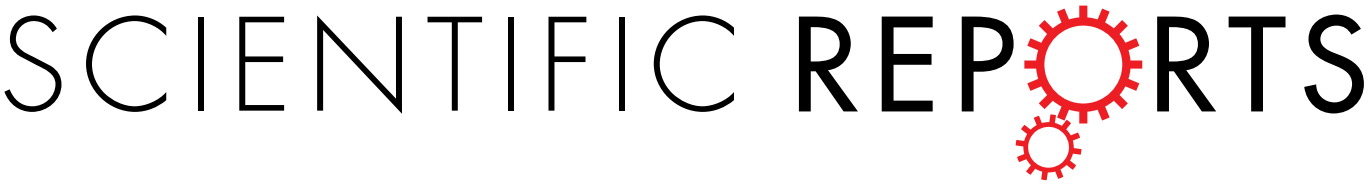

OPEN

Invisible Trojan-horse attack

Shihan Sajeed[1,2], Carter Minshull[1,3], Nitin Jain[4] & Vadim Makarov[3,1,2]

Received: 28 April 2017
Accepted: 10 July 2017
Published online: 21 August 2017

We demonstrate the experimental feasibility of a Trojan-horse attack that remains nearly invisible to the single-photon detectors employed in practical quantum key distribution (QKD) systems, such as Clavis2 from ID Quantique. We perform a detailed numerical comparison of the attack performance against Scarani-Acín-Ribordy-Gisin (SARG04) QKD protocol at 1924 nm versus that at 1536 nm. The attack strategy was proposed earlier but found to be unsuccessful at the latter wavelength, as reported in N. Jain *et al.*, New J. Phys. 16, 123030 (2014). However at 1924 nm, we show experimentally that the noise response of the detectors to bright pulses is greatly reduced, and show by modeling that the same attack will succeed. The invisible nature of the attack poses a threat to the security of practical QKD if proper countermeasures are not adopted.

Executive summary. A previous study in 2014 proposed a Trojan-horse attack against Clavis2 receiver (Bob) module; however the attack fell short of the performance level needed to breach the system security – by a large margin of roughly 100 times. Our present study shows that if an attacker resorts to using a longer wavelength (> 1900 nm) not ordinarily used in telecommunication, the same attack may breach the security. Although a complete eavesdropping apparatus is still quite challenging to build, it might be possible with today's or near-future technology. To prevent this, we have recommended the manufacturer to install a wavelength filter, which is a simple fiber-optic component that can be added just outside the installed system without having to recall it to the factory. For customers using ID Quantique's QKD products for critical data protection, we recommend that they inquire the manufacturer about this upgrade at the next convenient opportunity, such as a scheduled on-site maintenance. Not every installed system requires this upgrade: some systems are using protocols not vulnerable to this attack, and some may already have the wavelength filter included as part of network configuration. Since QKD cannot be attacked retroactively, security of customers' historical network transmissions is not affected by this study.

Introduction. Quantum cryptography allows two parties, Alice and Bob, to obtain random but correlated sequences of bits by exchanging quantum states[1–3]. The bit sequences can then be classically processed to get shorter but secret keys. The security of the key relies on the fact that an adversary Eve cannot eavesdrop on the exchange without introducing errors noticeable to Alice and Bob. This constitutes a solution to the problem of key distribution in cryptography, and is better known as quantum key distribution (QKD).

The security of keys distributed over the 'quantum channel' connecting Alice and Bob can be validated by a theoretical security proof. If the amount of errors observed by the two parties exceed a certain threshold, they abort the QKD protocol. Conversely, if the incurred quantum bit error rate (QBER) is below the abort threshold Q_{abort}, the protocol guarantees that Eve cannot know the secret key, except with a vanishingly small probability[3].

However, due to discrepancies between theory and practice, the operation of the QKD protocol may be manipulated by Eve in order to gain information about the key without introducing too many errors. Such discrepancies can arise due to imperfections in the physical devices used in the implementation and/or incorrect assumptions in the theoretical security proofs[3–5]. The field of 'quantum hacking' investigates practical QKD implementations to find such theory-practice deviations, demonstrate the resultant vulnerability via proof-of-principle attacks, and propose countermeasures to protect Alice and Bob from Eve. Over the years, many vulnerabilities have been discovered and attacks have been proposed and demonstrated on both commercial and laboratory QKD systems; see refs.[6–8] for reviews. In most cases, it was shown that under attack conditions, the QBER $Q \leq Q_{\text{abort}}$ but Eve's knowledge of the secret key was substantially larger than the predictions of the security proof.

In the so-called Trojan-horse attack[9] (introduced as a 'large pulse attack' a few years before[10]), Eve probes the properties of a component inside Alice or Bob by sending in a bright pulse and analyzing a suitable

[1]Institute for Quantum Computing, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. [2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. [3]Department of Physics and Astronomy, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. [4]Department of Physics, Technical University of Denmark, Fysikvej, Kongens Lyngby, 2800, Denmark. Correspondence and requests for materials should be addressed to S.S. (email: shihan.sajeed@gmail.com) or N.J. (email: nitinj@iitbombay.org)

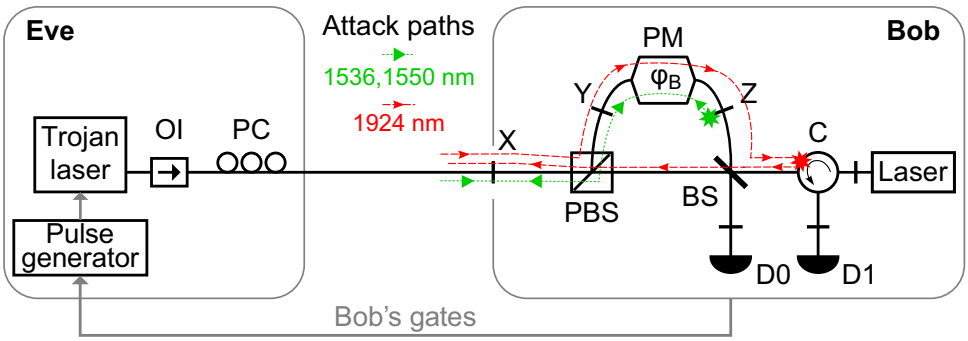

Figure 1. Basic experimental schematic and attack paths at $\lambda_s = 1536$ nm and $\lambda_l = 1924$ nm. The scheme and operation of Bob's setup is described in detail in ref.[13,17]. The stars indicate the back-reflection sources exploited in ref.[11] and in this work. Trojan laser models: Eblana Photonics EP1925-DM-B06-FA at λ_l and Alcatel 1905 LMI at λ_s. OI, optical isolator; PC, polarization controller; PBS, polarizing beamsplitter; BS, 50:50 beamsplitter; C, circulator; D, single-photon detectors; X, Y, Z, bulkhead fiber-optic connectors.

back-reflected pulse. This attack was recently demonstrated[11] with the intention to breach the security of the Scarani-Acín-Ribordy-Gisin QKD protocol (SARG04)[12] running on the commercial QKD system Clavis2 from ID Quantique[13]. SARG04 is a four-state protocol that is equivalent to the Bennett-Brassard QKD protocol (BB84)[1] in the quantum stage. Their difference comes in the classical processing stage: in SARG04, the bases selections of Bob are used for coding the secret bits, unlike in BB84 where they are publicly revealed. Therefore, if Eve surreptitiously gets information about Bob's bases selections at any time, she can compromise the security of the QKD system running SARG04. (In contrast, a Trojan-horse attack on Bob running the BB84 protocol is normally useless[10], unless it is combined with other attacks[14–16]).

In the attack demonstration[11], it was shown that getting the bases' information in a remote manner was indeed possible via homodyne measurement of the back-reflected photons. The path taken by these photons at 1550 nm, as depicted by the green dotted line in Fig. 1, traverses Bob's phase modulator (PM) twice. The homodyne measurement thus allowed discerning the phase applied by Bob, which is equivalent to knowing his basis selection. This 'phase readout' was accurate in >90% cases even when the mean photon number of the back-reflected pulses was ≈3.

Despite that, an overall attack on the QKD system did not have a chance to succeed owing to a side effect produced when the bright pulses went on to hit the detectors D0 and D1, as may be visualized in Fig. 1. To elaborate, the bright pulses result in a severe afterpulsing in these InGaAs/InP single-photon detectors (SPDs), which are operated in a gated mode. For a single bright pulse that hits D1, even if well outside a gate, the cumulative probability of a spurious detection event due to afterpulsing crosses 40% (which is ~4 times the detection probability of a single photon) in just 5 gate periods[18]. The resulting detection events (clicks) are accidental, i.e., erroneous in half of the cases. Hence, only a handful of Trojan-horse pulses (THPs) suffice to rapidly elevate the number of erroneous clicks and make the QBER surpass Q_{abort}, even though Eve's actual knowledge I_E^{act} of the key is still quite small. An elaborate attack strategy to improve I_E^{act} was proposed and numerically simulated, however, it could also not simultaneously satisfy $Q \leq Q_{\text{abort}}$ together with $I_E^{\text{act}} > I_E^{\text{est}}$, where I_E^{est} is the estimated (theoretical) security bound on Eve's knowledge that Clavis2 uses to produce the final secret key[11]. While ref.[11] did not prove that a better attack could not be constructed, the attack proposed failed in practice by a large margin.

In this Article, we provide experimental evidence that this Trojan-horse attack could however succeed if Eve were to craft bright pulses at a wavelength where the afterpulsing experienced by the SPDs is considerably lower. The underlying physics is that photons with energy lower than the bandgap of the SPD absorption layer material (InGaAs) mostly pass the material unabsorbed, thereby causing negligible afterpulsing. Indeed, we confirm experimentally that at a relatively longer wavelength $\lambda_l = 1924$ nm, the SPD has much less afterpulsing than at $\lambda_s = 1536$ nm (similar to the wavelength used in ref.[11]). We then perform a numerical comparison of the attack conditions and performance at λ_l with these at λ_s. By means of an optimized simulation that assumes fairly realistic conditions, we show that the actual attack at λ_l can break the security of Clavis2. The attack in itself is general enough to be potentially applicable to most discrete-variable QKD systems, and can be categorized with those that exploit vulnerabilities arising from the wavelength-dependence of optical components[19,20].

Experiment

While using $\lambda_l = 1924$ nm for the attack offers the benefit of reduced afterpulsing, the transmittance and reflectance properties of different optical components inside Bob vary greatly in comparison with those measured at $\lambda_s = 1536$ nm. Most relevant to the attack, the attenuation is generally higher; for instance, the optical loss through the PM at λ_l is $\gtrsim$20 dB higher than that at λ_s. Furthermore, the modulation itself varies with λ since the modulator's half-wave voltage is a function of wavelength. If Eve uses light at λ_l to estimate Bob's randomly modulated phase ($\varphi_B = 0$ or $\pi/2$ at λ_s) through the homodyne measurement of a pulse that made a single pass through the PM, the measurement outcomes will not be on orthogonal quadratures.

Altogether, it is thus likely that compared to ref.[11], Eve would not only need to inject a larger mean photon number $\mu_{E\to B}$ into Bob, but may also require a higher mean photon number $\mu_{B\to E}$ in the back-reflection for

Paths & points	Loss at λs (dB)	Loss at λ_l (dB)
X–Y	0.9	3.6
Y–Z	2.6	23.0
Z*	51.7	
Z–C*–X		58.4 to 65.8 (polarization-dependent)
X–D0	8.8 (via long arm)	15.5 (via short arm)
X–C–D1	9.2 (via long arm)	25.8 (via short arm)

Table 1. Comparison of optical losses in Bob at λs versus λ_l. See Fig. 1 for location of the paths and points. The loss during reflection $\Gamma_{Z^{\star}}$ was measured at 1550 nm[11], which we consider to be close enough to our $\lambda_s = 1536$ nm.

successful homodyne measurements. To calculate the efficacy of the attack, we experimentally quantify at λ_l (relative to λ_s) the following three aspects: increased attenuation, altered phase modulation, and decreased afterpulsing. Figure 1 shows a schematic of the experimental setup used for various measurements.

Increased attenuation. To gauge the increase in attenuation, we measured the optical loss of various components of Bob at both λ_s and λ_l. In Fig. 1, the dotted line (path X–Y–Z*–Y–X, where ☆ indicates the source of reflection) shows the attack path used in ref.[11]. Relevant loss values are given in the left column of Table 1. With a round trip loss of $L_{X-Y-Z^{\star}-Y-X}(\lambda_s) = 2L_{X-Y}(\lambda_s) + \Gamma_{Z^{\star}} + 2L_{Y-Z}(\lambda_s) = 58.7$ dB, Trojan-horse pulses injected with $\mu_{E\to B} \approx 2 \times 10^6$ photons yielded $\mu_{B\to E} \approx 4$ photons in the back-reflection from Bob. Here, $\Gamma_{Z^{\star}} = 51.7$ dB is the loss during reflection at Z, the fiber connector after Bob's PM.

For an attack at λ_l with Trojan-horse pulses traversing the same path, the round trip loss would be $L_{X-Y-Z^{\star}-Y-X}(\lambda_l) = 104.9$ dB (with the further assumption that $\Gamma_{Z^{\star}}$ is independent of wavelength). The attack pulses at λ_l would therefore face 46.2 dB more attenuation than at λ_s. A major contribution to this large attenuation is from the PM, which even gets doubled since the THPs travel through the PM twice.

However, since a single pass can also yield information about φ_B, Eve can opt for a different route where only either the input forward-traveling THP or the back-reflected pulse passes through Bob's PM. All Eve requires is a reasonably large source of reflection from any component after the 50:50 beamsplitter (BS). Indeed, during our loss measurements at λ_l we observed a large attenuation through the optical circulator (C), a part of which stems from a rather generous back-reflection. We estimated the loss $L_{Z-C^{\star}-X}(\lambda_l)$ for the path Z–C*–X (via BS twice and polarizing beamsplitter once) using a photon-counting method, described below.

We temporarily connected the polarization-controlled output of the 1924 nm laser at Z to send light towards the BS. The average power of the pulsed laser, operated at 5 MHz repetition rate, was $P_{avg} = 21.55\,\mu$W, corresponding to a mean photon number per pulse $\mu_Z = 4.14 \times 10^7$. An SPD was connected at X to detect the back-reflections from C. To prevent other back-reflections from contributing to the photon counts, Bob's laser and detectors D0 and D1 were disconnected, and the patchcords (with open connectors) were coiled on a pencil to strongly attenuate the propagating light.

Two counters (Stanford Research Systems SR620) were used to measure the number of optical pulses sent by the laser $N = 4.98 \times 10^6$ and the number of pulses received by the detector $n = 323$ maximized over input polarization at Z. The mean photon number per pulse at X was estimated as $\mu_X \approx 59.7$ from the relation,

$$\frac{n-d}{N} = 1 - e^{-\mu_X \eta_d} \approx \mu_X\, \eta_D, \tag{1}$$

where $d = 60$ is the number of dark counts and $\eta_D = 8.85 \times 10^{-7}$ is the single-photon detection efficiency at λ_l, which was estimated in a separate experiment similar to the one in ref.[20]. The ratio of the mean photon numbers μ_Z/μ_X provides the overall loss $L_{Z-C^{\star}-X}(\lambda_l) \approx 58.4$ dB. The dashed line in Fig. 1 shows the complete attack path. Eve's THPs from the quantum channel enter the long arm of Bob, pass through the modulator, and after a reflection from the BS, propagate to the circulator. Here, they get back-reflected and then take the short arm to exit Bob, passing through the BS again. Using Table 1, this path can be characterized by a total loss $L_{X-Y-Z-C^{\star}-X}(\lambda_l) = L_{X-Y}(\lambda_l) + L_{Y-Z}(\lambda_l) + L_{Z-C^{\star}-X}(\lambda_l) = 85.0$ dB.

As noted above, the value of μ_X was polarization-sensitive. For the worst input polarization, μ_X decreased by 7.4 dB, changing the overall loss to $L_{X-Y-Z-C^{\star}-X}(\lambda_l) = 92.4$ dB. For the rest of the paper, we shall assume the attack pulses to be in a polarization midway between the best and the worst, leading to a loss figure of $L_{X-Y-Z-C^{\star}-X}(\lambda_l) = 87.3$ dB used to decide Eve's photon budget. In terms of photon numbers, this implies that in order to get the same number of photons out from Bob (i.e., $\mu_{B\to E} \approx 4$), Eve needs to inject $\rho = 10^{(-58.7+87.3)/10} = 7.24 \times 10^2$ times more photons at λ_l than at λ_s.

Altered phase modulator response. We now explain an impact of the altered phase modulation experienced by Eve's THPs at λ_l as they travel through Bob's PM. As mentioned before, Bob randomly chooses between voltages $V_0 (=0)$ or $V_{\pi/2}$ to apply a phase $\varphi_B = 0$ or $\pi/2$ on Alice's incoming quantum signal at (or in the vicinity of) $\lambda_s = 1536$ nm. Eve's objective is to learn φ_B. The double pass through the PM in ref.[11] implied that Eve had to discriminate between a pair of coherent states with angle $\theta(\lambda_s) \equiv \theta_s = 2 \times \pi/2 = \pi$ between them, as illustrated in Fig. 2(a). At $\lambda_l = 1924$ nm, the phase modulator is expected to lose efficiency and provide less phase shift at the

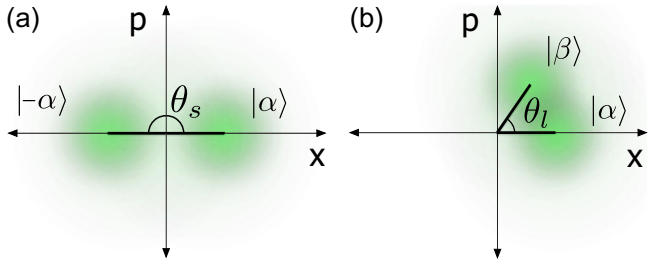

Figure 2. Illustrative phase space representation of the back-reflected states. Eve attempts to discern $\varphi_B = 0$ or $\pi/2$ by performing optimal detection on the back-reflected weak coherent states $|\alpha\rangle$ and $|\beta\rangle$ that have a non-zero overlap. (**a**) The complex amplitude $\beta = \alpha e^{i\theta_s} = -\alpha$, as a result of the double pass at the attack wavelength of λ_s. (**b**) $\beta = \alpha e^{i\theta_l}$, as a result of the single pass at λ_l through Bob's modulator.

same voltage. Furthermore, Eve's THP only traverses it once. Assuming a linear response of the PM, one can calculate the angle $\theta_l = [V_{\pi/2}(\lambda_s)/V_{\pi/2}(\lambda_l)] \times \pi/2$ between the coherent states available to Eve.

Since the half-wave voltage of the PM at 1924 nm was not specified by the manufacturer, we experimentally measured it. We constructed a balanced fiber-optic Mach-Zehnder interferometer, incorporating the path X–Z (Fig. 1) into one of its arms. We applied a square modulation voltage to the PM, and observed interference fringes at the output port of the interferometer. We adjusted the voltage amplitude until it was causing no light modulation at the output port, indicating an exact 2π phase shift. From this, we found that $V_{\pi/2}(\lambda_l) = 5.7$ V. By the same method with the 1536 nm laser, we found $V_{\pi/2}(\lambda_s) = 3.35$ V.

From this measurement, we calculated $\theta_l \approx 0.294\pi < \theta_s$. The increased overlap between the two states $|\alpha\rangle$ and $|\beta\rangle$ with $|\alpha| = |\beta|$, as depicted in Fig. 2(b), would make discrimination between Bob's choices of φ_B more difficult. Eve can however increase the brightness of the injected Trojan-horse pulse: this would elicit a higher mean photon number in the back-reflection, effectively translating the states farther from the origin to diminish the overlap. The increment factor that makes the distance between the states at λ_l equal to that at λ_s is given by

$$\nu = \frac{|\alpha - \beta|^2 \text{ at } \lambda_s}{|\alpha - \beta|^2 \text{ at } \lambda_l} = \frac{1 - \cos\theta_s}{1 - \cos\theta_l} = 5.04, \tag{2}$$

implying that a mean photon number $\mu_{B\to E} \approx 20$ at λ_l would ensure a close-to-unity probability in the phase readout[11].

Decreased afterpulse probability. To quantify the decrease in the afterpulse probabilities in Bob's detectors, we used the setup shown in Fig. 1. A single THP was synchronized to the first in a series of detection gates[11,18] of Bob, and the times at which clicks occurred in the onward gates were then recorded. The delay of the THP relative to the first gate was adjusted such that the pulses going through Bob's long arm hit the detectors just a few nanoseconds after the gate was applied by Bob. Although we did utilize a polarization controller, only a maximum of ~45% of the incoming optical power at λ_l could be routed through the long arm. The remaining light, after having suffered propagation losses through the short arm, hit D0 and D1 around 50 ns *before* the first gate (propagation time through the short arm is ≈50 ns faster than the long arm in Clavis2[17]). These light pulses before the gate were found to be the dominant cause for increased noise in the detectors.

Figure 3 shows the time distribution of counts recorded in detector D0 at the wavelengths λ_s and λ_l. Each of the histograms was prepared by recording 10^6 counts. To make the most of the limited number of histogram bins in the counter (SR620), each bin was 0.4 μs wide and included counts from two consecutive gates. This allowed us to cover a time range of >80 μs. THPs with mean photon numbers $\mu_s = 2.68 \times 10^4$ and $\mu_l = 8.32 \times 10^7$ were used for wavelengths λ_s and λ_l respectively. Despite $\mu_s \ll \mu_l$, the data acquisition for the latter took much longer, indicating that most of the clicks were actually (thermal) dark counts. The number of counts per bin settled down at a constant value, representing dark counts, after ~40 μs (right half of the histogram). The total number of thermal dark counts collected could then be calculated by multiplying this value by the total number of bins in the entire histogram. All remaining counts could then be attributed to afterpulsing. Table 2 lists these counts at the two wavelengths. The afterpulse counts (*ApC*) make the bulk of the counts at λ_s, while dark counts (*DC*) are in the majority at λ_l.

It can also be observed in Fig. 3 that afterpulsing decay profile at both wavelengths is roughly similar, however the ratio of longer to shorter lifetime components is slightly larger at λ_l. Although this would help our modeled attack[11], for simplicity we have conservatively assumed that the decay parameters at λ_l are the same as at λ_s[18], aside from different overall afterpulse probability. The decay parameters and $Z^\star$ were measured at 1550 nm[11,18], which we consider to be close enough at our wavelength $\lambda_s = 1536$ nm.

To compute a numerical factor γ that compares the afterpulsing noise induced at the two wavelengths, we first take the ratio (*ApC*/*DC*) at each wavelength. Then, assuming the dark count probability per detector gate stayed constant between the two measurements, we take a ratio of these ratios. We assume a linear scaling of the afterpulse probability with the energy of the THP, and further normalise for the dissimilar mean photon numbers μ_s and μ_l of the THPs. The numerical factor is then

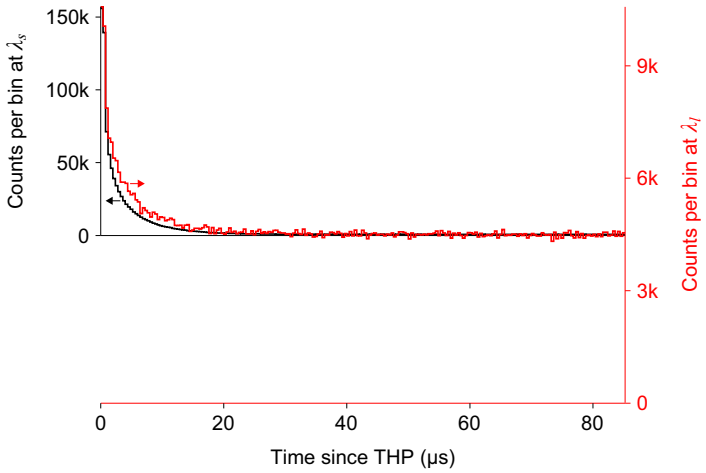

Figure 3. Afterpulse profiles at $\lambda_s = 1536$ nm and $\lambda_l = 1924$ nm. Note that the histograms are rescaled such that their peak counts and dark count rates match in the plot, making visual comparison of decay curves easy. The decay curves are similar but not identical. A total of 10^6 counts were histogrammed at each wavelength. The originally collected histogram data exhibited a saturation effect, in which count rate in later bins was slightly suppressed (by 6.4% for λ_s, 1.0% for λ_l) because of significant click probability in early bins. This has been corrected in the plotted histograms, increasing their total count number above 10^6.

λ (nm)	μ	*ApC*	*DC*
1536	2.68×10^4	867760	162854
1924	8.32×10^7	44981	962140

Table 2. Counts due to thermal dark noise (*DC*) and afterpulsing (*ApC*), extracted from Fig. 3 and corrected for the saturation effect. (*ApC* + *DC*) is greater than 10^6 owing to this correction.

$$\gamma = \frac{\mu_s}{\mu_l}\frac{(ApC_l/DC_l)}{(ApC_s/DC_s)} = 2.83 \times 10^{-6}. \tag{3}$$

In other words, a photon at λ_l is only 2.83×10^{-6} times as likely to cause an afterpulse as a photon at λ_s.

Attack modeling and discussion. Relative to λ_s, an attack at λ_l can thus effectively decrease the afterpulsing probability in D0 by

$$\delta_0 = \rho\nu\gamma = 1.03 \times 10^{-2}. \tag{4}$$

The factor $\rho\nu = 3.65 \times 10^3$ combines the results discussed previously on the aspects of increased attenuation and altered phase modulation, which required THPs injected into Bob at λ_l to be $\rho\nu$ times brighter than at λ_s to ensure optimal attack performance.

To calculate the afterpulsing probability for D1, one must also consider different losses from Bob's entrance to detectors D0 and D1 for the two attack paths (via the long arm at λ_s and short arm at λ_l, as shown in Fig. 1). We minimised $L_{X-Y}(\lambda_l)$ by adjusting input polarisation at X, then measured losses between X and the detectors through the short arm. $L_{X-C-D1}(\lambda_l)$ varied by a factor of 11 over the input polarization, while $L_{X-D0}(\lambda_l)$ unexpectedly was independent of the input polarization. Using the measured loss values (listed in the last two rows in Table 1), we calculate the effective decrease in the afterpulsing probability in D1

$$\begin{aligned}\delta_1 &= \delta_0 \times 10^{[L_{X-C-D1}(\lambda_s) - L_{X-D0}(\lambda_s) - L_{X-C-D1}(\lambda_l) + L_{X-D0}(\lambda_l)]/10} \\ &= 1.05 \times 10^{-3}.\end{aligned} \tag{5}$$

With afterpulsing amplitudes reduced by δ_0 and δ_1, we have repeated the simulation of the attack strategy proposed in ref.[11]. Let us first recap this strategy, in which Eve manipulates packets or 'frames'[13] of quantum signals traveling from Alice to Bob in the quantum channel. For instance, she may simply block the quantum signals for several contiguous time slots in a frame, thereby preventing any detection clicks (except those arising from dark counts) in Bob over a certain period of time. Conversely, she could substitute the quantum channel with a low-loss version to increase the detection probability in another group of slots. Such actions increase the efficacy of Eve's attack; they provide her some control over when inside a frame Bob's SPDs enter 'deadtime' – a period in which both D0 and D1 are insensitive to single photons and cannot register detection clicks. (In Clavis2, a 10 μs long deadtime is automatically triggered by a click in either of the detectors[18]). This is essentially done by

attacking in bursts, i.e., probing the phase modulator by sending bright THPs in a group of slots, thus making the SPDs enter deadtime as quickly as possible to let the afterpulses decay harmlessly and contribute as little as possible to the QBER. By balancing the usage of the low-loss line and the number of slots blocked per frame, Eve can also ensure that Bob does not notice any significant deviation of the observed detection rate (typically averaged over a large number of frames).

A numerical simulation modeling the above attack strategy during the operation of the QKD protocol is used to calculate Bob's incurred QBER Q and Eve's actual knowledge of the raw key I_E^{act}. This is performed for different attack combinations, i.e., by varying the number of slots that are blocked or simply passed via the low-loss line (with or without accompanying THPs). If for at least one combination, I_E^{act} exceeds the estimation I_E^{est} from the security proof but $Q < Q_{\text{abort}}$, the attack strategy is successful in breaching the security.

For an attack at λ_l, we have been able to find several such combinations for the given frame size of $N_f = 1075$ slots and a quantum channel transmittance $T = 0.25$. For instance, in one such combination, a total of 433 slots out of N_f are blocked by Eve. The remaining 642 slots pass from Alice to Bob via a low-loss line with transmittance $T_{LL} = 0.5$, and out of them only 334 slots–periodically distributed in 12 bursts of 28 slots each inside the frame–are accompanied by THPs to read the modulation. With this attack combination, we were able to obtain $I_E^{\text{act}} = 0.515 > I_E^{\text{est}} = 0.506$ (calculation based on Clavis2 parameters and the attack conditions[11]) and $Q = 7.8\% < Q_{\text{abort}} \approx 8\%$ (empirically determined in ref.[21]). We remark here that for a similar value of Q, the best optimized attacks at λ_s could not even yield $I_E^{\text{act}} \sim 0.080$. Furthermore, in contrast to the $T_{LL} = 0.9$ used in ref.[11], implementing the attack strategy with $T_{LL} = 0.5$ here makes the attack closer to be feasible in practice.

Note that in the simulation, we have mixed measurement results from two samples of Clavis2 system. The optical loss measurements at λ_l and the relative decrease in afterpulsing come from the system installed in Waterloo (Bob module serial number 08020F130), while the decay parameters of trap levels in avalanche photodiodes measured at λ_s come from the system in Erlangen (Bob module serial number 08008F130). The decay parameters and $Z^\star$ were measured at 1550 nm[11,18], which we consider to be close enough at our wavelength $\lambda_s = 1536$ nm. We further note that the latter figures vary significantly between D0 and D1, although the two avalanche photodiodes were of the same type and at the same temperature[18]. Therefore our simulation only gives a rough indication of attack performance. Results of the actual attack, if it is performed, will vary from sample to sample. However, also note that we have tested a single long wavelength of 1924; a different wavelength may well yield better attack performance. Finally, more recent commercial systems deploy SPDs with much better efficiencies and afterpulsing characteristics and, as noted in ref.[11], this benefits the eavesdropping strategy.

We expect homodyne detection at 1924 nm to be easy to implement by using p-i-n diodes with extended infrared response[22,23]. Based on the published specs, the latter should provide detection performance in our setting similar to that demonstrated at 1550 nm[11]. Separating Eve from Bob by some distance of fiber does not degrade the attack very fast; we have measured 7.5 dB/km loss at 1924 nm in a 16.5 cm diameter spool of Corning SMF-28e[24] fiber.

The easiest countermeasure to protect the QKD system from this attack is to properly filter the light entering the system[20,25]. E.g., adding a narrow-pass filter at Bob's entrance will force Eve to use the signal wavelength λ_s and reduce her attack performance to the original failure, provided poor detector afterpulsing properties are maintained in production[11]. Another countermeasure would be to use a QKD protocol that does not require the receiver's PM settings to be secret, such as BB84 with decoy states[3,10,26]. However, protecting the source's PM settings will still be required in most QKD protocols[25,27].

Conclusion

In conclusion, we have shown that despite the increased attenuation and sub-optimal phase modulation experienced around 1924 nm, the Trojan-horse attack performed at this wavelength has a very good chance of being invisible, because the afterpulsing experienced by Bob's detectors is extremely low. This attack is mostly implementable with commercial off-the-shelf components. Therefore, an urgent need exists to incorporate effective countermeasures into practical QKD systems to thwart such threats.

References

1. Bennett, C. H. & Brassard, G. Quantum cryptography: Public key distribution and coin tossing. In *Proc. IEEE International Conference on Computers, Systems, and Signal Processing (Bangalore, India)*, 175–179 (IEEE Press, New York, 1984).
2. Gisin, N., Ribordy, G., Tittel, W. & Zbinden, H. Quantum cryptography. *Rev. Mod. Phys.* **74**, 145–195 (2002).
3. Scarani, V. *et al.* The security of practical quantum key distribution. *Rev. Mod. Phys.* **81**, 1301–1350 (2009).
4. Makarov, V. Cracking quantum cryptography. In *CLEO/Europe and EQEC 2011 Conference Digest*, ED3_1 (Optical Society of America, 2011).
5. Scarani, V. & Kurtsiefer, C. The black paper of quantum cryptography: real implementation problems. *Theor. Comput. Sci.* **560**, 27–32 (2014).
6. Lo, H.-K., Curty, M. & Tamaki, K. Secure quantum key distribution. *Nat. Photonics* **8**, 595–604 (2014).
7. Jain, N. *et al.* Attacks on practical quantum key distribution systems (and how to prevent them). *Contemp. Phys.* **57**, 366–387 (2016).
8. Liang, L.-M., Sun, S.-H., Jiang, M.-S. & Li, C.-Y. Security analysis on some experimental quantum key distribution systems with imperfect optical and electrical devices. *Front. Phys.* **9**, 613–628 (2014).
9. Gisin, N., Fasel, S., Kraus, B., Zbinden, H. & Ribordy, G. Trojan-horse attacks on quantum-key-distribution systems. *Phys. Rev. A* **73**, 022320 (2006).
10. Vakhitov, A., Makarov, V. & Hjelme, D. R. Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography. *J. Mod. Opt.* **48**, 2023–2038 (2001).
11. Jain, N. *et al.* Trojan-horse attacks threaten the security of practical quantum cryptography. *New J. Phys.* **16**, 123030 (2014).
12. Scarani, V., Acín, A., Ribordy, G. & Gisin, N. Quantum cryptography protocols robust against photon number splitting attacks for weak laser pulse implementations. *Phys. Rev. Lett.* **92**, 057901 (2004).
13. Clavis2 specification sheet, http://www.idquantique.com/images/stories/PDF/clavis2-quantum-key-distribution/clavis2-specs.pdf, visited (16 Apr 2017).

14. Makarov, V., Anisimov, A. & Skaar, J. Effects of detector efficiency mismatch on security of quantum cryptosystems. *Phys. Rev. A* **74**, 022313 (2006). Erratum ibid. **78**, 019905 (2008).
15. Qi, B., Fung, C.-H. F., Lo, H.-K. & Ma, X. Time-shift attack in practical quantum cryptosystems. *Quant. Inf. Comp.* **7**, 73–82 (2007).
16. Lydersen, L. & Skaar, J. Security of quantum key distribution with bit and basis dependent detector flaws. *Quant. Inf. Comp.* **10**, 60–76 (2010).
17. Stucki, D., Gisin, N., Guinnard, O., Ribordy, G. & Zbinden, H. Quantum key distribution over 67 km with a plug&play system. *New J. Phys.* **4**, 41 (2002).
18. Wiechers, C. *et al.* After-gate attack on a quantum cryptosystem. *New J. Phys.* **13**, 013043 (2011).
19. Li, H.-W. *et al.* Attacking a practical quantum-key-distribution system with wavelength-dependent beam-splitter and multiwavelength sources. *Phys. Rev. A* **84**, 062308 (2011).
20. Jain, N. *et al.* Risk analysis of Trojan-horse attacks on practical quantum key distribution systems. *IEEE J. Sel. Top. Quantum Electron.* **21**, 6600710 (2015).
21. Jain, N. *et al.* Device calibration impacts security of quantum key distribution. *Phys. Rev. Lett.* **107**, 110501 (2011).
22. Extended InGaAs PIN photodiodes IG22-series, http://www.lasercomponents.com/us/product/ingaas-500-2600-nm-1/, visited (16 Apr 2017).
23. InGaAs PIN photodiodes G12182 series, http://www.hamamatsu.com/resources/pdf/ssd/g12182_series_kird1118e.pdf, visited (16 Apr 2017).
24. Corning SMF-28e optical fiber, http://www.princetel.com/datasheets/SMF28e.pdf, visited (16 Apr 2017).
25. Lucamarini, M. *et al.* Practical security bounds against the Trojan-horse attack in quantum key distribution. *Phys. Rev. X* **5**, 031030 (2015).
26. Hwang, W.-Y. Quantum key distribution with high loss: Toward global secure communication. *Phys. Rev. Lett.* **91**, 057901 (2003).
27. Sajeed, S. *et al.* Attacks exploiting deviation of mean photon number in quantum key distribution and coin tossing. *Phys. Rev. A* **91**, 032326 (2015).

Acknowledgements

We thank ID Quantique for cooperation, technical assistance, and providing us the QKD hardware. This work was funded by Industry Canada, CFI, NSERC (programs Discovery and CryptoWorks21), Ontario MRI, and the US Office of Naval Research. N.J. acknowledges the warm hospitality of the Institute for Quantum Computing, where this work was carried out.

Author Contributions

S.S. and C.M. performed the experiments. N.J. performed attack modeling and contributed to the experiments. V.M. supervised the study. All authors performed data analysis and contributed to writing the article.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Change History: A correction to this article has been published and is linked from the HTML version of this paper. The error has been fixed in the paper.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.