Imperfect preparation and Trojan attack on the phase modulator in the decoy-state BB84 protocol

ALEKSEI REUTOV

QRate, Moscow, Russia
Moscow Center for Advanced Studies, Moscow, Russia

Abstract

Quantum key distribution (QKD) provides a theoretically secure method for cryptographic key exchange by leveraging quantum mechanics, but practical implementations face vulnerabilities such as Trojan horse attack on phase modulators. This work analyzes the security of QKD systems under such attacks, considering both ideal and imperfect state preparation scenarios. The Trojan attack model is generalized to arbitrary states of probing pulses and conservative bounds of information leakage through side-channel of special form are introduced. The quantum coin imbalance, a critical security parameter, remains low (on the order of 10^{-7} for ideal state preparation and 10^{-5} for imperfect preparation) with this new approach and presence additional hardware passive countermeasures. Numerical simulations confirm nonzero secure key rate at distances over 100 km through optical fiber channel.

Introduction

Quantum key distribution (QKD) represents a main area of modern quantum cryptography, suggesting a theoretically secure method [1] for exchanging cryptographic keys between distant parties. By leveraging the principles of quantum mechanics, an ideal implementation of QKD ensures that any eavesdropping attempt on the key exchange process is detectable, thereby providing a level of security unattainable by classical cryptographic methods. However, the practical implementation of QKD systems is not without challenges [2]. Real-world devices often deviate from assumptions of idealized theoretical QKD schemes, introducing vulnerabilities that can be exploited by sophisticated attacks. Among these, the Trojan horse attack [3] on phase modulators stands out as a powerful threat, capable of compromising the security of QKD systems without leaving detectable influence on observables measured during QKD protocol.

This article addresses the BB84 decoy-state protocol [4,5] and the specifics of Trojan-horse attacks on phase modulators (PMs). Unlike some previous studies [6–8] that assume coherent states for the probing Trojan pulses, this article is appealed to a more general scenario where the eavesdropper can prepare arbitrary states for phase-modulator's probing. Such generalization allows to derive conservative bounds on the information leakage caused by Trojan attacks and Alice's side channels. Side channel of specific and low-dimensional form is proposed as upper bound for any Trojan probing by arbitrary (pure or mixed) state. This result is formalized in three Statements, one for asymptotic case and two other for finite-key effects with statistical corrections based on Chernoff bounds. The possible imperfect state preparations are also considered and the imperfect choice of bits and bases are modeled using a Gaussian approximation close to behavior of quantum-state sources in practical QKD systems.

Presented in this article analysis reveals that the quantum coin imbalance, a critical parameter affecting phase error, remains weakly varying in the presence of Trojan attack and without it. There are demonstrated for realistic case of QKD implementation with isolators and spectral filters that the quantum coin imbalance values are on the order of 10^{-7} for ideal state preparation and 10^{-5} for imperfect preparation. These findings are validated through numerical simulations of the secret key rate, which show that our conservative approach yields low changes in key rate and maximum transmission distance. Furthermore, the performance of the protocol under realistic conditions of practical schemes is evaluated and it is demonstrated that secure key distribution is achievable over distances higher than 100 km.

In the following sections, an exposition of the state preparation process (Appendices A and B contain more details about imperfect state preparation and Gaussian approximation for it) is provided, the theoretical framework for Trojan attack on PM is analyzed and the numerical results support article's conclusions. This findings underscore the importance of rigorous security analysis in the design and implementation of practical QKD systems, ensuring their resilience against both known and emerging threats.
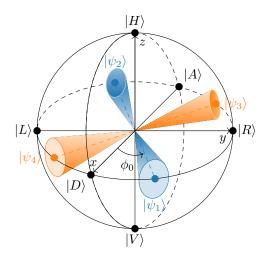
Figure 1: Polarization states on the Bloch sphere transmitted by Alice, reprinted from [9]. The blue and orange dots mark the perfectly prepared (2) states and the (4) states are schematically depicted as orange and blue areas on the sphere surface.

1 State preparation

The variation of BB84 protocol with decoy-state technique [4, 5] is used in this paper. BB84 [1], introduced by Bennett and Brassard in 1984, is a quantum key distribution (QKD) method that enables two distant parties, Alice and Bob, to securely generate a shared secret key using the principles of quantum mechanics. Alice sends qubits to Bob encoded in one of two non-orthogonal bases and Bob measures them in a randomly chosen basis, after which they publicly compare bases to discard mismatched results. The final key is distilled through error correction and privacy amplification providing information-theoretical security.

The decoy-state technique [10–13] is a method to enhance security and key generation rates in practical implementations using weak coherent pulses, which are susceptible to photon-number-splitting (PNS) attacks. By randomly adding decoy pulses of varying intensities to signal pulses, Alice and Bob can estimate the contribution of single-photon and multi-photon measurements. It allows two distant parties to detect and handle with eavesdropper's (Eve's) PNS attacks. This technique significantly improves the robustness and performance of QKD systems under realistic conditions.

In our implementation of decoy-state BB84 protocol [14] we use polarization encoding of prepared states. Any arbitrary pure polarization state can be described by spherical coordinates $\varphi \in [0, 2\pi)$ and $\theta \in [0, \pi]$ on the Bloch sphere,

$$|\psi(\varphi, \theta)\rangle = \cos\left(\frac{\theta}{2}\right)|H\rangle + e^{i\varphi}\sin\left(\frac{\theta}{2}\right)|V\rangle \, , \tag{1}$$

where $|H\rangle$ and $|V\rangle$ denote the horizontal and vertical polarization states respectively. The azimuth angle is set to $\theta = \pi/2$ in the protocol and the relative phase, which determines the basis and the bit value, is adjusted to $\phi_i \in \{0, \pi, \pi/2, 3\pi/2\}$. Thus, the azimuth angle of the ith state is $\varphi_i = \phi_0 + \phi_i$.

States in our practical scheme are prepared and measured (with assumption of ideal phase modulators) in elliptical polarization bases $X' : \{|\psi_1\rangle, |\psi_2\rangle\}$ and $Y' : \{|\psi_3\rangle, |\psi_4\rangle\}$ obtained by rotating the basis vectors $X : \{|D\rangle, |A\rangle\}$ and $Y : \{|R\rangle, |L\rangle\}$

around the z-axis by ϕ_0 (see Fig. 1):

$$|\psi_{1,2}^{\text{perfect}}\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle \pm e^{i\phi_0}|V\rangle\right), \quad |\psi_{3,4}^{\text{perfect}}\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle \pm ie^{i\phi_0}|V\rangle\right). \quad (2)$$

The density matrices of these states can be expressed as:

$$\rho_i = |\psi_i^{\text{perfect}}\rangle\langle\psi_i^{\text{perfect}}|. \quad (3)$$

Bob will perform measurements in the bases X' and Y' randomly applying one of the positive operator-valued measures $\{|\psi_i^{\text{perfect}}\rangle\langle\psi_i^{\text{perfect}}|\}$.

However, phase modulators a not ideal in practical schemes (e.g., due to imperfect voltage control or mechanical inaccuracy of connection between the optical components [9]). Thus, there is a deviation from $\theta = \pi/2$ on the Bloch sphere, as well as deviations of φ_i from the ideal $\varphi_i \in \{0, \pi, \pi/2, 3\pi/2\}$. As a result of all these imperfections, the ideal states (2) change and can be written as:

$$|\psi_i\rangle = |\psi(\varphi_i, \theta)\rangle, \quad (4)$$

where φ_i and θ are random variables with some probability distributions. Note that, these states are no longer mutually orthogonal in the corresponding basis and do not lie in the xy plane of the Bloch sphere. In Appendix A and B the case of imperfect preparation is considered in more details and the density matrices $\overline{\rho}_1$, $\overline{\rho}_2$, $\overline{\rho}_3$ and $\overline{\rho}_4$ are determined for the four polarization states of the protocol instead of $|\psi_i^{\text{perfect}}\rangle$.

2 Trojan attack on phase modulator

Trojan probing [3] of phase modulators in QKD schemes refers to an attack strategy where Eve actively interrogates the phase modulation components to extract sensitive information about the quantum states being transmitted. Eve can be remaining undetected by conventional security measures during this attack. Possible robust countermeasures includes real-time monitoring, anomaly detection, passive isolation of QKD devices by outer probing emission and modifications of theoretical protocol.

A number of papers [6–8] consider the Trojan attack under the assumption that the probe pulses are coherent states $|\alpha\rangle$ (the polarization dependencies are omitted for simplicity):

$$|\alpha\rangle = e^{-\frac{|\alpha^2|}{2}} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}}|n\rangle. \quad (5)$$

But Eve does not have to be limited by this assumption and theoretically can prepare any state of light. Moreover, infinite-dimensional coherent state complicates the fidelity calculation (especially for imperfect state preparation) needed to determine the phase error that affects the key rate. Therefore, this section provide framework for arbitrary probing state with simplification of fidelity calculation and, accordingly, determining the phase error.

Firstly, let's assume that Eve is able to prepare a Trojan probing with an arbitrary state as presented, for example, in [15]. But, unlike [15], this paper focus on detailed consideration Trojan probing of PM and corresponding usage of fidelity between quantum states. It is important to mention the interpretation of Alice behavior when she randomly choose *firstly* a number of photon in pulse [16, 17] (and

only then randomly choose a label of decoy or signal intensity). This paper introduce similar logic: there is assumed that Eve *firstly* decide to obtain a specific number of probing photons reflected from Alice devices and Eve's choice can be described as probability distribution P_n.

Let the eavesdropper emission reflected from Alice's devices be given by an arbitrary pure state $|\psi_{\text{out}}\rangle$:

$$|\psi_{\text{out}}\rangle = \sum_{n=0}^{\infty} \sqrt{P_n} \, |n\rangle \,. \tag{6}$$

Hence, the Trojan light can be described as a random variable, the outcome of which is the emission of n photons $|n\rangle$ with probability P_n. The eavesdropper attack is bounded in intensity from above by the value μ_{out} due to the maximum intensity permissible for the integrity of optical fiber [6,18] (i.e., up to critical intensity of light in optical fiber and optical devices after which the optical elements will be degradate and any transmission will be impossible). $|\psi_{\text{out}}\rangle$ has also 4 possible variations, depending on 4 settings on the transmitter phase modulator, which set the states $\overline{\rho}_i$ (33).

Assumption. *Let the protocol and its implementation be subject to the condition $\mu_{\text{out}} < 1$, where μ_{out} is the maximum intensity of the output Trojan emission.*

The choice of such a limitation is due to the fact that the presence of a non-zero key rate requires low values of the output probing intensity μ_{out} [6].

In the case of the implementation of an arbitrary Trojan emission (for now, it means Trojan pure states with arbitrary photon-number distribution), there are random sets of n_i probing photons and each set is obtained i-th times from Alice's setup. The total number of photons is limited by:

$$\sum_{i=1}^{N} n_i = N\mu_{\text{out}} \,, \tag{7}$$

where N is the number of pulses transmitted by Alice during the quantum key distribution. It is worth noting to say that $N\mu_{\text{out}} < N$ follows by Assumption, which means that it is impossible to randomly fill all Alice's pulses with non-zero Trojan photons. Each pulse with a non-zero Trojan photon means a pulse compromised by attack and this affects the final length of the secret key, but zero Trojan photons in the i-th pulse do not provide an information leakage. It follows that the more pulses are filled (and compromised by Trojan emission), the more information will be received by an eavesdropper through the Trojan side channel.

Consider the following side channel:

$$|\psi_{\text{s.ch.}}\rangle = \sqrt{1 - \mu_{\text{out}}} \, |\text{vac}\rangle + \sqrt{\mu_{\text{out}}} \, |1\,\text{ph.}\rangle \,, \tag{8}$$

where $|\text{vac}\rangle$ is the vacuum state, $|1\,\text{ph.}\rangle$ is the state with one photon. Such side channel will have an average intensity equal to μ_{out}. The proposed state corresponds to the situation when the Alice's pulses are filled either with one Trojan photon or remain empty. The total number of filled positions will be $N\mu_{\text{out}}$.

Statement 1. *Asymptotically (for $N \to +\infty$), any Trojan emission with arbitrary representation (6) and with average intensity no higher than μ_{out} gives the eavesdropper less equal or information than side channel (8).*

Proof. The amount of information received by the eavesdropper is determined by the Alice's pulses that contained Trojan photons. The more such filled pulses means the more information for the eavesdropper. Let us define the number of filled pulses $|\psi_{\text{out}}\rangle$ as K. The number of filled pulses K will be:

$$K = \sum_{n=1}^{\infty} N P_n \leq N \sum_{n=1}^{\infty} P_n \cdot n \leq N \sum_{n=0}^{\infty} P_n \cdot n = N \langle \psi_{\text{out}} | \hat{N} | \psi_{\text{out}} \rangle \leq N \mu_{\text{out}} \,, \quad (9)$$

where $\langle \psi_{\text{out}} | \hat{N} | \psi_{\text{out}} \rangle$ is average intensity of $|\psi_{\text{out}}\rangle$ and it is taken into account that $N \to +\infty$. $N\mu_{\text{out}}$ is the number of filled pulses for (8), i.e., due to the inequality (9), side channel (8) gives more or equal information to the eavesdropper than arbitrary Trojan probing (6). $\qquad\square$

However, in practical implementations of QKD, the statistics of measured observables are finite and the condition $N \to +\infty$ is not satisfied. In addition, due to channel losses and the usage of weak coherent states with significant vacuum component, only $M_1^L < N$ pulses are used to generate the secret key (where M_1^L is the decoy-state method estimation of a number of single-photon bits in the verified key). All other pulses do not provide information about the key and their leakage will not lead to the disclosure of bits of the verified key. Accordingly, the substitution $N \to M_1^L$ will be used in the proof of the following statement.

Statement 2. *With an accuracy of 2ε, any Trojan attack on PM of form (6) and with an intensity no higher than μ_{out} gives the eavesdropper less information than side channel with form:*

$$|\psi_{\text{s.ch.}}\rangle = \sqrt{1 - \mu'_{\text{out}}}\, |\text{vac}\rangle + \sqrt{\mu'_{\text{out}}}\, |1\,\text{ph.}\rangle \,, \quad (10)$$

where μ'_{out} is is determined through the implicit equation $M_1^L \mu'_{\text{out}} - \delta^L(M_1^L \mu'_{\text{out}}) = M_1^L \mu_{\text{out}} + \delta^U(M_1^L \mu_{\text{out}})$ and $\delta^{L,U}(x)$ is statistical corrections of the Chernoff bound, M_1^L is lower bound for single-photon clicks.

Remark. The upper bound K_1^U of the number of single-photon clicks compromised by a probing attack is defined similarly (9):

$$K_1^U = \mathbb{E}[K_1] + \delta^U(K_1) = M_1^L \sum_{n=1}^{\infty} P_n + \delta^U(\mathbb{E}[K_1]) \leq M_1^L \mu_{\text{out}} + \delta^U(\mathbb{E}[K_1]) \,, \quad (11)$$

where used

$$\sum_{n=1}^{\infty} P_n \leq \sum_{n=1}^{\infty} n P_n \quad (12)$$

and $\mu_{\text{out}} = \sum_{n=1}^{\infty} n P_n$ is upper bound for the average Trojan output intensity for $|\psi_{\text{out}}\rangle$ and $\delta^U(x)$ is provided by the Chernoff bound.

Proof. It can be shown for the Chernoff bounds that for two expectations $0 \leq A \leq B$:

$$f(A) \leq f(B) \quad (13)$$

where $f(x) = x + \delta^U(x)$, i.e. the inequality of expectations provides the inequality for the value bounds (the proof is given in the Appendix C in Lemma 1). Let $B = M_1^L \mu_{\text{out}}$ and $A = \mathbb{E}[K_1] \leq M_1^L \mu_{\text{out}}$, then the inequality (11) is rewritten as:

$$K_1^U \leq M_1^L \mu_{\text{out}} + \delta^U(M_1^L \mu_{\text{out}}) \,, \quad (14)$$

The expectation of compromised single-photon clicks for a side channel of the form (10) is defined as $\mathbb{E}[K_{s.ch.}] = M_1^L \mu'_{\text{out}}$. Similarly to (14), the lower bound is estimated as:

$$K_{s.ch.}^L \geq M_1^L \mu'_{\text{out}} - \delta^L(M_1^L \mu'_{\text{out}}),\tag{15}$$

and let this estimation be greater than the upper estimation of single-photon clicks compromised by the Trojan attack:

$$K_1^U \leq K_{s.ch.}^L.\tag{16}$$

This is satisfied by the following condition on μ'_{out}:

$$M_1^L \mu'_{\text{out}} - \delta^L(M_1^L \mu'_{\text{out}}) = M_1^L \mu_{\text{out}} + \delta^U(M_1^L \mu_{\text{out}}),\tag{17}$$

which proves the original statement with an accuracy of 2ε. $\qquad\square$

Still we assume pure states for Trojan probing. Now we write a generic (possibly mixed) state for Eve's probing system:

$$\rho_{\text{out}} = \sum_m p_m \left|\psi_m\right\rangle \left\langle\psi_m\right|,\tag{18}$$

where $\left|\psi_m\right\rangle$ can be not mutually orthogonal and has form close to (6):

$$\left|\psi_m\right\rangle = \sum_{n=0}^{\infty} \sqrt{P_{n,m}} \left|n\right\rangle.\tag{19}$$

Any ρ_{out} can be purified:

$$\left|\psi'_{\text{out}}\right\rangle = \sum_m \sqrt{p_m} \left|\psi_m\right\rangle \left|a_m\right\rangle = \sum_m \sum_{n=0}^{\infty} \sqrt{p_m P_{n,m}} \left|n\right\rangle \left|a_m\right\rangle,\tag{20}$$

where $\{\left|a_m\right\rangle\}$ is orthonormal basis of Eve's ancillary system. The number of vacuum Trojan pulses can be obtained as:

$$K_0 = N \sum_m p_m P_{0,m} = N - N \sum_m p_m \sum_{n=1}^{\infty} P_{N,m}\tag{21}$$

and filled Trojan pulses is

$$K = N \sum_m p_m \sum_{n=1}^{\infty} P_{N,m}\tag{22}$$

It is allow us to write equation similar to (9):

$$
\begin{aligned}
K = N \sum_{n=1}^{\infty} \sum_m p_m P_{N,m} &\leq N \sum_{n=1}^{\infty} n \cdot \sum_m p_m P_{n,m} \\
&\leq N \sum_{n=0}^{\infty} n \cdot \sum_m p_m P_{n,m} = N \left\langle\psi'_{\text{out}}\right| \hat{N} \left|\psi'_{\text{out}}\right\rangle = \text{Tr}[\hat{N}\rho_{\text{out}}] \leq N\mu_{\text{out}}.
\end{aligned}
\tag{23}
$$

Note, that operator $\hat{N}$ does not act on ancillary system of Eve. With the equation (23), Statement 1 and Statement 2, we can write final generalized statement.

Statement 3. *With an accuracy of 2ε, any Trojan attack on PM of arbitrary form (18) and with an intensity no higher than μ_{out} gives the eavesdropper less or equal information than side channel with form:*

$$|\psi_{\text{s.ch.}}\rangle = \sqrt{1 - \mu'_{\text{out}}}\,|\text{vac}\rangle + \sqrt{\mu'_{\text{out}}}\,|1\,\text{ph.}\rangle\,, \qquad (24)$$

where μ'_{out} is is determined through the implicit equation $M_1^L \mu'_{\text{out}} - \delta^L(M_1^L \mu'_{\text{out}}) = M_1^L \mu_{\text{out}} + \delta^U(M_1^L \mu_{\text{out}})$ and $\delta^{L,U}(x)$ is statistical correction of the Chernoff bound, M_1^L is lower bound for single-photon clicks.

Now we found that any Trojan attack with average intensity $\text{Tr}[\hat{N}\rho_{\text{out}}] \leq \mu_{\text{out}}$ can be conservatively bounded by a side channel of the form (10) with 2ε-accuracy. The side channel (10) can be represented in the case of polarization encoding as four density matrices as:

$$
\begin{aligned}
\rho_{E,1} &= \begin{pmatrix} \frac{\mu'_{\text{out}}}{2} & \frac{\mu'_{\text{out}}}{2} & \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ \frac{\mu'_{\text{out}}}{2} & \frac{\mu'_{\text{out}}}{2} & \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & 1-\mu'_{\text{out}} \end{pmatrix}, \\[6pt]
\rho_{E,2} &= \begin{pmatrix} \frac{\mu'_{\text{out}}}{2} & -\frac{\mu'_{\text{out}}}{2} & \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ -\frac{\mu'_{\text{out}}}{2} & \frac{\mu'_{\text{out}}}{2} & -\frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & -\frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & 1-\mu'_{\text{out}} \end{pmatrix}, \\[6pt]
\rho_{E,3} &= \begin{pmatrix} \frac{\mu'_{\text{out}}}{2} & -\frac{i\mu'_{\text{out}}}{2} & \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ \frac{i\mu'_{\text{out}}}{2} & \frac{\mu'_{\text{out}}}{2} & \frac{i}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & -\frac{i}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & 1-\mu'_{\text{out}} \end{pmatrix}, \\[6pt]
\rho_{E,4} &= \begin{pmatrix} \frac{\mu'_{\text{out}}}{2} & \frac{i\mu'_{\text{out}}}{2} & \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ -\frac{i\mu'_{\text{out}}}{2} & \frac{\mu'_{\text{out}}}{2} & -\frac{i}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} \\ \frac{1}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & \frac{i}{2}\sqrt{\mu'_{\text{out}}(1-\mu'_{\text{out}})} & 1-\mu'_{\text{out}} \end{pmatrix}.
\end{aligned}
\qquad (25)
$$

where, for example, the first matrix is obtained from $|\psi_{\text{s.ch.},1}\rangle\langle\psi_{\text{s.ch.},1}|$:

$$|\psi_{\text{s.ch.},1}\rangle = \sqrt{1-\mu'_{\text{out}}}\,|\text{vac}\rangle + \sqrt{\mu'_{\text{out}}}\,|D_{1\,\text{ph.}}\rangle \qquad (26)$$

and $|D_{1\,\text{ph.}}\rangle = (|H_{1\,\text{ph.}}\rangle + |V_{1\,\text{ph.}}\rangle)/\sqrt{2}$ denotes diagonal polarization with one photon. $|\psi_{\text{s.ch.},1}\rangle$ limits from above the Trojan light that came out of the Alice's setup when the Alice's pulse $\bar{\rho}_1$ is transmitted. All density matrices are found similarly to (26).

The density matrices (25) correspond to the side channel emission when the phase modulator is configured to generate states $|\psi_1\rangle$, $|\psi_2\rangle$, $|\psi_3\rangle$ and $|\psi_4\rangle$, respectively. Then the light in different polarization bases emitted from Alice's setup can be conservatively estimated from above as $\rho_{BE,X'}$ and $\rho_{BE,Y'}$:

$$
\begin{aligned}
\rho_{BE,X'} &= \bar{\rho}_1 \otimes \rho_{E,1} + \bar{\rho}_2 \otimes \rho_{E,2}\,, \\
\rho_{BE,Y'} &= \bar{\rho}_3 \otimes \rho_{E,3} + \bar{\rho}_4 \otimes \rho_{E,4}\,.
\end{aligned}
\qquad (27)
$$

(Here a generic form of the Alice's states $\bar{\rho}_i$ is choosed due to possible imperfect state preparation described in Appendices A and B.)

To quantify the difference between the phase error $E_1^{\text{ph},X'}$ and the bit error $E_1^{Y'}$, let use the concept of quantum coin introduced in [19] for an equivalent entanglement-based virtual protocol. Applying the complementarity argument [20]
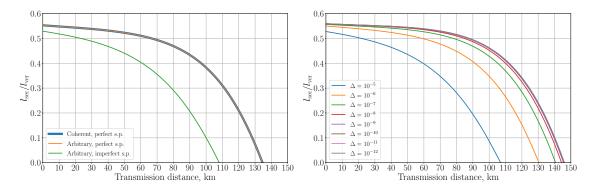
Figure 2: Left: the secret key rate per one verified click. Blue line indicates the scenario for the Trojan attack by probing with coherent pulses in the presence of perfect Alice's state preparation ($\Delta = 5 \times 10^{-7}$), orange and green are for Trojan probing by arbitrary pure states with perfect ($\Delta = 5 \times 10^{-7}$) and imperfect ($\Delta = 9.2 \times 10^{-6}$) Alice's state preparations respectively. Right: the secret key rate per one verified click for different values of $D\Delta$. Note, configurations of three isolators {28 dB, 28 dB, 48 dB} and {28 dB, 48 dB, 48 dB} provide (for the case of ideal state preparation) $\Delta \approx 10^{-9}$ and $\Delta \approx 10^{-11}$ respectively.

and the Bloch sphere bound [21] to the quantum coin yields the following inequality [22]:

$$\sqrt{F(\rho_{BE,X'}, \rho_{BE,Y'})} \le 1 - \mathcal{Y}_1 + \mathcal{Y}_1\left(\sqrt{E_1^{\mathrm{ph},X'} E_1^{Y'}} + \sqrt{(1 - E_1^{\mathrm{ph},X'})(1 - E_1^{Y'})}\right), \quad (28)$$

where the conditional probability of single-photon clicks $\mathcal{Y}_1 = (\mathcal{Y}_1^{X'} + \mathcal{Y}_1^{Y'})/2$ is introduced and the fidelity F between two states is defined as:

$$F(\rho_{BE,X'}, \rho_{BE,Y'}) \equiv \left[\mathrm{Tr}\left(\sqrt{\sqrt{\rho_{BE,X'}}\, \rho_{BE,Y'}\, \sqrt{\rho_{BE,X'}}}\right)\right]^2. \quad (29)$$

Here we have not taken into account possible transformations of all matrices by the same unitary operator U to Eve's and Alice's system, but it can be shown (through properties of square root of matrices and trace and properties of self-adjoint and unitary matrices) that substituting U into (29) will not change the value of $F(\rho_{BE,X'}, \rho_{BE,Y'})$. Solving (28), one can obtain the following upper bound on the single-photon phase errors:

$$E_1^{\mathrm{ph},X'} \le E_1^{Y'} + 4\Delta'(1 - \Delta')(1 - 2E_1^{Y'}) + 4(1 - 2\Delta')\sqrt{\Delta'(1 - \Delta')E_1^{Y'}(1 - E_1^{Y'})}, \quad (30)$$

$$\Delta' = \frac{1 - \sqrt{F(\rho_{BE,X'}, \rho_{BE,Y'})}}{2\mathcal{Y}_1} = \frac{\Delta}{\mathcal{Y}_1}, \quad (31)$$

where the value $\Delta = (1 - \sqrt{F})/2$ is usually called the quantum coin imbalance.

3   Results and discussion

The maximum input intensity of the Trojan probing is chosen as $I_{\mathrm{in}} = 2.5 \times 10^{12}$ photons per Alice's pulse and this value was determined for commercial QKD

scheme [18]. The total loss of the Trojan photons in the Alice's equipment is $\alpha_A = 172\,\mathrm{dB}$, where two optical isolators with $28\,\mathrm{dB}$ and $48\,\mathrm{dB}$ is used as passive countermeasure against Trojan-horse attack (see [18] for detailed description of the commercial scheme and experimental analysis of its Trojan-horse loophole). Authors determine the Trojan output intensity as $\mu_{out} = 10^{-\alpha_A/10} I_{in} = 1.5 \times 10^{-5}$. In this paper $\mu_{out} = 10^{-6}$ is chosen, which is realistic value for scheme with 2-3 isolators and corresponds previous investigations [6, 23]. However, the work [18] suggest to use an additional isolator for achieving value $\mu_{out} \leq 10^{-9}$. Therefore, the values up to $\mu_{out} = 10^{-12}$ are considered in the context of an influence on the secrete key rate.

Two cases of Alice's states are examined: the ideal state preparation (3) and the imperfect one with the Gaussian approximation (35). The imbalance value of the quantum coin were found as $\Delta = 5 \times 10^{-7}$ and $\Delta = 9.2 \times 10^{-6}$ by new approach and with the states (3) and (35), respectively. $\Delta = 5 \times 10^{-7}$ is for Trojan attack by coherent states [6] and perfectly prepared states. Simulation of the secret key rates gives the results shown in Fig. 2 (see Subsection 2.1 Experimental Setup and Appendix B in [9] for a description of the simulation methods and parameters).

It is worth noting to say that there are presented a rather conservative approach by replacing one side channel (Trojan attack with emission given by arbitrary statistics) with another side channel given by (10). However, the evaluation for the conservative side-channel (10) yields almost the same secrete key rate as in the scenario with the additional assumption of coherent Trojan states. The secrete key rate is also estimated for realistic setup parameters and imperfect state preparation (for simplicity described by a Gaussian model of imperfect polarization-state preparation) and the key transmission maximum range is found higher than 100 km (see left Fig. 2).

The main contribution to the quantum-coin imbalance value is derived from imperfect state preparation and it can be seen in significant difference between Δ's for ideal state preparation and modeled imperfect one. For example, the value $\mu_{out} = 10^{-100}$ provide $\Delta = 8.8 \times 10^{-6}$ for the proposed model of imperfectly prepared states. Nevertheless, after $\Delta < 10^{-8}$ (right Fig. 2) the influence on the key rate decrease rapidly. Consequently, the main leakage threat comes from imperfect preparation and conservative estimation of the Trojan-horse attack by side channel (10) (instead of previous approaches [6–8]) does not significantly reduce the secrete key rate.

4   Conclusions

This work provides a analysis of the security of quantum key distribution (QKD) systems against Trojan horse attack on phase modulators, considering both ideal and imperfect state preparation. By addressing imperfections in transmitted states and modeling Trojan probing by arbitrary generic (pure or mixed) states, a framework for enhancing QKD security in practical scenarios is developed. The Trojan attack model is generalized to arbitrary states, conservative bounds on information leakage are derived as leakage through single-photon side-channel states. This approach allows to more realistically assess the threat posed by Trojan horse attacks and it does not rely on restrictive assumptions about the nature of the eavesdropper's probing pulses.

The quantum coin imbalance remains low (5×10^{-7} for ideal and 9.2×10^{-6}

for imperfect preparation) even under Trojan attack on PM, demonstrating the resilience of practical QKD systems when equipped with proper countermeasures, such as isolators and spectral filters, which are essential for mitigating the risk of information leakage. Moreover, numerical simulations confirm secure key distribution over distances higher than 100 km. This work bridges theory and practice, offering tools to enhance QKD security against emerging threats for practical applications and increasing the feasibility of quantum-secured communication.

Acknowledgements

Appendix A   Imperfect state preparation

Density matrices $\rho_{X'} = \frac{1}{2}(|\psi_1\rangle\langle\psi_1| + |\psi_2\rangle\langle\psi_2|)$ and $\rho_{Y'} = \frac{1}{2}(|\psi_3\rangle\langle\psi_3| + |\psi_4\rangle\langle\psi_4|)$ can be written as follows in the case of imperfect preparation:

$$
\begin{aligned}
\rho_{X'} &= \frac{1}{2}\begin{pmatrix} 1 + \cos\theta & e^{-i\frac{\varphi_1+\varphi_2}{2}}\cos\left(\frac{\varphi_1-\varphi_2}{2}\right)\sin\theta \\ e^{i\frac{\varphi_1+\varphi_2}{2}}\cos\left(\frac{\varphi_1-\varphi_2}{2}\right)\sin\theta & 1 - \cos\theta \end{pmatrix}, \\
\rho_{Y'} &= \frac{1}{2}\begin{pmatrix} 1 + \cos\theta & e^{-i\frac{\varphi_3+\varphi_4}{2}}\cos\left(\frac{\varphi_3-\varphi_4}{2}\right)\sin\theta \\ e^{i\frac{\varphi_3+\varphi_4}{2}}\cos\left(\frac{\varphi_3-\varphi_4}{2}\right)\sin\theta & 1 - \cos\theta \end{pmatrix}.
\end{aligned}
\tag{32}
$$

The condition (2) $\rho_{X'} = \rho_{Y'}$ is satisfied for perfectly prepared states, i.e. the photon source is basis-independent. If $\rho_{X'} \neq \rho_{Y'}$ is true, then the values of single-photon phase errors $E_1^{\mathrm{ph},X'}$ in X'-basis are not estimated from the measured bit error $E_1^{Y'}$ in Y'-basis (or vice versa).

In the paper [24] devoted to fully passive QKD, the authors consider an equivalent protocol based on virtual entanglement. In this protocol, the source emits signal states with mixed polarization in the Z-basis. It is argued that Alice's imperfect preparation of states in the Z-basis is equivalent to Bob's imperfect measurement. This leads to one of the ideas in [24, 25] – the replacing of the source of randomly fluctuating on Bloch sphere pure states $\{|\psi_i\rangle\}$ with an equivalent source emitting mixed states $\{\overline{\rho}_i\}$,

$$
\overline{\rho}_i = \int_0^{2\pi} \int_0^{\pi} p_i(\varphi,\theta)|\psi(\varphi,\theta)\rangle\langle\psi(\varphi,\theta)|d\varphi d\theta\,,
\tag{33}
$$

where distributions $\{p_i(\varphi,\theta)\}$ is characterized directly from the test experimental setup measurements (a simple Gaussian approximation of this distributions is proposed in Appendix B). The density matrices (32) in this case are replaced by another one:

$$
\overline{\rho}_{X'} = \overline{\rho}_1 + \overline{\rho}_2\,, \quad \overline{\rho}_{Y'} = \overline{\rho}_3 + \overline{\rho}_4\,.
\tag{34}
$$

Appendix B   Gaussian distribution $p_i(\varphi, \theta)$

Using a simple Gaussian model $p_i(\varphi, \theta) = G\big(\varphi, \overline{\varphi}_i, \sigma_{\varphi_i}\big) G\big(\theta, \overline{\theta}, \sigma_\theta\big)$, the following analytical approximation can be obtained:

$$
\begin{aligned}
\overline{\rho}_i &\simeq \int_0^{2\pi} \int_0^{\pi} G\big(\varphi, \overline{\varphi}_i, \sigma_{\varphi_i}\big) G\big(\theta, \overline{\theta}, \sigma_\theta\big) |\psi(\varphi, \theta)\rangle \langle \psi(\varphi, \theta)| d\varphi d\theta \\
&\simeq \frac{1}{2}
\begin{pmatrix}
1 + e^{-\frac{\sigma_\theta^2}{2}} \cos \overline{\theta} & e^{-i\overline{\varphi}_i - \frac{1}{2}(\sigma_{\varphi_i}^2 + \sigma_\theta^2)} \sin \overline{\theta} \\
e^{i\overline{\varphi}_i - \frac{1}{2}(\sigma_{\varphi_i}^2 + \sigma_\theta^2)} \sin \overline{\theta} & 1 - e^{-\frac{\sigma_\theta^2}{2}} \cos \overline{\theta}
\end{pmatrix},
\end{aligned}
\tag{35}
$$

where for simplicity the probability density normalization coefficients are omited, since

$$
\int_0^{x_{\max}} G(x, \overline{x}, \sigma_x) dx = \frac{1}{2}\left[\operatorname{erf}\left(\frac{\overline{x}}{\sqrt{2}\sigma}\right) + \operatorname{erf}\left(\frac{x_{\max} - \overline{x}}{\sqrt{2}\sigma}\right) \right] \simeq 1,
\tag{36}
$$

for $\sigma_x \ll \overline{x}$ and $\sigma_x \ll x_{\max} - \overline{x}$.

Appendix C   Non-decreasing of Chernoff bound

Lemma 1. *The following holds:*

$$
f(x) > f(y),
\tag{37}
$$

for the Chernoff bound $f(x) = (1 + \delta)x$

$$
\left[\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right]^x = \varepsilon, \quad \delta > 0,\ x > 0,
\tag{38}
$$

and for $x > y$. The same statement holds for the lower Chernoff bound.

Proof. Let introduce the notation $f(x) = (1 + \delta)x = zx$ and rewrite (38):

$$
\frac{e^{z-1}}{z^z} = \varepsilon^{1/x}
\tag{39}
$$

Derivative with respect to x is:

$$
z' \frac{e^{z-1}}{z^z} \ln z = \frac{\varepsilon^{1/x}}{x^2} \ln(\varepsilon)
\tag{40}
$$

Substitute the left side of the equation (39) into the right side of (40):

$$
z' \frac{e^{z-1}}{z^z} \ln z = \frac{e^{z-1}}{z^z} \frac{1}{x^2} \ln(\varepsilon)
\tag{41}
$$

$$
z' \ln z = \frac{1}{x^2} \ln(\varepsilon)
\tag{42}
$$

Find $\ln(\varepsilon)$ using (39):

$$
\ln(\varepsilon) = x(z - 1 - z\ln(z))
\tag{43}
$$

and substitute in (42):

$$
z' \ln z = \frac{z - 1 - z\ln(z)}{x}
\tag{44}
$$

$$z' = \frac{z - 1 - z\ln(z)}{x\ln z} \tag{45}$$

Derivative $f'(x)$ is:

$$f'(x) = xz'(x) + z(x) = \frac{z - 1 - z\ln(z)}{\ln z} + z = \frac{z - 1}{\ln z} \tag{46}$$

Since $\delta > 0$, then $z = 1 + \delta > 1$ and $\ln(z) > 0$ which means $f'(x) = (z - 1)/\ln z > 0$. Consequently, $f(x)$ is an increasing function. Similarly, reversed inequality can be proven for the lower Chernoff bound. $\qquad\square$

Remark. Also the differential equation (42) gives an exact solution for $\delta(x) > 0$:

$$\delta(x) = e^{1 + W_0\left(-\frac{x + \ln(\varepsilon)}{ex}\right)} - 1, \tag{47}$$

where $W_0(y)$ is one of the Lambert W-function branches. Note, that other Chernoff bounds can be similarly rewritten in an explicit form with Lambert W-function branches.

References

[1] C. H. Bennett and G. Brassard, "Quantum cryptography: Public-key distribution and coin tossing", in *Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing (IEEE, New York, 1984)*, pp. 175–179. 1984.

[2] F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, "Secure quantum key distribution with realistic devices", *Reviews of Modern Physics* **92** (2020) no. 2, , `arXiv:1903.09051`.

[3] A. Vakhitov, V. Makarov, and D. R. Hjelme, "Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography", *Journal of Modern Optics* **48** (2001) no. 13, 2023–2038.

[4] A. Trushechkin, E. Kiktenko, and A. Fedorov, "Practical issues in decoy-state quantum key distribution based on the central limit theorem", *Phys. Rev. A* **96** (2017) 022316, `arXiv:1702.08531`.

[5] A. Trushechkin, E. Kiktenko, D. Kronberg, and A. Fedorov, "Security of the decoy state method for quantum key distribution", *Physics-Uspekhi* **64** (2021) no. 1, 88, `arXiv:2101.10128`.

[6] M. Lucamarini, I. Choi, M. Ward, J. Dynes, Z. Yuan, and A. Shields, "Practical Security Bounds Against the Trojan-Horse Attack in Quantum Key Distribution", *Physical Review X* **5** (2015) no. 3, , `arXiv:1506.01989`.

[7] W. Wang, K. Tamaki, and M. Curty, "Finite-key security analysis for quantum key distribution with leaky sources", *New. J. Phys.* **20** (2018) no. 8, 083027, `arXiv:1803.09508`.

[8] K. Tamaki, M. Curty, and M. Lucamarini, "Decoy-state quantum key distribution with a leaky source", *New. J. Phys.* **18** (2016) no. 6, 065008, `arXiv:1803.06045`.

[9] A. Reutov, A. Tayduganov, V. Mayboroda, and O. Fat'yanov, "Security of the decoy-state bb84 protocol with imperfect state preparation", *Entropy* **25** (2023) no. 11, 1556, `arXiv:2310.01610`.

[10] W.-Y. Hwang, "Quantum Key Distribution with High Loss: Toward Global Secure Communication", *Phys. Rev. Lett.* **91** (2003) 057901, `arXiv:quant-ph/0211153`.

[11] H.-K. Lo, X. Ma, and K. Chen, "Decoy State Quantum Key Distribution", *Phys. Rev. Lett.* **94** (2005) 230504, `arXiv:quant-ph/0411004`.

[12] X.-B. Wang, "Beating the Photon-Number-Splitting Attack in Practical Quantum Cryptography", *Phys. Rev. Lett.* **94** (2005) no. 23, , `arXiv:quant-ph/0410075`.

[13] X. Ma, B. Qi, Y. Zhao, and H.-K. Lo, "Practical decoy state for quantum key distribution", *Phys. Rev. A* **72** (2005) 012326, `arXiv:quant-ph/0503005`.

[14] A. Duplinskiy, V. Ustimchik, A. Kanapin, V. Kurochkin, and Y. Kurochkin, "Low loss qkd optical scheme for fast polarization encoding", *Opt. Express* **25** (2017) no. 23, , `arXiv:1709.06655`.

[15] M. Pereira, G. Currás-Lorenzo, A. Navarrete, A. Mizutani, G. Kato, M. Curty, and K. Tamaki, "Modified bb84 quantum key distribution protocol robust to source imperfections", *Phys. Rev. Res.* **5** (2023) 023065, `arXiv:2210.11754`.

[16] C. C. W. Lim, M. Curty, N. Walenta, F. Xu, and H. Zbinden, "Concise security bounds for practical decoy-state quantum key distribution", *Physical Review A* **89** (2014) no. 2, , `arXiv:1311.7129`.

[17] D. Tupkary, S. Nahar, P. Sinha, and N. Lütkenhaus, "Phase error rate estimation in qkd with imperfect detectors", *arXiv eprint* (2024) , `arXiv:2408.17349`.

[18] V. Makarov, A. Abrikosov, P. Chaiwongkhot, A. K. Fedorov, A. Huang, E. Kiktenko, M. Petrov, A. Ponosova, D. Ruzhitskaya, A. Tayduganov, D. Trefilov, and K. Zaitsev, "Preparing a commercial quantum key distribution system for certification against implementation loopholes", *Physical Review Applied* **22** (2024) no. 4, , `arXiv:2310.20107`.

[19] D. Gottesman, H.-K. Lo, N. Lüttkenhaus, and J. Preskill, "Security of quantum key distribution with imperfect devices", *Quant. Inf. Comput.* **4** (2004) no. 5, 325–360, `arXiv:quant-ph/0212066`.

[20] M. Koashi, "Simple security proof of quantum key distribution based on complementarity", *New J. Phys.* **11** (2009) no. 4, 045018.

[21] K. Tamaki, M. Koashi, and N. Imoto, "Unconditionally Secure Key Distribution Based on Two Nonorthogonal States", *Phys. Rev. Lett.* **90** (2003) 167904, `arXiv:quant-ph/0212162`.

[22] H.-K. Lo and J. Preskill, "Security of Quantum Key Distribution Using Weak Coherent States with Nonrandom Phases", *Quantum Info. Comput.* **7** (2007) no. 5, 431–458, `arXiv:quant-ph/0610203`.

[23] M. Pereira, M. Curty, and K. Tamaki, "Quantum key distribution with flawed and leaky sources", *npj Quantum Information* **5** (2019) , `arXiv:1902.02126`.

[24] W. Wang, R. Wang, C. Hu, V. Zapatero, L. Qian, B. Qi, M. Curty, and H.-K. Lo, "Fully Passive Quantum Key Distribution", *Phys. Rev. Lett.* **130** (2023) 220801, `arXiv:2207.05916`.

[25] V. Zapatero, W. Wang, and M. Curty, "A fully passive transmitter for decoy-state quantum key distribution", *Quantum Science and Technology* **8** (2023) no. 2, 025014, `arXiv:2208.12516`.