


## Counterfactual Trojan horse attack

Zheng-Hong Li <sup>1,2,3,\*</sup> Luojia Wang,<sup>4,5</sup> Jingping Xu,<sup>5</sup> Yaping Yang,<sup>5,†</sup> M. Al-Amri <sup>3,6,7,8</sup> and M. Suhail Zubairy<sup>3</sup>

<sup>1</sup>*Department of Physics, Shanghai University, Shanghai 200444, China*

<sup>2</sup>*Shanghai Key Laboratory of High Temperature Superconductors, Shanghai University, Shanghai 200444, China*

<sup>3</sup>*Institute for Quantum Science and Engineering (IQSE) and Department of Physics and Astronomy, Texas A&M University, College Station, Texas 77843-4242, USA*

<sup>4</sup>*State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China*

<sup>5</sup>*MOE Key Laboratory of Advanced Micro-structured Materials, School of Physics Science and Engineering, Tongji University, Shanghai 200092, China*

<sup>6</sup>*CQOQI, KACST, P.O. Box 6086, Riyadh 11442, Saudi Arabia*

<sup>7</sup>*National Center for Laser and Optoelectronics, KACST, Riyadh 11442, Saudi Arabia*

<sup>8</sup>*Department of Physics, KKU, P.O. Box 9004, Abha 61413, Saudi Arabia*



(Received 17 October 2019; accepted 5 February 2020; published 27 February 2020)

Trojan horse attack is a common eavesdropping strategy which can attack various quantum secure communication systems. Its basic idea is to send auxiliary photons into a legitimate communicator's apparatuses and steal information by analyzing the reflected photons. In this paper, we consider a different kind of Trojan horse attack, the so-called counterfactual Trojan horse attack, which has not been studied in detail so far. In such an attack, the eavesdropper may steal the secret information by "ghost" photons, which can spontaneously avoid being detected, even if the detector is an ideal one. We present the details and requirements of such an attack. We also illustrate our results by considering two protocols, the ping-pong protocol and the counterfactual quantum cryptography. Furthermore, we discuss the nature of the counterfactual Trojan horse attack and the strategy to successfully deal with it. Our results indicate that additional resources may be required for the protection against such an attack.

DOI: [10.1103/PhysRevA.101.022336](https://doi.org/10.1103/PhysRevA.101.022336)

### I. INTRODUCTION

The fundamental laws of quantum mechanics provide some novel ways to transmit messages securely. Since the earliest protocols BB84 [1,2] and E-91 [3] were published, quantum key distribution (QKD) has turned into one of the most mature quantum information techniques [4]. QKD provides a secure way for two remote legitimate users Alice and Bob to create a private key through a quantum communication channel and transmit a secret message that is encrypted by the key through a classical channel. There are other protocols that avoid the key-generation process, and messages are exchanged directly and securely through a quantum channel. Such quantum secure direct communication protocols have been extensively studied and show important application prospects [5,6]. Both QKD and quantum secure direct communication protocols rely on the quantum communication channel. In principle, the quantum physics, particularly the no-cloning theorem [7], guarantees the security of the transmitted qubits in the quantum channel. However, the quantum channel itself has left a "back door" for other classes of eavesdropping attacks such as Trojan horse attacks [8].

Trojan horse attacks have received a great deal of attention. In a typical Trojan horse attack, Eve has potential access to the apparatuses belonging to the legitimate users (Alice and Bob) such that she can insert her own photons to gain the secure information by probing the apparatuses [4,8,9]. A basic principle to counter such attacks is to improve the design of Alice and Bob's apparatuses in a way that Eve's eavesdropping photon can be filtered out and prevented from entering their apparatuses. In addition, in order to counter the known Trojan horse attacks, it is also important to expose the eavesdropper. This usually requires auxiliary monitoring detectors to actively search for the eavesdropper's photon [8].

There are two types of Trojan horse attack strategies that are commonly considered in order to avoid being detected. Both of them are based on the detector imperfections [10–14]. One of them is the invisible photon Trojan horse attack [10], which uses "invisible" eavesdropping photons that are insensitive to a single-photon detector. The invisibility condition is satisfied, for example, by carefully choosing the wavelength of the eavesdropping photon such that it is not within the measurement range of the detector. In this case, the photon cannot trigger the detector. The other kind of Trojan horse attack is the delay-photon Trojan horse attack [11,12], whose basic idea is to avoid the active time of the detector due to the fact that a single-photon detector may be blinded by the legitimate communication photon [13]. Using the above

\*refirefox@shu.edu.cn

†yang\_yaping@tongji.edu.cn

detector imperfection, in a delay-photon Trojan horse attack, the eavesdropping photon can be inserted following a legitimate photon with a short time delay so that the detector is unable to respond [12].

Obviously, these types of Trojan horse attacks exploit the imperfections of detector either in the wavelength domain or in the time domain. If the detector is ideal, then it is impossible to successfully perform both types of attacks simply because the eavesdropping photon is not truly invisible or undetectable.

In this paper, we introduce another kind of Trojan horse attack, the counterfactual Trojan horse attack, in which eavesdropping photons are like “ghosts” and cannot be seen even when the detector is an ideal one. As an eavesdropper, Eve is considered to have unlimited resources at her disposal. Then, based on the interaction-free measurement [15,16] and the quantum Zeno effect [17–19], we demonstrate that, when exposed to continuous measurement by Bob, an eavesdropping photon can stay in Eve’s device without leaking into the public transmission channel (between Alice and Bob). Thus, the photon never triggers Bob’s detector even the detector is an ideal one. It seems that eavesdropping photons can sense the presence of the detector in advance and avoid being detected accordingly. In that sense, we call this strategy of attack “counterfactual” [20–23]. With the counterfactual Trojan horse attack, we elaborate on how to distinguish different quantum operations such as the identity operator  $I$ , Pauli- $X$  gate  $\sigma_x$ , Pauli- $Y$  gate  $\sigma_y$  and Pauli- $Z$  gate  $\sigma_z$ , which can be used by Bob to encode information, under the condition that the probability of finding an eavesdropping photon in the public transmission channel by continuous measurement is close to zero.

It is worth noting that the concept of counterfactual quantum attack has already been proposed in Ref. [24] but without detailed description and analysis. In that paper, we show that a counterfactual quantum attack requires multiple interactions between the eavesdropping photon and Bob’s apparatus and suggest a double-chained Mach-Zehnder interferometers to distinguish operations only related to  $I$  and  $\sigma_z$ . However, in this paper, we show that the counterfactual Trojan horse attack can be implemented with an array of single-chained interferometers, which can drastically reduce the number of interactions and thus not only reduce the resources used by eavesdropping devices but also relax the condition of a successful counterfactual Trojan horse attack.

In the second part of the present paper, we discuss the general defense strategies to counter the counterfactual Trojan horse attack. We show that an important outcome of our present work is that the defense strategies against Trojan horse attacks must to be upgraded. However, we should point out that, in spite of successfully countering Eve’s attack, we may not, in all situations, find out whether Eve is there.

The structure of this paper is as follows. In Sec. II, we show two setups of the counterfactual Trojan horse attack. The first setup is for the single-cycle counterfactual Trojan horse attack, which is based on the single-chained version of the interaction-free measurement [16], but with necessary changes so that it can distinguish between different operations. For comparison, we also introduce the double-cycle counterfactual Trojan horse attack mentioned in Ref. [24],

which is based on double-chained interferometers [25]. We demonstrate the most important feature of the counterfactual Trojan horse attack is that it is hard to expose Eve. In Sec. III, we describe in detail how the counterfactual Trojan horse attack recognizes several different quantum operations ( $I$ ,  $\sigma_x$ ,  $\sigma_y$ ,  $\sigma_z$  and so on). In Secs. II and III, we also discuss the conditions for a successful counterfactual Trojan horse attack. In Sec. IV, we give two eavesdropping examples. One is the ping-pong protocol [5,26] and the other one is the counterfactual quantum cryptography protocol [27–29]. We show that, without a careful design of the communication apparatuses, the counterfactual Trojan horse attack could be an unnoticed threat to the security of quantum communication. In Sec. V, for the purpose of exposing the counterfactual Trojan horse attack, we introduce a possible scheme which may increase the chance of Bob finding Eve’s photon by using multiple random measurements. In Sec. VI, we present the concluding remarks.

## II. COUNTERFACTUAL TROJAN HORSE ATTACK SETUPS

In a conventional Trojan horse attack, Eve accesses the public transmission channel between Alice and Bob and sends her eavesdropping photons into one communicator’s (Bob) apparatus as shown in Fig. 1. Its main purpose is to determine the state of the apparatus by analyzing the reflected photons. If the information exchanged between Alice and Bob depends on Bob’s manipulations, and these manipulations correspond to distinguishable apparatus states, Eve can successfully steal the information.

There are three key points for a successful eavesdropping. The first is whether Eve can identify Bob’s manipulation, which could be a unitary operation or a measurement. The second is whether the eavesdropping photon can return to Eve. In principle, any component that can reflect photons in a communication apparatus can be exploited by Eve to carry out the attack, but the specific analysis depends on the actual system design. It is worth noting that the second condition is usually not difficult to satisfy in a round-trip communication protocol. In such a protocol, Bob returns Alice’s photons after manipulating them (to encode his information), which provides a window for Eve’s photons to escape also from Bob’s apparatus. Therefore, the round-trip-type communication is vulnerable to Trojan horse attack [10–12]. Here, for convenience, we only consider a typical round-trip-type communication as shown in Fig. 1, in which Bob uses a quantum device (QD, its details depend on a specific communication protocol) to manipulate Alice’s photons. In addition, there is a normal mirror (MR) at Bob’s end so that he can return incoming photons. The last key point of Trojan horse attack is how to avoid being discovered. Since extra photons are used, an intuitive idea for defense is to use auxiliary detection to find Eve’s photons. While most existing Trojan horse attacks exploit the defects of the actual detector to avoid being discovered, here we show that this may not be necessary.

In the following, we discuss the counterfactual Trojan horse attack. Here, Eve’s eavesdropping device is specially designed such that when Bob makes active measurements the probability of eavesdropping photons appearing in the public transmission channel approaches zero.

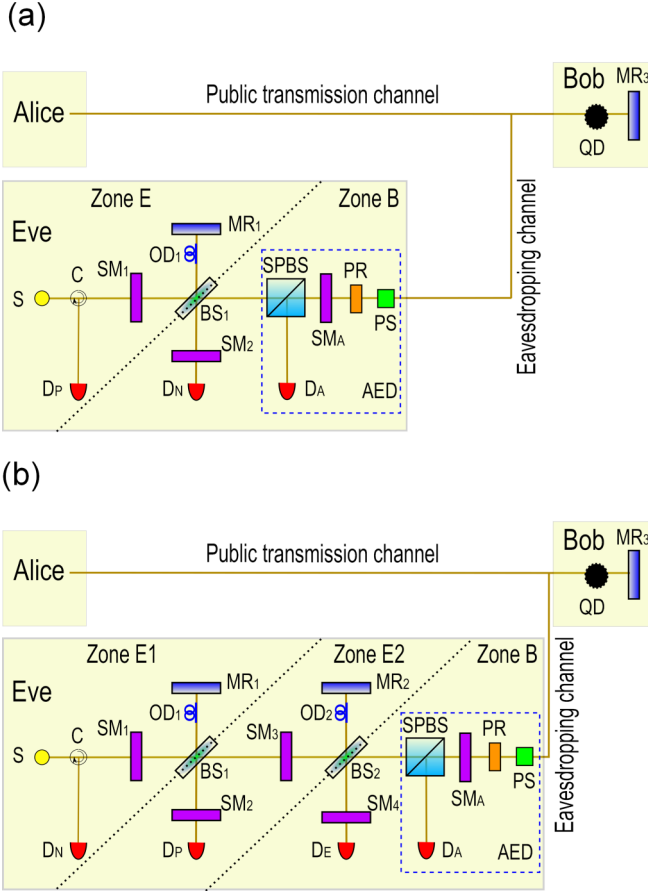


FIG. 1. A schematic diagram of Eve’s attack: (a) the setup of the single-cycle counterfactual Trojan horse attack and (b) the double-cycle counterfactual Trojan horse attack. In the figure, S stands for photon source, D for detector, C for optical circulator, BS for beam splitter, SPBS for switchable polarizing beam splitter, OD for optical delay, QD for quantum device, MR for normal mirror, PR for polarization rotator, PS for phase shifter, AED for auxiliary eavesdropping device, and SM for switchable mirror, which is transparent when it is turned off.

We present two possible setups at Eve’s end that can achieve the above objective. In Fig. 1(a), the setup is for the single-cycle counterfactual Trojan horse attack, while in Fig. 1(b), the setup is for the double-cycle counterfactual Trojan horse attack. The latter case is the one that was mentioned in Ref. [24]. In general, Eve’s single-photon source (S) prepares an eavesdropping photon either horizontally (H) or vertically (V) polarized. This photon may enter the eavesdropping channel, be manipulated, and then go back to trigger different detectors (D) at Eve’s end according to different Bob’s manipulations. In addition to the above-mentioned optical elements, BS stands for beam-splitter, OD stands for optical delay, and C stands for optical circulator. In particular, the blue dashed line box has an auxiliary eavesdropping device (AED), which consists of a switchable polarizing beam splitter (SPBS), a polarization rotator (PR), a phase shifter (PS), a switchable mirror (SM), and a detector. The AED helps Eve to distinguish Bob’s multiple manipulations (see Sec. III).

For the sake of convenience, in this section, we introduce only the most basic functions of the two setups. Thus, the

AED is assumed to be inactive (transparent). The basic functions of Eve’s eavesdropping device are as follows:

Working mode 1: When the eavesdropping photon sent to Bob is back to Eve without any changes (no phase change), the detector  $D_N$  clicks with a 100% probability.

Working mode 2: When the eavesdropping photon sent to Bob is back, but after undergoing a  $\pi$  phase shift, the detector  $D_P$  clicks with a 100% probability.

Working mode 3: When Bob uses detectors to search for the eavesdropping photon (so the eavesdropping channel is completely blocked), those detectors register the photon with a probability close to zero. If the photon is not found by Bob’s detectors,  $D_P$  clicks. We emphasize that this feature, i.e., avoiding Eve’s photons being found, is the key to the counterfactual Trojan horse attack.

### A. Single-cycle counterfactual Trojan horse attack

For the sake of clear description, we divide the structure of Fig. 1(a) into two zones by black dotted lines. The photon state  $|P_{E(B)}\rangle$  indicates Eve’s photon with a polarization state  $P = H, V$  appearing in zone E(B), while  $|0_{E(B)}\rangle$  indicates that there is no photon appearing in zone E(B). The process of a photon passing BS<sub>1</sub> can be described as

$$\begin{aligned}
 |P_E\rangle|0_B\rangle &\rightarrow \cos\frac{\pi}{4M}|P_E\rangle|0_B\rangle + \sin\frac{\pi}{4M}|0_E\rangle|P_B\rangle, \\
 |0_E\rangle|P_B\rangle &\rightarrow \cos\frac{\pi}{4M}|0_E\rangle|P_B\rangle - \sin\frac{\pi}{4M}|P_E\rangle|0_B\rangle, \quad (1)
 \end{aligned}$$

where  $M$  is a nonzero integer and  $|\cos\frac{\pi}{4M}|^2$  is the reflectivity of BS<sub>1</sub>. When SM<sub>1</sub> and SM<sub>2</sub> are turned on to work as ordinary mirrors, the photon can pass BS<sub>1</sub> multiple times. To ensure the interference, the optical lengths of SM<sub>1</sub> ↔ BS<sub>1</sub> and SM<sub>2</sub> ↔ BS<sub>1</sub> are designed to be the same. Likewise, it is assumed to be true for the optical lengths BS<sub>1</sub> ↔ MR<sub>1</sub> and BS<sub>1</sub> ↔ MR<sub>3</sub>. The delay line OD<sub>1</sub> is used to compensate for the optical path difference due to the public transmission channel contained in the path BS<sub>1</sub> ↔ MR<sub>3</sub>. We treat the process that the photon starts from SM<sub>1</sub> or SM<sub>2</sub>, passes BS<sub>1</sub>, gets reflected from MR<sub>1</sub> or MR<sub>3</sub>, and returns to SM<sub>1</sub> or SM<sub>2</sub> as a cycle.

In the single-cycle counterfactual Trojan horse attack, initially a  $P$  polarized photon is sent into the interferometer from Eve’s source. Once the photon passes the initially transparent SM<sub>1</sub>, the mirror is turned on, and so is SM<sub>2</sub>. After  $m$  cycles, i.e., the photon passing BS<sub>1</sub> for  $2m$  times, SM<sub>1</sub> and SM<sub>2</sub> are turned off (transparent) to output the photon. We analyze the evolution of Eve’s photon according to Bob’s manipulation as follows.

For working mode 1, the output photon state is

$$|P_E\rangle|0_B\rangle \xrightarrow{m} \cos\frac{m\pi}{2M}|P_E\rangle|0_B\rangle + \sin\frac{m\pi}{2M}|0_E\rangle|P_B\rangle. \quad (2)$$

If  $m = M$ , the detector  $D_N$  clicks with a unit probability. On the other hand, if  $m = 2M$ , the output state is  $-|P_E\rangle|0_B\rangle$  and  $D_P$  clicks with a unit probability.

For working mode 2, in each cycle, the photon state evolves into  $\cos\frac{\pi}{4M}|P_E\rangle|0_B\rangle + \sin\frac{\pi}{4M}|0_E\rangle|P_B\rangle$  for the first time Eve’s photon passes BS<sub>1</sub>, and turns into  $\cos\frac{\pi}{4M}|P_E\rangle|0_B\rangle - \sin\frac{\pi}{4M}|0_E\rangle|P_B\rangle$  due to Bob’s phase operation. It becomes  $|P_E\rangle|0_B\rangle$  after passing BS<sub>1</sub> for the second time. Finally,  $D_P$  clicks with a unit probability.

For working mode 3, the first cycle with Bob's measurement turns Eve's photon state into  $\cos \frac{\pi}{4M} (\cos \frac{\pi}{4M} |P_E\rangle|0_B\rangle + \sin \frac{\pi}{4M} |0_E\rangle|P_B\rangle)$ , if Bob does not get the photon. Under the same condition and after  $m$  cycles, the photon state is

$$\cos^{m-1} \frac{\pi}{2M} \cos \frac{\pi}{4M} \times \left( \cos \frac{\pi}{4M} |P_E\rangle|0_B\rangle + \sin \frac{\pi}{4M} |0_E\rangle|P_B\rangle \right). \quad (3)$$

Because of the quantum Zeno effect, the probability that Eve's photon stays in her device is approximately  $\cos^{2m} \frac{\pi}{2M}$ . As a consequence, the total probability that Bob detects Eve's photon tends to be

$$P_s = 1 - \cos^{2M} \frac{\pi}{2M} \approx \frac{\pi^2}{4M}, \quad (4)$$

when  $m = M \gg \pi^2$ . In this case, when  $M$  tends to infinite, the detector  $D_P$  clicks asymptotically with a unit probability, while Bob's chance of finding out Eve's photon is close to zero.

### B. Double-cycle counterfactual Trojan horse attack

For comparison, here we also elaborate on the counterfactual attack setup that was proposed in Ref. [24]. As shown in Fig. 1(b), there are two Michelson interferometers. The inner Michelson interferometer, composed of  $SM_{3,4}$ ,  $BS_2$ ,  $MR_{2,3}$ ,  $OD_2$ , and Bob's QD, is structurally nested in one arm of the outer Michelson interferometer, which also contains  $SM_{1,2}$ ,  $BS_1$ ,  $MR_{1,3}$ , and  $OD_1$ . Here,  $OD_{1(2)}$  is used to compensate for the optical path difference between  $BS_{1(2)} \leftrightarrow MR_{1(2)}$  and  $BS_{1(2)} \leftrightarrow MR_3$ . The whole structure could be divided into three zones based on  $BS_1$  and  $BS_2$ .  $|P_{E1}\rangle|0_B\rangle$ ,  $|P_{E2}\rangle|0_B\rangle$ , and  $|0_E\rangle|P_B\rangle$  indicate that  $P$ -polarized photon appears in zones E1, E2, and B, respectively. The function of  $BS_1$  can be described as

$$\begin{aligned} |P_{E1}\rangle|0_B\rangle &\rightarrow \cos \frac{\pi}{4M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{\pi}{4M'} |P_{E2}\rangle|0_B\rangle, \\ |P_{E2}\rangle|0_B\rangle &\rightarrow \cos \frac{\pi}{4M'} |P_{E2}\rangle|0_B\rangle - \sin \frac{\pi}{4M'} |P_{E1}\rangle|0_B\rangle, \end{aligned} \quad (5)$$

while the function of  $BS_2$  is

$$\begin{aligned} |P_{E2}\rangle|0_B\rangle &\rightarrow \cos \frac{\pi}{2N'} |P_{E2}\rangle|0_B\rangle + \sin \frac{\pi}{2N'} |0_E\rangle|P_B\rangle, \\ |0_E\rangle|P_B\rangle &\rightarrow \cos \frac{\pi}{2N'} |0_E\rangle|P_B\rangle - \sin \frac{\pi}{2N'} |P_{E2}\rangle|0_B\rangle, \end{aligned} \quad (6)$$

where  $|\cos \frac{\pi}{4M'}|^2$  and  $|\cos \frac{\pi}{2N'}|^2$  ( $M', N'$  are integers) are the reflectivity of  $BS_1$  and  $BS_2$ , respectively.

Initially Eve inputs a  $P$  polarized photon. After the photon enters the outer interferometer,  $SM_1$  and  $SM_2$  are turned on until the photon completes  $M'$  outer cycles [25,30]. In each outer cycle, the photon passes  $BS_1$  twice. Between the two passes, the photon enters the inner interferometer and leaves after  $N'$  inner cycles (the photon passing  $BS_2$  for  $2N'$  times) by Eve controlling  $SM_{3,4}$ .

According to the discussion for the single-cycle counterfactual Trojan horse attack, it is not difficult to find out that after the photon passes  $BS_2$  for the first time in the  $m$ th outer cycle and the  $n$ th inner cycle before Bob's operation, for

working mode 1, we have the photon state [30]

$$\begin{aligned} |P_{E1}\rangle|0_B\rangle &\rightarrow \cos \frac{\pi}{4M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{\pi}{4M'} \\ &\times \left[ \cos \frac{(2n-1)\pi}{2N'} |P_{E2}\rangle|0_B\rangle + \sin \frac{(2n-1)\pi}{2N'} |0_E\rangle|P_B\rangle \right]. \end{aligned} \quad (7)$$

We notice that the photon component inside the inner interferometer are either in the path from  $BS_2$  to  $MR_3$  or from  $BS_2$  to  $MR_2$ . It needs to pass through  $BS_2$  once more to complete the  $n$ th inner cycle. The corresponding photon state is  $\cos \frac{\pi}{4M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{\pi}{4M'} [\cos \frac{n\pi}{N'} |P_{E2}\rangle|0_B\rangle + \sin \frac{n\pi}{N'} |0_E\rangle|P_B\rangle]$ . In the case  $n = N'$ ,  $SM_3$  and  $SM_4$  are turned off so that the photon leaves the inner interferometer. Then, the photon passes through  $BS_1$ , and the  $m$ th outer cycle is completed, leading the photon state to become  $|P_{E1}\rangle|0_B\rangle$ . After the photon completes  $M'$  outer cycles, the photon state is still  $|P_{E1}\rangle|0_B\rangle$ , which leads  $D_N$  to click.

For working mode 2, in the  $m$ th outer cycles and the  $n$ th inner cycle, and before Bob's phase operation, the photon state is

$$\begin{aligned} |P_{E1}\rangle|0_B\rangle &\rightarrow \cos \frac{(2m-1)\pi}{4M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{(2m-1)\pi}{4M'} \\ &\times \left( \cos \frac{\pi}{2N'} |P_{E2}\rangle|0_B\rangle + \sin \frac{\pi}{2N'} |0_E\rangle|P_B\rangle \right). \end{aligned} \quad (8)$$

The photon state becomes  $\cos \frac{m\pi}{2M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{m\pi}{2M'} |P_{E2}\rangle|0_B\rangle$  after  $m$  outer cycles, and  $|P_{E2}\rangle|0_B\rangle$  after  $M'$  outer cycles. Then,  $D_P$  clicks.

For working mode 3, the photon state in the  $m$ th outer cycle and the  $n$ th inner cycle (before Bob's measurement) is

$$\begin{aligned} |P_{E1}\rangle|0_B\rangle &\rightarrow \cos \frac{(2m-1)\pi}{4M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{(2m-1)\pi}{4M'} \cos \frac{\pi}{2N'} \\ &\times \cos^{n-2} \frac{\pi}{N'} \left( \cos \frac{\pi}{N'} |P_{E2}\rangle|0_B\rangle + \sin \frac{\pi}{N'} |0_E\rangle|P_B\rangle \right), \end{aligned} \quad (9)$$

where we assume that  $N'$  is sufficiently large so that the interference in the outer cycles can be considered uninterrupted [20,30]. When the  $m$ th cycle is completed, the photon state is approximately  $\cos \frac{m\pi}{2M'} |P_{E1}\rangle|0_B\rangle + \sin \frac{m\pi}{2M'} |P_{E2}\rangle|0_B\rangle$  and tends to  $|P_{E2}\rangle|0_B\rangle$  when  $m = M'$ , which causes  $D_P$  to click. The total probability of Bob finding Eve's photon is approximately [30]

$$P_d = \sum_{m=1}^{M'} \sin^2 \left( \frac{m\pi}{2M'} \right) \left[ 1 - \cos^{2N'} \left( \frac{\pi}{N'} \right) \right] \approx \frac{M'\pi^2}{2N'}. \quad (10)$$

In summary, the output photon state can have one of these three forms: (i)  $|P_{E1}\rangle|0_B\rangle$ , which is registered by  $D_N$  for the working mode 1; (ii)  $|P_{E2}\rangle|0_B\rangle$  registered by  $D_P$  for the working mode 2; and (iii) approximately  $|P_{E2}\rangle|0_B\rangle$  for the working mode 3. In the last case, the chance of Bob finding Eve's photon is nearly zero if  $N' \gg \pi^2 M'$ .

### C. Implementation conditions and comparison of two counterfactual Trojan horse attacks

So far, we considered how Eve distinguishes Bob's 0-phase operation (the working mode 1) and  $\pi$ -phase



operation (the working mode 2) under the condition that the eavesdropping photon can hardly be detected (the working mode 3). Apparently, both the single-cycle and double-cycle counterfactual Trojan horse attacks satisfy our requirements. The condition for a successful attack is that Eve's eavesdropping photons must be manipulated multiple times using Bob's consistent operation (see Supplementary Material III of Ref. [24]). Therefore, Bob's apparatus needs to be accessible for a certain period of time, and the state of Bob's apparatus should remain unchanged during this access time window. Suppose the time Eve's eavesdropping photon spent inside Bob's apparatus is  $\tau$ ; then the minimum time required for Eve to perform a successful counterfactual attack is  $M\tau$  for the single-cycle counterfactual Trojan horse attack, and  $M'N'\tau$  for the double-cycle counterfactual Trojan horse attack. Moreover, according to Eqs. (4) and (10), if we want the eavesdropping photon in the double-cycle counterfactual Trojan horse attack to have smaller chance of being found compared to that with the single-cycle counterfactual Trojan horse attack, i.e.,  $P_s \geq P_d$ , we must have  $M \leq N'/(2M') < M'N'$ , which means that the access time window required by the double-cycle counterfactual Trojan horse attack is longer than the single-cycle counterfactual Trojan horse attack. It is worthwhile mentioning that the optical structure of the single-cycle counterfactual Trojan horse attack is also simpler than that of the double-cycle counterfactual Trojan horse attack. Therefore, the single-cycle counterfactual Trojan horse attack is less time-consuming and less demanding for experimental implementation, which is critical for the success of the counterfactual attack in practice.

Nevertheless, the above conclusion does not mean that the double-cycle counterfactual Trojan horse attack has no advantage. For example, in the case of working mode 1, the probability of the photon appearing in the public transmission channel approaches one at the later stage of the single-cycle counterfactual Trojan horse attack but remains tiny in each cycle of the double-cycle counterfactual Trojan horse attack. In fact, in all three working modes, as long as  $M'$  and  $N'$  tend to infinity, Bob's probability of finding Eve photons in any cycle of the double-cycle counterfactual Trojan horse attack tends to 0 [25,30].

### III. DISCRIMINATION OF BOB'S MULTIPLE MANIPULATIONS

In quantum information and quantum communication, quantum operations that are used to encode information can take different forms other than the ways described in the previous section. Here, we discuss how Eve distinguishes other commonly used quantum operations such as Pauli operators  $\sigma_x$ ,  $\sigma_y$ ,  $\sigma_z$  with the help of the AED under the condition that Bob finds nothing if he tries to detect Eve's eavesdropping photon. More specifically, assuming that Bob might perform several kinds of operations (all known to Eve) on photons, Eve needs to identify which one Bob used.

In order to facilitate the discussion, we characterize Pauli operators based on photon polarization, which are  $\sigma_x = |H_B\rangle\langle V_B| + |V_B\rangle\langle H_B|$ ,  $\sigma_y = -i|H_B\rangle\langle V_B| + i|V_B\rangle\langle H_B|$ , and  $\sigma_z = |H_B\rangle\langle H_B| - |V_B\rangle\langle V_B|$ . Similarly, we define the operator  $\sigma_I = |H_B\rangle\langle H_B| + |V_B\rangle\langle V_B|$ , which means that

Bob does not change the state of the photon passing through his  $QD$  as shown in Fig. 1. It is worth pointing out that the above operators just represent Bob's local operations since they have no contribution if a photon does not pass through Bob's  $QD$  (with state  $|0_B\rangle$ ). Regarding Eve's eavesdropping photon, Bob's operation can be generally described as  $U_B = |0\rangle\langle 0| + U_{PS}(\alpha_H, \alpha_V)U_{PR}(\beta)$ , where  $U_{PS}(\alpha_H, \alpha_V) = e^{i\alpha_H}|H\rangle\langle H| + e^{i\alpha_V}|V\rangle\langle V|$  represents that Bob gives a  $\alpha_H$  phase shift to the H photon and a  $\alpha_V$  phase shift to the V photon, while  $U_{PR}(\beta) = \cos\beta|H\rangle\langle H| + \cos\beta|V\rangle\langle V| + \sin\beta|V\rangle\langle H| - \sin\beta|H\rangle\langle V|$  represents that Bob rotates the photon polarization with angle  $\beta$ , i.e.,  $|H\rangle \rightarrow \cos\beta|H\rangle + \sin\beta|V\rangle$  and  $|V\rangle \rightarrow \cos\beta|V\rangle - \sin\beta|H\rangle$ . Here, we have omitted the subscript  $B$  for convenience. It is not difficult to see that when Bob's local operation is  $\sigma_I$  (i.e., working mode 1),  $U_{BI} = |0\rangle\langle 0| + U_{PS}(0, 0)U_{PR}(0) = I$ , where  $I$  is the identity operator. While for  $\sigma_x$ ,  $\sigma_y$ ,  $\sigma_z$ , and  $\sigma_z\sigma_x$ , Bob's operations can be represented as  $U_{Bx} = |0\rangle\langle 0| + U_{PS}(\pi, 0)U_{PR}(\pi/2)$ ,  $U_{By} = |0\rangle\langle 0| + U_{PS}(\pi/2, \pi/2)U_{PR}(\pi/2)$ ,  $U_{Bz} = |0\rangle\langle 0| + U_{PS}(0, \pi)U_{PR}(0)$ , and  $U_{Bzx} = |0\rangle\langle 0| + U_{PS}(\pi, \pi)U_{PR}(\pi/2)$ , respectively. In addition, one can easily obtain the following relations:

$$\begin{aligned} U_{PS}(\alpha_H, \alpha_V)U_{PS}(-\alpha'_H, -\alpha'_V) \\ = U_{PS}[(\alpha_H - \alpha'_H), (\alpha_V - \alpha'_V)], \end{aligned} \quad (11)$$

$$U_{PR}(\beta)U_{PR}(\beta') = U_{PR}(\beta + \beta'), \quad (12)$$

$$[U_{PR}(\beta), U_{PS}(\alpha_H, \alpha_H)] = [U_{PR}(\beta), U_{PS}(\alpha_V, \alpha_V)] = 0. \quad (13)$$

From Eqs. (11) and (12), we obtain  $[U_{PS}(\alpha_H, \alpha_V)]^L = U_{PS}(L\alpha_H, L\alpha_V)$  and  $[U_{PR}(\beta)]^L = U_{PR}(L\beta)$ .

#### A. Discrimination between two different Bob's operations

Suppose that Bob's two operations are  $U_B = U_{PS}(\alpha_H, \alpha_V)U_{PR}(\beta)$  and  $U_{B'} = U_{PS}(\alpha'_H, \alpha'_V)U_{PR}(\beta')$ , where  $\beta, \beta' \in \{0, \frac{\pi}{2}\}$  and  $\alpha_{H(V)}, \alpha'_{H(V)} \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ . All the parameters  $(\alpha, \beta, \alpha', \beta')$  are known to Eve, and she only needs to identify Bob's operation, but with the condition that Bob cannot find her photon. The basic idea is to utilize the three working modes of the previous section with the help of Eve's auxiliary operations. In general, in each cycle, Eve's photon needs to go through the AED before it enters the eavesdropping channel. At first, Eve turns the switchable polarization beam splitter (SPBS) and  $SM_A$  off (transparent). When Eve's photon (with a specific polarization state  $|P\rangle$  prepared by the photon source) passes through them,  $SM_A$  is turned on. Then, the polarization rotator (PR) rotates the photon polarization with  $U_{ER} = U_{PR}(-\beta')$ . After that, Eve's photon enters the eavesdropping channel, be operated by Bob ( $U_B$  or  $U_{B'}$ ) and sent back to the AED, where the phase shifter, PS, shifts the photon phase with the operator  $U_{EL} = U_{PS}(-\alpha'_H, -\alpha_V + \alpha_H - \alpha'_H)$ . Unless  $SM_A$  becomes transparent, Eve's photon will be reflected and travel back and forth between  $SM_A$  and Bob for  $L$  times. As a result, the total

operation performed by the AED and Bob can be represented as

$$O_B = [U_{EL}U_B U_{ER}]^L = |0\rangle\langle 0| + e^{iL(\alpha_H - \alpha'_H)} \\ \times [\cos L(\beta - \beta')|H\rangle\langle H| + \cos L(\beta - \beta')|V\rangle\langle V| \\ - \sin L(\beta - \beta')|H\rangle\langle V| + \sin L(\beta - \beta')|V\rangle\langle H|], \quad (14)$$

$$O'_B = [U_{EL}U_B U_{ER}]^L \\ = |0\rangle\langle 0| + |H\rangle\langle H| + e^{iL(\alpha_H - \alpha'_H + \alpha'_V - \alpha_V)}|V\rangle\langle V|. \quad (15)$$

When  $SM_A$  is turned off (making it transparent), the eavesdropping photon passes through it and reaches SPBS. Eve can either make SPBS transparent or reflect only V photons to detector  $D_A$ . Next we show how she identifies  $U_B$  and  $U_{B'}$  by setting  $|P\rangle$ ,  $L$ , and SPBS.

(i) When  $\beta \neq \beta'$ , Eve selects  $|P\rangle = |H\rangle$  and  $L = \lfloor \frac{(2k+1)\pi}{2(\beta - \beta')} \rfloor$ , ( $k$  is an integer), i.e.,  $O_B|H\rangle = \pm e^{iL(\alpha_H - \alpha'_H)}|V\rangle$  and  $O_{B'}|H\rangle = |H\rangle$ . In addition, SPBS is set to reflect V photons. Then,  $O_B$  causes Eve's device to be in working mode 3, while  $O_{B'}$  causes the device to be in working mode 1.

(ii) When  $\beta = \beta'$ ,  $\alpha'_H \neq \alpha_H$ , Eve selects  $|P\rangle = |H\rangle$  and  $L = \lfloor \frac{(2k+1)\pi}{\alpha_H - \alpha'_H} \rfloor$ , ( $k$  is an integer), i.e.,  $O_B|H\rangle = -|H\rangle$  and  $O_{B'}|H\rangle = |H\rangle$ . SPBS is set to be transparent. Then,  $O_B$  causes Eve's device to be in working mode 2, while  $O_{B'}$  causes Eve's device to be in working mode 1.

(iii) When  $\beta = \beta'$ ,  $\alpha'_H = \alpha_H$ ,  $\alpha'_V \neq \alpha_V$ , Eve selects  $|P\rangle = |V\rangle$  and  $L = \lfloor \frac{(2k+1)\pi}{\alpha'_V - \alpha_V} \rfloor$  ( $k$  is an integer), i.e.,  $O_B|V\rangle = |V\rangle$  and  $O_{B'}|V\rangle = -|V\rangle$ . SPBS is set to be transparent. Then,  $O_B$  causes Eve's device to be in working mode 1, while  $O_{B'}$  causes Eve's device to be in working mode 2.

Here, we need to emphasize that not arbitrary  $\alpha$  and  $\beta$  can be identified by the method described above since the expression of  $L$  may not be satisfied. However, once the expression is satisfied, the corresponding parameters must be distinguishable such as  $\beta, \beta' \in \{0, \frac{\pi}{2}\}$  and  $\alpha_{H(V)}, \alpha'_{H(V)} \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ . In addition, we also need to point out that the methods used to distinguish Bob's operations are not unique. The above is just a strategy that is easy to describe in a unified way.

Consequently, Eve is able to identify Bob's operation. The remaining question is, if Bob does the measurement, can he find Eve's photon? We notice that Eve's photon travels back and forth between  $SM_A$  and Bob for  $L$  times for each cycle of the counterfactual Trojan horse attack. Even when Bob makes only one measurement that results in blocking that cycle. Therefore, the results given in Eqs. (4) and (10) are still correct, and the condition for a successful counterfactual Trojan horse attack is the same as that presented in the previous section. Eve needs an access time window to complete the nonlocal interaction between her eavesdropping photon and Bob's apparatus many times. During this period, Bob's manipulation needs to be consistent.

### B. Identifying Bob's multiple operations

In the above discussion, we show how Eve can identify two of Bob's operations by only one "ghost" photon. Obviously, if Eve sends  $j$  ghost photons which have no interference

with each other, the probability of any photon being found by Bob is  $1 - (1 - P_{i=s,d})^j \approx jP_i$ . When  $P_i$  tends to 0, the probability also tends to 0. It is worth noting here that for those Eve's photons, they can be used to distinguish Bob's different operations. Hence, it is possible for Eve to recognize multiple Bob operations with multiple photons.

Next, we assume that Bob may take five operations  $\{U_{BI}, U_{Bx}, U_{By}, U_{Bz}, U_{Bzx}\}$  [26]. We show Eve can use four ghost photons (a, b, c, and d) to distinguish them. These photons can be sent to Bob either by separate eavesdropping devices or one by one using the same eavesdropping device so that Eve can treat them independently.

Regarding the "a" photon, its initial photon state is  $|H\rangle$  and SPBS is set to reflect V photon with  $L = 1$  and  $U_{EL} = U_{ER} = I$ . Then, for Bob's operations  $U_{BI}$  and  $U_{Bz}$ ,  $D_N$  clicks, while for  $U_{Bx}$ ,  $U_{By}$  and  $U_{Bzx}$ ,  $D_P$  clicks.

Regarding the "b" photon, it is used to distinguish between  $U_{BI}$  and  $U_{Bz}$ . Initially, its photon state is  $|V\rangle$ , while  $L = 1$ ,  $U_{EL} = U_{ER} = I$ , and SPBS is set to be transparent. Now, for  $U_{BI}$ , it is  $D_N$  clicks, while for  $U_{Bz}$ , it is  $D_P$  clicks.

Regarding the "c" photon, it is used to distinguish between  $U_{Bzx}$  and  $\{U_{Bx}, U_{By}\}$ . Initially, its photon state is  $|V\rangle$ , while  $L = 2$ ,  $U_{EL} = U_{ER} = I$ , and SPBS is set to be transparent. Here, for  $U_{Bzx}$ ,  $D_P$  clicks, while for  $U_{Bx}$  and  $U_{By}$ , it is  $D_N$  clicks.

Regarding the "d" photon, it is used to distinguish between  $U_{Bx}$  and  $U_{By}$ . Initially, its photon state is  $|H\rangle$ , while  $L = 2$ ,  $U_{EL} = U_{PS}(-\pi/2, \pi/2)$ ,  $U_{ER} = U_{PR}(-\pi/2)$ , and SPBS is set to be transparent.  $D_P$  clicks for  $U_{Bx}$ , while  $D_N$  clicks for  $U_{By}$ .

As a result, with four ghost photons, Eve can identify five Bob's operations without being noticed.

## IV. COUNTERFACTUAL TROJAN HORSE ATTACK ON TYPICAL QUANTUM SECURE COMMUNICATION PROTOCOLS AND ITS GENERAL DEFENSE STRATEGY

In this section, we show how to eavesdrop two quantum secure communication protocols. The general defense strategy is also discussed. We show that the counterfactual Trojan horse attack is not difficult to defeat, but it must be taken seriously. The system vulnerabilities that the counterfactual Trojan horse attack can exploit must be eliminated. This can be done by properly designing communication apparatuses and not just improving experimental conditions and equipment quality.

### A. Ping-pong protocol

First, we elaborate on how to attack the ping-pong protocol [5], which is famous as a deterministic secure direct communication protocol. In the ping-pong protocol, Alice prepares polarization-entangled EPR pairs in one of the four Bell states  $|\phi^+\rangle = (|H_A\rangle|V_B\rangle + |V_A\rangle|H_B\rangle)/\sqrt{2}$ , where A and B indicate the "home photon" kept by Alice and the "travel photon" sent to Bob, respectively. Bob randomly chooses from two modes, the message mode and the control mode. In the message mode, Bob uses local unitary operations  $\sigma_I$  or  $\sigma_z$  to encode his message on the travel photon. The resulting states are  $|\phi^+\rangle$  or  $|\phi^-\rangle = (-|H_A\rangle|V_B\rangle + |V_A\rangle|H_B\rangle)/\sqrt{2}$ . Then, Bob sends the travel photon back to Alice, who performs a Bell measurement on the two photons to read out the result. In the control mode,

Bob uses the travel photon for an eavesdropping check. He measures the travel photon polarization in arbitrary measuring basis. Then, he sends the measurement result along with the measuring basis to Alice through the public classical channel. To verify the security of the communication channel, Alice performs a measurement in the same measuring basis on her photon and compares the outcome with Bob.

The ping-pong protocol sparked an interest in its applications to direct secure quantum communication. An analysis in the related security issue indicates that the ping-pong protocol is not secure when considering quantum channel noise [31] and actual detector imperfections [32]. However, here we show that Eve can steal the information in the ideal situation without being exposed, because the auxiliary active detection in the control mode is not reliable for our eavesdropping photons.

As discussed in Sec. III, utilizing the setup proposed in Fig. 1, Eve sends a vertically polarized photon ( $|V\rangle$ ) to Bob's apparatus. In the meantime, she sets  $L = 1$ ,  $U_{EL} = U_{ER} = I$ , and SPBS to be transparent. Eve needs to complete  $M$  cycles through the Michelson interferometer within one access time window for the single-cycle counterfactual Trojan horse attack, and  $M'N'$  cycles for the double-cycle counterfactual Trojan horse attack. In the message mode, Bob's operation  $\sigma_I$  directly reflects back the incoming photon and causes a detection event at Eve's detector  $D_N$ . Bob's operation  $\sigma_z$  induces a  $\pi$  phase shift to Eve's photon and causes a detection event at  $D_P$ . Thus, Eve can distinguish Bob's phase operation. In the control mode, Bob measures all incoming photons, thus blocks Eve's eavesdropping channel. According to Sec. II, Eve's eavesdropping device is in working mode 3. Consequently, the probability of the eavesdropping photon being found in the transmission channel is nearly zero and Eve's  $D_P$  clicks. In other words, Bob cannot "see" Eve's photon. In addition, it is worth noting that Bob needs to announce his measuring basis in the control mode. This is equivalent to telling Eve that her  $D_P$  is triggered by Bob's measurement instead of Bob's phase operation.

In addition, we note that an upgraded version of ping-pong protocol is proposed in Ref. [26]. Here, in the message mode, Bob utilizes four operations which are  $\sigma_I$ ,  $\sigma_x$ ,  $\sigma_z$ , and  $\sigma_{zx}$  to encode two-bit information. However, this protocol is still insecure under the counterfactual Trojan horse attack. As discussed in Sec. III, Eve can send three ghost photons to distinguish these four operations.

### B. Counterfactual quantum cryptography

Next, we discuss the counterfactual Trojan horse attack implementation on the counterfactual quantum cryptography (N09) [27], which is a QKD protocol utilizing the interaction free measurement. The N09 protocol distributes a secret key relying on the mere possibility for signal particles to be transmitted without any particle carrying secret information in the transmission channel. The protocol can be described as follows. Alice launches a single photon either in state  $|H\rangle$  (representing the bit value 0) or  $|V\rangle$  (representing the bit value 1) at random into a Michelson-type interferometer. One arm of the interferometer is public, but the rest is only accessible by Alice including two detectors  $D_1$  and  $D_2$ . By controlling the

public interferometer arm, Bob has two choices to influence Alice's photon evolution. The first choice is that Bob measures the H photon and returns the V photon, which represents his bit value 0. The second choice is that Bob measures the V photon and returns the H photon, which represents his bit value 1. Bob's two choices can be achieved with a polarization selection device and the detector  $D_3$ . Then, if Alice's and Bob's bit values are equal, blocking in the public interferometer arm destroys the interference and causes  $D_3$  to click with a 50% probability, while Alice's two detectors ( $D_1$ ,  $D_2$ ) each have a 25% probability of being triggered. However, if Alice's and Bob's bit values are not equal, only  $D_2$  clicks due to the interference. After the detection Alice and Bob announce which of their detectors clicks. If  $D_2$  or  $D_3$  clicks, they also check the polarization state to detect Eve's intervention. A sifted key is obtained only when  $D_1$  clicks. The N09 protocol is assumed to provide security advantages in the photon-number-splitting attack [27], the normal intercept-resend attack, and the general Trojan horse attack when hiding the quantum channel in a quantum network (see Appendix in Ref. [33]). The security of N09 has been proved based on a perfect single-photon source [34,35] and a weak-coherent-laser source [36]. However, this protocol is no longer secure under Trojan horse attack when Bob's detector is not ideal [12]. We now discuss what happens if Eve uses counterfactual Trojan horse attack.

In principle, counterfactual quantum cryptography is insecure under the counterfactual Trojan horse attack since Bob's measurement can no longer register Eve's photon. To implement the counterfactual Trojan horse attack, each time Alice launches a photon, Eve sends a V polarized eavesdropping photon to Bob using one of two types of the counterfactual Trojan horse attack setups (here AED is not necessary and can be set to be transparent). If Bob chooses to block the V photon and return the H photon, this leads Eve's detector  $D_P$  to click with almost a unit probability. If Bob chooses to block the H photon and return the V photon, Eve's photon causes  $D_N$  to click. Thus, Eve knows exactly every move of Bob. In 2012, an experimental realization of the counterfactual quantum cryptography was reported [28], which can be successfully attacked by the counterfactual Trojan horse attack as described above.

However, we should point out here that in Ref. [28], the polarization selection device is different from Noh's original design. In Ref. [28], the polarization selection device is implemented by a half-wave plate and a polarization beam splitter, while in Noh's work [27,29], the device consists of an optical loop (OL), a PBS, and a high speed optical switch (SW). According to Noh's description, the polarization selection device has no special requirements, which is optional [27]. However, as explained below, the original design of the polarization selection device in Noh's work can defeat the counterfactual Trojan horse attack, but the device proposed in Ref. [28] cannot. Therefore, the specific design of the device will determine if the communication can survive the counterfactual Trojan horse attack. To defeat the counterfactual Trojan horse attack, there are additional restrictions on device design.

The basic idea of polarization selection device suggested in Ref. [27] is to separate the H photon and V photon by a short distance, which is determined by the length of OL. The timing of different polarized photons passing through the

SW is different. By accurately controlling the switching time, Bob can route specific polarized photon to  $D_3$ . Regarding the counterfactual Trojan horse attack, the key here is that the active time of SW is less than the interval time  $T$  between the H and V photons, while an eavesdropping photon takes more time in Bob's apparatus than  $T$  ( $T < \tau$ ). Therefore, there is not enough time for Eve to complete a successful counterfactual Trojan horse attack, unless Eve can bypass OL and detect SW directly. It is worth mentioning that in Ref. [24], the method (quantum multiuser authorization system) for defending the counterfactual Trojan horse attack has the same spirit. That work also points out that the best defense strategy against the counterfactual Trojan horse attack is to precisely control the apparatus access time window so that the counterfactual Trojan horse attack cannot be completed in time. Regarding Ref. [28], unfortunately, its polarization selection device cannot control the access time window, and thus the scheme is insecure under the counterfactual Trojan horse attack.

In a nutshell, we emphasize three points at the end of this section. First, in the above discussion, we assume that all eavesdropped object's apparatuses (Alice and Bob) are ideal. Our main purpose is to show that the use of detectors to detect eavesdropping photons is not always reliable, even the detector is an ideal one. Second, the defense against the counterfactual Trojan horse attack depends heavily on the design of the communication apparatus, as in the case of the counterfactual quantum cryptography. Thus, the counterfactual Trojan horse attack must be taken seriously. Third, although the counterfactual Trojan horse attack can be prevented by controlling access time, this does not mean that eavesdroppers can be exposed. In fact, an important feature of the counterfactual Trojan attack is that it is hard to expose Eve, which gives Eve the advantage of hiding herself. A special case worth explaining is that Bob makes his apparatus accessible only when Alice's photon arrives. In such case, to avoid exposure, Eve cannot interfere with Alice's photon but rather set her photon state to be orthogonal with Alice's photon state. Filtering out Eve's photons does not help Bob to confirm that Eve is there, because his measurements only force the eavesdropping photons to stay inside Eve's device. As for Eve, she can take this advantage to scan Bob's weakness by using photons that have different states without being noticed. Although Eve may not succeed, a hidden eavesdropper is always a threat, hence, in the next section we will discuss a possible scenario to increase the probability of exposing the eavesdropper, which has not been discussed in Ref. [24] or somewhere else.

**V. POSSIBLE SCHEME TO MAKE "GHOST" PHOTONS VISIBLE**

Here, we assume that Eve has unlimited resources. Under this condition, we discuss how Bob can improve the probability of finding Eve's "ghost" photons.

We notice that most counterfactual quantum communication protocols are very sensitive to the channel noise [20,30,37–39], which means that when considering each cycle, the transmission channel is randomly blocked. As the channel noise increases, the communication efficiency, i.e.,

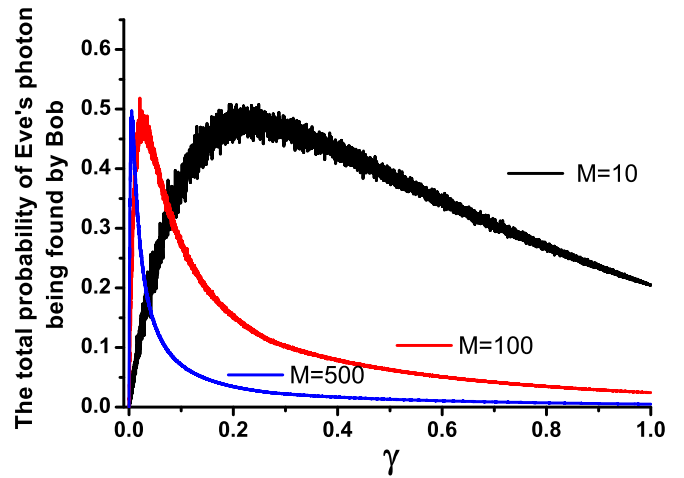


FIG. 2. The total probability of Eve's photon being detected in the transmission channel by Bob, versus the probability of Bob turning on his detector in each cycle,  $\gamma$ , for different cycle number  $M$ . Here, we assume that Eve performs the single-cycle counterfactual Trojan horse attack and Bob knows exactly Eve's attack strategy.

the probabilities of  $D_{N,P}$  registering a photon, decreases. Those researches also imply the probability of Eve's photon absorbed in the public transmission channel increases. There is no doubt that Bob can achieve similar results on purpose, i.e., for each cycle of the counterfactual Trojan horse attack, Bob can decide to block or not block the transmission channel at random. Then, he may improve the probability of finding Eve's photon.

In Figs. 2 and 3, we assume that Bob knows exactly Eve's attack strategy such that Bob could randomly choose to measure or pass the photon in each cycle of Eve's counterfactual Trojan horse attack. The probability that Bob performs the measurement is  $\gamma$ . We plot the total probability of Bob finding Eve's photon in the public transmission channel as a function

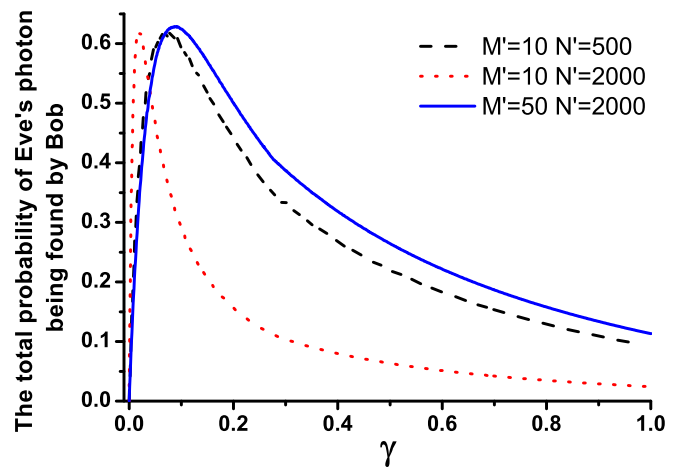


FIG. 3. The total probability of Eve's photon being detected in the transmission channel by Bob, versus the probability of Bob turning on his detector in each cycle,  $\gamma$ , for different cycle numbers  $M'$  and  $N'$ . Here, we assume that Eve performs the double-cycle counterfactual Trojan horse attack and Bob knows exactly Eve's attack strategy.



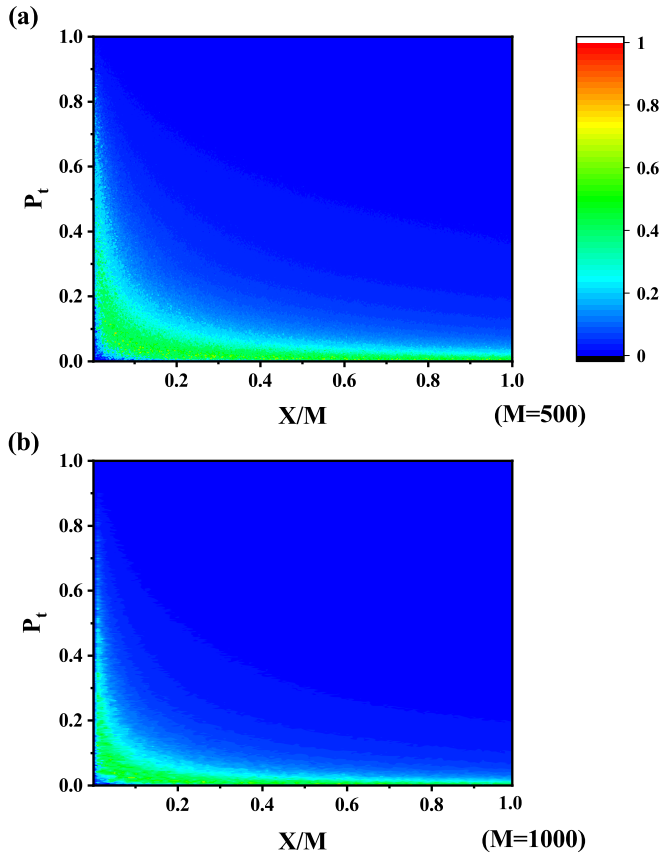


FIG. 4. We consider the single-cycle counterfactual Trojan horse attack and plot the total probability of Eve's photon being detected in the transmission channel by Bob vs  $P_t$  and  $X/M$ , for (a)  $M = 500$  and (b)  $M = 1000$ . Here,  $M$  is unknown to Bob.  $X$  is the total number of Bob's slots, while  $P_t$  is the probability of Bob turning on his detector in each slot.

of  $\gamma$  for various cycle numbers. Figure 2 is plotted for the single-cycle counterfactual Trojan horse attack, while Fig. 3 is plotted for the double-cycle counterfactual Trojan horse attack. Here we need to explain that since Bob's measurement process is random, we take multiple samples for each  $\gamma$  value and calculate the average probability. Therefore, Fig. 2 is not smooth. As shown in the figures, when Bob uses multiple random measurements, he has a large probability of finding Eve's photon at certain value of  $\gamma$ . Therefore, in principle, it is feasible for Bob to increase the probability of finding Eve by multiple random measurements. The next problem is, what if Bob does not know the details of Eve's counterfactual Trojan horse attack?

In the following, we assume that Bob does not know Eve's attack strategy. He splits the access time window for his one signal (bit value 0 or 1) into  $X$  time slots, each of which spans longer time than one cycle of Eve's counterfactual Trojan horse attack (otherwise, the situation is the same as in Figs. 2 and 3). In each slot, Bob chooses to continuously measure or pass the photon at random with measurement probability  $P_t$ . In Figs. 4 and 5, we plot the total probability of Bob finding Eve's photon in the public transmission channel as a function of  $X$  and  $P_t$ . Figure 4 is plotted for the single-cycle counterfactual Trojan horse attack with (a)  $M = 500$  and

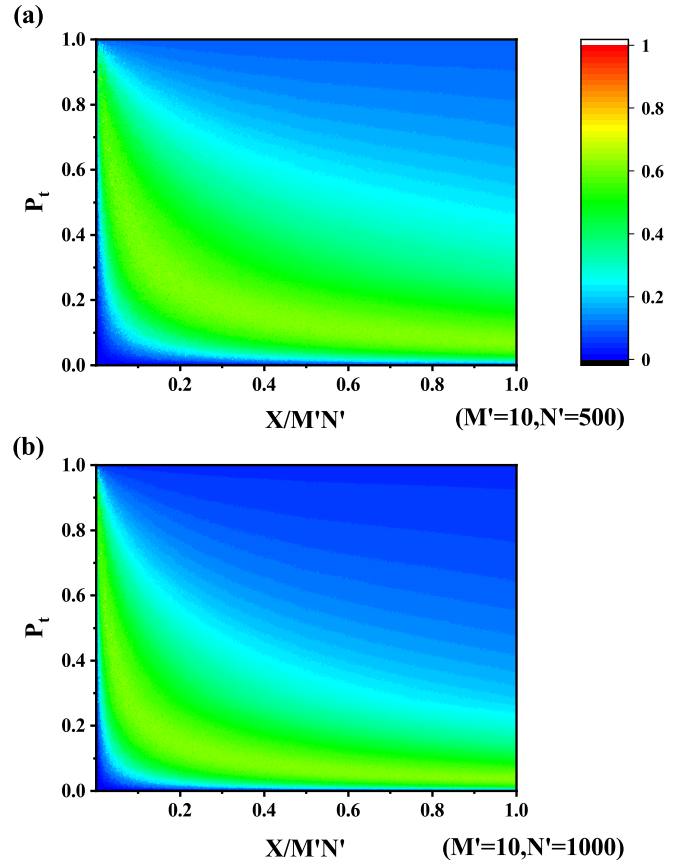


FIG. 5. We consider the double-cycle counterfactual Trojan horse attack and plot the total probability of Eve's photon being detected in the transmission channel by Bob vs  $P_t$  and  $X/M'N'$ , for (a)  $M' = 10$ ,  $N' = 500$  and (b)  $M' = 10$ ,  $N' = 1000$ . Here,  $M'$  and  $N'$  are unknown to Bob.  $X$  is the total number of Bob's slots, while  $P_t$  is the probability of Bob turning on his detector in each slot.

(b)  $M = 1000$ , while Fig. 5 is plotted for the double-cycle counterfactual Trojan horse attack with (a)  $M' = 10$ ,  $N' = 500$  and (b)  $M' = 10$ ,  $N' = 1000$ . According to Eqs. (4) and (10), we can calculate the total probability of Bob finding Eve's eavesdropping photon when Bob measures Eve's photon all the time. This probability is 0.5% for Fig. 4(a), 0.25% for Fig. 4(b), 10% for Fig. 5(a), and 5% for Fig. 5(b). Obviously, according to Figs. 4 and 5, with multiple random measurements, Bob can greatly enhance the probability of finding Eve's eavesdropping photon. The reason behind this behavior is that the counterfactual Trojan horse attack is highly dependent on the quantum Zeno effect. If Bob makes continuous measurement, the interference process in Eve's interferometer is interrupted, and hence Eve's photon is prevented from entering the transmission channel. However, when Bob stops measuring, the quantum Zeno effect is simply elapsed. The interference accelerates the speed at which photons enter the transmission channel. This can roughly explain why when Bob takes intermittent measurements, the probability of him finding Eve's photon increases.

Finally, it is desirable to emphasize three points. First, from Figs. 4 and 5, we can see that the single-cycle counterfactual Trojan horse attack saves Eve's resources (fewer cycles but

lower probability of being detected) and is more resistant to multiple random measurements than the double-cycle counterfactual Trojan horse attack (less green area). Second, the above simulations are based on the fact that all Eve's cycles are evenly distributed under an access time window. Even under this condition, Bob needs to choose the appropriate parameters ( $X$  and  $P_t$ ) to maximize the probability of finding Eve photons. However, if Eve chooses a different strategy, Bob's parameters may no longer be optimal. For example, as shown in the figure, for small  $X$ , Bob has a better chance of finding Eve's photon. A smaller  $X$  means that a slot has a longer duration. Apparently, if Eve compresses the time it takes her to complete an attack, she will reduce the probability of being disturbed by multiple random measurements. Therefore, the success of the defense strategy based on multiple random measurements depends on a specific Eve's attack strategy. Third, one can argue that specific designed multiple measurements rather than randomness can be used to increase the probability of finding Eve photons. For example, for the single-cycle counterfactual Trojan horse attack, we can unblock the transmission channel until the very last cycle and then measure Eve's photon. The probability of finding Eve photon is close to 100%. However, this defense strategy is still based on information about Eve's attack strategy. Under the conditions that Eve's strategy is unknown, multiple random measurements provide more robust choice.

## VI. CONCLUSION

In this paper, we discuss the basic principle and advantages of the counterfactual Trojan horse attack and introduce

two types of setups. Using the ideas of the interaction free measurement and the quantum Zeno effect, the counterfactual Trojan horse attack can completely steal secret information through "ghost" photons, which is capable of avoiding being detected spontaneously even if the detector is an ideal one. We demonstrate the effect of the counterfactual Trojan horse attack through examples of the ping-pong protocol and the counterfactual quantum cryptography protocol. We show that the eavesdropping checking technique utilizing auxiliary detector to measure eavesdropping photon may not be foolproof. In principle, the counterfactual Trojan horse attack can threaten any communication apparatus containing components that can reflect photons. Hence, more sophisticated defense strategy must be used when designing an actual secure communication apparatus. In addition, we present the properties and requirements that are needed for a successful counterfactual Trojan horse attack. One of the main requirements is to keep Bob's manipulation unchanged during the transmission of his one signal. We also briefly discuss the general defense strategy and the possible way to expose the counterfactual Trojan horse attack, which can be achieved by adding multiple random measurements.

## ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (NSFC) (Grants No. 11704241, No. 11474221, and No. 11874287) and Shanghai Science and Technology Committee (Grant No. 18JC1410900); This work is also supported by a grant from the King Abdulaziz City for Science and Technology (KACST).

- 
- [1] C. H. Bennett, and G. Brassard, in *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing, Bangalore, India* (IEEE, New York, 1984), p. 175.
  - [2] C. H. Bennett and G. Brassard, *IBM Tech. Discl. Bull.* **28**, 3153 (1985).
  - [3] A. K. Ekert, *Phys. Rev. Lett.* **67**, 661 (1991).
  - [4] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, *Rev. Mod. Phys.* **74**, 145 (2002).
  - [5] K. Boström and T. Felbinger, *Phys. Rev. Lett.* **89**, 187902 (2002).
  - [6] G.-L. Long, F.-G. Deng, C. Wang, X.-H. Li, K. Wen, and W.-Y. Wang, *Front. Phys. China* **2**, 251 (2007).
  - [7] W. K. Wootters and W. H. Zurek, *Nature (London)* **299**, 802 (1982).
  - [8] N. Gisin, S. Fasel, B. Kraus, H. Zbinden, and G. Ribordy, *Phys. Rev. A* **73**, 022320 (2006).
  - [9] N. Jain, E. Anisimova, I. Khan, V. Makarov, C. Marquardt, and G. Leuchs, *New J. Phys.* **16**, 123030 (2014).
  - [10] Q.-Y. Cai, *Phys. Lett. A* **351**, 23 (2006).
  - [11] F.-G. Deng, P. Zhou, X.-H. Li, C.-Y. Li, and H.-Y. Zhou, [arXiv:quant-ph/0508168](https://arxiv.org/abs/quant-ph/0508168).
  - [12] X. Q. Yang, K. J. Wei, H. Q. Ma, S. H. Sun, Y. G. Du, and L. A. Wu, *Phys. Lett. A* **380**, 1589 (2016).
  - [13] H. Weier, H. Krauss, M. Rau, M. Fürst, S. Nauerth, and H. Weinfurter, *New J. Phys.* **13**, 073024 (2011).
  - [14] X.-H. Li, F.-G. Deng, and H.-Y. Zhou, *Phys. Rev. A* **74**, 054302 (2006).
  - [15] A. C. Elitzur and L. Vaidman, *Found. Phys.* **23**, 987 (1993).
  - [16] P. G. Kwiat, H. Weinfurter, T. Herzog, A. Zeilinger, and M. A. Kasevich, *Phys. Rev. Lett.* **74**, 4763 (1995).
  - [17] P. G. Kwiat, A. G. White, J. R. Mitchell, O. Nairz, G. Weihs, H. Weinfurter, and A. Zeilinger, *Phys. Rev. Lett.* **83**, 4725 (1999).
  - [18] O. Hosten, M. T. Rakher, J. T. Barreiro, N. A. Peters, and P. G. Kwiat, *Nature (London)* **439**, 949 (2006).
  - [19] X.-S. Ma, X. Guo, C. Schuck, K. Y. Fong, L. Jiang, and H. X. Tang, *Phys. Rev. A* **90**, 042109 (2014).
  - [20] H. Salih, Z.-H. Li, M. Al-Amri, and M. S. Zubairy, *Phys. Rev. Lett.* **110**, 170502 (2013).
  - [21] Z.-H. Li, M. Al-Amri, and M. S. Zubairy, *Phys. Rev. A* **88**, 046102 (2013).
  - [22] Y. Cao, Y. H. Li, Z. Cao, J. Yin, Y. A. Chen, H. L. Yin, T. Y. Chen, X. F. Ma, C. Z. Peng, and J. W. Pan, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4920 (2017).
  - [23] C. Liu, J.-H. Liu, J.-X. Zhang, and S.-Y. Zhu, *Sci. Rep.* **7**, 10875 (2017).
  - [24] Z.-H. Li, M. Al-Amri, and M. S. Zubairy, *Sci. Rep.* **8**, 3899 (2018).
  - [25] Z.-H. Li, M. Al-Amri, and M. S. Zubairy, *Phys. Rev. A* **89**, 052334 (2014).
  - [26] Q.-Y. Cai and B.-W. Li, *Phys. Rev. A* **69**, 054301 (2004).

- [27] T.-G. Noh, *Phys. Rev. Lett.* **103**, 230501 (2009).
- [28] G. Brida, A. Cavanna, I. P. Degiovanni, M. Genovese, and P. Traina, *Laser Phys. Lett.* **9**, 247 (2012).
- [29] Y. Liu, L. Ju, X.-L. Liang, S.-B. Tang, G.-L. Shen Tu, L. Zhou, C.-Z. Peng, K. Chen, T.-Y. Chen, Z.-B. Chen, and J.-W. Pan, *Phys. Rev. Lett.* **109**, 030501 (2012).
- [30] L. J. Wang, Z.-H. Li, J. P. Xu, Y. P. Yang, M. Al-Amri, and M. S. Zubairy, *Opt. Express* **27**, 20525 (2019).
- [31] A. Wójcik, *Phys. Rev. Lett.* **90**, 157901 (2003).
- [32] F.-G. Deng, X.-H. Li, C.-Y. Li, P. Zhou, and H.-Y. Zhou, *Chin. Phys.* **16**, 277 (2007).
- [33] T.-G. Noh, [arXiv:0809.3979v2](https://arxiv.org/abs/0809.3979v2).
- [34] Z.-Q. Yin, H.-W. Li, W. Chen, Z.-F. Han, and G.-C. Guo, *Phys. Rev. A* **82**, 042335 (2010).
- [35] S. Zhang, J. Wang, and C.-J. Tang, *Chin. Phys. B* **21**, 060303 (2012).
- [36] Z.-Q. Yin, H.-W. Li, Y. Yao, C.-M. Zhang, S. Wang, W. Chen, G.-C. Guo, and Z.-F. Han, *Phys. Rev. A* **86**, 022313 (2012).
- [37] Q. Guo, L. Y. Cheng, L. Chen, H. F. Wang, and S. Zhang, *Sci. Rep.* **5**, 8416 (2015).
- [38] F. Li, J.-X. Zhang, and S.-Y. Zhu, *J. Phys. B: At. Mol. Opt. Phys.* **48**, 115506 (2015).
- [39] Z.-H. Li, M. Al-Amri, X. H. Yang, and M. S. Zubairy, *Phys. Rev. A* **100**, 022110 (2019).