Plug-and-play attack on a quantum key distribution system as black box

Anqi Huang,[1,*] Qingquan Peng,[1,†] Xialong Yuan,[1,†] Junxuan Liu,[1] Yichen Liu,[1] Zihao Chen,[1] and Vadim Makarov[2,3,4]

[1] *College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China*
[2] *Russian Quantum Center, Skolkovo, Moscow 121205, Russia*
[3] *Vigo Quantum Communication Center, University of Vigo, Vigo E-36310, Spain*
[4] *NTI Center for Quantum Communications, National University of Science and Technology MISiS, Moscow 119049, Russia*

(Dated: April 11, 2025)

Quantum key distribution (QKD) protocol has been proven to be informationally-theoretical security. Unfortunately, due to device imperfections in practice, QKD systems have exposed various vulnerabilities that are exploited by an eavesdropper to conduct quantum hackings, such as laser-seeding attacks, blinding attacks, etc. Most of these attacks currently remain only at the stage of possibility verification or white-box testing. In this paper, we propose and implemented plug-and-play attack on a QKD system as a black box, whose interface and access for the public are the only known information. Through this attack, we actively modified the gate positions and synchronization parameters of the QKD system during the calibration procedure, allowing the attack operate during the whole lifetime of the system running without being noticed. Furthermore, the implemented hacking system only connects to the quantum channel but has no access to the inside of QKD engine, which takes minutes to optimize the hacking parameters to start the eavesdropping. This work illustrates Eve's capability to successfully eavesdrop on keys from QKD systems under current conditions in a more intuitive and concrete way.

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, QKD has matured to implementation that cover distances up to 1002 km [2], commercial QKD systems are available [3–6]. Although the QKD protocol provides informationally-theoretical security based on the principles of quantum mechanics [7], in practical applications, the deviation between hardware equipment and the ideal model remains an urgent problem to be solved. Various attacks exploit imperfections in components of the QKD system, such as photon sources [8], detectors [9], and quantum-state modulators [10], to eavesdrop on secure keys. Most of these attacks are white-box attacks, requiring the adversary to know the internal details of the QKD system and spend hours fine-tuning complex parameters to execute a successful attack [11].

This work proposes and implements a plug-and-play attack on a QKD system as a black box with no access to the inside of the system under test. In this attack, Eve wiretaps classical channels and employs reverse engineering to decode communications between Alice and Bob. By analyzing public documentation and the classical information exchanged between Alice and Bob, we identify the vulnerability during the calibration procedure, in which Eve can actively modified the gate position and synchronization parameters only by hacking through the quantum channel. As a result, the attack creates the mismatch between gate positions of the paired single-photon detectors (SPDs) in a measurement basis, helping Eve to learn the information of secret key. Moreover, we implement the plug-and-play hacking (PPH) system to automatically run the attack and optimize the hacking parameters in minutes, while no intercept-and-resend operation is needed. Here, we mainly focus on the parts of the whole plug-and-play attack related to quantum aspect.

QKD system under test. The QKD system implements the polarization-encoding decoy-state BB84 QKD protocol, in which Alice sends quantum-state pulses at the repetition frequency of 40 MHz, and Bob employs the passive basis selection, as shown in Fig. 1 (a). This system operates in three stages – authentication, calibration, and key distribution. During authentication, Alice and Bob authenticate their devices via the classical channel. Calibration involves three steps as follows. First, the delay scanning, in which Alice adjusts her attenuator based on Bob's click counts, while Bob determines each SPD's gate position respectively by detecting the maximal number of counts. Second, the polarization correction, in which Alice sends single-polarization states, and Bob adjusts its polarization controller until H:V and P:N click ratio reaches 99:1. Third, the synchronization alignment, in which Bob aligns the first frame position with Alice via synchronization signals. In the key distribution stage, Alice sends four polarization states, $|H\rangle$, $|V\rangle$, $|P\rangle$, and $|N\rangle$. Bob randomly choses either the X or Z basis for detection via a scheme of passive basis selection and announces the basis of click event. Alice and Bob repeat this process to accumulate enough raw keys. Then they conduct sifting to obtain the sifted key and quantum bit error rate (QBER). If QBER is less than 3%, the system operates post processing, error correction and privacy amplification. Otherwise, the system runs polar-

* angelhuang.hn@gmail.com
† These authors contributed equally

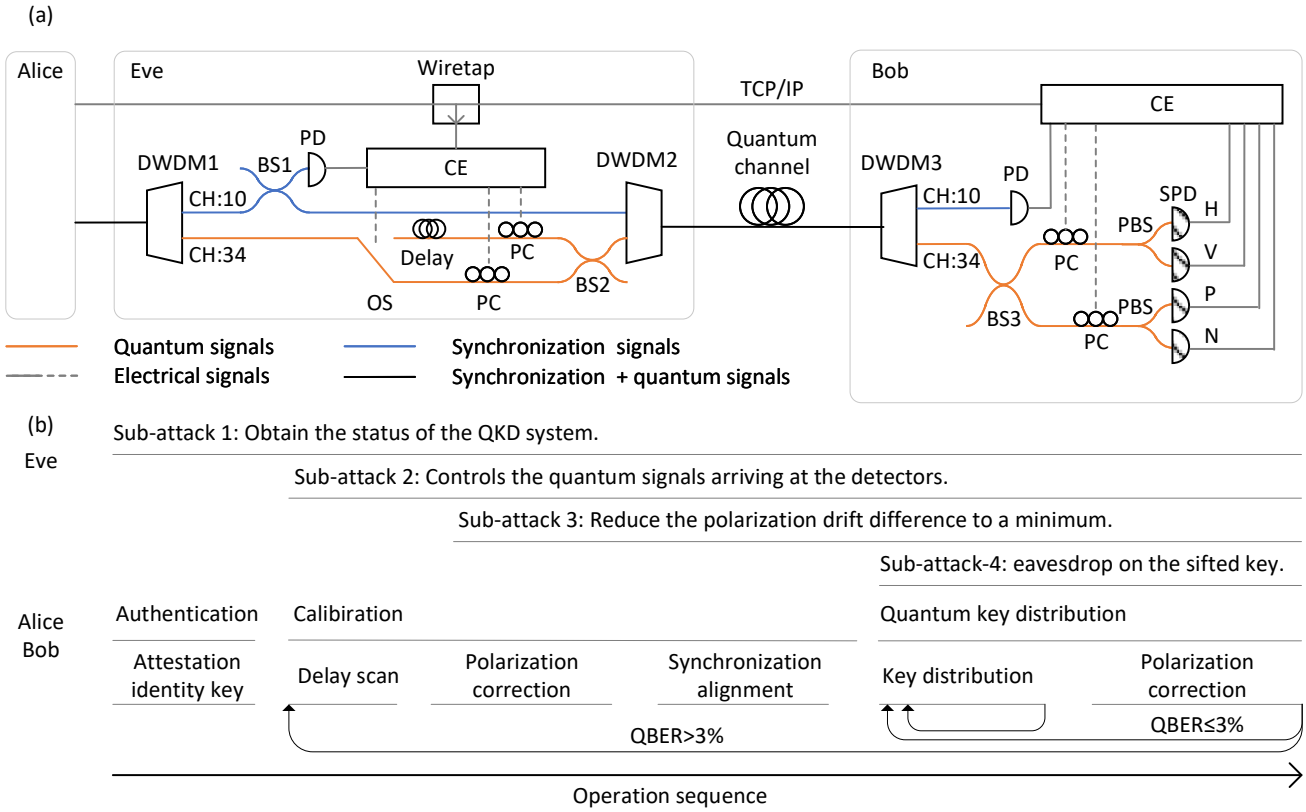

FIG. 1. Plug-and-play hacking experiment. (a) Diagram of the connection between the Eve and the black-box QKD system. Alice and Bob are black boxes, whereas Bob's internal structure is a speculative diagram. DWDM, dense-wavelength-division multiplexer; BS, beam splitter; PBS, polarization beam splitter; OS, optical switch; PD, photoelectric detector; SPD, single-photon detector; CE, control electronics; PC, polarization controller. (b) The timing of the four subattacks in a PPH system.

ization correction, trying to reduce the QBER. If no key is generated within 10 minutes, the QKD system restarts from the beginning of calibration.

Plug-and-play hacking system The PPH system is independent of the encoding degrees of freedom the system under test, and its structure is shown in Fig. 1(a) as Eve who controls the quantum and classical channels. She separates Alice's signal into 1550.12 nm quantum-state signal and 1569.59 nm synchronization signal via dense-wavelength division multiplexer 1 (DWDM1). The synchronization signal is then split by a beam splitter (BS1), whose one part enters DWDM2 and the other part is detected by a photodiode (PD) generating a trigger signal for the control electronics (CE). The CE directs the quantum-state signal through an optical switch (OS) to either a long or short path, controlling its arrival time at Bob. Both paths have polarization controllers (PCs) installed to compensate for polarization drift. The quantum-state and synchronization signals are then combined via BS2 and DWDM2 before being sent to Bob. The PPH system operates automatically throughout the entire QKD process, executing four distinct sub-attacks at different stages as shown in Fig. 1(b). The specific procedure of each sub-attack is presented as follows.

Sub-attack 1 starts from authentication. Eve wiretaps the classical channel to decode data and obtain system status in realtime, providing operational parameters for the subsequent attacks. Since this attack assists to the quantum hacking and is confined to the computer software aspect, it is not analyzed in depth here.

Sub-attack 2 manipulates the arriving time of quantum-state signals during the delay-scanning phase, causing inconsistent gate positions for paired SPDs in the same basis. In this sub-attack, Eve intercepts the synchronization light through DWDM1 and BS1, which triggers the CE to generate an electrical pulse sequence with a 40-μs period and 50% duty cycle, thereby controlling the OS. The OS forces quantum states to pass the short path during the first half of each 40-μs and pass the long path during the other half. Figure. 2 (a) shows that without eavesdropping, the signal light exhibits a uniform periodic pattern, whereas when eavesdropping occurs, the signal light is offset. Figure 2 (b) provides an enlarged view of Fig. 2 (a), specifically showing that when quantum-state pulses travel exclusively through the long path, only one pulse is detected within a 25-ns period. By contrast, when part of the pulses travels through the long path while another part takes the short path, two pulses are detected within the same period. During the delay scanning step, Bob divides one signal period into

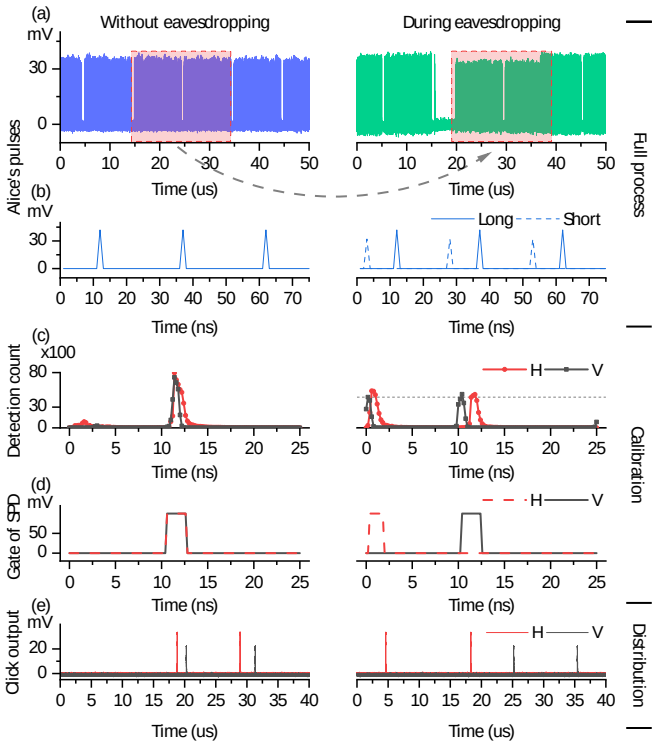

FIG. 2. (a) Experimental data of the delay scan signal. (b) Timing sequences of Alice's pulses under various transmission paths.

(c) Maximum counts detected at different gate positions of SPDs. (d) Gate positions of SPDs. (e) Detection clicks during key distribution.

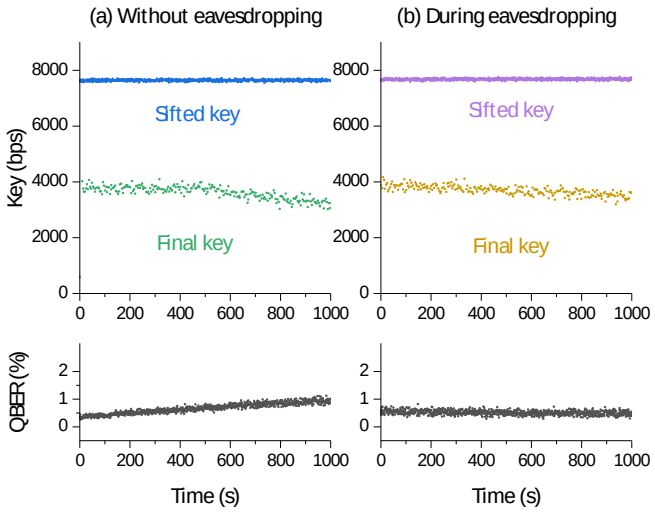

FIG. 3. Comparison of QKD performance with and without eavesdropping. (a) Performance without an eavesdropper. (b) Performance with an eavesdropper. In both scenarios, the traces from top to bottom correspond to the sifted key rate, the final key rate after error correction and privacy amplification, and the QBER.

125 detection intervals and determines the gate positions based on peak detection counts. Without eavesdropping (Fig. 2(c) and (d) on the left), the peak counts for both H and V SPDs are centered near the 11th nanosecond. However, during eavesdropping (Fig. 2(c) and (d) on the right), the maximum count for the H SPD shifts to the first nanosecond, while the V SPD remains at 11 ns. Thus, this attack causes a mismatch in the gate positions of the paired SPDs in the HV basis. Similarly, this occurs in the PN basis as well. Once the gate positions are set during calibration, they remain to be fixed. Figure 2(e) shows that under normal conditions, both SPDs may click during the whole period of 40 μs. However, under eavesdropping, the H state is only detected in the first half of the period, while the V state is detected in the second half due to the mismatch between the fixed gate positions and the signal's arrival times.

Sub-attack 3 corrects polarization drift starting from the polarization correction phase. Due to the principles of quantum physics, Eve cannot directly and accurately measure the transmitted quantum state. Nevertheless, she can iteratively adjust the PCs of her short and long paths based on the QBER feedback provided by Bob, which begins during the polarization correction phase and continues until the system ceases operation. By employing machine learning, Eve can compensate for the polarization drift difference caused by quantum-state transmission via the long and short paths within minutes, thereby better concealing her existence.

Sub-attack 4 infers the sifted key by cross-referencing information from Sub-attacks 1 and 2 during key distribution. In this sub-attack, Eve synchronizes her clock with Bob's, and then uses the control information from the OS to determine whether the quantum state in a click event comes from the long or short path, thereby precisely identifying the quantum state information detected by Bob. Using this information, Eve can reconstruct the sifted key shared between Alice and Bob.

Attack performance. We conducted multiple 10-30 min QKD sessions over a few hours, with and without attack. Results from these two scenarios illustrated in Fig. 3 show that Eve's attack does not significantly affect key rates. This is because, under ideal conditions, the attack causes no QBER fluctuations that would alert Alice and Bob. Consequently, the QBER keeps in the acceptable range ($< 3\%$) in both cases. Since Eve has the same sifted key as Bob, she can theoretically apply similar processing to generate the same final key.

To summarize, we have successfully demonstrated an online plug-and-play attack on a QKD system. 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. Furthermore, the attack is independent of the encoding degree of freedom, making it applicable to other similar BB84-based QKD systems.

[1] H.-K. Lo, M. Curty, and K. Tamaki, Secure quantum key distribution, Nature Photonics **8**, 595 (2014).

[2] Y. Liu, W.-J. Zhang, C. Jiang, J.-P. Chen, C. Zhang, W.-X. Pan, D. Ma, H. Dong, J.-M. Xiong, C.-J. Zhang, H. Li, R.-C. Wang, J. Wu, T.-Y. Chen, L. You, X.-B. Wang, Q. Zhang, and J.-W. Pan, Experimental twin-field quantum key distribution over 1000 km fiber distance, Physical Review Letters **130**, 210801 (2023).

[3] Quantumctek, `https://www.quantum-info.com`.

[4] Toshiba, `https://www.global.toshiba`.

[5] Qreta, `https://goqrate.com`.

[6] Id quantique, `https://www.idquantique.com`.

[7] D. Gottesman, H.-K. Lo, N. Lutkenhaus, and J. Preskill, Security of quantum key distribution with imperfect devices, *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, Proceedings 10.1109/ISIT.2004.1365172 (2004).

[8] S. Nauerth, M. Fürst, T. Schmitt-Manderbach, H. Weier, and H. Weinfurter, Information leakage via side channels in freespace bb84 quantum cryptography, New Journal of Physics **11**, 065001 (2009).

[9] C. Kurtsiefer, P. Zarda, S. Mayer, and H. Weinfurter, The breakdown flash of silicon avalanche photodiodes-back door for eavesdropper attacks?, Journal of Modern Optics **48**, 2039 (2001).

[10] A. Huang, Á. Navarrete, S.-H. Sun, P. Chaiwongkhot, M. Curty, and V. Makarov, Laser-seeding attack in quantum key distribution, Physical Review Applied **12**, 064043 (2019).

[11] N. Gisin, Quantum cryptography: where do we stand?