An advanced Eve of QKD: breaking a security assumption and hacking a black box

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*

Theoretical security proof of quantum key distribution (QKD) is based on certain assumptions about realistic devices. Once the assumptions are discrepant with the behavior of devices in the real life, unconditional security of the entire QKD system cannot be guaranteed. This work presents a concrete example of disproving an assumption that Bob's detection probability under blinding attack [2] cannot be proportional to his single-photon detection efficiency, on which the theoretical security analysis in Ref. 3 relies. This talk is partly based on our recent preprint [1].

The countermeasure implementation [3] that we analyze is an attempt to secure existing QKD schemes against detector-control attacks [2, 4, 5]. This is more practical in short-term than replacing the entire scheme with a device-independent one. Furthermore, this countermeasure is deployed in the most advanced currently available commercial QKD system Clavis2 from ID Quantique. Unfortunately, our testing of the patched system shows a gap between academia and industry: the implemented countermeasure is not as effective as the academia expected.

After showing the feasibility of our attack, we will consider a practically interesting question how robust is the attack for a future demonstration in a *black-box setting*, when Eve only has access to the public communication lines [6]. A general strategy of hacking a black box will be proposed as well.

Ref. 3 claims that a countermeasure with two non-zero decoy detection efficiencies is effective against the blinding attack [2], since Eve cannot mimic these two non-zero detection efficiencies $\eta_1 < \eta_2$ after blinding and controlling Bob's two detectors D0, D1. However, our testing results show that Eve can match the expected detection probabilities by adjusting the energy or timing of her trigger pulse which is sent during the gated time to the blinded detectors. As shown in Fig. 1, we measure the relation between the energy of trigger pulse and click probability for lower and higher detection efficiency levels. The position of trigger pulse is fixed in the middle of gate signal. For detector D0, if trigger pulse energy E_1 is chosen, D0 always clicks, while at E_2, the detector only clicks if higher bias voltage is applied (corresponding to higher efficiency). To match the lower efficiency η_1 for D0, Eve selects trigger pulse energy E_1 with probability q_1 to satisfy $\eta_1 = q_1$. Then, for higher efficiency η_2, Eve selects energy level E_2 with probability q_2 to satisfy

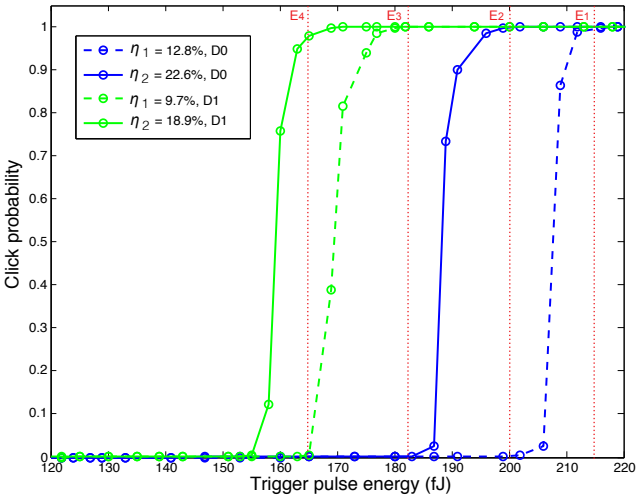

FIG. 1. Click probabilities under blinding attack versus energy of trigger pulse. The blinding power is 0.38 mW and the timing of trigger pulse is aligned to the middle of the gate.

$\eta_2 = q_1 + q_2$. The same method is used for detector D1. Therefore, Eve reproduces correct detection probabilities for blinded detectors as the protocol requires.

We have also tested the relation between time-shift of trigger pulse and click probability of both efficiency levels (Fig. 2). The trigger pulse energy is fixed in this case. For D0, Eve can always trigger a click by choosing time-shift T_1, but only trigger a click at high bias voltage by choosing T_2. When T_1 is selected with probability q_1, lower efficiency can be matched as $\eta_1 = q_1$. T_2 is selected with probability q_2 to match $\eta_2 = q_1 + q_2$ for higher efficiency. The same strategy is utilized for D1. In this way, Eve also hacks Clavis2 system tracelessly.

Generally, a finite set of decoy detection efficiency levels $\eta_1 < \eta_2 < \eta_3 < ... < \eta_n$ can be hacked by properly setting probabilities of different attacking energy levels or time-shifts. We take energy levels of trigger pulse as an example. According to the result in Fig. 1, it is reasonable to extrapolate that we can find n distinct levels of trigger pulse energy $E_1 > E_2 > E_3 > ... > E_n$ in this situation. Then Eve can apply E_k $(k = 1, ..., n)$ with probability q_k to satisfy $\eta_k = \sum_{i=1}^{k} q_i$. This reproduces every expected value of η_k and hacks the system by breaking the assumption.

In the realistic scenario of hacking a black box, a com-

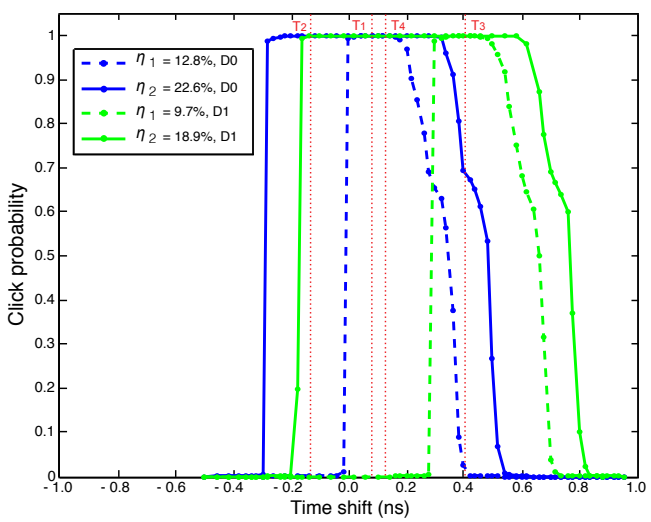

FIG. 2. Click probabilities under blinding attack versus relative time-shift of trigger pulse. The blinding power is 0.38 mW. The energy of trigger pulse for D0 is 0.22 pJ and for D1 is 0.19 pJ. These energy levels are marked as red × in Fig. 3.

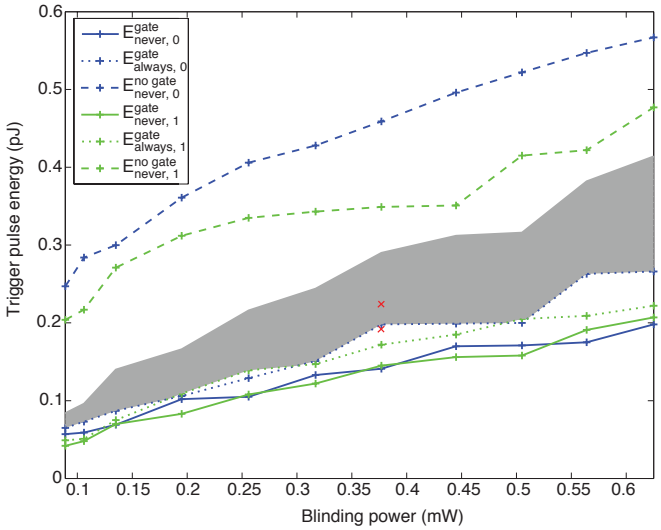

FIG. 3. Energy thresholds of trigger pulse versus continuous-wave blinding power. Shaded area shows the range of trigger pulse energies of the perfect attack [1, 2].

plete attack plan needs to consider all relevant practicalities, which we will discuss in the talk. Eve may first purchase 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 as a black box. Although the latter sample can be of the same model and design, it will generally have different values of internal analog parameters, such as energy thresholds E_k. Eve may need several attempts to find correct values of attacking parameters. A full demonstration of our attacks in this scenario remains to be tested.

The first simplified version of countermeasure implementation currently deployed in Clavis2 [1] is likely to be hacked as the following analysis shows. The simplified implementation does not use two non-zero efficiency levels, but instead suppresses gates randomly, corresponding to *zero* expected efficiency. In this setting it will be of utmost importance for Eve to avoid triggering clicks in the absence of the gate, because this would risk revealing her attack attempt. Our 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. 3). It also tolerates some fluctuation in experimental parameters for detector control, which makes it robust against reasonably expected fluctuations and imprecision of the system parameters.

A full two non-zero efficiencies 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 will depend on the actual industrial implementation of the full countermeasure. Demonstrating the full attack can be a future study.

The breakability of assumption indicates that the model of a practical detector should be more precise in security analysis, if one wishes to close the detector loopholes without resorting to measurement-device-independent QKD. The failure of countermeasure evidences the necessity of security certification for quantum communication equipment, which unfortunately the entire community still lacks. Therefore, we should keep investigating security issues in QKD. Only if the actual threat from eavesdropper is well recognized and constrained, QKD protocol design and implementation can be closer to unconditional security.

Acknowledgements. We thank Nicolas Gisin and Charles Ci Wen Lim for discussions.

* angelhuang.hn@gmail.com

[1] A. Huang, S. Sajeed, P. Chaiwongkhot, M. Soucarros, M. Legre, and V. Makarov, arXiv:1601.00993 [quant-ph].
[2] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Nat. Photonics **4**, 686 (2010).
[3] C. C. W. Lim, N. Walenta, M. Legré, N. Gisin, and H. Zbinden, IEEE J. Sel. Top. Quantum Electron. **21**, 6601305 (2015).
[4] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Opt. Express **18**, 27938 (2010).
[5] C. Wiechers, L. Lydersen, C. Wittmann, D. Elser, J. Skaar, C. Marquardt, V. Makarov, and G. Leuchs, New J. Phys. **13**, 013043 (2011).
[6] N. Gisin, abstract of keynote talk at QCrypt 2015, Tokyo, September 28 – October 2, 2015, arXiv:1508.00341 [quant-ph].