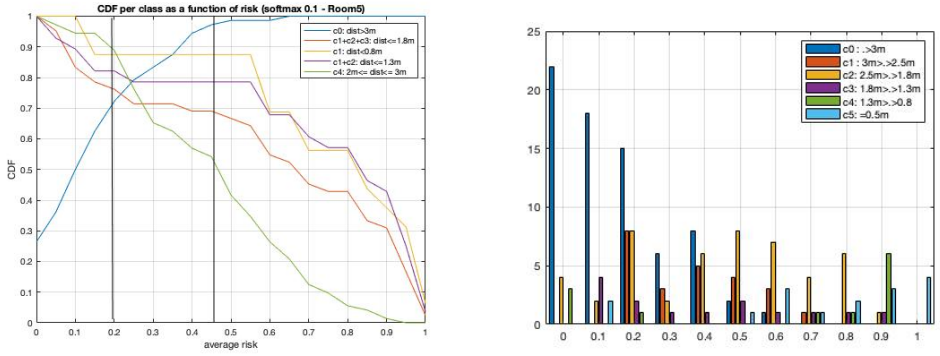


Figure 14: Distribution of score estimation as a function of the ground truth score class (based on distance), with the weighted softmax (55).

room, the analysis of the data exhibits a received power that is $10dB$ lower than in the other, in average.

8 Conclusion

This document has been written to formalize a method for the infectiousness risk computation from BLE RSSI values, relying extensively on the work done in the StopCovid consortium and on the results produced by the PEPP-PT project. Our objective was to propose a robust algorithm to compute an infectiousness probability, from BLE measurements and to make it compatible with the ROBERT protocol.

We wrote this report to make it accessible by people interested to understand the parameters used in StopCovid. We also provide theoretical elements to justify the model. We keep the model as simple as possible to avoid over-fitting on the dataset and to make it as flexible as possible.

There is no revolution in the model. Many parts of the algorithm were proposed in the PEPP-PT group. The difference with the algorithm proposed by the German group, is that in our approach we took care of physical phenomena (fading, shadowing) to derive estimation rules rather than using a machine learning approach. We believe that it is a good way to produce a reference model, generalisable to other setup.

It is worth that this preliminary report exploits a lot the PEPP-PT dataset and we thank the authors of these works for sharing their data, thoughts and models. The next report will show how this model can be applied to other dataset, especially those that have been obtained on the dataset acquired by the StopCovid consortium. These data are really complementary to the data from the German group, and will provide a better understanding of the performance of our algorithms in various situations. As the algorithm is rather flexible and has few parameters to tune, the algorithm will be updated according to the observation of these data.

A Raw traces

The traces given in Fig.15 are example of the trace acquired in the $4mn$-rounds as described in [3]. The raw RSSI values are represented a a function of time. These traces have been acquired with an average rate of 1 packet each two seconds. Therefore, for a window of $2mn$, we can expect to get in average 60 measures. We observed a limited precision loss if the number of

samples is reduced down to 16 points, meaning that a rate of 1 HELLO packet each 8 seconds could be reasonably sufficient to compute a reliable average power. The use of the linear domain average (according to Eq.(8)) allows to significantly reduce the fading effect, and the dominant effect is the shadowing.

Because the scenarios generated in this experimentation are relatively static, a static shadowing may appear and lasts a whole slot. This fact makes this dataset challenging, since for some situations, we may have a strong shadowing, and thus a low signal even for two soldiers standing less than one meter to each other.

References

[1] Steffen Meyer. *Distance Measurements and Classification*. Tech. rep. confidential, Proximity Tracing App project (PEPP-PT PNTK System). Fraunhofer IIS, Apr. 2020.

[2] Thomas Wiegand. *Test report from the 1. Measurement Campaign*. Tech. rep. Proximity Tracing App project (PEPP-PT PNTK System). Fraunhofer HHI, Apr. 2020.

[3] Jackie Ma et al. *Report from the Measurement Campaign*. Tech. rep. Proximity Tracing App project (PEPP-PT PNTK System). Fraunhofer HHI, Apr. 2020.

[4] Patrick Wagner Thomas Wiegand and Felix Sattler. *Assessment of Corona Virus infection risk from proximity time series*. Tech. rep. Proximity Tracing App project (PEPP-PT PNTK System). Fraunhofer HHI, Apr. 2020.

[5] Steffen Meyer. *On-Device contact risk estimation from RSS Data, implementation details*. Tech. rep. Proximity Tracing App project (PEPP-PT PNTK System). Fraunhofer IIS, Apr. 2020.

[6] Yu-Chi Pu and Pei-Chun You. "Indoor positioning system based on BLE location fingerprinting with classification approach". In: *Applied Mathematical Modelling* 62 (2018), pp. 654–663.

[7] Douglas J Leith and Stephen Farrell. "Coronavirus Contact Tracing: Evaluating The Potential Of Using Bluetooth Received Signal Strength For Proximity Detection". In: (2020).

[8] Simon R Saunders and Alejandro Aragón-Zavala. *Antennas and propagation for wireless communication systems*. John Wiley & Sons, 2007.

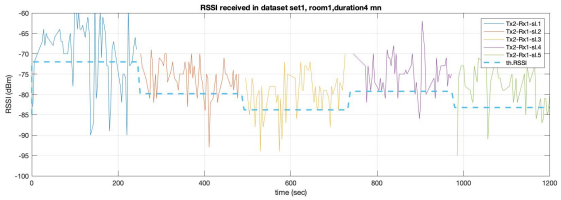
[9] Meiling Luo et al. "Realistic prediction of BER and AMC for indoor wireless transmissions". In: *IEEE Antennas and Wireless Propagation Letters* 11 (2012), pp. 1084–1087.

[10] Paul Ferrand, Jean-Marie Gorce, and Claire Goursaud. "On the packet error rate of correlated shadowing links in body-area networks". In: *Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP)*. IEEE. 2011, pp. 3094–3098.

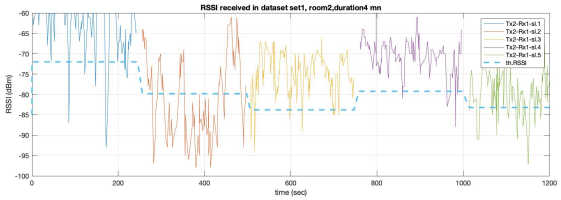
[11] Matthieu Lauzier et al. "Full mesh channel measurements on Body Area Networks under walking scenarios". In: *2013 7th european conference on Antennas and propagation (EuCAP)*. IEEE. 2013, pp. 3508–3512.

[12] Marvin K Simon and Mohamed-Slim Alouini. "Digital communications over fading channels (mk simon and ms alouini; 2005)[book review]". In: *IEEE Transactions on Information Theory* 54.7 (2008), pp. 3369–3370.

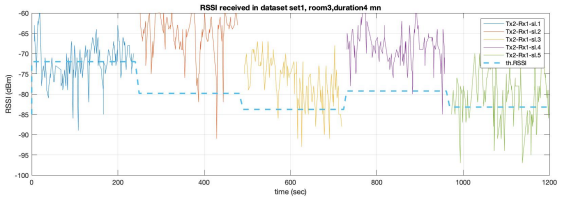
[13] Young-Chai Ko and M-S Alouini. "Estimation of the local mean power over Nakagami fading channels". In: *12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. PIMRC 2001. Proceedings (Cat. No. 01TH8598)*. Vol. 1. IEEE. 2001, pp. C–C.



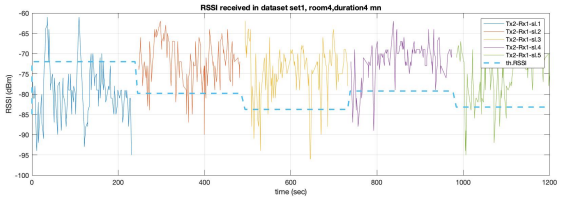
(a) room 1



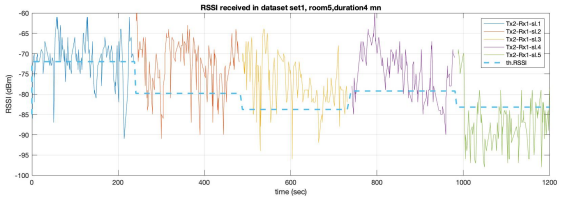
(b) room 2



(c) room 3

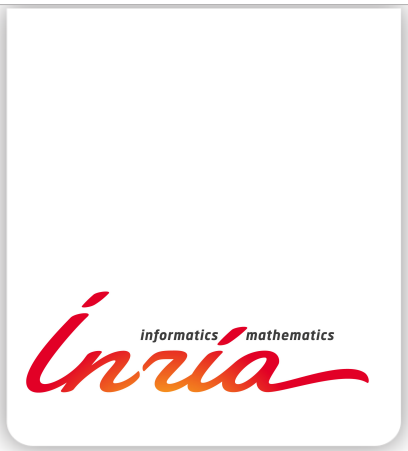


(d) room 4



(e) room 5

Figure 15: RSSI traces received by $Rx2$ from $Tx1$ during the $4mn$ scenario, i.e. where the soldiers stand in each position for $4mn$. The reference curves are the theoretical average RSSI expected from the model derived in the first section, with $PL_L = -85dBm$ at $d_L = 5m$, and $n = 1.3$.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399