

INFORMATION SCIENCE

Penetration testing of quantum key distribution system as a black box

Anqi Huang^{1,2,3,†,*}, Qingquan Peng^{3,†}, Junxuan Liu^{3,†}, Xialong Yuan^{3,*}, Cheng Peng³, Zhengyuan Yang³, Hansen Wu³, Zihao Chen³, Guangfu Sun^{1,2}, Feixue Wang^{1,2,*} and Vadim Makarov^{4,5,6}

ABSTRACT

Quantum key distribution (QKD) establishes a shared secret between remote parties and is proven unbreakable in theory. Unfortunately, practical implementations of QKD have device imperfections leading to security vulnerabilities. Most of these have been verified in a white-box testing scenario, when one has access to the system hardware for its analysis. Here we implement an automated penetration testing on a QKD system as a black box, using only its public communication lines and limited operator's manual. Our implementation parses information transmitted in the classical communication line and toggles an optical delay in the quantum communication line. This allows it to tamper with timing settings of detector gates in the QKD system during its calibration procedure and passively eavesdrop 98.97% of the sifted key. The entire testing is fully automated and takes minutes to begin the eavesdropping. Our work paves the way for automated penetration testing of QKD installations as security verification.

Keywords: Quantum key distribution, Penetration testing, Black box

INTRODUCTION

Quantum key distribution (QKD) is one of the promising technologies in quantum cryptography, allowing two parties to securely share a random and secret key through a public channel whose security is guaranteed by the principles of quantum physics [1]. Technically, commercial QKD systems are available [2–8], and advanced QKD implementations reach distances up to 1002 km [9]. Although the QKD protocol provides information-theoretic security in principle [10], the deviation of hardware equipment from the ideal model in practice remains a vital problem to be solved. Various attacks exploit imperfections in components of the QKD system, such as photon sources [11–15], quantum-state modulators [12,16–19], and detectors [20–27], to eavesdrop on secure keys. So far, all these known attacks are white-box ones, requiring the adversary to know the internal details of the QKD system and finely tune complex parameters to execute a successful attack [21,28,29]. Certification methodology for QKD is also built around the white-box scenario [30,31]. Meanwhile,

penetration testing with limited or no knowledge of the target system is standard in the security industry [32,33]. This demand becomes more urgent as quantum communication moves towards scalable networked architectures [34,35].

In this work, we propose and implement a plug-and-play attack on a decoy-state BB84 [36–38] QKD system as a black box. That is, Eve is only permitted to launch attacks via the quantum and classical channels with no access to the inside of the QKD system. Instead, she only obtains the public information about the tested QKD system from the user manual and the software interface, such as the protocol employed, repetition frequency, optical wavelength, and system operating process, but has no prior knowledge about the existence of any vulnerability. By analyzing public documentation and the classical information exchanged between Alice and Bob, we identify the vulnerability during a certain calibration procedure, in which Eve actively induces a timing mismatch between the gate positions of paired single-photon detectors (SPDs) within a measurement basis [39–41].

¹College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China;

²Center for Cryptologic Research, National University of Defense Technology, Changsha 410073, China;

³College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China;

⁴Russian Quantum Center, Skolkovo, Moscow 121205, Russia;

⁵Vigo Quantum Communication Center, University of Vigo, Vigo E-36310, Spain;

⁶NTI Center for Quantum Communications, National University of Science and Technology MISiS, Moscow 119049, Russia

*Corresponding authors.

Email: angelhuang_hn@gmail.com; 2350312617@qq.com; fxwang@nudt.edu.cn.

[†]Equally contributed to this work..

Received: XX XX Year;

Revised: XX XX Year;

Accepted: XX XX Year

Subsequently, in the raw key exchange phase, Eve exploits this mismatch to eavesdrop on the secret key information by continuously manipulating the arrival time of quantum states. Notably, this attack is automated and operates independently of the encoding degrees of freedom utilized in the QKD system under test. Moreover, it manipulates the transmission time and path of quantum states without requiring any intercept-resend operations [42,43].

Based on this security vulnerability, we implement a plug-and-play hacking (PPH) apparatus in realistic conditions over channels between Alice and Bob. Our apparatus comprises both hardware and software elements. Its hardware allows Eve to adjust the length of quantum channel and wiretaps information from the classical channel. Meanwhile, the software interprets all data in the classical channel and dynamically tunes attack parameters in real time based on the status of the QKD system under test. Through this process, it successfully obtains Bob's basis choices and calculates his partial quantum bit error rate (QBER) in both bases. By correlating this information with the length switch between short and long paths, Eve deduces Bob's sifted key. The capability of automatically executing attacks and optimizing hacking parameters within minutes is demonstrated on multiple QKD sessions lasting a few hours, successfully eavesdropping 98.97% sifted key.

RESULTS

System under test

The engineering-validated QKD prototype served as the black-box system under test. From the public documentation, it is known that this QKD system features four detectors and employs a decoy-state BB84 QKD protocol with polarization coding and passive basis selection. The system, integrating both fully functional hardware and software, operates automatically to ensure reliable generation of secure keys.

Figure 1a depicts the schematic of the hardware in the QKD system. Alice transmits quantum-state pulses at the repetition frequency of 40 MHz, which are then sent over a single-mode fiber together with synchronization pulses by dense wavelength-division-multiplexing (DWDM) technique. After receiving the combined optical signals from Alice, Bob uses his DWDM3 module to demultiplex the synchronization pulses and quantum-state pulses. The former are detected by a photoelectric detector (PD2) to synchronize Bob's

clock with Alice's. As for the quantum signals, they are split into two equal parts via a 50:50 beam splitter (BS3). Each part individually undergoes polarization correction before being directed through a polarization beam splitter (PBS). It separates orthogonally polarized quantum states, which are then detected by two SPDs.

The software operation sequence of the QKD system is shown in Fig. 1b, which is observed from the user interface. It has three stages: initialization, calibration, and quantum key distribution. During initialization, the software is started and the connection between Alice and Bob is established. Calibration involves three phases as follows. First, the delay scanning, in which Alice sends four polarizations, namely $|H\rangle$ (horizontal polarization state), $|P\rangle$ (45° polarization state), $|V\rangle$ (vertical polarization state), and $|N\rangle$ (135° polarization state). Bob then determines each SPD's gate position by detecting the maximum number of counts.

Second, the polarization correction. Alice initially transmits $|H\rangle$ states, and Bob adjusts his polarization controller (PC3) until his click ratio $|H\rangle:|V\rangle$ reaches 99:1. Then, Alice sends $|P\rangle$ states, and Bob adjusts PC4 using the same procedure to achieve a similar click ratio. Third, the synchronization alignment, in which Bob locks his clock with Alice's using synchronization pulses.

At the quantum key distribution stage, Alice randomly sends four polarization states $|H\rangle$, $|V\rangle$, $|P\rangle$, and $|N\rangle$. Bob randomly chooses either X or Z basis for detection via his scheme of passive basis selection. Alice and Bob repeat this process to accumulate a sufficiently long raw key over a single raw key exchange period. Then they announce the basis of click event to obtain the sifted key and reveal a small portion of sifted key to calculate the QBER. If QBER is less than 3%, the system operates post-processing, such as error correction and privacy amplification. Otherwise, the system runs polarization correction, trying to reduce the QBER. If no key is generated within 10 minutes, the QKD system restarts from the beginning of calibration.

Vulnerabilities of black-box QKD system

Eve wiretaps classical channel to intercept the classical communications between the transmitter, Alice, and the receiver, Bob. By analyzing the public documentation and the classical information exchanged between Alice and Bob, we identify the vulnerability during the calibration procedure. We infer that Eve can actively modify the gate position of detectors and synchro-

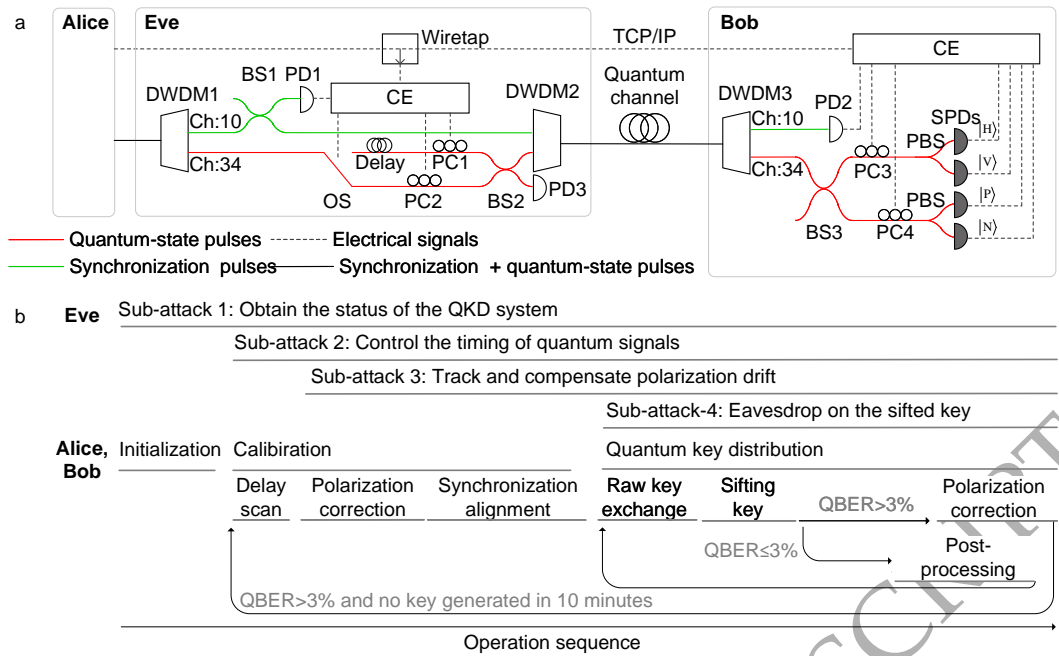


Figure 1. Plug-and-play hacking experiment. (a) Scheme of Eve and the black-box QKD system. Alice and Bob are black boxes, whereas Bob's internal structure is a speculative diagram. DWDM, dense-wavelength-division multiplexer; BS, beam splitter; PBS, polarization beam splitter; OS, optical switch; PD, photoelectric detector; SPD, single-photon detector; CE, control electronics; PC, programmable polarization controller; Ch, wavelength channel; $|H\rangle$, $|V\rangle$, $|P\rangle$, $|N\rangle$, SPD output signals corresponding to the four polarization states. (b) Timing of the four subattacks in the PPH apparatus.

nization parameters only by hacking the QKD system through both the quantum channel and classical channel. Her detailed workflow is as follows.

In the QKD system, Bob independently adjusts each SPD's gate position to achieve optimal detection efficiency during the calibration procedure. This process allows Eve to manipulate the time of quantum state arriving at Bob by choosing a long or a short path of the quantum channel, thereby independently controlling the gate position of each SPD [41]. As a result, the mismatch between gate positions of the paired SPDs in a measurement basis is actively created by Eve. Then, during the process of raw key exchange, Eve can exploit this mismatch to learn the information of secret key by continuously controlling the arrival time of quantum state to Bob, just as she does during the calibration.

Implementation of plug-and-play attack

Eve's PPH apparatus (Fig. 1a) is independent of the encoding degrees of freedom the system under test. It controls the quantum and classical channels. Eve separates Alice's signal into 1550.12 nm (wavelength channel 34) quantum signal and 1569.59 nm (wavelength channel 10) synchronization signal via DWDM1. The synchronization signal is then split by BS1. Its one

part enters DWDM2 and the other part is detected by PD1 that generates a trigger signal for the control electronics (CE). The latter routes the quantum signal through an optical switch (OS) to either a long or short path, thereby controlling its arrival time at Bob. Both paths are equipped with programmable PCs to compensate for polarization drift. The quantum-state pulses and synchronization pulses are subsequently combined via BS2 and DWDM2 before being sent to Bob. Eve's apparatus operates automatically throughout the entire QKD process, executing four distinct sub-attacks at different stages, as shown in Fig. 1b. The specific procedure of each sub-attack is as follows.

Sub-attack 1 starts monitoring the information transmitted by the QKD system on the public channel from the authentication stage. Specifically, Eve wiretaps the classical channel to decode data and obtain system status in real time, extracting operational parameters for the subsequent attacks. This sub-attack allows the PPH apparatus to operate without requiring an in-depth understanding of the internal structure of the QKD system. In other words, it does not need comprehensive and detailed knowledge of the internal implementation specifics of the system. Instead, as long as the basic communication protocols and interfaces of the system are understood, the PPH apparatus can be inserted

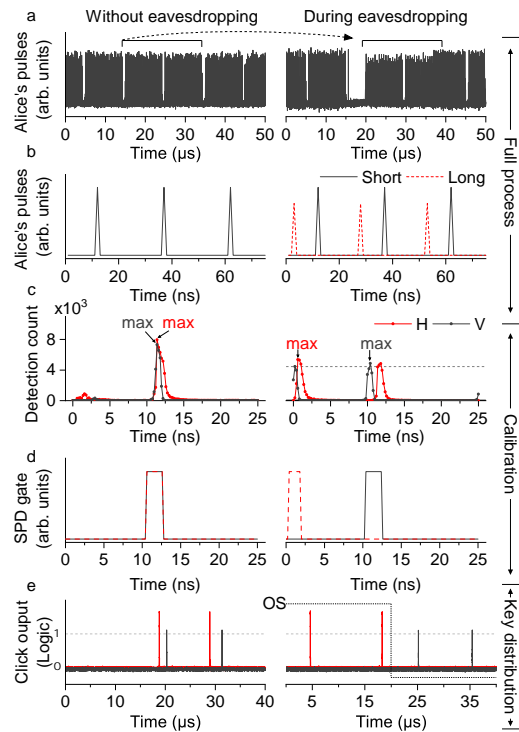


Figure 2. Experimental data from QKD system. (a) The optical pulses emitted from Alice's monitor port are detected by PD3 during the delay scanning phases. (b) Timing sequences of signals detected by SPDs under various transmission paths. (c) Counts detected at different gate positions of SPDs. (d) Gate positions of SPDs. (e) Detection signals during key distribution. The switch-path time of the OS on a period is represented by a dotted line.

into the system and launch an attack. It should be noted that the tested QKD system encrypts classical communication using a private protocol. However, Eve has conducted reverse engineering at the algorithm level, which is purely a classical software work that we do not present in detail here.

Sub-attack 2 manipulates the arrival time of quantum signal starting the delay-scanning phase, causing inconsistent gate positions for paired SPDs in the same basis. In this sub-attack, the timing at which Eve switches paths depends on the synchronization period of the tested system. Specifically, she taps the synchronization light via PD1, which triggers her CE to generate an electrical pulse sequence with a 40- μ s period (This value is correlated with the synchronization rate of the tested QKD system) and 50% duty cycle to control the OS. The OS forces quantum states to pass through the short path during the first half of each 40 μ s period and the long path during the second half. This operation ensures that each Bob's SPD exhibits two highly comparable peaks within a 25 ns period. To visually demonstrate the effect of the PPH apparatus

on Alice's quantum-state pulses, an unattenuated light emitted from the monitor port on Alice's box is connected to the PPH apparatus and detected by PD3. Figure 2a illustrates that without eavesdropping, the quantum-state pulses exhibit a uniform periodic pattern, whereas when eavesdropping occurs, the quantum-state pulses shift. Notably, thanks to synchronization alignment stage, the QKD system still maintains clock synchronization between Alice and Bob even under the significant shift. In the following test, Eve only shifts the quantum-state pulse within a 25-ns period to introduce as small a loss as possible. Figure 2b provides an enlarged view to specifically show that when quantum-state pulses travel exclusively through the long path, only one pulse is detected within the 25-ns period. By contrast, when some pulses travel through the long path while others take the short path, two pulses about 10 ns apart are statistically detected within the same period.

During the delay scanning phase, Bob divides one signal period into 125 detection intervals of 0.2 ns and determines the gate positions based on peak detection counts. Without eavesdropping (Fig. 2c and d on the left), the peak counts of both SPDs for $|H\rangle$ and $|V\rangle$ are centered near the 11th nanosecond. However, during eavesdropping (Fig. 2c and d on the right), the maximum count of the SPD for $|H\rangle$ shifts to the first nanosecond, while that of the SPD for $|V\rangle$ remains at the original position (When the maximum counts of two detection peaks are comparable for each SPD, one of these peaks is then randomly selected. If Eve fails to diverge the gate positions, she disconnects the channel, forcing Bob to restart from calibration). This attack causes a mismatch in the gate positions of the paired SPDs in Z basis. A similar situation occurs in X basis. It is worth noting that the mismatch does not trigger any alerts from the QKD system. Once the gate positions are set during calibration, they remain fixed. This fixed configuration ensures that during the raw key exchange stage, the four SPDs can respond promptly and accurately upon receiving the corresponding quantum states. As shown on the left side of Fig. 2e, both SPDs for the $|H\rangle$ and $|V\rangle$ may click during the whole period of 40 μ s. However, in the presence of eavesdropping, the timing distribution of the responses may be altered, as shown on the right side of Fig. 2e. In this scenario, the $|H\rangle$ state is detected exclusively during the first half of the 40- μ s switching period, while the $|V\rangle$ state is detected solely during the the second half. This is because, dur-

ing the first half of the period, the SPD for $|H\rangle$ is operational due to the alignment between the gate position and the quantum-state arrival time, while the SPD for $|V\rangle$ remains non-responsive due to the misalignment of its gate position. Vice versa for the second half of the period, since the OS switches about every 20 μs . This phenomenon demonstrates that the PPH apparatus is independent of the encoding degree of freedom in the QKD system and can be adapted to other BB84 QKD systems by fine-tuning the attack parameters to match the specific target one.

Sub-attack 3 corrects polarization drift starting from the polarization correction phase. The transmission of quantum states through the long and short paths in sub-attack 2 can be characterized by the linear operators \hat{E}_l (long path) and \hat{E}_s (short path), respectively. For any input polarization state $|\psi\rangle \in \{|H\rangle, |V\rangle, |P\rangle, |N\rangle\}$, the output states after transmission are given by

$$\begin{aligned}\hat{E}_l|\psi\rangle &= \sqrt{1-\epsilon}|\psi\rangle + e^{i\theta}\sqrt{\epsilon}|\psi_{\perp}\rangle, \\ \hat{E}_s|\psi\rangle &= \sqrt{1-\xi}|\psi\rangle + e^{i\theta'}\sqrt{\xi}|\psi_{\perp}\rangle,\end{aligned}\quad (1)$$

where $|\psi_{\perp}\rangle$ denotes the polarization state orthogonal to $|\psi\rangle$, ϵ and ξ is the probability of polarization drift, and θ and θ' is the phase offset associated with the drift component. In ideal polarization correction for a single path, Bob can compensate for such drift using his PCs in each measurement basis (Fig. 1a). However, the path-dependent drifts in sub-attack 2 result in $\hat{E}_l \neq \hat{E}_s$. To mitigate this inconsistency, Eve deploys PCs in both the long and short paths. She iteratively adjusts these PCs using real-time QBER feedback from Bob to align the drift operators such that $\epsilon \approx \xi$ and $\theta \approx \theta'$, keeping QBER $< 3\%$ under the attack. This process begins during polarization correction and continues until system termination. It is worth noting that fitting algorithm enable the PPH apparatus to compensate for the polarization drift difference caused by quantum-state transmission via the long and short paths within minutes, thereby better concealing her existence. Additionally, she can adaptively correct polarization drift caused by environmental changes in real time. As a result, the eavesdropping attack becomes the plug-and-play one. Further details are given in the Online Methods.

Sub-attack 4 infers the sifted key by cross-referencing information from sub-attacks 1 and 2 during key distribution. Eve needs to meet two additional conditions to deduce the sifted key, beyond the mismatch between gate positions of the paired SPDs. First, she needs to synchronize her clock with Bob's in order to determine

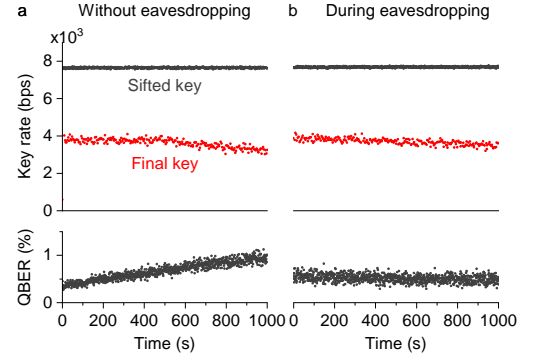


Figure 3. QKD performance with and without eavesdropping as measured by Alice and Bob. (a) Sessions without Eve in the fiber-optic line. (b) Eve eavesdrops on the fiber-optic line during the QKD sessions. The traces in the graphs from top to bottom correspond to the sifted key rate, the final key rate after error correction and privacy amplification, and the QBER.

which path the photon takes corresponding to which SPD of Bob clicks in the announced basis. Second, she must be certain about the bit-value mapping of Bob's SPDs. Fortunately, through sub-attack 1, Eve can obtain all the data transmitted by Alice and Bob over the classical channel, enabling her to meet the above two conditions. Specifically, according to the public information in the first raw key exchange period: the detection slots and his corresponding basis choices announced by Bob combining with the sifted slots and the portion of the sifted key bits revealed by Alice, Eve aligns Bob's detection click slots with the 40- μs switching cycle of the OS. By further analyzing the bit values and the QBER disclosed by Alice and Bob, Eve is able to deduce the bit-value mapping corresponding to Bob's four SPDs. Consequently, Eve exploits these timing and mapping relationships to infer the sifted key in each subsequent key distribution round. In our experiment, Eve successfully deduced 98.97% of the sifted key generated by the QKD system. Detailed calculations and analyses are provided in the Online Methods .

These above-mentioned sub-attacks work in concert, enabling Eve to remotely deploy the PPH apparatus and helping her eavesdrop on the secret key. It should be noted that, since the PPH apparatus wiretaps information from the classical channel and changes the length of the public quantum channel, it neither breaks the security assumptions in the BB84 QKD protocol nor violates the principles of quantum physics. As a result, it remains undetected by the QKD system.

Attack performance

We conducted multiple 10–30 min QKD sessions over a few hours, with and without attack. During these sessions, we recorded performance statistics, all public communication data between Alice and Bob, and the key generation rate. In the key distribution stage, the legitimate parties continuously monitored the key rates and QBER to assess the security of the transmission line. Figure 3 presents results from two representative sessions—one with eavesdropping by Eve and one without. They show that Eve's attack does not substantially affect the key rates. This occurs because the attack introduces no detectable QBER fluctuations that would alert Alice and Bob to Eve's presence. The slight increase in QBER without eavesdropping are caused by normal polarization drift during the operation of this QKD system. However, during eavesdropping sessions, the QBER becomes more stable due to the PPH apparatus's real-time polarization compensation of environmental changes. Consequently, QBER remains in the acceptable range ($< 3\%$) in both cases. Since Eve has almost the same sifted key as Bob, she can theoretically apply post-processing similar to that of Alice and Bob, and generate the same final key.

DISCUSSION AND CONCLUSION

In this work, we demonstrate that during calibration, Eve can induce the mismatch in the gate positions of Bob's SPDs by controlling the different arrival times of quantum states at Bob. This makes our proposed attack be independent of encoding degree of freedom. As a result, the countermeasure of closing the SPD's gate mismatch through randomization of base choices [41] in calibration is not applicable to this attack.

The QKD system needs be equipped with an intrinsic mechanism as countermeasure to eliminate this vulnerability. A free-running SPD may be an alternative, but its effectiveness against this calibration attack remains to be verified in future work. For example, a key consideration could be how to determine the effective counting time window of a free-running SPD. An alternative approach could involve a time-correlated randomness test module in the QKD system [44,45], enhancing the system's ability to detect Eve's presence. Notably, to counter this attack, Bob also can employ a "four-state measurement" scheme [31,39]. In this scheme, he randomly selects both his measurement basis and the mapping of each SPD outcome to the logical bits 0

and 1. During the sifting process, Bob can deduce Alice's encoded information by integrating his own SPD results, his chosen states, and the bases announced by Alice. The mapping of each SPD may be flipped, but it remains private and is never disclosed publicly. As a result, even if Eve knows which SPDs click, she cannot infer the actual bit value. However, this scheme still has a potential loophole. For example, Eve may attempt to inject a strong pulse in order to read out Bob's detector assignments, similar to the strategy employed in a Trojan-horse attack [17, 46,47]. More importantly, we recommend that QKD systems employ the measurement-device-independent QKD protocol [48] and its derivative ones, such as twin-field QKD [49–51], sending-or-not-sending twin-field QKD [52], and side-channel-security QKD [53,54]. In these protocols, the entire measurement unit is treated as an untrusted third party, Charlie. As a result, it theoretically guarantee the security of a QKD system, regardless of Eve's attack on the detection side.

To summarize, we have successfully implemented a penetration testing on a QKD system as a black box through an online plug-and-play attack. This attack allows Eve to manipulate gate positions and synchronization parameters with no access to the internal components of the QKD engine, learning the information of secret key. Notably, this vulnerability and attack exploit the fact that the gate position of each SPD is calibrated separately. This means that this attack should be independent of the encoding degree of freedom, making it potentially applicable to other BB84-based QKD systems. Furthermore, this plug-and-play attack provides a testbed for online penetration testing, which advances methodology and techniques for security certification of QKD.

METHODS

Sub-attack 3

The goal of compensating the polarization drift is to minimize the polarization difference between the long path and the short path, obtaining the similar QBERs for both quantum states in Z/X basis. Thus, the objective of sub-attack 3 is to compensate for the polarization difference between the long and short paths, i.e., achieving $\epsilon \approx \xi$ and $\theta \approx \theta'$, based on reducing the deviation among QBERs for different quantum states Δ_{error} ,

$$\Delta_{error} = |\Delta_H - \Delta_V| + |\Delta_P - \Delta_N|, \quad (2)$$

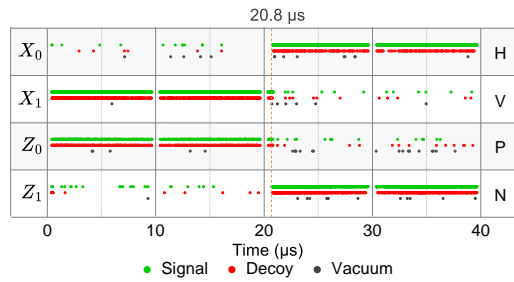


Figure 4. Alice publicly announces a portion of the valid click events and bit values at the first raw key exchange period.

where Δ_H , Δ_V , Δ_P , and Δ_N represent the QBERs of the signal states corresponding to $|H\rangle$, $|V\rangle$, $|P\rangle$, and $|N\rangle$, respectively. This QBER information is obtained from sub-attack 1.

To minimize the QBERs, Eve can control her PC by adjusting the angles of polarization rotation in the X, Y, and Z directions. Here, we fit the data using a curve function and set the values of the X, Y, and Z axes to minimize Δ_{error} . Sub-attack 3 starts from polarization correction and remains active throughout the QKD process, continually refining its estimations. As data accumulates, the attack adjusts the polarization offset corrections more accurately. Consequently, the system's QBER stabilizes at a value lower than 3%.

Sub-attack 4

In Sub-attack 4, Eve chooses the first raw key exchange period to establish synchronization and correlations among the basis, bit values, and SPDs' click. Then, Eve is able to obtain the sifted key shared between Alice and Bob in the following raw key exchange period based on these correlations. Subsequently, Eve uses error correction and privacy amplification to obtain the final key, which is identical to the one shared between Alice and Bob. Next, this section elaborates on how Eve obtain the sifted key.

At the end of each raw key exchange period, Bob communicates to Alice the details of all valid click events, including the basis he chose and the precise timing slot of these clicks. Subsequently, Alice publicly announces the instances in which she and Bob used the same basis at corresponding time sequences. This procedure is known as the sifting process in QKD systems. At the same time, Alice also publicly reveals a small subset of the sifted key bits so that Bob can calculate the QBER information. Following this, Bob provides Alice with detailed QBER information, including the total number of valid events and errors of each

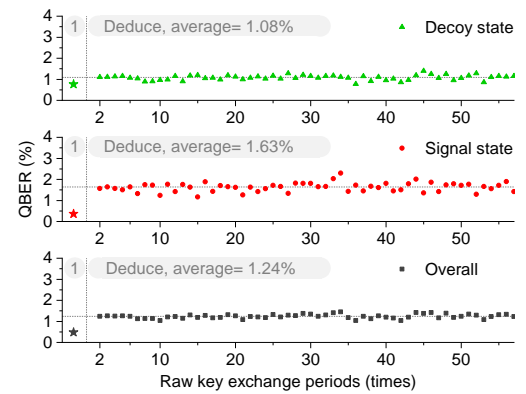


Figure 5. The QBER value calculated for 60 key exchange periods. The first value represents QBER between Alice and Bob, marked with stars. The subsequent values represent QBER between Alice and Eve.

quantum state. In the PPH apparatus, the above transmission data will be wiretapped by the sub-attack 1. Then, sub-attack 4 divides all the valid click events within a raw key exchange period into units of a 40- μs switching period, as shown in Fig. 4. It can be seen that the majority of response events in the initial 20.8- μs interval originate from quantum states corresponding to $|V\rangle$ and $|P\rangle$, whereas responses in the subsequent 19.2- μs interval are attributed to $|H\rangle$ and $|N\rangle$ quantum states. Therefore, Eve can synchronize her clock with Bob's by leveraging the timing distribution of Bob's SPD click events, without resorting to complex physical means such as time-to-digital converter devices.

After Bob has completed the basis comparison with Alice, he will contrast his own data with the data that Alice has publicly disclosed. Then, Bob shares part of the sifted key with Alice to calculate the QBER, as illustrated in Table 1, and this subset of the sifted key is discarded from the following post-processing. Similarly, this QBER information is wiretapped by the sub-attack 1 in the PPH apparatus. Then, sub-attack 4 analyzes the error counts and the total quantities of each quantum state in the QBER information. Despite the imbalance in the counts between different measurement bases, Eve can mitigate it by increasing the OS modulation speed or introducing higher channel attenuation. Eve infers that the click events of Z_0 , Z_1 , X_0 , and X_1 , correspond to Bob's $|H\rangle$, $|V\rangle$, $|P\rangle$, and $|N\rangle$ SPDs respectively. The relationship between the bit and the SPDs in each basis is marked on the right side of Fig. 4. Consequently, when Bob announces the commencement of the following raw key exchange period, Eve employs the information from sub-attack 2 to ascertain whether Bob is responding to the $|H\rangle/|N\rangle$ state of the short path or the

Table 1. The information disclosed during the first round of QBER estimation.

Bob's SPD	Signal state			Decoy state			Vacuum state
	errors	total	QBER	errors	total	QBER	
H	12	1042	1.15%	11	809	1.36%	14
V	6	2405	0.25%	1	1039	0.01%	7
P	5	2083	0.24%	5	308	1.62%	22
N	5	2245	0.22%	6	871	0.69%	12

$|V\rangle/|P\rangle$ state of the long path at each click slot. Using this information, Eve can reconstruct the sifted key shared between Alice and Bob.

Furthermore, to verify the accuracy of Eve's sifted key, we compared Eve's bit values with those disclosed by Alice and Bob to estimate the QBER in the next 60 rounds of the key exchange. Figure 5 shows the QBER of the signal state, decoy state, and overall bit values between Eve and Alice. It can be seen that Eve's average overall error rate is 1.24%, where the minimum is 1.03%. This enables Eve to implement error correction and privacy amplification to eavesdrop on the final key. It should be noted that, considering factors such as the leakage of the optical switch, dark counts p_d , polarization drift, and other factors that cause Bob's SPDs to click, Eve is unable to reconstruct some quantum bit data for events she cannot control. For instance, the error rate $e_0 = p_d$ for Eve, while for Bob it is $e_0 = 0.5 \times p_d$.

FUNDING

This work was funded by the Quantum Science and Technology—National Science and Technology Major Project (2021ZD0300704), the National Natural Science Foundation of China (62371459, 62531025 and 62061136011), and the Postdoctoral Fellowship Program of CPSF (GZC20252817). V.M. was funded by the Galician Regional Government (consolidation of Research Units: AtlantTIC and own funding through the "Planes Complementarios de I+D+I con las Comunidades Autónomas" in Quantum Communication), MICIN with funding from the European Union NextGenerationEU (PRTR-C17.I1), and the "Hub Nacional de Excelencia en Comunicaciones Cuánticas" funded by the Spanish Ministry for Digital Transformation and the Public Service and the European Union NextGenerationEU.

AUTHOR CONTRIBUTIONS

A.H. conceived the research. A.H., Q.P., J.L., and X.Y. design the experiment and completed data validation. Q.P. was responsible for the development and implementation of the automated attack program. C.P., Z.Y., H.W. and Z.C. performed the extraction of the sifted key. A.H. and F.W. supervise the research. All authors participated in data analysis, and jointly

drafted, revised, and finalized the manuscript.

Conflict of interest statement. None declared.

REFERENCES

- Lo HK, Curty M, Tamaki K. Secure quantum key distribution. *Nat Photonics* 2014; **8**: 595–604.
- ID Quantique. *Quantum-Safe Security & Quantum Detection Systems*. <http://www.idquantique.com> (16 March 2026, date last accessed).
- QuantumCTek. *QuantumCTek-Quantum Secures Every Bit*. <http://www.quantum-info.com> (16 March 2026, date last accessed).
- Qasky. *Anhui ASKY quantum Technology CO.,LTD*. <http://www.qasky.com> (16 March 2026, date last accessed).
- QRate. *QRate: Quantum encryption. Security guaranteed by the laws of physics*. <https://goqr.com> (16 March 2026, date last accessed).
- Toshiba. *Global Top Page*. <https://www.global.toshiba/ww/top.html> (16 March 2026, date last accessed).
- LuxQuanta. *Quantum Key Distribution - LuxQuanta*. <https://www.luxquanta.com> (16 March 2026, date last accessed).
- QTI. *QTI Quantum Telecommunications Italy*. <https://www.qticompany.com> (16 March 2026, date last accessed).
- Liu Y, Zhang WJ, Jiang C *et al*. Experimental twin-field quantum key distribution over 1000 km fiber distance. *Phys Rev Lett* 2023; **130**: 210801.
- Gottesman D, Lo HK, Lutkenhaus N *et al*. Security of quantum key distribution with imperfect devices. *Quantum Inf Comput* 2004; **4**: 325–60.
- Nauerth S, Fürst M, Schmitt-Manderbach T *et al*. Information leakage via side channels in freespace BB84 quantum cryptography. *New J Phys* 2009; **11**: 065001.
- Huang A, Navarrete Á, Sun SH *et al*. Laser-seeding attack in quantum key distribution. *Phys Rev Appl* 2019; **12**: 064043.
- Huang A, Li R, Egorov V *et al*. Laser-damage attack against optical attenuators in quantum key distribution. *Phys Rev Appl* 2020; **13**: 034017.
- Ponosova A, Ruzhitskaya D, Chaiwongkhot P *et al*. Protecting fiber-optic quantum key distribution sources against light-injection attacks. *PRX Quantum* 2022; **3**: 040307.

15. Peng Q, Chen JP, Xing T *et al.* Practical security of twin-field quantum key distribution with optical phase-locked loop under wavelength-switching attack. *npj Quantum Inf.* 2025; **11**: 7.
16. Gisin N, Fasel S, Kraus B *et al.* Trojan-horse attacks on quantum-key-distribution systems. *Phys Rev A* 2006; **73**: 022320.
17. Jain N, Anisimova E, Khan I *et al.* Trojan-horse attacks threaten the security of practical quantum cryptography. *New J Phys* 2014; **16**: 123030.
18. Lu FY, Ye P, Wang ZH *et al.* Hacking measurement-device-independent quantum key distribution. *Optica* 2023; **10**: 520–27.
19. Ye P, Chen W, Zhang GW *et al.* Induced-photonrefraction attack against quantum key distribution. *Phys Rev Applied* 2023; **19**: 054052.
20. Kurtsiefer C, Zarda P, Mayer S *et al.* The breakdown flash of silicon avalanche photodiodes—back door for eavesdropper attacks? *J Mod Opt* 2001; **48**: 2039–47.
21. Lydersen L, Wiechers C, Wittmann C *et al.* Hacking commercial quantum cryptography systems by tailored bright illumination. *Nat Photonics* 2010; **4**: 686–89.
22. Bugge AN, Sauge S, Ghazali A *et al.* Laser damage helps the eavesdropper in quantum cryptography. *Phys Rev Lett* 2014; **112**: 070503.
23. Huang A, Sajeed S, Chaiwongkhot P *et al.* Testing random-detector-efficiency countermeasure in a commercial system reveals a breakable unrealistic assumption. *IEEE J Quantum Electron* 2016; **52**: 8000211.
24. Wu Z, Huang A, Chen H *et al.* Hacking single-photon avalanche detectors in quantum key distribution via pulse illumination. *Opt Express* 2020; **28**: 25574–90.
25. Gao B, Wu Z, Shi W *et al.* Ability of strong-pulse illumination to hack self-differencing avalanche photodiode detectors in a high-speed quantum-key-distribution system. *Phys Rev A* 2022; **106**: 033713.
26. Su J, Chen J, Lu F *et al.* Security analysis of orthogonal state attack on a high-speed quantum key distribution system. ArXiv:2506.03718v3.
27. Kang X, Chen JL, Wang ZH *et al.* Hacking high-speed quantum-key-distribution systems by tailoring the blinding pulse. *Phys Rev Applied* 2026; **25**: 024053.
28. Gerhardt I, Liu Q, Lamas-Linares A *et al.* Full-field implementation of a perfect eavesdropper on a quantum cryptography system. *Nat Commun* 2011; **2**: 349.
29. Gisin N. Quantum cryptography: where do we stand? *Qcrypt Tokyo* 2015; .
30. ISO/IEC 23837-2:2023(en). *Information security — Security requirements, test and evaluation methods for quantum key distribution — Part 2: Evaluation and testing methods.* <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:23837:-2:ed-1:v1:en> (16 March 2026, date last accessed).
31. Makarov V, Abrikosov A, Chaiwongkhot P *et al.* Preparing a commercial quantum key distribution system for certification against implementation loopholes. *Phys Rev Appl* 2024; **22**: 044076.
32. Alhamed M and Rahman MMH. A systematic literature review on penetration testing in networks: future research directions. *Appl Sci* 2023; **13**: 6986.
33. Bacudio AG, Yuan X, Chu BTB *et al.* An overview of penetration testing. *Int J Netw Secur Appl* 2011; **3**: 19.
34. Lu FY, Wang ZH, Zhou Y *et al.* Fully heterogeneous prepare-and-measure quantum network for the next stage of quantum internet. *Nat. Commun.* 2025; **16**: 11487.
35. Zheng Y, Wang H, Jia X *et al.* Large-scale quantum communication networks with integrated photonics. *Nature* 2026; **651**: 68–75.
36. Wang XB. Beating the photon-number-splitting attack in practical quantum cryptography. *Phys Rev Lett* 2005; **94**:230503.
37. Lo HK, Ma X, Chen K. Decoy state quantum key distribution. *Phys Rev Lett* 2005; **94**: 230504.
38. Ma X, Qi B, Zhao Y *et al.* Practical decoy state for quantum key distribution. *Phys Rev A* 2005; **72**: 012326.
39. Qi B, Fung CHF, Lo HK *et al.* Time-shift attack in practical quantum cryptosystems. *Quantum Inf Comput* 2007; **7**: 73–82.
40. Zhao Y, Fung CHF, Qi B *et al.* Quantum hacking: Experimental demonstration of time-shift attack against practical quantum-key-distribution systems. *Phys Rev A* 2008; **78**:042333.
41. Jain N, Wittmann C, Lydersen L *et al.* Device calibration impacts security of quantum key distribution. *Phys Rev Lett* 2011; **107**: 110501.
42. Bennett CH and Brassard G. Quantum cryptography: Public key distribution and coin tossing. In: *Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing* 1984; 175-9.
43. Bennett CH, Bessette F, Brassard G *et al.* Experimental quantum cryptography. *J Cryptology* 1992; **5**: 3–28.
44. Woollerton L, Brown P, Colbeck R. Tight analytic bound on the trade-off between device-independent randomness and nonlocality. *Phys Rev Lett* 2022; **129**: 150403.
45. Hangleiter D and Eisert J. Computational advantage of quantum random sampling. *Rev Mod Phys* 2023; **95**: 035001.
46. Vakhitov A, Makarov V, Hjelme DR. Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography. *J Mod Opt* 2001; **48**: 2023–38.
47. Sajeed S, Minshull C, Jain N *et al.* Invisible trojan-horse attack. *Sci Rep* 2017; **7**: 8403.
48. Lo HK, Curty M, Qi B. Measurement-device-independent quantum key distribution. *Phys Rev Lett* 2012; **108**: 130503.
49. Wang XB, Yu ZW, Hu XL. Twin-field quantum key distribution with large misalignment error. *Phys Rev A* 2018; **98**: 062323.
50. Lucamarini M, Yuan ZL, Dynes JF *et al.* Overcoming the rate–distance limit of quantum key distribution without quantum repeaters. *Nature* 2018; **557**: 400–3.
51. Wang S, Yin ZQ, He DY *et al.* Twin-field quantum key distribution over 830-km fibre. *Nat Photonics* 2022; **16**: 154–61.

52. Jiang C, Yu ZW, Hu XL *et al.* Unconditional security of sending or not sending twin-field quantum key distribution with finite pulses. *Phys Rev Appl* 2019; **12**: 024061.
53. Wang XB, Hu XL, Yu ZW. Practical long-distance side-channel-free quantum key distribution. *Phys Rev Appl* 2019; **12**: 054034.
54. Jiang C, Hu XL, Yu ZW *et al.* Side-channel security of practical quantum key distribution. *Phys Rev Res* 2024; **6**: 013266.

ORIGINAL UNEDITED MANUSCRIPT