Gaps between industrial and academic solutions to implementation loopholes in QKD: testing random-detector-efficiency countermeasure in a commercial system

Anqi Huang,[1, 2, *] Shihan Sajeed,[1, 2] Poompong Chaiwongkhot,[1, 3]
Mathilde Soucarros,[4] Matthieu Legré,[4] and Vadim Makarov[1, 3, 2]

[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] *ID Quantique SA, Chemin de la Marbrerie 3, 1227 Carouge, Geneva, Switzerland*

(Dated: 5 January 2016)

In the last decade, efforts have been made to reconcile theoretical security with realistic imperfect implementations of quantum key distribution (QKD). However, in the process gaps have recently emerged between academic and industrial approaches to closing loopholes created by implementation imperfections. In academic research labs, many practical security problems appear to be reliably solved, in principle, by advanced schemes and protocols. Meanwhile the industry prefers practical and easier solutions, even without security verification in some cases. In this paper, we present a concrete example of ID Quantique's random-detector-efficiency countermeasure against detector blinding attacks. As a third-party tester, we have found that the first industrial implementation of this countermeasure is effective against the original blinding attack, but not immune to a modified blinding attack. Moreover, we show experimentally that the full countermeasure as in academic proposal [C. C. W. Lim *et al.*, IEEE J. Sel. Top. Quantum Electron. **21**, 6601305 (2015)] is still vulnerable against the modified blinding attack. Our testing results show several specific disparities between the industrial practical solution and the academic perfect solutions. Our work illustrates that forming an implementation-and-testing closed loop is necessary to bridge the gaps and improve the practical security of QKD systems.

I. INTRODUCTION

Currently, applied cryptography systems rely on the hardness of certain mathematical assumptions, which only provides computational security [1, 2]. Once an eavesdropper has enough computing power, such as a quantum computer, the security of these classical encryption algorithms will be broken [3, 4]. However, quantum key distribution (QKD) allows two parties, Alice and Bob, to share a secret key based on the laws of quantum mechanics [5–8]. Because of no-cloning theorem [9], an eavesdropper with arbitrary computing power cannot copy the information sent by Alice without leaving any trace, which guarantees the unconditional security of communication [10–15].

For this gradually maturing technology, practical QKD systems have been realised in laboratories [16–19] and several companies have provided commercial QKD systems to general customers [20]. However, imperfect components used in the implementations lead to security issues that have attracted an increasing attention in the last decade [21–30]. Since increasing number of quantum attacks have been demonstrated, academic community is already aware of the security threat from practical loopholes. Therefore, the next step is to come up with loophole-free countermeasures.

Unfortunately, while the entire community is trying to bridge the gap between theoretical and practical security, another issue is impending. Gaps have emerged between the academic and industrial countermeasures. The first gap is between the ideal proposals and the practical technology. The academia prefers reliable device-independent QKD protocols that eliminate most of the side channels by design [31, 32]. For example, for an important set of attacks that exploit various imperfections in single-photon detectors [22, 24–26, 28, 30, 33, 34], a measurement-device-independent QKD protocol eliminates the detectors from the trusted part of the system and thus makes it independent of all detector imperfections [32]. However, the industry currently attempts to patch loopholes in existing protocols [35–37], because the measurement-device-independent protocol is too challenging to implement into a customer-friendly product with stable performance [38].

These patches for existing protocols seem to be practical and realizable. However, the second gap is exposed when engineers implement these countermeasures into actual commercial QKD systems. Limitations of hardware do not always fully allow the original countermeasure, which forces companies to further simplify the countermeasure to patch existing systems. Moreover, engineers sometimes make implementation mistakes, owing to an incomplete understanding of the academic countermeasure and practical security of QKD in general. These factors lead to disparity between the final implementations and the original proposals.

Aside from the technical difficulties, a more critical gap is the mindset. Ideally, the industry should be able to implement robust fixes to the security problems in commercial products, and get them tested either by themselves

[*] anqi.huang@uwaterloo.ca

or by an independent organization to verify the security. However, the industry is not quite there yet. A few companies (such as ID Quantique and SeQureNet) are trying to break the mold by letting third-party labs examine their solutions to previously discovered loopholes [39, 40]. Meanwhile, it seems that most of the industry has their heads firmly planted into sand, and does not allow independent testers to examine their 'countermeasure implementations' hands-on. As a result, the existing gaps between the academia and the QKD industry are increasing day by day.

In this paper, the current scenario is illustrated with an example. We examine ID Quantique's attempted countermeasure to earlier discovered bright-light detector control attacks [26, 33, 34] that were demonstrated 5 years ago on ID Quantique's and MagiQ Technologies' QKD products. The countermeasure is to randomly remove some detector gates to force the effective detection efficiency to zero during those slots. The idea is that when an eavesdropper is performing the blinding attack, she will produce click during these removed gates thus get caught. Our experimental results show that although this countermeasure is effective against the original detector blinding attack [26], it is no longer effective if the eavesdropper modifies her attack slightly. We note here that this countermeasure implemented by ID Quantique is a simplification of the original countermeasure proposal [37]. Hence, we have gone further ahead and manually implemented a full version of the countermeasure using two non-zero detection efficiency levels as proposed in [37], and tested it. Our testing shows that even the full countermeasure is vulnerable to the modified blinding attack. The countermeasure is based on the assumption that the detection probability under blinding attack cannot be proportional to the photon detection efficiency, which we disprove experimentally.

The paper is organized as follows. Section II reviews a hacking-and-patching timeline of ID Quantique's Clavis2 QKD system and introduces the countermeasure. In Section III, testing results of ID Quantique's first countermeasure implementation are reported and our modified blinding attack is introduced. Section IV theoretically analyses conditions of a successful attack and shows that the modified blinding attack satisfies them. Moreover, in Section V, based on certain assumptions about a future implementation of the full countermeasure [37], we demonstrate two possible methods to hack this full version implementation. We discuss the practicality of our attacks against installed commercial QKD lines in Section VI and conclude in Section VII.

II. FROM LOOPHOLE DISCOVERY TO COUNTERMEASURE IMPLEMENTATION

In 2009, the vulnerability of the commercial QKD system Clavis2 [41] to detector blinding attacks was identified and a confidential report was submitted to ID Quantique (the work was published shortly afterwards [26]). After this, ID Quantique has been trying to figure out an experimental countermeasure against these attacks. The timeline of this security problem is shown in Fig. 1. In 2010, ID Quantique proposed a countermeasure that randomizes the efficiency of a gated avalanche photodiode (APD) by randomly choosing one out of two different gate voltages, and filed this idea for a patent [42]. In this way, an eavesdropper Eve does not know the exact efficiency of Bob in every gated slot and thus cannot maintain his detection statistics. At the sifting phase, if the observed detection rates differ from the expected values, Alice and Bob would be aware of Eve's presence and discard their raw keys.

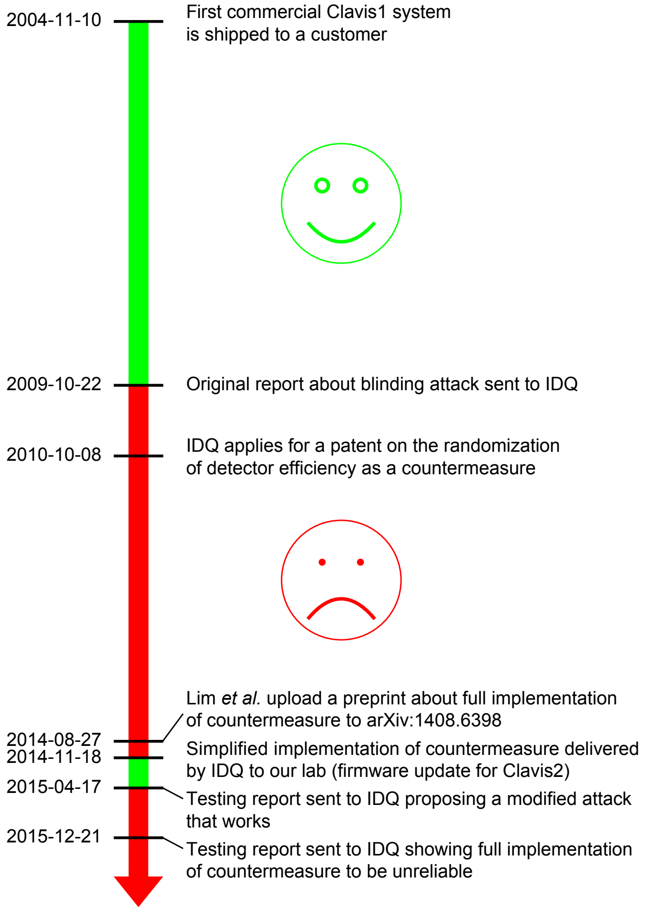
2004-11-10
First commercial Clavis1 system is shipped to a customer
2009-10-22
Original report about blinding attack sent to IDQ
2010-10-08
IDQ applies for a patent on the randomization of detector efficiency as a countermeasure
Lim *et al.* upload a preprint about full implementation of countermeasure to arXiv:1408.6398
2014-08-27
2014-11-18
Simplified implementation of countermeasure delivered by IDQ to our lab (firmware update for Clavis2)
2015-04-17
Testing report sent to IDQ proposing a modified attack that works
2015-12-21
Testing report sent to IDQ showing full implementation of countermeasure to be unreliable

FIG. 1. Timeline of hacking-countermeasure-hacking for the bright-light detector control class of attacks.

In 2014, Lim *et al.* proposed a specific protocol to realize this countermeasure [37], which takes blinding attack into account and analyses the security mathematically. In the protocol, Bob randomly applies two non-zero detection efficiencies $\eta_1 > \eta_2 > 0$, and measures detection rates R_1 and R_2 conditioned on these efficiencies. The effect of detector blinding attack is accounted via the factor $(\eta_1 R_2 - \eta_2 R_1) / (\eta_1 - \eta_2)$. Without the blinding attack, the detection rate is proportional to the efficiency, making this factor zero. Under attack, the factor will

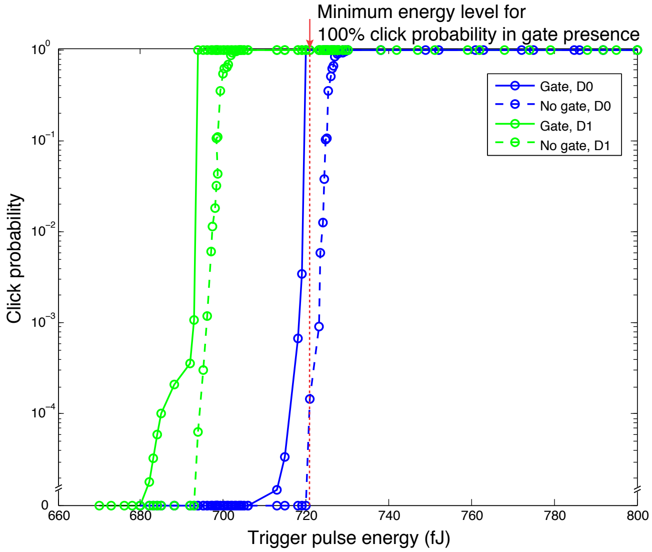

FIG. 2. Click probability under original blinding attack [26] versus energy of trigger pulse. The blinding power is 1.08 mW, as the same as the power used in the published original attack [26]. The timing of trigger pulse is 0.7 ns long, 3 ns after the centre of the gate signal, which should roughly reproduce the original attack [26].

be greater than zero, and reduces the secure key rate. This solution intends to introduce an information gap between Eve and Bob, for Eve has no information about Bob's random efficiency choice.

Later in 2014, ID Quantique implemented the countermeasure as a firmware patch. The hardware in Clavis2 is not capable of generating two nonzero efficiency levels that switch randomly between adjacent detector gates. As a result, implementation is in a simplified form by suppressing gates randomly with 2% probability. The suppressed gates represent zero efficiency $\eta_2 = 0$, while the rest of the gates represent calibrated efficiency $\eta_1 = \eta$. Ideally, in the updated system, there should be no click in the absence of the gate. In practice, transient electromagnetic interference may extremely infrequently lead to a click without a gate. Therefore, an alarm counter is used with the system lifetime limit of 15 clicks in the absence of the gate. If this limit is reached, it triggers the firmware to brick the system and require factory maintenance. In case of blinding attack [26], click probability should not depend on the gate voltage and the attack should therefore cause clicks at the slots of gate absence.

III. TESTING THE COUNTERMEASURE

In this section, we demonstrate that the countermeasure presently implemented by ID Quantique is effective against the original blinding attack [26], but not sufficient against the general class of attacks attempting to take control of Bob's single-photon detectors.

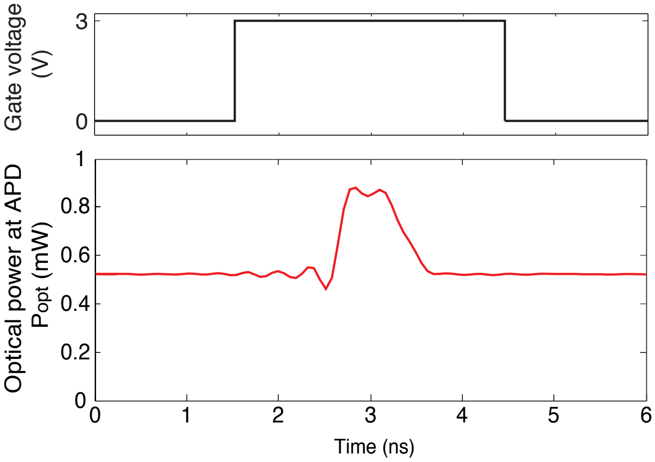

FIG. 3. Idealized APD gate signal and real oscillogram of optical trigger pulse. Relative time between the gate voltage transitions and the optical pulse is approximate. The c.w. signal is generated by a 1536 nm laser diode; the trigger pulse signal is obtained by modulating pump current of a separate 1551 nm laser diode, using an electrical pulse generator [26].

For the original blinding attack, Eve sends bright-light continuous-wave (c.w.) laser light to blind Bob's detectors. Then a trigger pulse is sent slightly after the gate to make a click. We repeat this attack for improved Clavis2 system and test the amount of energy to trigger a click which is shown in Fig. 2. From Fig. 2, we can see the trigger pulse energy for gate presence (solid curves) is lower than that for gate absence (dashed curves), because minute electrical fluctuations of APD voltage following the gate signal lower the click threshold slightly.

However, if Eve tries to trigger a click with 100% probability when the gate is applied, this amount of trigger pulse energy (marked by a dotted vertical line in Fig. 2) also might trigger a click with non-zero probability when the gate is suppressed, which is monitored and results in an alarm. Therefore, Eve cannot hack the system with full controllability. To avoid clicks in slots of gate suppression, Eve could in theory decrease the level of trigger pulse energy to trigger a click sometimes with gate presence, but never with gate absence. This also satisfies a necessary condition of a successful attack which we will discuss in Section IV later. Unfortunately, in practice, our testing result shows the amount of trigger pulse energy required to trigger D0 without the gate is about 710 fJ, which is only 1.5% less than the amount of energy for 100% click (720 fJ) when the gate is present. The 1.5% difference of these two energy levels is likely not big enough to achieve a reliable attack operation that avoids triggering the countermeasure. Also, D1 will always trigger at these energy levels, revealing the attack. Eve could target D1 using a slightly lower energy level, but the relative precision required is similar there. Routine fluctuations of temperature and other equipment parameters may lead to some instability of these trigger pulse energy

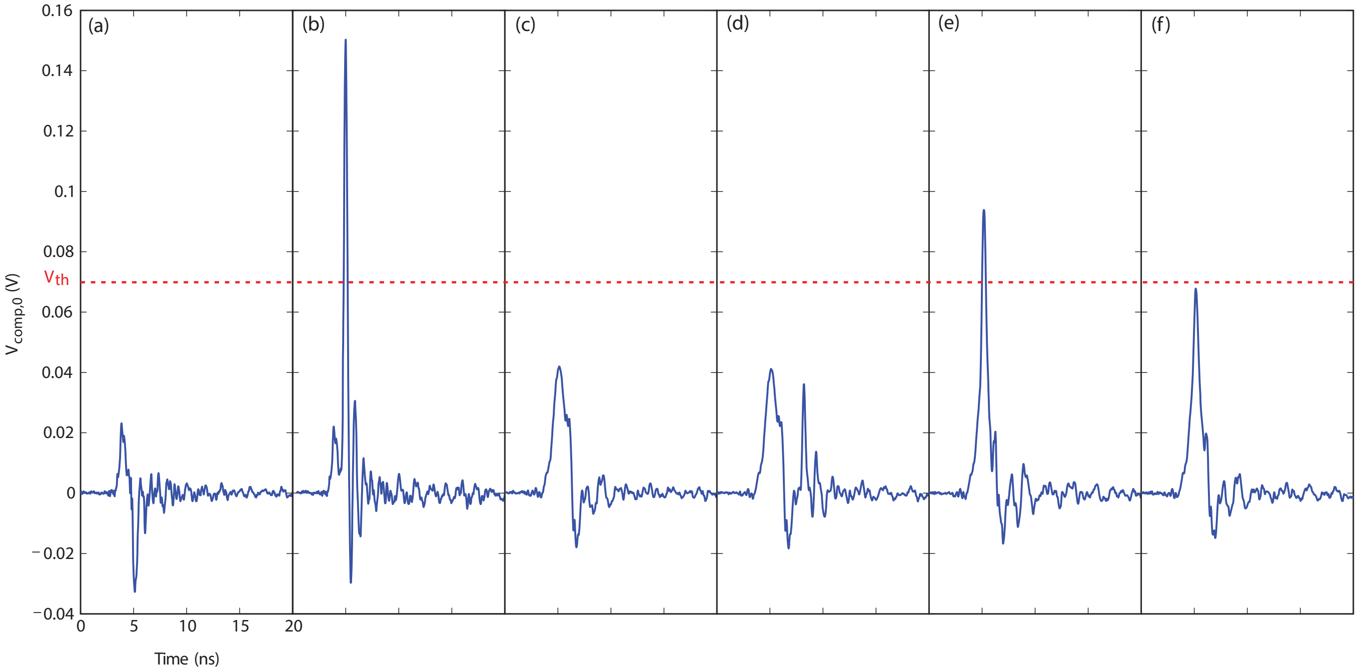

FIG. 4. Oscillograms at comparator input in the detector circuit, proportional to APD current. (a) Geiger mode. The small positive and negative pulses are due to gate signal leakage through the APD capacitance of ~ 1 pF. (b) Geiger mode, single-photon avalanche. (c–f) The detector is blinded with 0.56 mW c.w. illumination, with (c) no trigger pulse applied, (d) 0.32 pJ trigger pulse applied 5 ns after the gate, (e) 0.32 pJ trigger pulse applied in the gate, and (f) 0.16 pJ trigger pulse applied in the gate.

levels, causing a risk for Eve to trigger a few clicks in the gate absence and brick the system being attacked. From this point of view, we think this first implementation of countermeasure is effective against the original blinding attack.

We can slightly modify our blinding attack to break the security of this countermeasure. Similarly to the original blinding attack, Bob's detectors are blinded by a bright-light laser first. Then, instead of sending a trigger pulse slightly after the gate as in the original attacks [26], we send a 0.7 ns long trigger pulse on top of the c.w. illumination *during the detector gate*, as shown in Fig. 3. This trigger pulse produces a click in one of Bob's two detectors only if Bob applies the gate and his basis choice matches that of Eve; otherwise there is no click.

To explain why this modified attack succeeds, let us remind the reader the normal operation of an avalanche photodiode (APD). The detectors in Clavis2 are gated APDs. When the gate signal is applied, the voltage across the APD V_{APD} is greater than its breakdown voltage V_{br}. If a single photon comes during the gated time, an avalanche happens and causes a large current. This current is converted into a voltage by the detector electronic circuit. If the peak voltage is larger than a threshold $V_{\mathrm{th}} = 70$ mV, the detector registers a photon detection (a 'click'). Fig. 4(a) and (b) show the cases of no photon coming and a photon introducing an avalanche. Appendix A explains more details of the detector operation principle and the blinding attack.

A bright laser is able to blind the APDs. Under c.w. illumination, the APD produces constant photocurrent that overloads the high-voltage supply and lowers V_{APD}. Then, even when the gate signal is applied, V_{APD} does not exceed V_{br} and the APD remains in the linear mode as a classical photodetector that is no longer sensitive to single photons. This means the detectors become blinded.

Under the blinding attack, Fig. 4(c–e) shows the detector voltages in different cases: when no trigger pulse is applied and when the trigger pulse is applied either after or in the gate. Since in the linear mode the gain factor of secondary electron-hole pairs generation in the APD depends on the voltage across it, the 3 V gate applied to the APD increases the gain factor. This larger gain during the gated time assists the APD in generating a larger photocurrent than the photocurrent outside the gate. Therefore the gate signal causes a positive pulse as shown in Fig. 4(c). The trigger pulse applied after the gate produces a second pulse, but the peak voltages of neither pulses exceed V_{th} [Fig. 4(d)]. However, when the trigger pulse is shifted inside the gate, the two pulse amplitudes add up, reach V_{th} and produce a detector click [Fig. 4(e)]. If Bob chooses a different measurement basis than Eve, only half of the trigger pulse energy arrives at each detector [26]. In this case, the peak voltage does not reach V_{th} [Fig. 4(f)]. Overall, only when the trigger pulse is applied during the gate time and Bob chooses the same basis as Eve, the detector under the blinding

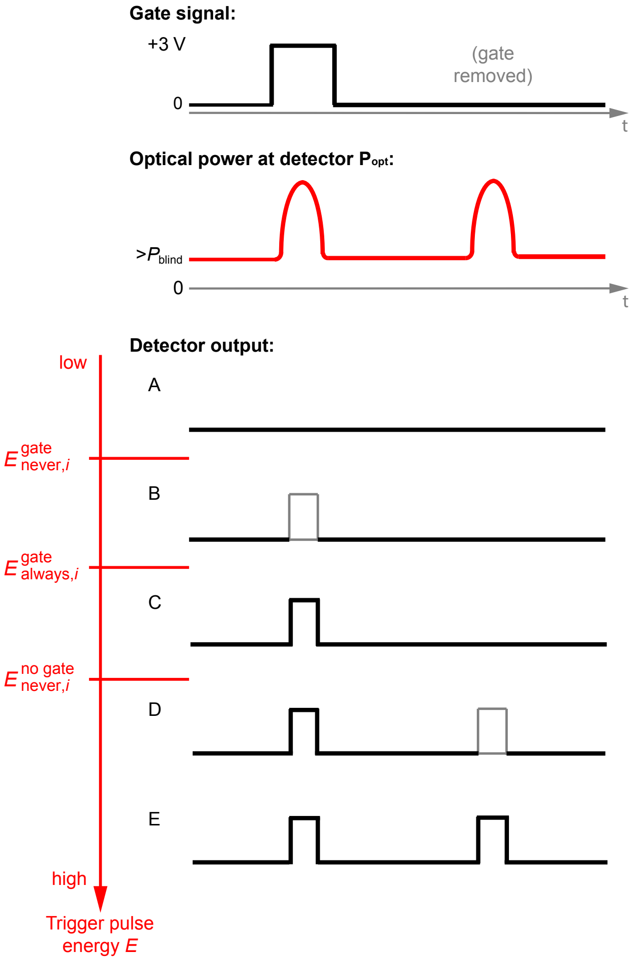
Gate signal:
+3 V
0
(gate removed)
t
Optical power at detector P_{opt}:
>P_{blind}
0
t
Detector output:
low
A
$E^{\text{gate}}_{\text{never},i}$
B
$E^{\text{gate}}_{\text{always},i}$
C
$E^{\text{no gate}}_{\text{never},i}$
D
E
high
Trigger pulse energy E

FIG. 5. Output of a blinded detector in Clavis2 under control of trigger pulses of different energy. The top graph shows a gate applied at the first slot, but suppressed at the second slot. However, an optical trigger pulse is sent to the detector in both slots. Graphs A–E show detector output versus trigger pulse energy E. In graph A, the energy is insufficient to produce a click. As the energy is increased above $E^{\text{gate}}_{\text{never},i}$, clicks intermittently appear in the presence of the gate, as shown in graph B. At the energy level above $E^{\text{gate}}_{\text{always},i}$, the gate always has a click, as shown in graph C. However, there is never a click when there is no gate. At a higher energy level above $E^{\text{no gate}}_{\text{never},i}$, clicks in the gate absence appear intermittently (graph D) or always (graph E).

attack clicks. As a result, Eve can control Bob's detectors to make Bob obtain the same measurement result as her, and does not introduce extra errors [26].

Contrary to most of previously demonstrated attacks attempting to take control of single-photon detectors [26, 28, 33], in the present demonstration the timing of the trigger pulse has to be aligned with the gate. Besides timing alignment, another important factor of the attack is the trigger pulse energy E. To test the effect of different trigger pulse energy, we gradually increase it and observe the detection outcomes. Figure 5 shows schematically in which order clicks appear in Clavis2 as E is increased.

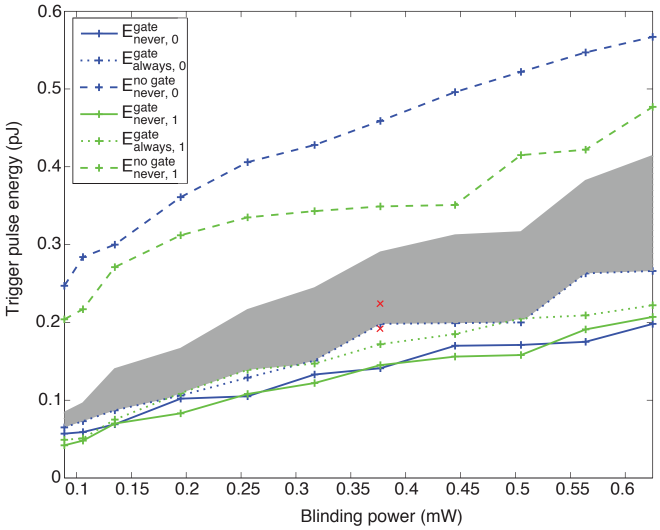

FIG. 6. Energy thresholds of trigger pulse versus c.w. blinding power. Shaded area shows the range of trigger pulse energies of the perfect attack.

We observe three thresholds.

- If $E \leq E^{\text{gate}}_{\text{never},i}$ (where $i \in \{0, 1\}$ is detector number), the detector never clicks when the gate is applied.
- If $E \geq E^{\text{gate}}_{\text{always},i}$, the detector always clicks when the gate is applied.
- If $E \leq E^{\text{no gate}}_{\text{never},i}$, the detector never clicks when the gate is suppressed.

Figure 6 shows these detection thresholds measured for a range of c.w. blinding powers. All the thresholds rise with the blinding power, because higher blinding power leads to a larger photocurrent and lower V_{APD}. The decreased V_{APD} leads to smaller gain and thus lower sensitivity to the trigger pulse. (Appendix B contains a more detailed investigation of the processes inside the detector.) As can be seen, for any given blinding power, $E^{\text{no gate}}_{\text{never},i}$ is much higher than the other click thresholds. This easily allows the original detector control attack [26] to proceed undetected by the countermeasure. A more formal analysis will be stated in the next section.

IV. CONDITIONS OF A SUCCESSFUL ATTACK

Experimental result of the previous section shows that the attack of Ref. 26 is possible in Clavis2. However, general conditions for a successful attack should be analysed theoretically. In this section, we first consider *strong conditions* for a perfect attack, in which Eve induces a click in Bob with 100% probability if their bases match and the gate is applied, and 0% probability otherwise. These conditions are definitely sufficient for a successful attack [26]. However, as we remark later in this section, even if these strong conditions are not satisfied, an attack may still be possible.

Strong conditions. If the detection outcome varies as Fig. 5 with the increase of trigger pulse energy, the order of the three thresholds is:

$$E^{\text{no gate}}_{\text{never},i} > E^{\text{gate}}_{\text{always},i} > E^{\text{gate}}_{\text{never},i}. \tag{1}$$

If Eve and Bob select opposite bases, half of the energy of trigger pulse goes to each Bob's detector. In this case, none of the detectors should click despite the gate presence. This is achieved if [26]

$$\frac{1}{2}\max_i \left\{E^{\text{gate}}_{\text{always},i}\right\} < \left(\min_i \left\{E^{\text{gate}}_{\text{never},i}\right\}\right). \tag{2}$$

The random gate suppression imposes additional conditions. In case of basis mismatch, half of the trigger pulse energy is arriving at the wrong detector. It should not induce a click when the gate signal is absent, which is achieved if

$$\frac{1}{2}E^{\text{gate}}_{\text{always},i} < E^{\text{no gate}}_{\text{never},i\oplus 1}. \tag{3}$$

If the bases match, we need to make sure there is no click when the gate is suppressed, but always a click in the expected detector in the gate presence. This is achieved if $E^{\text{gate}}_{\text{always},i} < E^{\text{no gate}}_{\text{never},i}$, which is already included in inequality (1). Although inequality (3) has a physical meaning, it mathematically follows from inequalities (1) and (2). Thus satisfying inequalities (1) and (2) represents the strong attack conditions and guarantees the same performance as in Ref. 26. The shaded area in Fig. 6 indicates a range of the trigger pulse energies Eve can apply for the perfect attack. The range is sufficiently wide to allow for a robust implementation, only requiring Eve to set correct energy with about ±15% precision.

Necessary condition. An attack may still be possible even if Eve's trigger pulse does not always cause a click in Bob when their bases match, and/or sometimes causes a click when their bases do not match [43]. The latter introduces some additional QBER but as long as it's below the protocol abort threshold, Alice and Bob may still produce key. The random gate removal countermeasure imposes the condition

$$E^{\text{no gate}}_{\text{never},i} > E^{\text{gate}}_{\text{never},i}, \tag{4}$$

which means Eve should be able to at least sometimes cause a click in the gate while never causing a click without the gate (lest the alarm counter is increased). This is a necessary condition for an attack. As the present paper details, there are strong engineering reasons why this condition is likely to be satisfied in a detector. Additional conditions will depend on exact system characteristics [43].

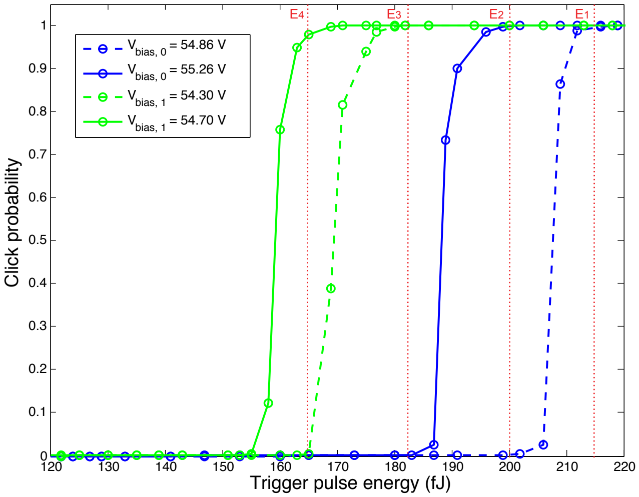

FIG. 7. Click probabilities under blinding attack versus energy of trigger pulse. Solid curves show the energy of trigger pulse for original V_{bias}, while dashed curves for reduced V_{bias} lowering photon detection efficiency by about a factor of 2. The blinding power is 0.38 mW and the timing of trigger pulse is aligned in the middle of the gate by minimizing its energy required to make a click.

V. WILL A FULL IMPLEMENTATION OF THE COUNTERMEASURE BE ROBUST?

We have proved so far that the current countermeasure with gate suppression cannot defeat the detector blinding attack. However, Lim's paper claims that the full version of countermeasure with two non-zero detection efficiencies is effective against detector side-channel attacks [37]. Even though this full countermeasure has not been implemented by ID Quantique, we have tested some properties of the detectors in Clavis2 to show two possible methods to hack the full countermeasure, based on certain assumptions about a future implementation.

Bob could choose randomly between $P/2$ and P detection efficiency by changing either gate voltage amplitude V_{gate} or high-voltage supply V_{bias} [37]. Since in Clavis2 hardware V_{gate} is fixed (see Appendix A), we assume an engineer will change V_{bias} to achieve different non-zero detection efficiencies. To achieve half of original detection efficiency, we lower V_{bias} manually. When $V_{\text{bias},0}$ of D0 drops from −55.26 V to −54.86 V, the detection efficiency P_0 reduces from 22.6% to 12.8%. Similarly, we decrease $V_{\text{bias},1}$ of D1 from −54.70 V to −54.40 V, leading to the detection efficiency P_1 reduction from 18.9% to 9.7%. After that, we test Eve's controllability of these two detectors.

First, we blind the detectors and then measure the relation between the energy of trigger pulse and probability to cause a click. The position of trigger pulse is fixed in the middle of gate signal. Figure 7 shows the testing result which indicates there is a transition range between 0% and 100% click probability.

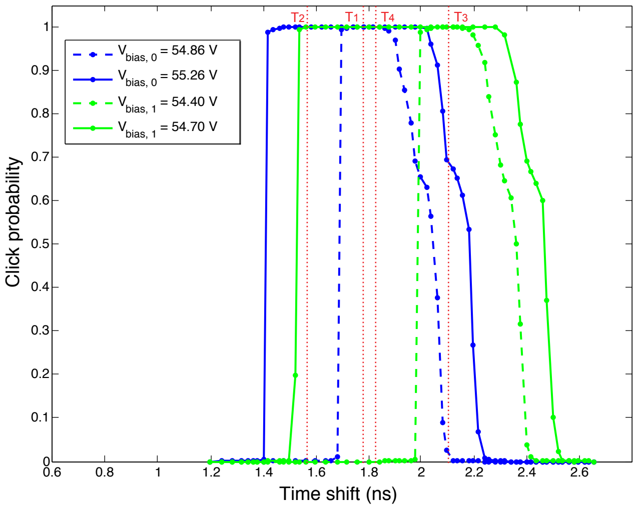

FIG. 8. Click probabilities under blinding attack versus relative time shift of trigger pulse. Solid curves give the detection probability at the original V_{bias}, and dashed curves give the detection probability at lower V_{bias}. Note that the latter extends over a relatively narrower time window. The blinding power is 0.38 mW. The energy of trigger pulse for D0 is 0.22 pJ and for D1 is 0.19 pJ. These energy levels are marked as red × in Fig. 6.

From the measurement result, Eve can randomly select different levels of trigger pulse energy (shown as dotted lines in Fig. 7) to attack the full version of countermeasure. As we know, only when Bob chooses the same measurement basis as Eve, all the energy of trigger pulse arrives targeted detector and achieves a click. For target D0, if trigger pulse energy E_1 is chosen, D0 always clicks, while at E_2, the detector only clicks if higher V_{bias} is applied. When E_1 and E_2 are chosen randomly with the same probability $P_0/2$, the detection probability for higher V_{bias} is P_0 and the detection probability for lower V_{bias} is only $P_0/2$. Therefore, the attack reproduces correct detection probabilities as the protocol requires. Similarly, for target D1, Eve can choose E_3 to trigger click always and choose E_4 to get a click only if higher V_{bias} is applies. This reproduces correct detection probabilities, $P_1/2$ and P_1. At the same time, E_1 and E_3 remain safely below $E^{\text{no gate}}_{\text{never } 0,1}$. This allow Eve to hack the countermeasure tracelessly.

Second, we test the correlation between time shift of trigger pulse and click probability of blinded detector. The trigger pulse energy we use in this test for D1 is slightly lower than that of D0, but both levels of energy are above $E^{\text{gate}}_{\text{always, 0 or 1}}$ in Fig. 6 marked as red ×. The measurement result is shown in Fig. 8.

This testing result illustrates another method to attack the countermeasure: randomly adjusting the time shift of the trigger pulse. For D0, after fixing the suitable energy level of trigger pulse, Eve can always trigger a click by choosing time shift T_1, but only trigger a click at higher V_{bias} by choosing T_2. Similarly, if target detector is D1, the detector always clicks at T_3, but only clicks at higher V_{bias} at T_4. Then, when Eve sends trigger pulse to control D0, she randomly selects T_1 and T_2 with equal probability $P_0/2$ to reproduce the correct detection efficiencies of D0. Eve utilizes the same strategy for D1 to achieve correct detection probabilities, $P_1/2$ and P_1. In this way, Eve also hacks Clavis2 system tracelessly.

From the above testing and analysis of the implementation that changes V_{bias}, we can guess that an alternative implementation that changes V_{gate} may leave a similar loophole. The reason for this practical loophole is a wrong assumption made in Lim's paper [37]. They assume Eve cannot generate faked states that trigger detections with probabilities that are *proportional* to the original photon detection efficiency. Here we have proved this is in fact possible. Therefore, the model of a practical detector should be more precise in security analysis, if one wishes to close the detector control loophole without resorting to measurement-device-independent QKD.

VI. OUR ATTACKS IN A BLACK-BOX SETTING

According to Kerckhoffs' principle [44], Eve always knows everything about the algorithms and hardware of Alice's and Bob's boxes, including the precise values of equipment parameters. The classical security community practices Kerckhoffs' principle since 1970's, and widely agrees that this is a good approach to implementation security [1]. This is supported by many examples of cryptographic systems that did not follow this principle and were compromised [45]. The quantum academic community certainly agrees that QKD should be made secure in this setting, which is necessary for QKD being unconditionally secure [10–15].

However, it is also a practically interesting question if any proposed attack can be mounted on today's commercial QKD systems in a black-box setting, when Eve only has access to the public communication lines but cannot directly measure signals and values of analog parameters inside Alice's and Bob's boxes [46]. In this realistic scenario, Eve may purchase (or acquire by other means) a sample of the system hardware, open it, make internal measurements and rehearse her attacks on it. Then she has to eavesdrop on her actual target, an installed system sample in which she has not had physical access to the boxes. Although the latter sample can be of the same model and design, it will generally have different values of internal analog parameters, owing to sample-to-sample variation in system components. A full implementation of our attacks in this scenario remains to be tested. In this setting it will be of utmost importance for Eve to avoid triggering clicks in the absence of the gate, because this would very quickly brick the system and risk revealing her attack attempt. The original blinding attack that applies the trigger after the gate becomes very sensitive to precise values of thresholds in the presence of the first version of countermeasure (Fig. 2). For this reason we think the countermeasure will likely be triggered by the original attack in the realistic black-box setting.

Our modified attack that applies the trigger inside the gate will likely avoid triggering the alarm, because the no-gate threshold energies are much higher that the energies required for detector control (Fig. 6). It also tolerates some fluctuation in experimental parameters for detector control. For example, when Eve applies 0.38 mW blinding power, 252 fJ trigger pulse energy, and times her trigger pulse at the middle of the gate, we have verified that the attack still works perfectly for up to ±21% change in the trigger energy (see Fig. 6) or up to ±1.3 ns change in the trigger timing. This makes it robust against reasonably expected fluctuations and imprecision of the system parameters. In particular, the timing accuracy required for our attack in much coarser than the several tens of picoseconds precision Alice and Bob use in normal operation [47]. The trigger energy setting precision is similar to the original attack that required ±16% [26].

Eve may need a few attempts to set a correct trigger energy when attacking a new copy of the system. She can do this by starting at a low trigger energy and attempting several increasing values of energy while watching the classical traffic Alice-Bob for the success or failure of the QKD session she has attacked [48]. A QKD session that fails because of too low detection efficiency is a naturally occurring event that is part of normal system operation, does not raise an alarm and is recovered from automatically in Clavis2 [47, 49].

A full two-level implementation of the countermeasure may require Eve to run more attempts, because of a finer degree of control required over the trigger pulse energy and timing. Yet, similarly to the first countermeasure implementation, the no-gate trigger energy that would raise alarm remains safely well above the energies required for detector control. The practicality of attack in the black-box setting is thus difficult to predict without having the actual industrial implementation of the full countermeasure, and actually demonstrating the full attack, which can be a future study.

VII. CONCLUSION

We have tested the first implementation of the countermeasure against the blinding attack in the commercial QKD system Clavis2. Our testing result demonstrates that presently implemented countermeasure is effective against the original blinding attack but not effective against a modified blinding attack. The modified attack fully controls Bob's single-photon detectors but does not trigger the security alarm. The modified attack is similar to the original detector blinding attack [26] with the only difference that the trigger pulses are time-aligned to coincide with the detector gates, instead of following it. We argue that this attack should be implementable in practice against an installed QKD communication line where Eve does not have physical access to characterising Alice and Bob, however such full demonstration has not yet been done, to our knowledge.

We have also tested the full proposed implementation of countermeasure with two non-zero efficiency levels, and found its security to be unreliable despite predictions of the theory proposal [37]. From the current testing results, bright-pulse triggering probabilities of the blinded detectors depend on several factors including V_{bias}, timing and energy of the trigger pulse (see Section V). This in principle allows Eve to compromise the full countermeasure implementation.

The improvement of Clavis illustrates a development process of QKD implementation in which the work of implementers and that of testers were widely separated in time. When the first generation commercial QKD system, Clavis1, was implemented, this system design was considered to be secure. However, several years later practical detector loopholes were discovered and the original blinding attack broke the system security. Even though patching these loopholes clearly required a massive replacement of the system concept and hardware, the company and researchers still tried and spent several more years proposing and implementing the academic software-only countermeasure. Then, as the third party, we again evaluate the practical security of the updated system. According to our testing result, this countermeasure is not as reliable as would be expected in a high-security environment of QKD. Although an ideal industrial countermeasure has not been achieved, everybody now has a more clear concept about the detector loopholes. This procedure emphasizes the necessity of security testing every time practical QKD systems are developed or updated. We only can reach the final practical security of any QKD system after several iterations of implementation development and testing verification.

Our work shows gaps between academia and industry. The academic community proposed a perfect solution, measurement-device-independent QKD protocol, to remove all detector loopholes, but the industry pursued an easier solution based on the existing system hardware. Our countermeasure testing illustrates that patching a loophole is still time-consuming and difficult. However, addressing practical vulnerabilities at the design stage of a QKD system is both cheaper and less messy than trying to retrofit patches on an existing deployed solution. Addressing security at the design stage should be the goal whenever possible.

ACKNOWLEDGMENTS

We thank C. C. W. Lim for discussions. This work was supported by Industry Canada, CFI, Ontario MRI, NSERC, CryptoWorks21, US Office of Naval Research, ID Quantique, European Commission's FET QICT SIQS and EMPIR 14IND05 MIQC2 projects. P.C. was supported from Thai DPST scholarship.

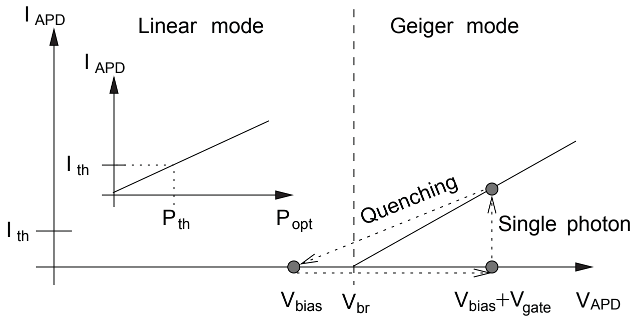

FIG. 9. Linear-mode and Geiger-mode APD operation (reprinted from [26]).

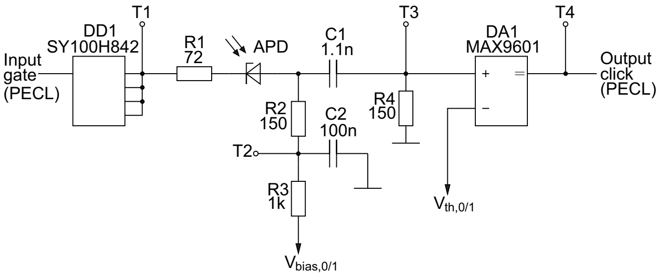

FIG. 10. Equivalent detector bias and comparator circuit, as implemented in Clavis2 (reprinted from [26]).

Appendix A: Background

In this section, we recap the operating principle of the single-photon detector, its implementation in Clavis2, and the original blinding attack [26]. Most available single-photon detectors are APDs operating in Geiger mode, in which they are sensitive to single photons [50]. As shown in Fig. 9, when the APD is reverse-biased above its breakdown voltage V_{br}, a single photon can cause a large current I_{APD}. If this current exceeds the threshold I_{th}, electronics registers this as a photon detection (a ‘click’). After that, an external circuit quenches the avalanche by lowering the bias voltage V_{APD} below V_{br}, and the APD comes into a linear mode. If the APD is illuminated by bright light (which does not happen in normal single-photon operation but can happen during an eavesdropping attack), I_{APD} in the linear mode is proportional to the incident bright optical power P_{opt}. I_{th} then becomes a threshold on the incident optical power P_{th} that makes a click.

From an engineering view, the detector can be analyzed by its circuit. Figure 10 shows an equivalent circuit diagram of the two detectors used in Clavis2. When no gate signal is applied, the APDs are biased slightly below their V_{br} by the negative high-voltage supply $V_{\mathrm{bias},0} = -55.26$ V, $V_{\mathrm{bias},1} = -54.70$ V [51]. To bring the APD into Geiger mode, an additional 3 V high, 2.8 ns long pulse is applied through a logic level converter DD1. The anode of the APD is AC-coupled to a fast comparator DA1. Since the capacitor C1 blocks the DC component, only when the current flowing through the APD changes, it generates a pulse as the input of DA1. If the peak voltage of this pulse is greater than the positive threshold $V_{\mathrm{th}} = 70$ mV, the comparator produces a logic output signal indicating a click. Once a click in either of the two Bob’s detectors is registered, the next 50 gates will not be applied to both detectors, which constitutes a deadtime to reduce afterpulsing.

If Eve sends a bright c.w. illumination to the gated detectors, the bright light makes the APD generate a significant photocurrent that monotonically increases with the optical power P_{opt}. When we consider effects of this current on the whole detector circuit (Fig. 10), the most useful one is a reduction of the voltage across the APD V_{APD}. Although the high-voltage supply V_{bias} stays constant, the photocurrent causes a significant voltage across R3 = 1 kΩ, thus V_{APD} drops. If we apply enough illumination power, V_{APD} will be less than V_{br} even inside the gate, and the APD then always stays in the linear mode. The detector becomes blind to single photons. In our testing, we measure the voltage at test point T2 V_{T2} in Fig. 10 and refer to this voltage as V_{APD} in the text. V_{T2} is close to real V_{APD}, because $\mathrm{R1} + \mathrm{R2} \ll R3$ [precisely, $V_{\mathrm{APD}} = V_{\mathrm{T2}} + (V_{\mathrm{T2}} - V_{\mathrm{bias}})(\mathrm{R1} + \mathrm{R2})/\mathrm{R3}$].

After blinding Bob’s detectors, Eve can conduct a faked-state attack. Eve first intercepts all photons sent by Alice. Whenever Eve detects a photon, she sends the same state to Bob via a bright trigger pulse of a certain energy, superimposed on her blinding illumination. Only if Bob chooses the same measurement basis as Eve and applies the gate, one of Bob’s detectors will click and he will get the same bit value as Eve. Otherwise, there is no click at Bob’s side. During the sifting procedure, Alice and Bob keep the bit values when they have chosen the same basis, and so does Eve. Therefore Eve has identical bit values with Bob, introduces no extra QBER, and does not increase the alarm counter. Eve then listens to the public communication between Alice and Bob and performs the same error correction and privacy amplification procedures as them, to obtain an identical copy of their secret key [26].

Appendix B: Analysis of processes in the detector

For further understanding of the detector behaviour under successful blinding attack, we attempt to quantitatively model electrical and thermal processes in it. As we mentioned previously, the bias voltage decreases when the blinding power is applied. A measured relationship between V_{APD} and continuous blinding power is shown in Fig. 11. Detector 0 is blinded at $P_{\mathrm{opt}} > P_{\mathrm{blind},0} = 73.4$ µW and detector 1 is blinded at $P_{\mathrm{opt}} > P_{\mathrm{blind},1} = 64.3$ µW. Higher blinding illumination leads to lower bias voltage. This is consistent with the same measurement done for the original blinding attack [26].

In a detector blinded by c.w. laser illumination, the

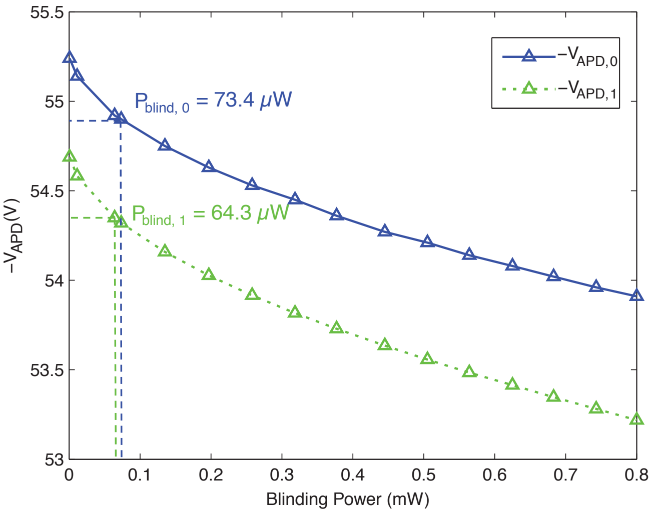

FIG. 11. Bias voltage of APDs versus c.w. blinding power.

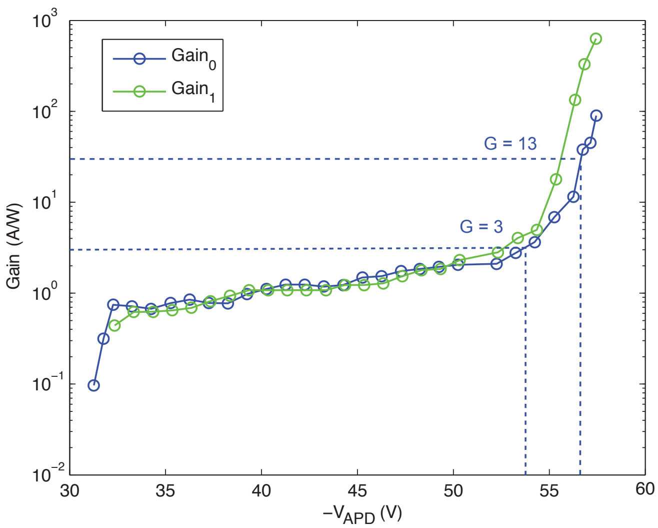

FIG. 12. Gain versus APD bias voltage. Values of gain for bias voltages below 31 V were negligibly low for a practical attack, and below the sensitivity of our measurement method.

gain factor is affected by not only the power of blinding laser, but also the gate signal. When the APD is blinded and forced to work in the linear mode, it can be treated as an ordinary photodiode with a finite internal gain. Photoelectrons and holes are accelerated by a high electric field and initiate a chain of impact ionizations that generates secondary electron-hole pairs. Thus, the APD has an internal multiplication gain factor $M > 1$, since one photon can yield many electrons of photocurrent flowing in the circuit. When $V_{\rm APD}$ is much lower than $V_{\rm br}$, M will be close to 1. However, the APD may not have any significant photosensitivity below so-called punch-through voltage, below which the electrical field does not extend into the absorption layer of InGaAs/InP heterostructure [52].

We have done a measurement of small-signal gain G of the APDs in Clavis2 by measuring their photocurrent response to a short optical pulse input. The results are shown in Fig. 12. There is virtually no photosensitivity below the punch-through voltage of about 31 V. Above that voltage G starts at $\sim$ 0.7 A/W (corresponding to $\sim$ 60% quantum efficiency assuming $M = 1$), then rises above 100 A/W closer to $V_{\rm br}$. The gain values measured at $V_{\rm br} - 2$ V are $\sim$ 7 and $\sim$ 10 A/W, which is consistent with values from data sheets of commercial APDs. From the above measurements, we know that Eve can vary the amount of blinding power to the detectors to control the bias voltage and thus the gain factor.

After we blind Bob's detectors in Clavis2, the gain factor is greater during the 2.8 ns gate duration, because the gate signal raises $V_{\rm APD}$. Thus the electrical charge generated by the APD in response to a trigger pulse applied in the gate is greater than when it's applied outside the gate. For example, in Fig. 4(c), the gate pulse alone contributes 1.053 pC extra charge on top of the current that would be generated without the gate. When the trigger pulse is applied after the gate [Fig. 4(d)], the total charge of the two pulses is 1.467 pC; however, when the trigger pulse is moved into the gate [Fig. 4(e)], the total charge rises to 1.613 pC. Therefore, a greater gain factor during the gated time helps the pulse to cross the threshold.

We have attempted to model the increased gain due to the gate. In our model, we consider a thermal effect and an internal resistance of the APD. On the one hand, an increased temperature raises $V_{\rm br}$ [53]. Electrical heating ($V_{\rm APD} \cdot I_{\rm APD}$) and the absorption of the blinding power result in a heat dissipation: 61.2 mW for detector 0 and 66.03 mW for detector 1 [54]. Then, an estimated 190 K/W thermal resistance [33] between each APD chip and the cold plate converts the power dissipation into the increased temperature. The temperature-dependent breakdown voltage increases with the coefficient of about 0.1 V/K [33]. As a result, $V_{\rm br}$ increases by 1.16 V (1.25 V) for detector 0 (1). Figure 12 shows the relation between gain factor and the actual $V_{\rm APD}$ in the linear mode. When $V_{\rm APD}$ is close to $V_{\rm br}$, the gain factor increases rapidly. On the other hand, we suppose the APD has a passive internal resistance, so the internal bias voltage across the ideal photodiode is less than the value of $V_{\rm APD}$ we test. By measuring the voltage of a stable avalanche pulse and calculating the current trough the detector circuit when avalanche happens, we obtain the internal resistance of 330 Ω in detector 0 and 275 Ω in detector 1. Therefore, the real bias voltage under blinding attack shown in Fig. 4(c–f) is 53.77 V, which corresponds to $G = 3$ A/W in detector 0 as shown in Fig. 12. When 3 V gate is applied, the bias voltage becomes 56.77 V which corresponds to $G = 13$ A/W in Fig. 12. However, the measured charges in Fig. 4(d) and (e) illustrate much less gain change: $G = 1.3$ A/W at 53.77 V and $G = 1.76$ A/W at 56.77 V [55]. The discrepancy may be explained by a larger actual thermal resistance between the APD and the cold plate than we estimate, which should be verified in future research.

[1] M. Naor, in *Advances in Cryptology – CRYPTO 2003* (Springer, Berlin, 2003) pp. 96–109.

[2] *ETSI white paper no. 8: Quantum safe cryptography and security* (ETSI, Sophia Antipolis, France, 2015).

[3] C. H. Bennett and D. P. DiVincenzo, Nature **404**, 247 (2000).

[4] P. W. Shor, in *Proceedings of 35th Annual Symposium on Foundations of Computer Science* (IEEE, 1994) pp. 124–134.

[5] C. H. Bennett and G. Brassard, in *Proceedings of IEEE International Conference on Computers, Systems, and Signal Processing, Bangalore, India* (IEEE Press, New York, 1984) pp. 175–179.

[6] A. K. Ekert, Phys. Rev. Lett. **67**, 661 (1991).

[7] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Rev. Mod. Phys. **74**, 145 (2002).

[8] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, Rev. Mod. Phys. **81**, 1301 (2009).

[9] W. K. Wootters and W. H. Zurek, Nature **299**, 802 (1982).

[10] H.-K. Lo and H. F. Chau, Science **283**, 2050 (1999).

[11] P. W. Shor and J. Preskill, Phys. Rev. Lett. **85**, 441 (2000).

[12] N. Lütkenhaus, Phys. Rev. A **61**, 052304 (2000).

[13] D. Mayers, J. ACM **48**, 351 (2001).

[14] D. Gottesman, H.-K. Lo, N. Lütkenhaus, and J. Preskill, Quant. Inf. Comp. **4**, 325 (2004).

[15] R. Renner, N. Gisin, and B. Kraus, Phys. Rev. A **72**, 012332 (2005).

[16] C. H. Bennett, Phys. Rev. Lett. **68**, 3121 (1992).

[17] T. Schmitt-Manderbach, H. Weier, M. Fürst, R. Ursin, F. Tiefenbacher, T. Scheidl, J. Perdigues, Z. Sodnik, C. Kurtsiefer, J. G. Rarity, A. Zeilinger, and H. Weinfurter, Phys. Rev. Lett. **98**, 010504 (2007).

[18] D. Stucki, N. Walenta, F. Vannel, R. T. Thew, N. Gisin, H. Zbinden, S. Gray, C. R. Towery, and S. Ten, New J. Phys. **11**, 075003 (2009).

[19] Y.-L. Tang, H.-L. Yin, S.-J. Chen, Y. Liu, W.-J. Zhang, X. Jiang, L. Zhang, J. Wang, L.-X. You, J.-Y. Guan, D.-X. Yang, Z. Wang, H. Liang, Z. Zhang, N. Zhou, X. Ma, T.-Y. Chen, Q. Zhang, and J.-W. Pan, IEEE J. Sel. Top. Quantum Electron. **21**, 1 (2015).

[20] Several companies sell QKD systems: ID Quantique (Switzerland), `http://www.idquantique.com/`; SeQureNet (France), `http://www.sequrenet.com/`; the Austrian Institute of Technology (Austria), `http://www.ait.ac.at/`; and QuantumCTek (China), `http://www.quantum-info.com/`.

[21] A. Vakhitov, V. Makarov, and D. R. Hjelme, J. Mod. Opt. **48**, 2023 (2001).

[22] V. Makarov, A. Anisimov, and J. Skaar, Phys. Rev. A **74**, 022313 (2006), erratum ibid. **78**, 019905 (2008).

[23] N. Gisin, S. Fasel, B. Kraus, H. Zbinden, and G. Ribordy, Phys. Rev. A **73**, 022320 (2006).

[24] B. Qi, C.-H. F. Fung, H.-K. Lo, and X. Ma, Quant. Inf. Comp. **7**, 73 (2007).

[25] Y. Zhao, C.-H. F. Fung, B. Qi, C. Chen, and H.-K. Lo, Phys. Rev. A **78**, 042333 (2008).

[26] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Nat. Photonics **4**, 686 (2010).

[27] S.-H. Sun, M.-S. Jiang, and L.-M. Liang, Phys. Rev. A **83**, 062331 (2011).

[28] L. Lydersen, M. K. Akhlaghi, A. H. Majedi, J. Skaar, and V. Makarov, New J. Phys. **13**, 113042 (2011).

[29] P. Jouguet, S. Kunz-Jacques, and E. Diamanti, Phys. Rev. A **87**, 062313 (2013).

[30] S. Sajeed, P. Chaiwongkhot, J.-P. Bourgoin, T. Jennewein, N. Lütkenhaus, and V. Makarov, Phys. Rev. A **91**, 062301 (2015).

[31] A. Acín, N. Brunner, N. Gisin, S. Massar, S. Pironio, and V. Scarani, Phys. Rev. Lett. **98**, 230501 (2007).

[32] H.-K. Lo, M. Curty, and B. Qi, Phys. Rev. Lett. **108**, 130503 (2012).

[33] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Opt. Express **18**, 27938 (2010).

[34] C. Wiechers, L. Lydersen, C. Wittmann, D. Elser, J. Skaar, C. Marquardt, V. Makarov, and G. Leuchs, New J. Phys. **13**, 013043 (2011).

[35] Z. L. Yuan, J. F. Dynes, and A. J. Shields, Appl. Phys. Lett. **98**, 231104 (2011).

[36] T. Honjo, M. Fujiwara, K. Shimizu, K. Tamaki, S. Miki, T. Yamashita, H. Terai, Z. Wang, and M. Sasaki, Opt. Express **21**, 2667 (2013).

[37] C. C. W. Lim, N. Walenta, M. Legré, N. Gisin, and H. Zbinden, IEEE J. Sel. Top. Quantum Electron. **21**, 6601305 (2015).

[38] J. Cartwright, "Quantum cryptography is safe again", Science News, 29 August 2013, `http://news.sciencemag.org/physics/2013/08/quantum-cryptography-safe-again`.

[39] H. Qin, R. Kumar, and R. Alléaume, Proc. SPIE **8899** (2013).

[40] S. Sajeed, I. Radchenko, S. Kaiser, J.-P. Bourgoin, A. Pappa, L. Monat, M. Legré, and V. Makarov, Phys. Rev. A **91**, 032326 (2015).

[41] Clavis2 specification sheet, `http://www.idquantique.com/images/stories/PDF/clavis2-quantum-key-distribution/clavis2-specs.pdf`.

[42] M. Legre and G. Ribordy, "Apparatus and method for the detection of attacks taking control of the single photon detectors of a quantum cryptography apparatus by randomly changing their efficiency", international patent appl. WO 2012/046135 A2 (filed 2010-10-10, published 2012-04-12).

[43] L. Lydersen, N. Jain, C. Wittmann, Ø. Marøy, J. Skaar, C. Marquardt, V. Makarov, and G. Leuchs, Phys. Rev. A **84**, 032320 (2011).

[44] A. Kerckhoffs, J. des Sciences Militaires **IX**, 5 (1883).

[45] S. Singh, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography* (Four Estate, London, 1999).

[46] N. Gisin, abstract of keynote talk at Qcrypt 2015, Tokyo, September 28 – October 2, 2015, arXiv:1508.00341 [quant-ph].

[47] N. Jain, C. Wittmann, L. Lydersen, C. Wiechers, D. Elser, C. Marquardt, V. Makarov, and G. Leuchs, Phys. Rev. Lett. **107**, 110501 (2011).

[48] V. Makarov and D. R. Hjelme, J. Mod. Opt. **52**, 691 (2005).

[49] V. Makarov, J.-P. Bourgoin, P. Chaiwongkhot,

M. Gagné, T. Jennewein, S. Kaiser, R. Kashyap, M. Legré, C. Minshull, and S. Sajeed, arXiv:1510.03148 [quant-ph].

[50] S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa, J. Mod. Opt. **51**, 1267 (2004).

[51] Using values from the sample of Clavis2 tested in our present study at the University of Waterloo, which is a different sample than in Refs. 26, 33, and 34.

[52] P. A. Hiskett, G. S. Buller, A. Y. Loudon, J. M. Smith, I. Gontijo, A. C. Walker, P. D. Townsend, and M. J. Robertson, Appl. Opt. **39**, 6818 (2000).

[53] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices* (Wiley-Interscience, 2007).

[54] Under 0.564 mW blinding power, $V_{\mathrm{APD},0} = 54.14$ V, $I_{\mathrm{APD},0} = 1.12$ mA. Heat dissipation of detector 0: $54.14\ \mathrm{V} \cdot 1.12\ \mathrm{mA} + 0.564\ \mathrm{mW} = 61.2\ \mathrm{mW}$; $V_{\mathrm{APD},1} = 53.484$ V, $I_{\mathrm{APD},1} = 1.224$ mA, Heat dissipation of detector 1: $53.484\ \mathrm{V} \cdot 1.224\ \mathrm{mA} + 0.564\ \mathrm{mW} = 66.03\ \mathrm{mW}$.

[55] When we apply a 0.32 pJ trigger pulse after the gate, this single trigger pulse contributes 0.414 pC charge which is the difference between the total charges in Fig. 4(c) and (d). $G = 0.414\ \mathrm{pC}/0.32\ \mathrm{pJ} = 1.3\ \mathrm{A/W}$. When we apply a 0.32 pJ trigger pulse during the gate, this single trigger pulse contributes 0.56 pC charge which is the difference between the total charges in Fig. 4(c) and (e). $G = 0.56\ \mathrm{pC}/0.32\ \mathrm{pJ} = 1.76\ \mathrm{A/W}$.