Security against Trojan-horse attacks on a self-compensating all-fiber polarization modulator

Alberto De Toni,[1] Aynur Cemre Aka,[1] Costantino Agnesi,[1] Davide Giacomo Marangon,[1] Giuseppe Vallone,[1,2] and Paolo Villoresi[1,2]

[1]*Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova, via Gradenigo 6B, IT-35131 Padova, Italy*

[2]*Padua Quantum Technologies Research Center, Università degli Studi di Padova, via Gradenigo 6A, IT-35131 Padova, Italy*

(Received 24 October 2025; accepted 22 April 2026; published 11 May 2026)

Quantum key distribution (QKD) leverages the principles of quantum mechanics to exchange a secret key between two parties. Unlike classical cryptographic systems, the security of QKD is not reliant on computational assumptions but is instead rooted in the fundamental laws of physics. In a QKD protocol, any attempt by an eavesdropper to intercept the key is detectable: this provides an unprecedented level of security, making QKD an attractive solution for secure communication in an era increasingly threatened by the advent of quantum computers and their potential to break classical cryptographic systems. However, QKD also faces several practical challenges such as transmission loss and noise in quantum channels, finite key size effects, and implementation flaws in QKD devices. Addressing these issues is crucial for the large-scale deployment of QKD and the realization of a global quantum Internet. A whole body of research is dedicated to the hacking of the quantum state source, for example using *Trojan-horse attacks* (THAs), where the eavesdropper injects light into the system and analyzes the back-reflected signal. In this paper, we study the vulnerabilities against THAs of the iPOGNAC encoder, first introduced in [M. Avesani *et al.*, Opt. Lett. **45**, 4706 (2020)], to propose adapted countermeasures that can mitigate such attacks.

DOI: 10.1103/bgdd-1vl9

I. INTRODUCTION

Quantum key distribution (QKD) represents one of the most mature applications of quantum information science, offering unconditional security for key exchange based on the fundamental laws of quantum mechanics [1,2]. Protocols such as BB84 and its decoy-state variants have been implemented over increasing distances and integrated into field-deployable systems [3,4]. However, the practical security of QKD systems depends critically on the faithful implementation of their underlying quantum protocols. Any deviation or imperfection in components can introduce vulnerabilities that are not covered by theoretical security proofs [5].

One such vulnerability is posed by *Trojan-horse attacks* (THAs) [6,7], in which an eavesdropper (Eve) injects bright light into the QKD apparatus, typically at the sender's (Alice's) side, and then analyzes the back-reflected signal to gain information about secret settings, such as basis choices or intensity levels [6,8,9]. These attacks exploit the physical layers of the system and, if undetected, can significantly compromise the security of the generated keys without introducing errors that are detectable by Alice and Bob (the receiver) [10].

A growing body of research has focused on both theoretical models of THAs and practical countermeasures, including the use of optical isolators, monitoring detectors, and new security proofs that account for side-channel leakage [6,11,12]. Despite these efforts, fully characterizing and mitigating THAs remains an open and crucial challenge in the secure deployment of QKD networks.

In this paper, we analyze the practical impact of THAs on the iPOGNAC, a self-compensating, all-fiber polarization modulation scheme, first proposed by Avesani *et al.* [13]. This analysis is of relevance since many different QKD systems have adopted the iPOGNAC, both for *discrete-variable* QKD, for polarization modulation [14] and time-bin modulation [15], and for *continuous-variable* QKD [16], showcasing the reliability, robustness, and effectiveness of the polarization encoder. Other works previously analyzed the impact of THAs on this type of encoder, like Ref. [17], but without addressing the possibility of Eve directly exploiting the Sagnac loop to gain information on the encoded symbol, a scenario in which she can get most of her optical power back.

In this work, we analyze the eavesdropper's capability of correctly guessing the generated state varying the energy on the output. We then evaluate the mutual information in different real-life implementations of the attack: continuous and pulsed laser, high and low mean photon number. We investigate the countermeasures that Alice can adopt in her system to drastically reduce the information leakage at Eve's side, give estimates of the limitations of such defenses, and analyze the performances of the different types of THAs in those conditions.

II. METHODS

Naïve implementation of Sagnac-loop based modulation schemes can be problematic since a significant portion of injected light through a THA can return to the eavesdropper containing an information leakage that can undermine the

security of the implemented quantum communication protocol. In fact, the light injected by the eavesdropper traverses the same modulation loop Alice uses to encode her information, meaning that it potentially experiences the same modulation and carries the same information as the QKD signal, allowing the eavesdropper to gain knowledge on the forthcoming symbol.

To analyze the performances of the THA on the iPOGNAC, we divided our work into three stages: an attack in the strong-light regime, which comprises a continuous-wave laser attack (CWLA) and a pulsed laser attack (PLA) (both of them make use of photo-diodes as light detectors), a pulsed laser attack in the weak-light regime exploiting Geiger-mode detectors such as *superconducting-nanowire single-photon detectors* (SNSPDs), and an evaluation and analysis on the countermeasures that Alice can take to counteract these attacks. The reasons we chose to study these cases is because high-power attacks are simpler and cheaper to implement, whereas single-photon-level attacks allow for a wider range of countermeasure levels. In the meantime, pulsed attacks permit one to concentrate more photons in a single time span, while CW attacks require less stringent synchronization (more on the motivations for each attack in Sec. VI). The stages are reported in the following sections, after a brief description of the setup.

Description of the system

The iPOGNAC polarization encoder is a device designed for stable, low-error, and calibration-free polarization modulation, particularly suited for QKD and quantum communications (see Ref. [13] for a complete description of the device). Its core innovation lies in its use of a Sagnac interferometer loop, which inherently compensates for environmental disturbances such as temperature fluctuations and phase drifts, ensuring long-term operational stability.

Its working principle begins with linearly polarized optical pulses injected into the iPOGNAC. A half-wave plate rotates the polarization to a diagonal state, after which a beam splitter separates the input and output streams. The transmitted light is coupled into a polarization-maintaining fiber (PMF). The light then enters a fiber-based polarization beam splitter (PBS), which initiates the Sagnac loop. Here, the two orthogonal polarization components (horizontal and vertical) travel in opposite directions—clockwise (CW) and counterclockwise (CCW)—along the slow axis of the PMF. Each component passes through a phase modulator at different times, allowing independent phase control and generation of states of the form $|\Delta\phi\rangle = \frac{1}{\sqrt{2}}(|H\rangle + e^{i\Delta\phi}|V\rangle)$ where $\Delta\phi = \phi_{\mathrm{CW}} - \phi_{\mathrm{CCW}}$, and ϕ_{CW} and ϕ_{CCW} are the phases applied by the phase modulator to the pulses that travel the Sagnac loop clockwise and counterclockwise respectively. By applying π or $\pi/2$ phase shifts to the CW and CCW pulses, the iPOGNAC can generate all the polarization states required for the BB84 QKD protocol, including diagonal $|D\rangle = \frac{1}{\sqrt{2}}(|H\rangle + |V\rangle)$, antidiagonal $|A\rangle = \frac{1}{\sqrt{2}}(|H\rangle - |V\rangle)$, circular left $|L\rangle = \frac{1}{\sqrt{2}}(|H\rangle + i|V\rangle)$, and circular right $|R\rangle = \frac{1}{\sqrt{2}}(|H\rangle - i|V\rangle)$ polarizations. Experimental results have demonstrated the iPOGNAC's ability to maintain a low quantum bit error rate over extended periods, both in laboratory and field-trial settings. The device's versatility is further highlighted by its ability to be reconfigured for tasks such as time-bin encoding [15], continous variable modulation [16], and intensity modulation [18], supporting the development of robust and flexible quantum networks [19,20]. Regarding the attack, at Eve's side, a 1560-nm laser shines through a *polarization controller* (PC) that controls the polarization in input at Alice's side. The light that comes back from Alice's system is redirected using a *circulator* to another PC and a PBS that allows the projective measurement on a *Si photodiode* (PD) or on the SNSPDs. The output of the PD is displayed using an *oscilloscope*, which is also used to store the data to analyze. Alice's setup is composed of a *variable optical attenuator* (VOA) to counteract the THA, and the asymmetric iPOGNAC, namely a *beam splitter* (BS) connected to Alice's source on one side to a Sagnac-loop composed of a PBS and a *phase modulator* on the other side. Alice's source emits diagonal polarized photons that enter the BS from the left side. For simplification purposes, the whole system (Alice + Eve) is controlled through the same *system on a chip* (SoC), a *field-programmable gate array* (FPGA) that allows one to generate electrical pulses and synchronize everything to the same clock rate [21]: this holds on the generalized assumption that Eve has access to the system of Alice (or can copy the classical signals coming out of it) to exploit the same synchronization signals she uses, or that she can use clock-recovery algorithms to extract the information on the synchronization [22]. The setup of the experiment is depicted in its entirety in Fig. 1. The modulation signals employed work by introducing a delay between the $|H\rangle$ and $|V\rangle$ symbols on the FPGA side: in this sense the Sagnac loop enables one to encode the polarization states on the optical signal by fine-tuning this delay (see Fig. 4). We point out that since Alice is using an attenuator as a countermeasure, the light of the THA coming out from her system will experience two times the attenuation set on the VOA. In our setup, the

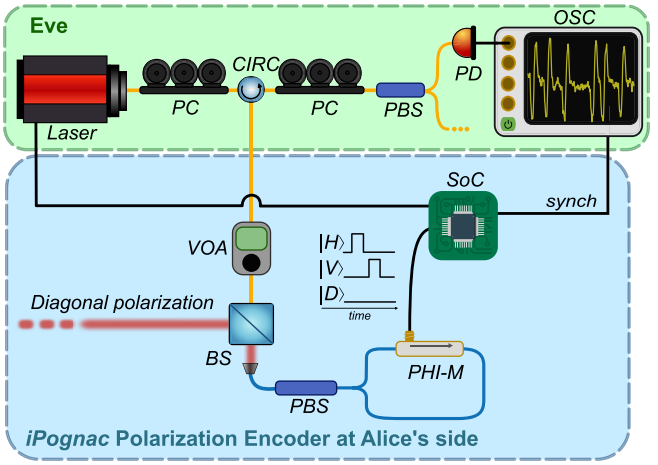

FIG. 1. Scheme of the setup for the THA on the iPOGNAC. SM fibers are shown in yellow, PM fibers are shown in blue, and electrical connections are shown in black. The optical power entering inside Alice's setup is indicated as μ_{in}, while the one coming out (subject to the attenuations of Alice's system and therefore smaller than μ_{in}) is indicated as μ_{out}.

total attenuation experienced by the light coming out from the system A_{tot} can be described by the equation

$$(A_{\text{tot}})_{\text{dB}} = 2(A_{\text{VOA}} + \Delta A) + 6 + E \tag{1}$$

where A_{VOA} is the attenuation manually set on the VOA (the term we mostly refer to in this paper when talking about countermeasures), ΔA is an inefficiency term given by the characterization of the VOA (in our case 4 dB), 6 dB is the attenuation given by the double crossing of the 50:50 beam splitter, and E are extra losses from coupling inefficiencies, defective components, and so on, in our case corresponding to $\approx$1 dB.

III. THEORETICAL MODEL

This section is dedicated to the calculation of the guessing probability in the function of the output mean photon number μ_{out}. We will cover three cases: using an optimal positive operator-valued measure (POVM) for the measurement, which gives an upper bound on the maximum information that can be extracted from the system, while using a fixed POVM, photon-number resolving detectors in the second case, and Geiger-mode photodetectors in the third case. All of them are reported in the following sections. Despite the theoretical model being valid both for weak and strong light regimes, the main limitations of the latter mostly come from the limited performances of the devices in use, like the noise floor of the detectors or the oscilloscopes (see Secs. IV C and VI). In this work we will concentrate on a three-state BB84 protocol, since it is more practical and widely used in real-world scenarios [23,24]. If the system were to implement a four-state protocol, the higher dimensionality would lead to more relaxed security bounds (Eve's guessing probability would decrease up to a minimum of 1/4) and therefore it would not represent a worst-case scenario from Alice's point of view.

A. Optimal POVM

When implementing a THA, Eve will manage to obtain a state that depends on the symbol modulated by Alice. With no lack of generality, we therefore assume that the states that Eve will receive are

$$\begin{aligned}
|\psi_1\rangle &= |\sqrt{\mu}\rangle_H \otimes |0\rangle_V && \text{if Alice sends } |H\rangle,\\
|\psi_2\rangle &= |0\rangle_H \otimes |\sqrt{\mu}\rangle_V && \text{if Alice sends } |V\rangle,\\
|\psi_3\rangle &= \left|\sqrt{\frac{\mu}{2}}\right\rangle_H \otimes \left|\sqrt{\frac{\mu}{2}}\right\rangle_V && \text{if Alice sends } |D\rangle.
\end{aligned} \tag{2}$$

Optimizing on the POVM F_e employed by Eve, it is possible to extract an upper bound on the quantity of mutual information that is accessible and can be extracted from the system, the so-called accessible information:

$$I_{\text{acc}}(A:E) = \max_{\{F_e\}} \sum_{a,e} p(a,e) \log_2 \frac{p(a,e)}{p(a)p(e)} \tag{3}$$

with $p(e|a) = \text{Tr}[F_e \rho_a]$ and ρ_a the density matrix of the received state $|\psi_a\rangle$. Instead, the average probability of correctly guessing a symbol p_g, in this scenario, is given by

$$p_g = \sum_a p(a)p(e=a|a). \tag{4}$$

Assuming the system is symmetric, namely $p(a) = 1/3$, $p(e=a|a)$, and $p(e\neq a|a)$ do not depend on a, then $p(e) = 1/3$ and

$$p_g^* = \text{Tr}[F_a^* \rho_a] \quad \forall a \tag{5}$$

where F_a^* is the optimal POVM. Therefore the accessible information can be related to the guessing probability by

$$I_{\text{acc}}(A:E) = p_g \log_2(3p_g) + (1-p_g)\log_2 \frac{3(1-p_g)}{2}. \tag{6}$$

While the maximization in (3) is in general hard to solve because there is not an explicit closed form for the optimal POVM, the accessible information can be upper bounded by the so-called Holevo bound [25], which expresses the amount of classical information that can be extracted from a quantum system, obtained as a mixture $\rho = \sum_a p_a \rho_a$:

$$I_{\text{acc}}(A:E) \leqslant \chi(\rho) = S(\rho) - \sum_a p_a S(\rho_a) = S(\rho). \tag{7}$$

The last equality follows from the fact that the state ρ is given by $\rho = \frac{1}{3}\sum_{a=1}^{3} |\psi_a\rangle\langle\psi_a|$ and the von Neumann entropy of a pure state is vanishing (i.e., $S(|\psi_a\rangle\langle\psi_a|) = 0$). Then, it follows that Eq. (3) is upper bounded by $S(\rho)$.

Solving this equation can yield an upper bound for the p_g. As a first step we therefore have to calculate $S(\rho)$, that is equal to the Shannon entropy of the eigenvalues λ_i of the matrix given by $\rho_{ij} = \frac{1}{3}\langle\psi_i|\psi_j\rangle$ (see the Appendix):

$$S(\rho) = H(\mu) := -\sum_{i=1}^{3} \lambda_i(\mu) \log_2[\lambda_i(\mu)]. \tag{8}$$

By using Eq. (7), an upper bound on the guessing probability p_g can be extracted in the function of μ by the implicit relation

$$I_{\text{acc}}(A:E) \leqslant H(\mu) \tag{9}$$

which gives the results reported in Fig. 2 as a red line.

The results just found depict a loose upper bound on the actual, tighter, quantum limit for the distinguishability of a set of states, which is given by the Helstrom bound [25,26] (originally defined for bipartite systems), which quantifies the maximum probability p_g^* of identifying the correct state. To compute this limit for the states identified by the density matrices $\{\rho_i\}$, it is possible to express the primal-form optimization problem:

$$\begin{aligned}
p_g^* &= \frac{1}{3}\max_{F_a} \sum_a \text{Tr}[F_a^* \rho_a],\\
F_a &\succeq 0 \quad \forall\, a \quad \text{and} \quad \sum_a F_a = \mathbb{I}.
\end{aligned} \tag{10}$$

This semidefinite program (SDP) is strictly feasible [27] (for instance, the trivial measurement $F_a = \mathbb{I}/3$ satisfies the constraints) which means that the Slater condition for convex optimization is satisfied [28], granting that strong duality holds and we can write this as a dual-problem SDP [29]. This method allows us to optimize over all generalized quantum measurements to find the optimal strategy by finding a Hermitian operator K that solves the following minimization problem (in general computationally simpler to solve):

$$p_g = \min_K \{\text{Tr}(K) | K \succeq p_i \rho_i, \quad \forall i\} \tag{11}$$

where $p_i = \frac{1}{3}$ in our case.

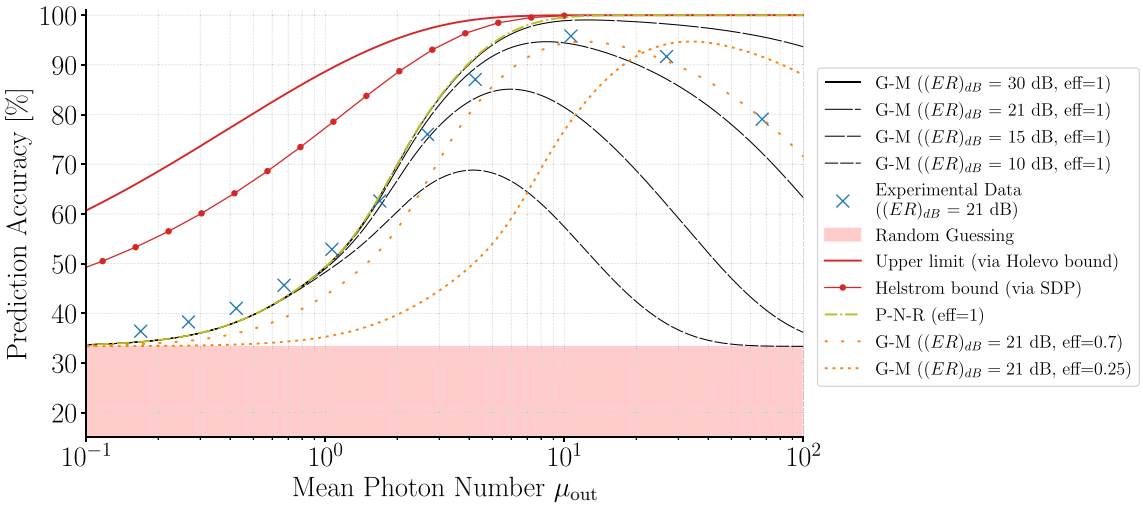

FIG. 2. Theoretical prediction accuracy with respect to the mean photon number μ_{out}, varying different POVMs, extinction ratios of the projective measurements, efficiencies, and experimental data for comparison.

The solution to this program allows one to relax the bound on μ with which an eavesdropper can get useful information on the system, and is reported in Fig. 2 as a red line with dots.

B. Fixed POVM

For this attack, we wanted to minimize the amount of detectors while still being able to distinguish the symbols, which corresponds to a minimum of two channels, as seen from the truth table of the detections reported in Table I.

Based on the table, we can already predict that the minimum number of photons per symbol required to distinguish all the three symbols is 2 (therefore the two channels, to be able to distinguish $|D\rangle$). Knowing the probability of a channel to *not click* is described by a Poissonian distribution [$\mathcal{P}_\mu(0) = \frac{\mu^0}{0!}e^{-\mu} = e^{-\mu}$], we can describe the guessing probabilities for each symbol in the following way:

$$\begin{aligned} P(|H\rangle_{\text{out}} \,\|\, |H\rangle_{\text{in}}) &= P(CH_1 \cap \overline{CH_2}) \\ &\overset{\text{i.i.d.}}{=} P(CH_1)P(\overline{CH_2}) \\ &= (1 - e^{-\mu_H \eta})e^{-\mu_H \eta R_{\text{ext}}} \end{aligned} \tag{12}$$

and similarly for $|V\rangle$ and $|D\rangle$

$$\begin{aligned} P(|V\rangle_{\text{out}} \,\|\, |V\rangle_{\text{in}}) &= e^{-\mu_V \eta R_{\text{ext}}}(1 - e^{-\mu_V \eta}), \\ P(|D\rangle_{\text{out}} \,\|\, |D\rangle_{\text{in}}) &= (1 - e^{-\mu_D \eta})^2 \end{aligned} \tag{13}$$

TABLE I. Truth table of the detections (clicks) of the symbols in a Geiger-mode detector using two channels.

	Detector	
Input	CH_1	CH_2
$\|H\rangle$	Click	No-click
$\|V\rangle$	No-click	Click
$\|D\rangle$	Click	Click

where η is the efficiency of the detectors and μ_H, μ_V, and μ_D are, respectively, the mean photon numbers when sending $|H\rangle$ and projecting the measurement on CH_1, sending $|V\rangle$ and projecting on CH_2, and sending $|D\rangle$—the projecting detector should be irrelevant. Assuming the detectors are equally balanced, i.e., $\mu = \mu_H = \mu_V = 2\mu_D$, and the efficiency is maximum, i.e., $\eta = 1$, the equations can be written as

$$\begin{aligned} P(|H\rangle_{\text{out}} \,\|\, |H\rangle_{\text{in}}) &= P(|V\rangle_{\text{out}} \,\|\, |V\rangle_{\text{in}}) \\ &= (1 - e^{-\mu})e^{-\mu R_{\text{ext}}}, \\ P(|D\rangle_{\text{out}} \,\|\, |D\rangle_{\text{in}}) &= (1 - e^{-\mu/2})^2. \end{aligned} \tag{14}$$

Summing the three cases, we obtain the total detection probability we can expect from a certain extinction ratio (i.e., prediction accuracy). The dependency of the prediction accuracy from mean photon number μ and extinction ratio R_{ext} (defined in linear scale as the ratio of $|\alpha|^2$ and $|\beta|^2$ in $|\psi\rangle = \alpha\,|H\rangle + \beta\,|V\rangle$, and $R_{\text{ext}} = 10^{-(R_{\text{ext}})_{\text{dB}}/10}$) can be graphically visualized in Figs. 2 and 3.

Let us suppose Eve is in possession of a perfect photon-number-resolving detector with optimal detection ef-

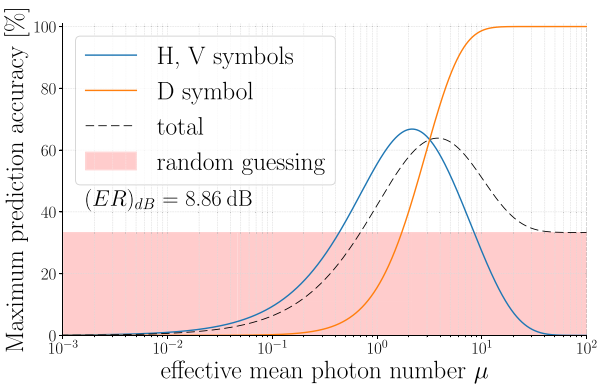

FIG. 3. Theoretical detection probabilities for each state and the total with respect to the mean photon number μ_{out} in the use case of only two detectors ($R_{\text{ext}} = 8.86$ dB).

TABLE II. Detection probabilities for each (Eve|Alice) case using G-M and P-N-R photodetectors. We considered a more realistic scenario for G-M detectors, introducing also a suboptimal extinction ratio (R_{ext}, in linear scale) coming from the projecting device at Eve's side. We identified as "vac" the vacuum state, where no detector clicks.

Probability	G-M	P-N-R
$\Pr(H\|H)$	$c(\mu_{\text{out}})\bar{c}(\mu_{\text{out}}R_{\text{ext}})$	$c(\mu_{\text{out}})$
$\Pr(V\|H)$	$c(\mu_{\text{out}}R_{\text{ext}})\bar{c}(\mu_{\text{out}})$	0
$\Pr(D\|H)$	$c(\mu_{\text{out}})c(\mu_{\text{out}}R_{\text{ext}})$	0
$\Pr(\text{vac}\|H)$	$\bar{c}(\mu_{\text{out}})\bar{c}(\mu_{\text{out}}R_{\text{ext}})$	$\bar{c}(\mu_{\text{out}})$
$\Pr(H\|V)$	$=\Pr(V\|H)$	0
$\Pr(V\|V)$	$=\Pr(H\|H)$	$c(\mu_{\text{out}})$
$\Pr(D\|V)$	$=\Pr(D\|H)$	0
$\Pr(\text{vac}\|V)$	$=\Pr(\text{vac}\|H)$	$\bar{c}(\mu_{\text{out}})$
$\Pr(H\|D)$	$c(\mu_{\text{out}}/2)\bar{c}(\mu_{\text{out}}/2)$	$c(\mu_{\text{out}}/2)\bar{c}(\mu_{\text{out}}/2)$
$\Pr(V\|D)$	$=\Pr(H\|D)$	$=\Pr(H\|D)$
$\Pr(D\|D)$	$c(\mu_{\text{out}}/2)c(\mu_{\text{out}}/2)$	$c(\mu_{\text{out}}/2)c(\mu_{\text{out}}/2)$
$\Pr(\text{vac}\|D)$	$\bar{c}(\mu_{\text{out}}/2)\bar{c}(\mu_{\text{out}}/2)$	$\bar{c}(\mu_{\text{out}}/2)\bar{c}(\mu_{\text{out}}/2)$

ficiency. We can model Eve's approach by describing the back-reflected photons as coherent states of the form

$$|\Psi_E\rangle = \frac{1}{\sqrt{2}}\left(|\mu^H_{\text{out}}\rangle + e^{i\varphi}|\mu^V_{\text{out}}\rangle\right) \tag{15}$$

where μ_{out} is the mean photon number coming back from Alice's system and φ holds the information on the encoding state given by the phase modulator. At $\mu_{\text{out}} = 0$ Eve is forced to randomly guess the symbol, therefore her $P^E_{\text{guess}}(0) = 1/3$. Already from $\mu_{\text{out}} = 1$, it can be seen that for Eve it is more convenient to use two channels for detection, because her $P^E_{\text{guess}}(1) = 2/3$, given that if she detects a photon on $|H\rangle$ or $|V\rangle$ her best strategy is to keep the received state, as she will guess correctly two times out of three. Even though she knows that the state she received could also be a $|D\rangle$ state, a random guess between these two symbols will yield a $P^E_{\text{guess}} = \frac{1}{2} < \frac{2}{3}$. Intuitively, increasing the number of detectors will just cause an increase in the uncertainty. For $\mu_{\text{out}} \geqslant 2$, we fall into the case described in Table I, therefore her $P^E_{\text{guess}}(\geqslant 2)$ will depend on the detection probabilities for each channel, namely, $\frac{1}{3}[\Pr(H|H) + \Pr(V|V) + \Pr(D|D)]$. In Table II we listed the detection probabilities Pr(Eve|Alice) both for photon-number-resolving and imperfect Geiger-mode photodetectors, noting as $c(\nu) = (1 - e^{-\nu})$ and $\bar{c}(\nu) = e^{-\nu}$ the functions, respectively, indicating the probability for a detector to click and not click. Given the aforementioned probabilities that Eve correctly guesses a symbol at a certain μ_{out}, we can derive the calculation of the total $P^E_{\text{guess}}(\mu_{\text{out}})$ as follows:

$$\begin{aligned}
&P^E_{\text{guess}}(\mu_{\text{out}}) \\
&\quad= \tfrac{1}{3}\Pr(\langle 0|\mu_{\text{out}}\rangle) + \tfrac{2}{3}\Pr(\langle 1|\mu_{\text{out}}\rangle) \\
&\qquad + P^E_{\text{guess}}(\geqslant 2)[1 - \Pr(\langle 0|\mu_{\text{out}}\rangle) - \Pr(\langle 1|\mu_{\text{out}}\rangle)] \\
&\quad= \tfrac{1}{3}e^{-\mu_{\text{out}}(\mu_{\text{out}}+1)}\big[\mu^2_{\text{out}} + e^{\mu^2_{\text{out}}}(-2e^{\mu_{\text{out}}/2} + 3e^{\mu_{\text{out}}} - 1) \\
&\qquad + 2e^{\mu_{\text{out}}/2}\left(\mu^2_{\text{out}} + 1\right) - e^{\mu_{\text{out}}}\left(\mu^2_{\text{out}} + 2\right) + 1\big].
\end{aligned} \tag{16}$$

The resulting guessing probability P^E_{guess} is displayed for the ideal photon-number-resolving detector with optimal extinction ratio as a green dot-dashed line, and for the Geiger-mode detector with different detection efficiencies and extinction ratios, in Fig. 2.

IV. STRONG LIGHT REGIME

The attack in the strong light regime has been performed employing commercially available photodiodes. These not only have bandwidths that allow for a full resolution of the spectral response, but can also estimate proportionally the quantity of incoming light, admitting a wide range of attacks, in contrast to the weak light regime where Geiger-mode-type photodetectors must be employed, limiting the possibilities. We acknowledge the fact that the methods hereby presented are suboptimal, and do not use all the information available from the total shape of the outputs. Machine-learning classifiers can be adopted to improve the results. For the scope of this paper, more centered on the countermeasures to these types of attacks, we will leave the introduction of machine-learning methods as a task for future works.

A. CW laser attack

For the hacking in the continuous regime, we made sure to fine-tune the first PC to enter Alice's setup with the $|D\rangle = \frac{1}{\sqrt{2}}(|H\rangle + |V\rangle)$ state. The second, analogously, serves to project the $|D\rangle$ state on the $|H\rangle$ and $|V\rangle$ state by means of the PBS and measure them with the photodiode. The resulting waveform should be equally split above and below the average, as it is visible on the screen of the oscilloscope reported in Fig. 5.

By taking that same plot and performing a "modulo-period" operation (in this case the period being 20 ns, since the repetition rate is 50 MHz), we obtain a superposition of the $|H\rangle$ and $|V\rangle$ symbols in the exact location in time where the symbols are (in the time span of one period) in the original waveform. We perform a two-dimensional histogram of these data, as for higher attenuations the shapes of the waveform get lost in the noise, obtaining a waveform characterized by the previously mentioned down-up behavior, like the one depicted in Fig. 4. From there we can easily estimate the location of the first symbol in the sequence, which automatically gives us all the following symbol locations by simply increasingly adding the period (20 ns).

After estimating all symbol locations, each symbol is classified as $|H\rangle$, $|V\rangle$, or $|D\rangle$ using thresholds, as shown in Fig. 5. The thresholds are chosen modeling the arrival of the heights of the symbols as Gaussian distributions with certain mean ν and variance σ^2 and using the *Bayes decision rule for minimum error*, which guarantees the lowest possible probability of error under known distributions and priors [30] and yields that the optimal threshold t^* minimizing the total error is

$$t^* = \arg\min_t E(t)$$

where $E(t)$ is the error function. In this sense, the best threshold happens to be found in the intersection point between the two normal distributions, because that is where the likelihood of observing the value x is the same for both the pair

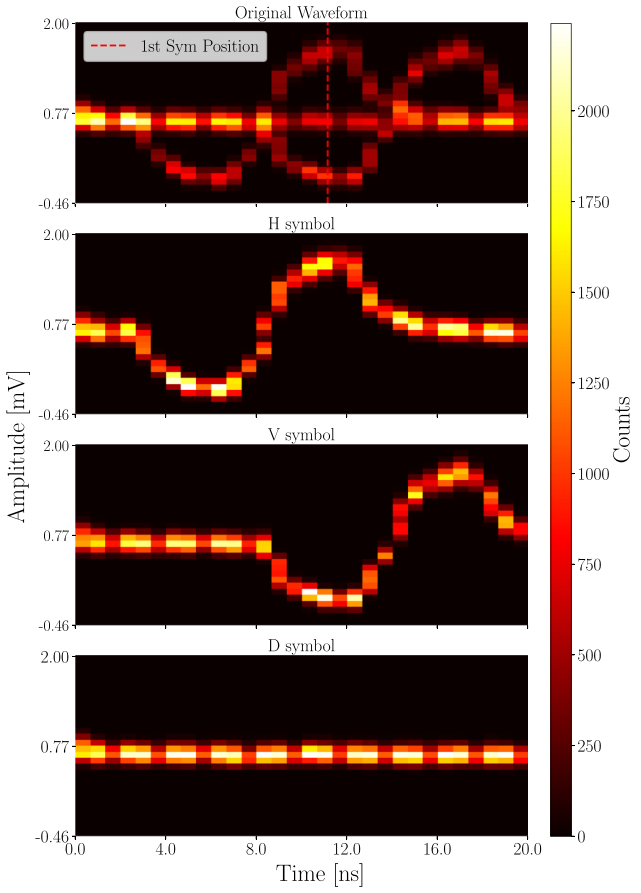

FIG. 4. Two-dimensional histogram of the waveform resulting from the oscilloscope, the result of the optical modulation in the continuous wavefront by the iPOGNAC during the THA, and $|H\rangle$, $|V\rangle$, and $|D\rangle$ symbols separately. All the waveforms presented are calculated by performing a modulo-period operation. In the top plot, the red vertical line shows the detected position of the first symbol in the original waveform.

distribution functions. In our scenario, the distributions of the $|H\rangle$, $|V\rangle$, and $|D\rangle$ symbols have very similar variances $\sigma_1^2 \simeq \sigma_2^2 = \sigma^2$. Therefore, the optimal threshold t^* can be obtained as the middle point of their averages ν_i:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t^* - \nu_1)^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t^* - \nu_2)^2}{2\sigma^2}\right).$$

Therefore $t^* = \frac{\nu_1+\nu_2}{2}$.

B. Pulsed laser attack

For hacking in the pulsed regime, following a method similar to that used in the previous section, a modulo-period operation is applied to the waveform, resulting in a superposition of the $|H\rangle$, $|V\rangle$, and $|D\rangle$ symbols, as illustrated in Fig. 6 (inside the pulse the three heights are visible, $|V\rangle$ at 0%, $|D\rangle$ at 50%, and $|H\rangle$ at 100%). Once the position of the first symbol is estimated by locating the maximum, all symbol positions are determined by iteratively adding the period.

As shown in Fig. 7, each symbol can be classified as $|H\rangle$, $|V\rangle$, or $|D\rangle$ using thresholds, after all the symbol locations have been determined. Thresholds are chosen using Gaussian

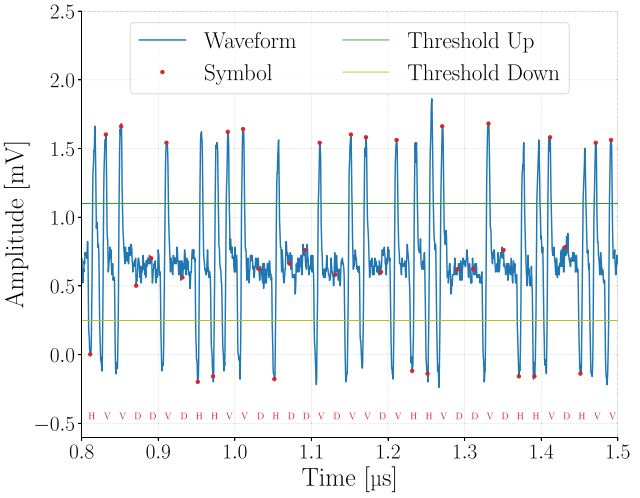

FIG. 5. Waveform with calculated symbol positions and classified symbols for the CWLA. Red dots show the estimated symbol locations, while thresholds are used to classify the symbol as $|H\rangle$, $|V\rangle$, or $|D\rangle$ (if the symbol location is above the threshold up, the symbol is classified as $|V\rangle$; if it is between the thresholds up and down, it is classified as $|D\rangle$; and if it is below the threshold down, it is classified as $|H\rangle$).

distributions of the symbols and the Bayes decision rule for minimum error, mentioned in the previous section.

C. Results for the strong light regime

The result of optimal prediction accuracy for continuous and pulsed laser attacks is shown in Fig. 8.

For the CW laser, we observed that the sequence can be fully reconstructed when the μ_{out} is larger than $\approx 10^7$, which corresponds to up to 8 dB of attenuation. The noise-floor

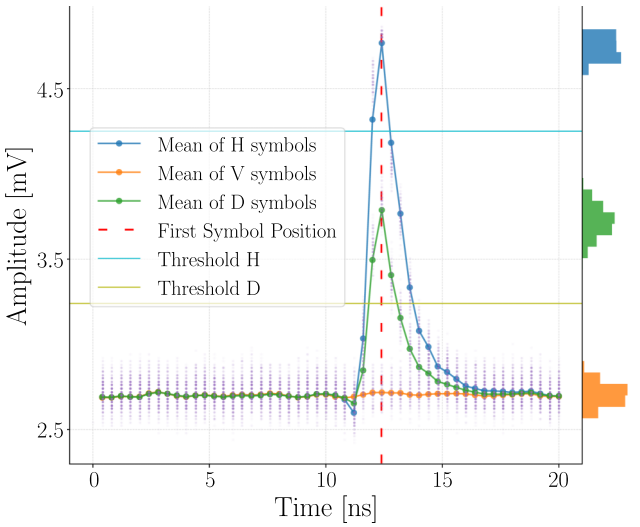

FIG. 6. Mean of $|H\rangle$, $|V\rangle$, and $|D\rangle$ symbols in the time-modulo period, the result of the optical modulation in the pulsed regime by the iPOGNAC during the THA. Data points are reported in purple. The red vertical line shows the detected position of the first symbol in the original waveform. Statistical distributions of these symbols are displayed on the right.

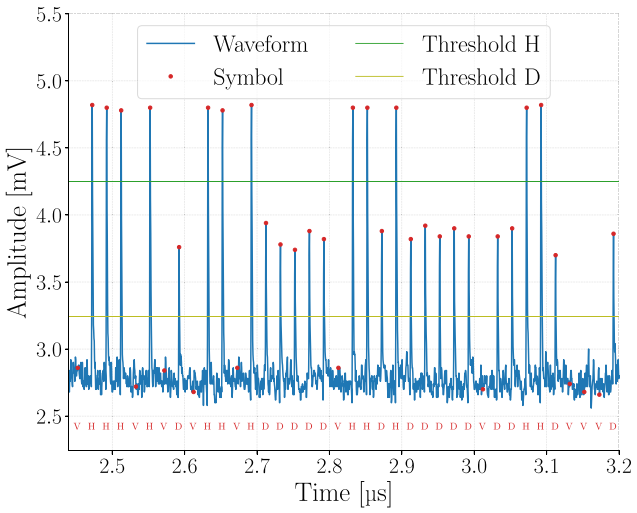

FIG. 7. Waveform with calculated symbol positions and classified symbols for the pulsed laser attack. Red dots show estimated symbol locations; thresholds are used to classify the symbol as $|H\rangle$, $|V\rangle$, or $|D\rangle$ (if the symbol is above the threshold H, the symbol is classified as $|H\rangle$; between the thresholds H and D, it is classified as $|D\rangle$; otherwise it is classified as $|V\rangle$).

limits are calculated by taking into account the sensitivities and noise floors of the equipment in use, e.g., photodiode and oscilloscope. From the setup scheme shown in Fig. 1, the laser from Eve's side passes through the VOA twice. Consequently, the effective attenuation on Eve's laser is doubled. Notably, for this test we used an *erbium-doped fiber amplifier* (EDFA) to increase the optical signal power in the case of the pulsed laser attack, which led to a noticeably higher average power of 60 mW. When compared to the CW case, this amplification allows for a gain of roughly ≈16.5 dB in the pulsed scenario, which increases Eve's capacity to compromise the transmission.

The mean photon number variation based on the attenuation set on the VOA is shown in Fig. 9. The theoretical output is calculated on the basis of the losses that occur in our setup. The total noise floor is calculated by the *root sum square* of the oscilloscope and photodiode standard deviations, namely,

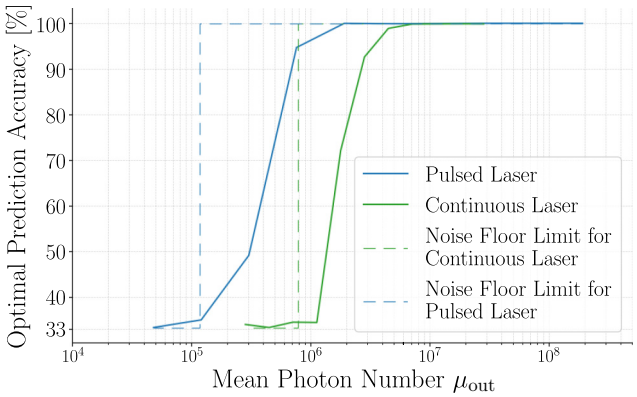

FIG. 8. Results on optimal prediction accuracy based on the mean photon number μ_{out} for the continuous and pulsed laser attack.

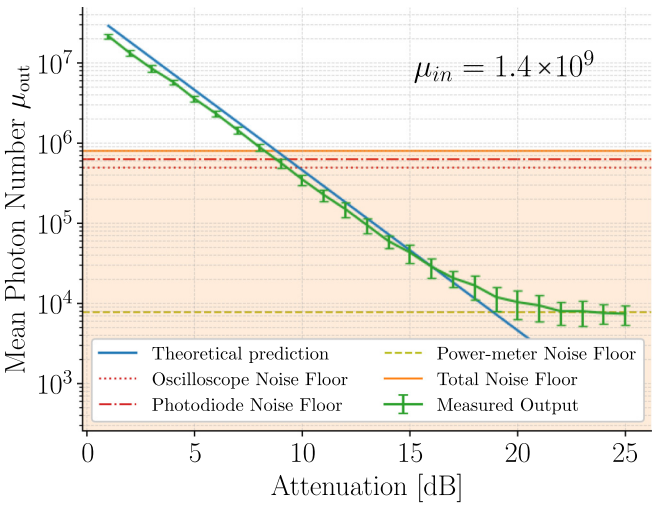

FIG. 9. Mean photon number μ_{out} at different attenuations set on the VOA. Noise floors thresholds are indicated with horizontal lines. The plateau after 15 dB is due to the noise floor of the power meter we used to characterize the output power (50 nW).

σ_{osc} and σ_{PD} as follows:

$$\sigma_{\text{total}} = \sqrt{\sigma_{\text{osc}}^2 + \sigma_{\text{PD}}^2} \simeq 5\ \mu W. \tag{17}$$

As seen in Fig. 9, around 8 dB, the output power of the CW laser is below the total noise floor, implying that the sequence cannot be predicted after this bound, consistent with the theoretical bound for the continuous laser shown in Fig. 8.

V. WEAK LIGHT REGIME

In Geiger-mode photodetectors like SNSPDs, the duration of time required for the current to fully return to the nanowire or for the voltage to drop to the baseline noise level is known as the detector's dead (reset) time [31]. The detectors we used have a dead time T_{dead} of ≈20 ns, therefore the maximum achievable repetition rate f_{max} for the pulsed signal can be

$$f_{\text{max}} \approx \frac{1}{T_{\text{dead}}} \approx 50\ \text{MHz}. \tag{18}$$

Given that during the period of dead time the detectors are unable to detect the incoming photons, there is a strong upper bound on the maximum achievable repetition rate, that is not only given by Eq. (18), but also by the saturation of the SPDs. In fact, a repetition rate of 50 MHz would yield meaningful results (i.e., at least two photons per pulse) only in a situation of maximum saturation of the single channels of the SNSPDs. Since this could lead to irreparable damages to the equipment, we opted to lower the repetition rate to 1 MHz to account for the dead time of the SNSPDs and lower the expected counts.

An alternative to counteract this problem is to exploit multiple channels of the SNSPDs and a $1 \times N$ beam splitter per PBS output. This could lead to an increase in the count rate, at the cost of increasing the number of channels used and at a required temporal alignment at the beginning, hypothetically doubling the maximum frequency with N. Since our analysis wants to tackle the issue with the least components possible,

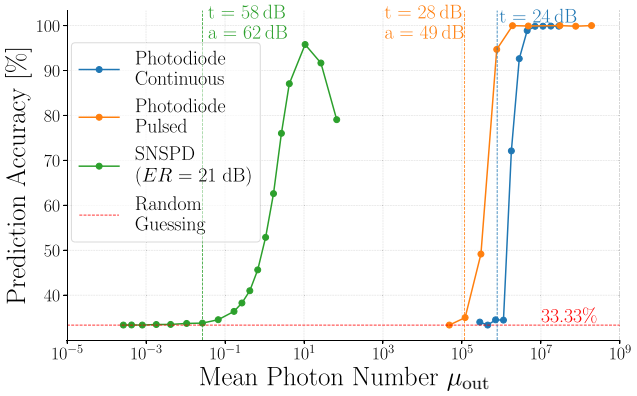

FIG. 10. Prediction accuracy in relation to the mean photon number μ_{out}, after normalizing the average power to 10 W for thermal damages (t) and 1 MW for ablation damages (a), for which the respective countermeasure attenuation is reported at the top.

we will not delve into this possibility, leaving it for future analyses.

Results for the weak light regime

Regarding the THA using SNSPDs as detectors, we measured the extinction ratio $(R_{\text{ext}})_{\text{dB}}$ of the PBS at our disposal (measuring ≈21 dB), and predicted the sequence for different levels of attenuation. In Fig. 2, not only the experimental results are visible, but also how different R_{ext}s highly influence the trend of the theoretical prediction accuracy curve in variation of μ. Our results show that a maximum of ≈95% prediction accuracy was achieved in near-perfect conditions of μ.

All these tests were performed in normal conditions of average power to give an estimate of the prediction accuracies. Potentially, real-world attackers can increase the power of their laser up to the maximum possible ratings of the components at their disposal (e.g., the maximum power a fiber can handle, which normally is around 10 W for the continuous regime and 1 MW for the pulsed regime [32,33]). For this purpose, we normalized the average powers of the results to a higher value to show the attenuation levels that are required by Alice to cope with such intensities. The results are collected and shown in Fig. 10, which shows that an attenuation of ≈60 dB involved in a THA will not lend meaningful results to Eve, even when the eavesdropper is employing single-photon detectors in their setup.

VI. COMPARISON OF THE ATTACKS

We studied a comparison of the attacks proposed in this paper, and collected them based on their complexities in Table III.

The continuous-wave attack does not require the alignment of the pulses with Alice's setup, differently from the pulsed-laser one. Once Alice and Eve are synchronized, achieved either having an electrical connection to Alice's SoC or exploiting clock-recovery algorithms like in Ref. [22], the continuous stream of photons provides all the information on the symbols being transmitted. Yet, this attack is the least

TABLE III. Use cases of the attacks proposed in this paper.

	Complexity		
Regime	Low	Medium	High
Strong	PD-CWLA	PD-PLA	
Weak			G-M

resistant to countermeasures, given that a considerable amount of optical power is wasted in useless information between two subsequent symbols. On the other hand, the pulsed-regime attack requires precise alignment of Eve's pulses with Alice's modulation at the beginning of the transmission. Once that is achieved, though, the stream is straightforward and the use of photodiodes (that can have bandwidths in the order of 2–20 GHz) guarantees every symbol of the sequence can be identified. It is important to acknowledge that all the results of the strong light regime are dependent on the total noise floor shown in Eq. (17) and Fig. 9, the sum of the ones of the photodiode and the oscilloscope. Improving the performances of these two components, it is possible to counteract more attenuation and reconstruct the sequence with better accuracies in the strong light regime. As shown in Fig. 10, we might conclude that the pulsed-laser attack yields meaningful results for just some dB of attenuation more than the continuous-wave attack. In reality, achieving higher peak powers is much easier than achieving higher average powers, for example using an EDFA, or by reducing the duty cycle of the pulses. Another advantage is that using a pulsed laser one can work closer to the thermal upper limit imposed by silica SM fibers, which is 1 MW for pulsed lasers [32,33]. In comparison to the photodiode used in the strong light scenario, attacks using SNSPDs are notably more complex, as reported in Table III. The main reason for this complexity is the interaction of many parameters that control the reliability and efficiency of SNSPDs. Moreover, key performance indicators such as the extinction ratio after the states are measured (the dependency of which is shown in Fig. 2), dark count rate, jitter time, and reset time must be carefully optimized in tandem to ensure accurate and efficient single-photon detection [34]. Better results on the prediction accuracy, using the same POVM, can be obtained by exploiting photon-number-resolving single-photon detectors, as we discussed in Sec. III.

VII. COUNTERMEASURES

Up to now we only considered passive countermeasures that involved attenuation at the entrance of Alice's setup using a VOA. This is because we investigated different types of THA counterattacks and categorized the most common defenses in (1) passive (a) filtering (using a *wavelength division multiplexer* filter in Alice's transmission band), (b) isolation (using an isolator or a circulator at Alice's output), and (c) attenuation (using optical attenuators at Alice's output) and (2) active *watchdog* (WD) detectors (exploiting a circulator connected to a detector at Alice's output).

Passive countermeasures can be all traced back to the attenuation type, since all of them can be overcome sending enough optical power [8,35]. By measuring input and output

power on different off-the-shelf components, we calculated for passive filtering and isolation to correspond to an approximate ≈60 dB of optical attenuation. The results that we reported confirm that even using SNSPDs as a means of attack is not enough to get useful results on a real-world-scenario QKD setup, since the transmission rates are generally higher (transmission frequencies higher than 50–100 MHz are sufficient with state-of-the-art equipment) and an isolator is often sufficient to block the incoming light in Alice's system. Furthermore, reducing back-reflections is a proactive measure that enhances overall security. This can be achieved by using angle-polished connectors instead of flat connectors, eliminating open ports, and fusing connections where possible. These methods can limit attack opportunities regardless of the spectral characteristics of individual components [9].

An active countermeasure system can also be used, exploiting what are usually known as watchdog detectors. These are devices that measure the light impinging in Alice's setup by means of a circulator situated at the output of the system. These are particularly helpful because an attacker can be immediately spotted by measuring its optical power (that is generally high). There are three cases a WD and an eavesdropper detector (ED) can face.

(1) *The noise floor of the WD is lower than the one of the ED*: in this case the attacker is caught and the communication terminates.

(2) *The noise floor of the WD is higher than the one of the ED*: in this case Alice could not perceive the presence of the attacker and can be hacked.

(3) *The noise floors of the WD and of the ED are comparable*: in this case the attacker is surely spotted, since it needs to shoot much higher power to cope with Alice's setup's internal attenuations.

Therefore, an eavesdropper should aim at using detectors with very low noise level (such as single-photon detectors) to counteract the use of watchdog detectors. Despite that, the effectiveness of both monitoring detectors (as well as the other passive countermeasures) can be limited by their spectral responsiveness, potentially allowing attacks at wavelengths where their sensitivity is negligible [9]. In this sense, having Eve sending short pulses could limit the visibility of the watchdog detector, not triggering the power measure. While a watchdog detector can be an effective defense, it is also susceptible to attacks. We suggest to place the detector behind the attenuation and filtration stage so it is more protected from blinding attacks. A "dual-threshold" monitoring strategy can also be implemented, where Alice monitors both the ac component (to detect pulses) and the dc baseline (to detect blinding or saturation attempts), ensuring that the countermeasure's attenuation remains an effective barrier even under high-power illumination. Thermal power meters, which have extremely high tolerances in the orders of 10 kW of power, can also be used as watchdog detector, even in the dual-threshold scheme.

In conclusion, a near-optimal countermeasure to a THA can be achieved by mixing all of the proposed solutions above.

We now try to estimate a lower bound on the attenuation required to counteract these attacks based on the assumption that the *laser-induced damage threshold* of the system is set at 10 W for thermal damages and 1 MW for ablation damages.

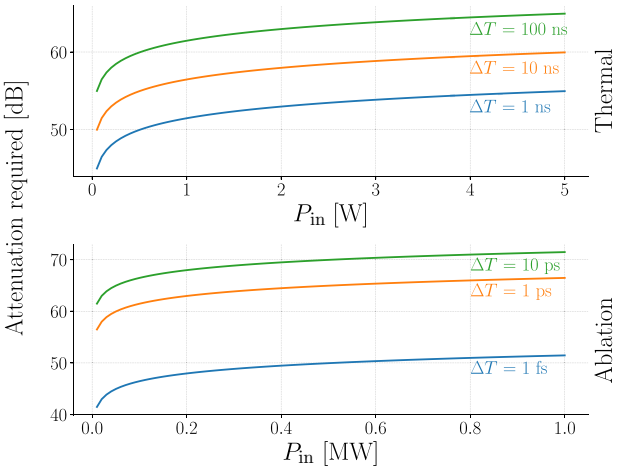

FIG. 11. Attenuation required to counteract the THA presented in this work in different scenarios of input peak power P_{in} and pulse width ΔT, both for thermal (top) and ablation (bottom) damage limits.

As shown in Fig. 2, a mean photon number of $\mu_{\text{out}} \approx 5$–$8$ photons per symbol is sufficient for Eve to obtain a successful attack with a prediction accuracy higher than 80% in good conditions of the extinction ratio. In order to have full security on the iPOGNAC, a stricter bound can be placed at $\mu_{\text{out}} < 0.1$ photons per symbol to get Eve's prediction accuracy closer to the random guessing limit of 33%. Therefore, we can give an estimate on the countermeasures that can be taken in these conditions. Assuming a total loss inside the iPOGNAC (excluded the VOA) in the best case (not accounting for coupling losses or imperfections in the components) of $\Delta P \simeq 6$ dB, we can derive the amount of attenuation at the output that can counteract the attack:

$$A_{\text{dB}} = \frac{1}{2}\left[10\log_{10}\left(\frac{\mu_{\text{in}}}{\mu_{\text{out}}}\right) - \Delta P\right] \tag{19}$$

where the $1/2$ is because of the optical pulse crossing two times the attenuation imposed by the VOA, and μ_{in} is solely dependent on parameters of the attacker:

$$\mu_{\text{in}} = \frac{\lambda P_{\text{in}} \Delta T}{hc} \tag{20}$$

where h is Planck's constant (6.6261×10^{-34} J s), c is the speed of light in vacuum (2.9979×10^{8} m/s), while P_{in}, λ, and ΔT are respectively the optical peak power in watts, the wavelength, and the width of the pulse of Eve's laser in seconds. In Fig. 11 the required levels of attenuation required to counteract different levels of P_{in} are shown. With ≈65–70 dB of attenuation, i.e., ≈130–140 dB of isolation, at the output, we can consider the iPOGNAC to be secure against the types of THA presented in this paper, which is in line with what is presented by Lucamarini *et al.* in Ref. [11] and Luo *et al.* in Ref. [17]. We want to underline the fact that the bound on the mean photon number μ_{out} is valid for any encoding setup subject to Trojan-horse attacks as it poses a fundamental limit on the mutual information shared between Alice and Eve. In contrast, the bound on the attenuation is unique from setup to

setup and therefore nonretroreflective ones might require less attenuation than what we found in this analysis.

On average, the output power coming out of current QKD systems employing the iPOGNAC revolves around $\mu \simeq 10^5$ photons per symbol, therefore already requiring an attenuation of around 50 dB to reach a typical mean photon number of 0.6 photons per symbol. As we have shown with our work, this is already a reasonable amount of countermeasure that deprives Eve of a lot of information. In order to be completely secure from this type of attacks, adding an isolator at the output (which when traveling in the opposite direction adds an attenuation of at least 30 dB) can be a good habit for state-of-the-art QKD systems in general.

VIII. CONCLUSION

We analyzed different methodologies to approach Trojan-horse attacks that can be executed on a QKD setup that exploits three-state efficient BB84 and a modulation scheme like the one of the iPOGNAC, a self-compensating all-fiber polarization encoder first proposed in Ref. [13]. We proposed two different categories of attack, in strong and and weak light regimes, where in the former the detection can be done exploiting photodiodes while in the latter higher-efficiency single-photon detectors must be used. Even though the attack types are different in the two regimes (utilizing a more deterministic approach in one and a more probabilistic one depending on the mean photon number in the other) we estimated that the weak regime can be considered successful with countermeasures that comprise attenuations ≈30 dB higher than the ones required in the strong light regime, allowing one to detect the sequence even in conditions of extremely low average power. However, the prediction accuracy is not as stable and deterministic as in the strong regime because of the detector types (the ones we adopted are Si photodiodes). Using photodiodes, in fact, the prediction accuracy follows a shape similar to an "inversed sigmoid function," from 100 to 33.33% (equivalent to a random guessing), while for single-photon detectors the prediction accuracy is highly limited by the saturation of the detectors and the extinction ratio of the polarization states.

We have shown that an effective and immediate Trojan-horse attack is possible in the continuous waveform regime (without having to adjust the delays between a light pulse and the modulation of Alice), that the same kind of attack in the pulsed regime can overcome more easily higher defense layers, and that attacks are also possible at higher attenuation levels using single-photon detectors with high sensitivity like SNSPDs. We investigated different types of countermeasures that can be applied to the setup that can limit the leaked information at Eve's side with ease. In a broader sense, we have estimated that having more than ≈65–70 dB of attenuation (≈130–140 dB of isolation) at Alice's output is in many cases a good habit to limit the effects of Trojan-horse attacks like the ones proposed, and that this can be improved in combination with wavelength filtering or optical isolation. We gave an estimate of the theoretical mutual information that can be extracted by the attack in the weak regime, both in a hypothetical scenario using an optimal POVM using perfect photon-number-resolving detectors and in a more realistic scenario using SNSPDs. Comparing our work to Luo *et al.*'s [17], the same conclusions regarding the amount of attenuation required are reached, but the methods are different. Our work has an overall more complete analysis of why these bounds need to be taken into account when designing a QKD source, giving at the same time both an upper bound on the mutual information on the eavesdropper side and a practical example on how these attacks can be conducted using state-of-the-art technologies. For future works, we are planning the implementation of a machine-learning classifier to distinguish the states in the strong light regime.

ACKNOWLEDGMENTS

A.D.T. acknowledges the financial support of Concessioni Autostradali Venete (CAV) S.p.A. in the framework of the doctoral scholarship agreement 38° Ciclo between CAV and the University of Padova. This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under the project "Quantum Secure Networks Partnership" (Grant Agreement No. 101114043). The University of Padova is participating in the EC funded project Nostradamus, Project No. CNECT/2023/OP/0032, in the role of contractor. The authors thank the whole project team for the support and valuable exchange. The authors would like also to thank the project Q-SecGround Space of the Italian Space Agency for providing the preliminary research basis for this work.

A.C.A., A.D.T., C.A., and D.G.M. contributed to the experimental design, tests, and analyses. A.D.T., C.A., and G.V. contributed in the development of the theoretical models of the attack. The manuscript was drafted and written by A.C.A. and A.D.T., and reviewed by C.A., G.V., and P.V.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

APPENDIX: CALCULATION OF THE VON NEUMANN ENTROPY

In this section, we show how to compute the von Neumann entropy of the state ρ received by Eve, where $\rho = \frac{1}{3}\sum_{j=1}^{3} |\psi_j\rangle\langle\psi_j|$ and the states $|\psi_j\rangle$ are defined in Eq. (2). The von Neumann entropy of ρ is equal to the Shannon entropy of its eigenvalues. Therefore, we need to evaluate the eigenvalues of ρ.

Since the three states $\{|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle\}$ are linearly independent they can be considered as the first three elements of the basis. Since the Hilbert space is infinite dimensional, we can complete the basis with orthonormal vectors $\{|\psi_4\rangle \ldots |\psi_\infty\rangle\}$ that are orthogonal to the subspace defined by $|\psi_j\rangle$ with $(j = 1, 2, 3)$.

Since ρ has support only in the three-dimensional subspace spanned by $\{|\psi_j\rangle\}$, we can restrict our attention to such sub-

space, in which the basis elements ρ_{jk} of ρ are given by the following equation:

$$\rho\,|\psi_k\rangle = \sum_j \rho_{jk}\,|\psi_j\rangle \tag{A1}$$

and since $\rho = \frac{1}{3}\sum_{j=1}^{3} |\psi_j\rangle\langle\psi_j|$ we have that

$$\rho\,|\psi_k\rangle = \frac{1}{3}\sum_{j=1}^{3} |\psi_j\rangle\langle\psi_j|\psi_k\rangle = \sum_{j=1}^{3}\left(\frac{1}{3}\langle\psi_j|\psi_k\rangle\right)|\psi_j\rangle\,.$$

Comparing with (A1) it is clear that

$$\rho_{jk} = \tfrac{1}{3}\langle\psi_j|\psi_k\rangle.$$

Therefore, the matrix ρ in the $\{|\psi_j\rangle\}$ basis is given by

$$\rho = \frac{1}{3}\begin{pmatrix} 1 & e^{-\mu} & e^{-\mu(1-\frac{1}{\sqrt{2}})} \\ e^{-\mu} & 1 & e^{-\mu(1-\frac{1}{\sqrt{2}})} \\ e^{-\mu(1-\frac{1}{\sqrt{2}})} & e^{-\mu(1-\frac{1}{\sqrt{2}})} & 1 \end{pmatrix} \tag{A2}$$

where the diagonal elements $\langle\psi_i|\psi_i\rangle$ correspond to the squared norms of the vectors. Since coherent states are normalized, these are all equal to 1. The components ρ_{ij} can be evaluated by recalling that the general scalar product between two single-mode coherent states is

$$\langle\alpha|\beta\rangle = \exp\left(\alpha^*\beta - \frac{|\alpha|^2}{2} - \frac{|\beta|^2}{2}\right). \tag{A3}$$

Therefore,

$$\begin{aligned}
\langle\psi_1|\psi_2\rangle &= \langle\sqrt{\mu}|0\rangle_H\langle 0|\sqrt{\mu}\rangle_V = e^{-\mu}, \\
\langle\psi_1|\psi_3\rangle &= \left\langle\sqrt{\mu}\middle|\sqrt{\frac{\mu}{2}}\right\rangle_H\left\langle 0\middle|\sqrt{\frac{\mu}{2}}\right\rangle_V = e^{-\mu(1-\frac{1}{\sqrt{2}})}, \\
\langle\psi_2|\psi_3\rangle &= \left\langle 0\middle|\sqrt{\frac{\mu}{2}}\right\rangle_H\left\langle\sqrt{\mu}\middle|\sqrt{\frac{\mu}{2}}\right\rangle_V = e^{-\mu(1-\frac{1}{\sqrt{2}})}
\end{aligned} \tag{A4}$$

and the other terms are evaluated by using the relation $\langle\psi_j|\psi_i\rangle = \langle\psi_i|\psi_j\rangle^*$.

Calculating the eigenvalues of the matrix ρ, we obtain

$$\lambda_1(\mu) = \frac{1}{3} - \frac{e^{-\mu}}{3}, \tag{A5}$$

$$\lambda_{2,3}(\mu) = \frac{1}{3} + \frac{e^{-\mu}}{6}(1 \pm \sqrt{1 + 8e^{\sqrt{2}\mu}}). \tag{A6}$$

Using the eigenvalues just found, we can calculate the Shannon entropy of the ensemble state ρ as

$$H(\mu) := -\sum_{i=1}^{3}\lambda_i(\mu)\log_2[\lambda_i(\mu)]. \tag{A7}$$

[1] C. H. Bennett and G. Brassard, Quantum cryptography: Public key distribution and coin tossing, Theor. Comput. Sci. **560**, 7 (2014).

[2] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Quantum cryptography, Rev. Mod. Phys. **74**, 145 (2002).

[3] M. Lucamarini, Z. L. Yuan, J. F. Dynes, and A. J. Shields, Overcoming the rate-distance limit of quantum key distribution without quantum repeaters, Nature (London) **557**, 400 (2018).

[4] S. Wang, Z.-Q. Yin, D.-Y. He, W. Chen, R.-Q. Wang, P. Ye, Y. Zhou, G.-J. Fan-Yuan, F.-X. Wang, W. Chen, Y.-G. Zhu, P. V. Morozov, A. V. Divochiy, Z. Zhou, G.-C. Guo, and Z.-F. Han, Twin-field quantum key distribution over 830-km fibre, Nat. Photon. **16**, 154 (2022).

[5] G. Currás-Lorenzo, M. Pereira, G. Kato, M. Curty, and K. Tamaki, Security framework for quantum key distribution with imperfect sources, Optica Quantum **3**, 525 (2025).

[6] N. Gisin, S. Fasel, B. Kraus, H. Zbinden, and G. Ribordy, Trojan-horse attacks on quantum-key-distribution systems, Phys. Rev. A **73**, 022320 (2006).

[7] N. Jain, E. Anisimova, I. Khan, V. Makarov, C. Marquardt, and G. Leuchs, Trojan-horse attacks threaten the security of practical quantum cryptography, New J. Phys. **16**, 123030 (2014).

[8] A. Vakhitov, V. Makarov, and D. R. Hjelme, Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography, J. Mod. Opt. **48**, 2023 (2001).

[9] N. Jain, B. Stiller, I. Khan, V. Makarov, C. Marquardt, and G. Leuchs, Risk analysis of Trojan-horse attacks on practical quantum key distribution systems, IEEE J. Sel. Top. Quantum Electron. **21**, 168 (2015).

[10] B. Qi, Trustworthiness of detectors in quantum key distribution with untrusted detectors, Phys. Rev. A **91**, 020303(R) (2015).

[11] M. Lucamarini, I. Choi, M. Ward, J. Dynes, Z. Yuan, and A. Shields, Practical security bounds against the Trojan-horse attack in quantum key distribution, Phys. Rev. X **5**, 031030 (2015).

[12] S. Sajeed, I. Radchenko, S. Kaiser, J.-P. Bourgoin, A. Pappa, L. Monat, M. Legré, and V. Makarov, Attacks exploiting deviation of mean photon number in quantum key distribution and coin tossing, Phys. Rev. A **91**, 032326 (2015).

[13] M. Avesani, C. Agnesi, A. Stanco, G. Vallone, and P. Villoresi, Stable, low-error, and calibration-free polarization encoder for free-space quantum communication, Opt. Lett. **45**, 4706 (2020).

[14] M. Avesani, G. Foletto, M. Padovan, L. Calderaro, C. Agnesi, E. Bazzani, F. Berra, T. Bertapelle, F. Picciariello, F. B. L. Santagiustina, D. Scalcon, A. Scriminich, A. Stanco, F. Vedovato, G. Vallone, and P. Villoresi, Deployment-ready quantum key distribution over a classical network infrastructure in Padua, J. Lightwave Technol. **40**, 1658 (2022).

[15] D. Scalcon, C. Agnesi, M. Avesani, L. Calderaro, G. Foletto, A. Stanco, G. Vallone, and P. Villoresi, Cross-encoded quantum key distribution exploiting time-bin and polarization states with qubit-based synchronization, Adv. Quantum Technol. **5**, 2200051 (2022).

[16] M. Sabatini, T. Bertapelle, P. Villoresi, G. Vallone, and M. Avesani, Hybrid encoder for discrete and continuous variable QKD, Adv. Quantum Technol. **8**, 2400522 (2025).

[17] T. Luo, Q. Liu, X. Sun, C. Huang, Y. Chen, Z. Zhang, and K. Wei, Security analysis against the Trojan horse attack on prac-

tical polarization-encoding quantum key distribution systems, Phys. Rev. A **109**, 042608 (2024).
[18] F. Berra, C. Agnesi, A. Stanco, M. Avesani, S. Cocchi, P. Villoresi, and G. Vallone, Modular source for near-infrared quantum communication, EPJ Quantum Technol. **10**, 27 (2023).
[19] M. Avesani, L. Calderaro, G. Foletto, C. Agnesi, F. Picciariello, F. B. L. Santagiustina, A. Scriminich, A. Stanco, F. Vedovato, M. Zahidy, G. Vallone, and P. Villoresi, Resource-effective quantum key distribution: A field trial in Padua city center, Opt. Lett. **46**, 2848 (2021).
[20] C. Agnesi, M. Giacomin, D. Sartorato, S. Artuso, G. Vallone, and P. Villoresi, In-field comparison between G.652 and G.655 optical fibres for polarisation-based quantum key distribution, IET Quantum Comm. **5**, 567 (2024).
[21] E. Monmasson and M. N. Cirstea, FPGA design methodology for industrial control systems—A review, IEEE Trans. Ind. Electron. **54**, 1824 (2007).
[22] L. Calderaro, A. Stanco, C. Agnesi, M. Avesani, D. Dequal, P. Villoresi, and G. Vallone, Fast and simple qubit-based synchronization for quantum key distribution, Phys. Rev. Appl. **13**, 054041 (2020).
[23] F. Grünenfelder, A. Boaron, D. Rusca, A. Martin, and H. Zbinden, Simple and high-speed polarization-based QKD, Appl. Phys. Lett. **112**, 051108 (2018).
[24] D. Bacco, I. Vagniluca, B. Da Lio, N. Biagi, A. Della Frera, D. Calonico, C. Toninelli, F. S. Cataliotti, M. Bellini, L. K. Oxenløwe, and A. Zavatta, Field trial of a three-state quantum key distribution scheme in the Florence metropolitan area, EPJ Quantum Technol. **6**, 5 (2019).
[25] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, Gaussian quantum information, Rev. Mod. Phys. **84**, 621 (2012).
[26] C. W. Helstrom, Quantum detection and estimation theory, J. Stat. Phys. **1**, 231 (1969).
[27] A. Tavakoli, A. Pozas-Kerstjens, P. Brown, and M. Araújo, Semidefinite programming relaxations for quantum correlations, Rev. Mod. Phys. **96**, 045006 (2024).
[28] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications* (SIAM, Philadelphia, PA, 2001).
[29] A. Assalini, G. Cariolaro, and G. Pierobon, Efficient optimal minimum error discrimination of symmetric quantum states, Phys. Rev. A **81**, 012315 (2010).
[30] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 2000).
[31] C. Autebert, G. Gras, E. Amri, M. Perrenoud, M. Caloz, H. Zbinden, and F. Bussières, Direct measurement of the recovery time of superconducting nanowire single-photon detectors, J. Appl. Phys. **128**, 074504 (2020).
[32] P. Peterka, D. Pugliese, B. Jiříčková, N. G. Boetti, H. Turčičová, I. Mirza, A. Borodkin, and D. Milanese, High-power laser testing of calcium-phosphate-based bioresorbable optical fibers, Opt. Mater. Express **11**, 2049 (2021).
[33] A. V. Smith, B. T. Do, G. R. Hadley, and R. L. Farrow, Optical damage limits to pulse energy from fibers, IEEE J. Sel. Top. Quantum Electron. **15**, 153 (2009).
[34] S. Tripathy, K. Tyagi, and P. Pratap, A comprehensive study of various superconductors for superconducting nanowire single photon detectors applications, iScience **27**, 110779 (2024).
[35] A. Ponosova, D. Ruzhitskaya, P. Chaiwongkhot, V. Egorov, V. Makarov, and A. Huang, Protecting fiber-optic quantum key distribution sources against light-injection attacks, PRX Quantum **3**, 040307 (2022).