

This is a repository copy of *Experimental vulnerability analysis of QKD based on attack ratings*.

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/175116/>

Version: Published Version

Article:

Kumar, Rupesh, Mazzoncini, Francesco, Qin, Hao et al. (1 more author) (2021)
Experimental vulnerability analysis of QKD based on attack ratings. Scientific Reports.
9564. ISSN 2045-2322

<https://doi.org/10.1038/s41598-021-87574-4>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



OPEN

Experimental vulnerability analysis of QKD based on attack ratings

Rupesh Kumar¹, Francesco Mazzoncini², Hao Qin³ & Romain Alléaume²✉

Inspired by the methodology used for classical cryptographic hardware, we consider the use of attack ratings in the context of QKD security evaluation. To illustrate the relevance of this approach, we conduct an experimental vulnerability assessment of CV-QKD against saturation attacks, for two different attack strategies. The first strategy relies on inducing detector saturation by performing a large coherent displacement. This strategy is experimentally challenging and therefore translates into a high attack rating. We also propose and experimentally demonstrate a second attack strategy that simply consists in saturating the detector with an external laser. The low rating we obtain indicates that this attack constitutes a primary threat for practical CV-QKD systems. These results highlight the benefits of combining theoretical security considerations with vulnerability analysis based on attack ratings, in order to guide the design and engineering of practical QKD systems towards the highest possible security standards.

Quantum Key Distribution (QKD) is a beautiful idea^{1,2}, that allows two legitimate users, the sender Alice and the receiver Bob, to establish a secret key with information-theoretic security (ITS) even against a quantum attacker, Eve.

QKD systems with increasing performances and reliability have been engineered over the past 25 years^{3–5}. This has placed QKD among the quantum technologies with the highest maturity level. It has also paved the way for large-scale deployments and for the effective demonstration of QKD in relevant real-world application contexts, requiring the ability to provide long-term security for data at flight or at rest⁶, such as private clouds or critical infrastructures for government, defense or health data management⁷. A necessary condition for the industrial take-off of QKD will, however, reside not only in the ability to engineer cost-effective QKD systems, but also in the capacity to provide solid guarantees regarding their security.

Theoretical security proofs^{8–10} constitute a strong conceptual framework to capture the security properties of QKD protocols, based on a model. QKD implementations may, however, not fully comply with the model used in the security proof, leading to security vulnerabilities and the possibility to launch side-channel attacks⁵. Optimizing the real-world security of a QKD system hence requires to consider not only the *theoretical security* of the QKD protocol, but also the *practical security* related to its implementation. As a matter of fact, engineering constraints may impose stringent limits to the security level that a QKD system can reach. Such constraints can indeed lead to significant deviations between the theoretical security level that could be expected with an idealized implementation and the security level that can be reached in practice by the real QKD system.

As system design and security evaluation are in practice almost always limited by resources, attacks that are easier to implement should be prioritized, as they represent the greatest threats. For instance, some attacks on QKD can be realized with a relatively simple procedure and inexpensive hardware, such as detector blinding attack¹¹ that has even been demonstrated on a live QKD connection¹². Some other attacks, on the other hand, such as the photon number splitting attack¹³ and more generally collective and coherent attacks on QKD⁸, have played a fundamental role in our understanding of QKD theory. Yet, their implementation requires the ability to store and retrieve single photons from a quantum memory, potentially over ms or larger timescales, which is currently out of reach, given the limitations of existing quantum memory technology¹⁴. Hence, to guarantee a very high security level for QKD, forward-looking methods and standards in quantum cryptography implementation security shall be adopted, following a methodology similar to the one used to certify the security of classical crypto-systems¹⁵, called Common Criteria.

The Common Criteria methodology that we introduce is applicable both for discrete-variable DV-QKD and continuous-variable (CV)-QKD, and also in principle to any practical quantum cryptographic system, provided it has reached a sufficient maturity level (necessary to make testing and vulnerability analysis meaningful).

¹Quantum Communications Hub, Department of Physics, University of York, Heslington YO10 5DD, UK. ²Télécom Paris-LTCl, Institut Polytechnique de Paris, 19 Place Marguerite Perey, 91120 Palaiseau, France. ³CAS Quantum Network Co., Ltd, No. 99 Xiupu Rd, Pudong New District, Shanghai 201315, China. ✉email: romain.alleaume@telecom-paris.fr

Rating	AP range
Basic	0–10
Moderate	11–15
High	16–19
Beyond high	20–∞

Table 1. Semi-qualitative scale for attack rating. This scaling is adapted with respect to the Common Evaluation Methodology²⁵, taking into account the fact that we consider 4 out of 5 factors in our analysis. For a more detailed discussion on the attack rating factors see Methods.

Moreover, we demonstrate in this article the practical relevance of this methodology by conducting vulnerability analysis on an experimental CV-QKD system.

In CV-QKD^{16,17} a few attacks have been proposed to exploit the security loopholes^{18–21}, however most of them have been demonstrated off-line. Demonstrating an active attack on a live CV-QKD system is challenging. It requires to overcome several difficulties such as the technical complexity of the attack strategy itself, or of the optical phase recovery. In this work, we demonstrate an active attack on a live CV-QKD system running the Gaussian modulated coherent state (GMCS) protocol¹⁶. Exploiting the non-linear response of the homodyne detector near its detection limit, an eavesdropper, Eve, can launch an attack called Saturation Attack^{22,23}.

We have considered two practical methods to mount the saturation attack in CV-QKD. The first and most challenging one is the coherent attack strategy²², where Eve resends coherent displaced signal²⁴ to induce the detector saturation. Our second attack strategy consists in the incoherent saturation attack²³, where we shine an independent laser towards Bob's coherent receiver. The implementation of this attack is considerably simpler and it constitutes a dangerous threat to practical CV-QKD systems. Inspired by the Common Criteria Common Evaluation Methodology v3.1 (CEM)²⁵, we introduce a metric called Attack Potential to QKD, and we evaluate the two aforementioned saturation attack strategies against this metric.

Although the Attack Potential approach to a practical security evaluation is not new in the world of IT security, its introduction in the context of QKD brings fresh perspectives. It has the ability to strengthen the security rationale associated with QKD system design and to accelerate the evolution towards a QKD industry capable of manufacturing QKD devices with high security assurance.

Results

Attack rating. One crucial part of the complex methodology for this security evaluation is the process of identifying, classifying and prioritizing threats associated to vulnerabilities in QKD systems. A comprehensive methodology offers general guidance and a metrics to rate the possible attacks against the assets. It also considers both the likelihood that a threat agent may successfully perform the attack and the magnitude of the impact that this attack has on the assets. In our rating procedure we shall focus on the likelihood of an attack, evaluating the total effort required to successfully mount the attack, called the Attack Potential: the higher the Attack Potential, the lower the chances of the attack being performed are. To determine the Attack Potential we consider different factors (such as the type of equipment required); for each of them we assign a numerical value, the sum of them is the actual Attack Potential. In Table 1 we define the semi-qualitative correspondence between rating and attack difficulty. Attack paths with an Attack Potential (AP) between 0 and 10 are for example rated as *Basic*. Such attacks can be implemented with little effort and therefore constitute very serious threats. On the other hand, attacks with an extremely high Attack Potential, rated here *Beyond High*, are extremely difficult to implement and therefore constitute less pressing threats. (Our rating scale is slightly different from the usual rating scale (Table 4 of the CEM²⁵). The main reason for that is that we had to reshape the boundary values with a factor 4/5, since in our attack potential evaluation we do not consider elapsed time as a factor (see Methods). We have also simplified the table a little, by merging *basic* and *enhanced-basic*).

The saturation attack. Saturation attack on CV-QKD consists in biasing the excess noise estimation by actively inducing the saturation of the homodyne detectors. This attack can be powerful²²: it can be combined with simple attack strategies by Eve (such as the intercept-resend attack²⁶) and lead to a full security breach.

In a CV-QKD system that implements the GMCS protocol, Alice prepares coherent states of quadratures $\{X_A, P_A\}$, modulates each quadrature according to a Gaussian distribution of variance V_A and centered on zero, and sends the modulated coherent state to Bob through the quantum channel. Bob randomly measures one of the quadratures using a balanced homodyne detector. This results in quadrature measurements X_B and P_B , with variance V_B . By correlating sent and measured quadrature values on a fraction of their data, the users estimate channel transmittance T and then the excess noise ξ . If these values are within the limit for validating the security of the key, they proceed to key distillation on rest of the data, if not the QKD protocol aborts.

The balanced homodyne detector is a crucial part of a CV-QKD system. The linearity range of the homodyne detector response is in general characterized offline, as part of the detector calibration process. Assuming it is the same for X and P quadrature, we can designate this linearity range as a quadrature interval $[\alpha_1, \alpha_2]$ ($\alpha_1 < 0 < \alpha_2$).

Balancing the homodyne detector prior to protocol run ensures that the mean of the homodyne output values remain close to zero, and therefore that the homodyne receiver is operated in its linear range, except if signals with very large quadrature values are received at Bob side. In case signals with $X_B \ll \alpha_1$ (respectively $X_B \gg \alpha_2$)

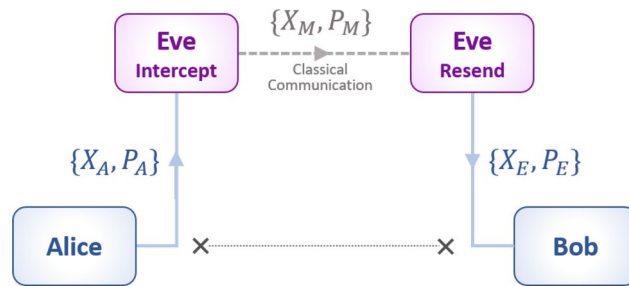


Figure 1. Scheme for saturation attack. $Eve_{intercept}$ intercepts Alice's Gaussian modulated signal of quadratures $\{X_A, P_A\}$ and shares her measurement results $\{X_M, P_M\}$ through the classical channel to Eve_{resend} . The resent and displaced signal of quadrature $\{X_E, P_E\}$ is measured by Bob homodyne detector.

are received, then the detector will saturate and output $X_B = \alpha_1$ (respectively $X_B = \alpha_2$). Close or beyond the detection limits, response of the detector becomes non-linear to the input signal quadrature, which will effectively reduce the measured variance, compared to the actual quadrature variance of the received optical signals.

In the saturation attack, in order to reach the non-linear regime of the detector response, Eve performs an intercept-resend attack and displaces the mean of the Gaussian quadrature modulation of the resent signals from zero to a value $\Delta = \sqrt{\Delta_X^2 + \Delta_P^2}$. Without loss of generality, we can set Δ_X equal to Δ_P such that the effect of displacement at Bob is identical for both quadrature measurements. To perform the saturation attack Eve sets the amount of displacement Δ such that quadrature of the coherent state received by Bob overpassed the linear range. As a consequence, the saturation affects the measured quadrature variance by Bob. Additionally, Eve may not only displace the quadrature value of the coherent state she resends to Bob, she may also apply some amplification factor G on the resent signal primarily to compensate the 3 dB loss occurring from her heterodyne measurement during IR attack.

Now, from the quadrature measurement data obtained from the saturated induced detector, Alice and Bob estimate channel parameters T_{sat} and ξ_{sat} both are influenced by displacement Δ and amplification factor G . For a given value of Alice's quadrature modulation variance V_A , Eve can optimize Δ and G such that excess noise ξ_{sat} drops below the null key noise threshold and eavesdropping remains undetected. In such case, even though Eve has mounted an entanglement-breaking intercept-resend attack (which should lead the QKD protocol to abort due to a too high excess noise, generating no key) the attack is not detected, due to saturation, and Alice and Bob generate key, that is however insecure: this constitutes a characterized security break.

To characterize the attack, and in particular its impact on key rate, we have defined the following conditions.

- The attacker, Eve, performs the saturation attack: Intercept-resend attack combined with displacement.
- The channel transmission estimation is unaffected ($T_{sat} = T$, where T is the channel transmission in absence of attack).
- Alice and Bob obtain a positive key rate from their estimated parameter T_{sat} and ξ_{sat} .

The actual realization of the saturation attack comprises of two steps: intercepting Alice's signal and resending a newly prepared signal to Bob with displacement Δ and gain G . We can consider that two cooperating eavesdroppers are involved in the attack: $Eve_{intercept}$, located near Alice intercepts the signals of quadratures $\{X_A, P_A\}$ and classically communicates the measurement results $\{X_M, P_M\}$ to Eve_{resend} located near to Bob as shown in Fig. 1. Due to the technical restrictions imposed by the laboratory equipment, we experimentally demonstrate only the resend step of the attack and model the impact of the measurement associated with the intercept step. $\{X_M, P_M\}$ is deduced from $\{X_A, P_A\}$ by simulating a heterodyne measurement, i.e. 3 dB loss factor and also the addition of a random Gaussian noise of variance 2 shot noise²⁶.

Efficient countermeasures against the saturation are known. As detailed in²⁷ active monitoring of the linearity of Bob's coherent detection can provide a robust countermeasure against saturation attack. This countermeasure, however, requires dedicated hardware (additional amplitude modulator at Bob side). In²² a second countermeasure is proposed, that can be implemented without using additional hardware: it relies on the pre-characterization of the detector linearity range, and consists in post-selecting measurement data, based on the fact that these quadrature measurements fall, within high confidence, in the linearity range of the detector. This countermeasure, that relies purely on software, has a small marginal cost.

Coherent attack strategy. The signal of quadrature $\{X_E, P_E\}$ is resent by Eve_{resend} , that we will from here onwards label as Eve. It is experimentally generated, using a setup built around a Sagnac interferometer, represented on Fig. 7, and whose functioning is detailed in Methods. The role of this set-up is to generate, knowing the in values $\{X_M, P_M\}$, a displaced coherent state of quadrature $\{X_E, P_E\}$ that are correspond to the encoding of $\{X_M, P_M\}$ on a coherent states, to which is applied a coherent gain $\sqrt{G/2}$ in amplitude, and a controlled coherent displacement by a value $\Delta = \Delta_X = \Delta_P$. The Sagnac loop offers a high phase stability which allows to precisely control Δ and therefore minimize the noise. Receiving the displaced coherent state $\{X_E, P_E\}$, Bob randomly measures one of the quadratures with a balanced homodyne detector, cf Fig. 6, hence obtaining X_B or

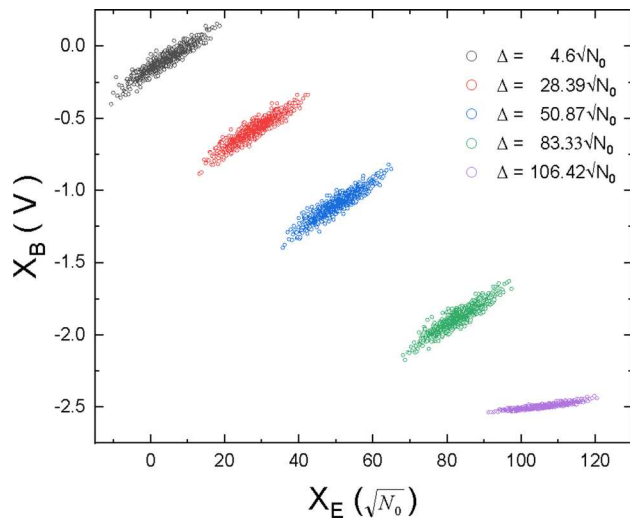


Figure 2. Response of homodyne output due to displacement. Input signal sent by Eve, with quadrature variance $\text{Var}(X_E) = 22N_0$ with 5 different displacement Δ values equals to $4.6\sqrt{N_0}$ (black), $28.39\sqrt{N_0}$ (red), $50.87\sqrt{N_0}$ (blue), $83.33\sqrt{N_0}$ (green) and saturation at $106.42\sqrt{N_0}$ (magenta). Displacement shifts Bob's quadrature measurement X_B (expressed here in volts). Large displacement value can lead to saturation (that occurs when X_B reaches -2.5 Volts).

P_B . Depending on the value of Δ , this quadrature measurement will be obtained in the linear or in the saturated regime.

Figure 2 shows the effect of the displacement Δ on Bob's experimental quadrature measurements. The mean value of the homodyne output X_B can be shifted towards one of the detection limit $\alpha_1 = -2.5V$ of the detector, for a given displacement setting. Selecting displacement angle to 225 degree would direct the shift towards the other limit of the linear range $\alpha_2 = 3.3V$. As can be seen on Fig. 2, when the detector is operated close to its linear range limit, then saturation occurs and quadrature variance reduces drastically.

The coherent displacement set-up demands an active feedback routine to compensate the relative phase drifts between the displaced signal and the local oscillator. Even though Sagnac loop provides a high stability, as illustrated by the level of control obtained on Fig. 2, we could not lower the residual quadrature noise due to imperfect phase drift compensation below the null key threshold. For example, considering that 2π phase drift occurring in 1 second, a $500 \mu\text{s}$ latency in the feedback loop creates about 0.2 degrees of phase error. This in turn results in $0.23\sqrt{N_0}$ fluctuations in homodyne output and generates excess noise of about $5N_0$. This implies that the excess noise ξ_{sat} is above the null key noise threshold value, and prevents the generation of key. In other words, in the current setup, Alice and Bob would easily detect attack based on coherent displacement. Reducing the feedback latency such that phase drift remains negligible within the feedback intervals, could however bring this attacking strategy to meet the attack success conditions.

Incoherent attack strategy. In order to overcome the implementation difficulties of the coherent displacement strategy, we have conceived and tested a much simpler strategy, based on incoherent laser pulse injection²³. Saturating the homodyne detector with external laser pulse indeed presents several operational advantages over the previous strategy. First, since it is incoherent with the local oscillator, an external laser adds only its own shot noise to the excess noise. More importantly, relative phase drift compensation is not required for keeping the homodyne in the saturated region. This greatly simplifies Eve's resent setup as shown in Fig. 3.

In this strategy, saturation is induced by an intense incoherent laser pulse sent along with the resent coherent state. The equivalence of the intensity I of the incoherent laser pulse to the displacement Δ can be given by $\Delta = \sqrt{\eta_b/I_{lo}(1 - 2T_{bs})}I$, where η_b is the efficiency of Bob, I_{lo} is the local oscillator intensity and T_{bs} is the effective transmittance applied to the incoherent laser pulse due to asymmetry of beam splitting ratio and the attenuator (shown in Fig. 6). In our case $T_{bs} \approx 49\%$. To bring uniformity in the description of both experimental strategies, since the primary requirement is to induce saturation, we also call "displacement" the effect of this incoherent shift of the measured quadratures.

Since optical phase drift compensation is not needed, saturation attack with incoherent laser pulse can achieve comparatively much performance in terms of quadrature stability and noise, and can meet success conditions, provided the channel loss is not too small (low channel loss make it more difficult for Eve to succeed in the intercept-resend attack). The results in terms of excess noise are given in Fig. 4a. The excess noise at Alice has been calculated from the variance of saturated homodyne output experimental data, at various transmission distance between Alice and Bob. It can be seen that the excess noise is below the null key threshold, which indicates Eve's intercept-resend attack remains untraceable. Figure 4b shows the maximal value of final key rate per pulse, estimated under collective attacks. Note that the condition $T_{sat} = T$ cannot be met for distance below 35 km—see noise model in section Method. A relaxed attack success condition, where Eve does not maintain $T_{sat} = T$ is given in supplementary information.

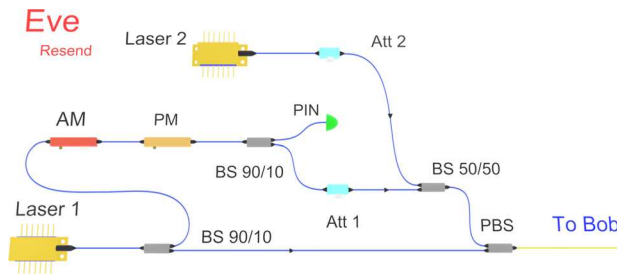


Figure 3. Setup for incoherent coherent attack strategy, relying on pulse injection from an external incoherent laser to induce saturation. AM: Amplitude Modulator, PM: phase Modulator, BS: BeamSplitter, PBS: Polarization BeamSplitter, Att: Variable Attenuator.

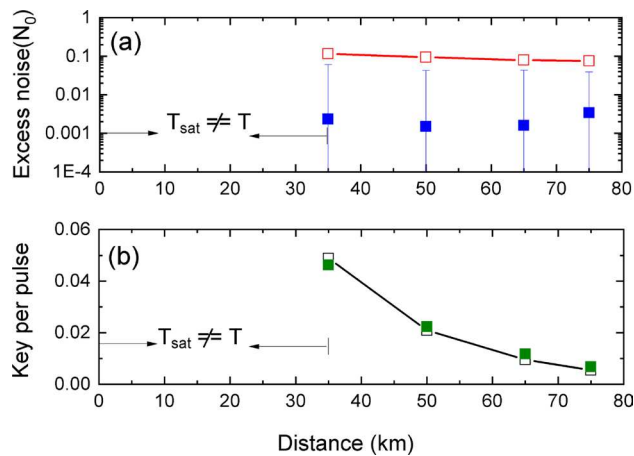


Figure 4. Results:- attack with incoherent light. (a) excess noise at Alice. Red squares indicate the null key noise threshold and blue squares the estimated values of ξ_{sat} . (b) key rate. Black squares are simulated values of final key per pulse while Green squares are from the experiment. Error bars are one standard deviation of fluctuations among ten smaller data block of size 10^7 . Success condition of $T_{sat} = T$ can not be fulfilled below 35 km.

Rating of the two attacks. Now that we have defined, and studied the two possible attacks paths to exploiting the saturation vulnerability of the homodyne receiver, we are ready to use Table 3 to evaluate their Attack Potential. We assume that the hacker Eve tries to obtain as much information as possible about the Target of Evaluation (TOE) design, i.e. we need to assume that Eve has a good knowledge about the specifications of the main components of the QKD system. Part of this information can indeed be found easily online. However, some important details might be system-specific or protected by a non-disclosure agreement between the vendor and the owner of the QKD system. For this reason, for both attacks, the Knowledge factor for the TOE factor is evaluated as *restricted*.

Both attacks rely on the intercept-resend strategy and can in principle be launched in real time. However, such online implementations of the attacks require to evaluate the optimal value of the displacement Δ and of the gain G (see methods): this can be obtained by manually tuning Eve's setup and measure the excess noise due to displacement, as in Fig. 8. Assuming a frequent trusted evaluation of the channel loss, this tuning might be quite challenging, especially in the case of the coherent attack, where the tuning precision is inevitably limited by the accuracy of the phase locking. As a result, for the coherent attack the Windows of Opportunity can be chosen as *difficult*, while *moderate* for the incoherent attack. The main differences between the two attack paths are related to the requirements in terms of equipment and expertise. As previously explained, the coherent attack requires Eve to resend coherent displaced signal while being successfully phase locked with Alice and Bob. To achieve this, Eve needs to be an *expert* in coherent optical communications, able to control noise at the quantum level and to have access to *bespoke* equipment. On the other hand, the incoherent attack only requires Eve to send an incoherent signal, without worrying about being phase locked with Alice and Bob: this is reflected in a simplified setup (Equipment *specialized*) and in a lower level of required technical expertise for Eve (Expertise *proficient*). From Table 3 we hence obtain an Attack Potential of 26 and 14 for coherent and incoherent attack respectively. As expected, the coherent attack is rated as *beyond high*, while the incoherent attack is only rated as *moderate*.

Discussion

Our experimental vulnerability analysis work, reported in the Results section, focuses on CV-QKD system. The attack rating methodology can apply equivalently to DV-QKD systems. Let us consider for example the 2011 DV-blinding attack²⁸, i.e. the first full-field attack against a commercial DV-QKD system. This DV blinding attack exploits a vulnerability of the single photon detectors: by injecting bright continuous-wave light and performing a full faked-state attack, an attacker Eve can take full control of Bob's detection outcomes. The similarities between the DV-blinding attack and our CV-incoherent attack strategy are reflected in the evaluation of their attack ratings. In the DV-blinding attack, Eve is required to know the specifications of the main components of the system, in particular Bob's detection part, and to tune her setup accordingly to Bob's. She moreover needs to align her polarization reference frame to match Bob's one. In both attack strategies, we can assign a restricted level for the Knowledge of the TOE and a moderate level for the Window of Opportunity. To successfully perform the DV-blinding attack, Eve needs to emulate Bob in the detection of the quantum state, using a duplicate of Bob's setup, and to deploy a faked-state generator (FSG) to force him to have her same detection values. The sophistication of the equipment and the expertise required is comparable to the CV-incoherent attack strategy: specialized equipment and proficient expertise. Summing all terms, the blinding attack rating would be moderate, with an Attack Potential of 14.

Quantifying the level of security assurance of a QKD device is a complex undertaking that should be based on a sound and largely recognized methodology. This objective translates into the requirement of standards for QKD security evaluation and certification. Such standards have recently started to be actively developed within international standardization groups^{29,30} and are now subject to an intense international effort.

The practical ability to evaluate QKD security will also require to set up evaluation lab facilities and to train "QKD security evaluation engineers", able to conduct penetration testing on QKD systems, both in terms of software and hardware. In that perspective, the experience accumulated in the context of classical secure hardware, and in particular the use of the Common Criteria methodology by the smartcard industry³¹, is invaluable.

The main message of this article is to point at the relevance of calculating Attack Potential to rate attacks against QKD device, following a methodology already in place to evaluate the security of classical cryptographic hardware. Moreover, Attack Potential can be used as a metric in order to balance the effort invested at QKD system design stage and at countermeasure development stage to thwart attacks, allowing to prioritize the attacks that constitute the most serious threats in practice.

One might however wonder whether this message should be read negatively, from a QKD viewpoint. Does it imply that the security QKD can provide essentially compares to the security that can be reached with classical hardware crypto-systems?

We want to argue that this not the case, for fundamental reasons: quantum crypto-systems strongly differ from their classical counterparts and provide a security advantage that is not only related to the information-theoretic security versus computational security paradigm. Quantum cryptography is moreover intrinsically based on models where the inner details of the physical layer are tied to information-theoretic measurable quantities. For instance, a functional QKD system is by definition sensitive to losses or errors occurring at single photon level. This is in strong contrast with classical systems, where information is typically encoded over a very large number of particles, such as classical optical pulses containing many photons. As a consequence, a classical system is oblivious to leakage occurring at the level of single quanta and cannot match the security level that can be provided, at least in principle, by a quantum crypto-system.

Considering the interplay between QKD implementation complexity and security also leads to an important reassessment. Theoretical security and practical security of a given QKD system may indeed significantly differ, notably when practical security is limited by engineering constraints. This calls to reconsider the absolute security claims sometimes associated with QKD and to adopt a more balanced viewpoint taking implementation complexity into consideration. We have depicted on Fig. 5 an case illustrating this situation: we consider two QKD protocols, (The term "protocol" has to be understood here in a holistic sense, ranging from the specifications of the physical implementation of the QKD devices, to the detail of the logical operations that they perform in order to establish a key.) P1 and P2, where protocol P1 has a more advanced security proof than protocol P2, therefore allowing to claim a higher security level in theory. However, it is possible that the protocol P2 has a lower implementation complexity than P1 and that, for the practically reachable implementation complexity corresponding to the engineering threshold, the practical security, i.e. the security that can be reached in practice by the QKD system, is larger with P2 than with P1.

Finally, the use of Attack Potential in QKD has also implications regarding the security that can be targeted. In particular, optimizing the security level of a given QKD device requires to first thwart attacks with the lowest Attack Potential before focusing on more complex ones. We have moreover explicitly demonstrated, on a live CV-QKD system, how different attacks related to the same theoretical vulnerability - i.e. the non-linear response of the homodyne receiver - can lead to different Attack Potentials. For a first attack path, detector saturation is reached using a coherent displacement. However, the practicality of this attack is limited due to noise generated from the imperfect phase drift compensation. The second attack path is on the other hand much more dangerous in practice: shining a simple external incoherent laser, it allows to drive the homodyne detector in the non-linear region of its characteristics and to precisely control the excess noise generated from Eve's intercept-resend attack, while meeting the conditions defined for the success of the attack.

Adapting existing criteria from IT security to the context of quantum cryptography is certainly a long and challenging path, but it is essential if we aim to make quantum devices relevant in the context of cyber security. We have summarized in Table 2 the results of our security evaluation procedure of two attack paths on CV-QKD, in terms of Attack Potential, illustrating that this methodology can constitute a useful step towards establishing forward-looking standards for the vulnerability assessment of QKD devices.

	Attack potential					Rating	Experimental results
	Exp	KoT	WoO	Equ	AP		
Coherent attack	6	3	10	7	26	Beyond high	✓ Noise model experimentally characterized
							× Attack not feasible under noise model
Incoherent attack	3	3	4	4	14	Moderate	✓ Attack experimentally demonstrated

Table 2. Summary of the analysis on the two attacks to the homodyne detection. We have reported the values for each factor of the Attack Potential, namely: Exp. stands for Expertise, KoT for Knowledge of the TOE, WoO for Window of Opportunity and Equ for Equipment. The factors chosen for the analysis are from Common Criteria²⁵.

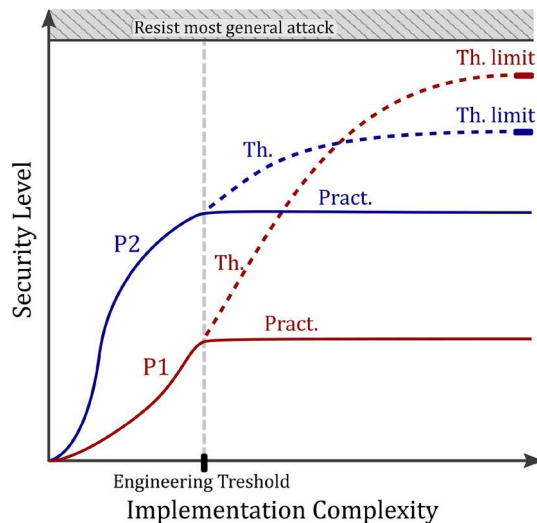


Figure 5. Pictorial representation of the possible divergence between theoretical (th.) and practical (pract.) QKD security. A QKD protocol P1 may have a stronger theoretical security (reachable for a perfect implementation) than another QKD protocol P2. Yet, in practice, QKD protocols can only be operated below a certain implementation complexity level materialized by the engineering threshold, and protocol P2 provides a stronger practical security than protocol P1.

Methods

Common criteria and attack rating. Common Criteria (CC) is the set of internationally recognized technical standards and configurations for security evaluations of Information Technology (IT) products and technology. The terminology and the concepts deployed in the CC aim to be as general as possible. Indeed, they are not intended to restrict the class of IT security problems of which CC is applicable, making them well suited to be extended for quantum communication devices, such as a QKD system. In simple terms, this comprehensive methodology aims at supporting the needs of three groups with a general interest in evaluation of the security properties of a certain Target of Evaluation (TOE): owners, developers and evaluators. In particular, what the owner of the TOE of the device wants is to protect his assets (any possible entity that he places value upon) from possible threat agents, i.e. someone or something that can abuse these assets against the interests of the owner.

In this context, we seek to offer a standardized framework for evaluating the risks associated to different threats and the effectiveness of the implemented countermeasures for quantum communication devices. The rating procedure consists in attributing a numeric value to the Attack Potential. In the Common Criteria framework, rating is performed by considering the following factors:

- Expertise
- Knowledge of the TOE
- Window of opportunity
- Equipment
- Elapsed time

Expertise refers to the level of technical expertise required to successfully perform the attack. Clearly an attack that can be mounted by a person with a regular level of education without an advanced knowledge in any specific field should be prioritized. The Knowledge of the TOE involves, instead, the amount of knowledge of the

Expertise	
Laymen	0
Proficient	3
Expert	6
Multiple experts	8
Knowledge of TOE	
Public	0
Restricted	3
Sensitive	7
Critical	11
Window of Opportunity	
Unnecessary / unlimited access	0
Easy	1
Moderate	4
Difficult	10
Equipment	
Standard	0
Specialized	4
Bespoke	7
Multiple bespoke	9

Table 3. Table for the evaluation of the Attack Potential used in the article. Elapsed time factor has not been considered: see text for explanations. For a complete guide on how evaluate those factors refer to the Common Evaluation Methodology version 3.1²⁵.

TOE design and operation required: retrieving detailed specifications about the device, for example, might be challenging for an attacker, leading to a higher Attack Potential. Regardless of the information acquired about the TOE, it is possible that, to successfully mount the attack, a previous tuning of the hacker's setup is needed. This aspect is considered in the Window of Opportunity, together with possible difficulties on getting access to the TOE. One last remarkable factor is the level of sophistication of the equipment used in the attack: an attack using equipment easy to obtain and simple to operate is obviously more dangerous than another attack that would require more advanced equipment. Quantum cryptography (QC) even often consider that the attacker Eve might have access to technology not available today such as large quantum computers or long-term quantum memories, and it is clear strength of QC to enable to prove security even in this context. From attack rating viewpoint, such "Quantum equipment" would be considered to have infinite rating, and is such not included in Table 3). To coherently consider these different factors and evaluate their contribution to the final Attack Potential we assign at each factor one numerical value, following Table 3.

In order to guarantee a consistent evaluation with respect to attacks to other TOEs, the different levels for each factor and their numerical values come from the Common Evaluation Methodology version 3.1²⁵ section B.4, where there is a full description of each possible level. Indeed, their description is particularly generic, to guarantee a meaningful characterization of an attack to a wide range of possible TOEs (such as a QKD device).

In the Common Criteria an additional factor is considered to rate the attacks: the elapsed time, i.e. the time taken to identify a certain vulnerability and to successfully mount the attack. To fix a correct timescale, this quantity needs to be compared with the usual time needed for a countermeasure to be applied. The vulnerability analysis that we report about in this article, has been performed on a laboratory QKD system. In this context, the main drivers of the elapsed time such as the product revision lifecycle, or the time during which an attacker could access the QKD system, cannot be meaningfully defined. For these reasons, we did not consider the elapsed time factor in our evaluation. We should however state that this factor will become well defined when considering the security of QKD products deployed on real-world networks and should hence be taken into account in future security evaluations of QKD, in conformity with Common Criteria.

In a complete vulnerability analysis, attack rating is sometimes split in two steps, for example in the case of smartcards³¹. The identification step is related to the effort required to create and apply the attack to the TOE for the first time. The exploitation step is then related to the effort required to apply the attack to the TOE knowing the techniques developed in the identification step. Both steps lead to a rating, based on the different factors. For the sake of simplicity, we have not distinguished these two steps in the present article, but we have just followed the general ground rules provided by Common Criteria. This can moreover be justified by the fact that the operational context associated to the exploitation step is essentially missing in the context of laboratory QKD prototype. We hence have rated attacks in a single step, based on the four factors out of five mentioned above (Expertise, Knowledge of the TOE, Window of opportunity and Equipment). This has lead us to adapt the rating methodology and the severity scale presented in Tables 1 and 3 with respect to the original tables from the Common Evaluation Methodology²⁵.

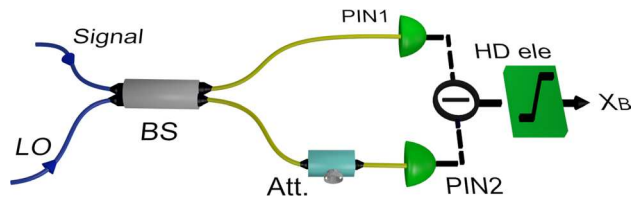


Figure 6. Balanced homodyne detection. An attenuator (Att) in one of the output port of 50/50 beamsplitter (BS) balances the photo current generated from photodiodes PIN1 and PIN2. Homodyne electronics circuit amplifies the subtracted photo currents.

Saturation of balanced homodyne detector. In a balanced homodyne detection, signal is mixed with intense local oscillator on a 50/50 beam splitter. Quadrature information is retrieved by subtracting the photocurrents generated from two photodiodes of identical detection efficiency, connected at the output ports of the beamsplitter. Due to imperfect splitting ratio of beam splitter, as well as efficiency mismatch of photodiodes, it is necessary to add appropriate attenuation at the respective output port of the beam splitter, see Fig. 6.

This equalizes the photocurrents and hence sets the mean of the output voltage of the homodyne detection close to zero. This is referred as balancing the homodyne or more precisely “balancing the homodyne with respect to the local oscillator port”. It has been shown that such balancing is essential to reduce the excess noise due to local oscillator intensity fluctuation³². In case of imperfect balancing one of the photodiodes generates more current than the other. As a result, the value of the homodyne output shifts towards the detection limit and this may lead to saturation.

The reason for saturation is due to the limited amplification factor of homodyne electronic circuitry. In our case circuit is made around Amptek A250 charge amplifier, powered by $\pm 5V$ power supply, that exhibits detection limit α_1 at $-2.5V$ in the negative DC level and α_2 at $+3.3V$ in the positive DC level (which is observed while interchanging photodiodes). Saturation behavior is also observed while setting low dynamic range of data acquisition card (say, $\pm 2V$) that is used to acquire homodyne output for post processing. In this work, we have set the data acquisition card range at $\pm 5V$, thus the linear range is limited solely by the homodyne electronic circuitry.

Parameter estimation. In CV-QKD system that uses GMCS protocol¹⁶, the quantum channel is characterized by its transmission T and its excess noise ξ . These parameters are estimated from Alice and Bob modulated and measured quadratures. Under saturation attack, these parameters are modified into T_{sat} and ξ_{sat} .

During the intercept-resend (IR) attack, the quadrature measured by Eve is: $X_M = X_A + X_0 + X'_0$, where X_0 is the vacuum noise quadrature due to Alice preparation and X'_0 is due to 3dB loss from Eve’s heterodyne measurement. The resent signal takes the form: $X_E = \sqrt{\frac{G}{2}}(X_M + X_{N_{A,E}}) + \Delta_X + X''_0$. Here, G is the amplification factor to compensate the loss from the heterodyne detection, $X_{N_{A,E}}$ accounts the technical noise from Alice and Eve, X''_0 is due to coherent state preparation by Eve. The term Δ_X determines the amount of shift in the mean value of quadrature. The same formalism holds true also for P quadrature. The parameter estimation takes the form²²:

$$T_{sat} = 2\langle X_A X_{B_{sat}} \rangle^2 / (G\eta_B V_A^2)$$

$$\xi_{sat} = \frac{2}{G\eta_B T_{sat}} \left(V_{B_{sat}} - G \frac{\eta_B T_{sat}}{2} V_A - N_0 - v_{ele} \right) \quad (1)$$

where $X_{B_{sat}}$ and $V_{B_{sat}}$ denote quadrature and its variance measured under saturation attack. One aspect of the attack worth mentioning here is that we assume that Eve does not tamper with the shot noise calibration phase.

Setup for coherent displacement. The experimental setup shown in Fig. 7 implements the resend session of the saturation attack. We have implemented CV-QKD *Eve_{resend}* system³³ using a Sagnac loop realized with variable beamsplitter (VBS). This allows to generate displaced GMCS signals as explained in Results. We have used a 1530.12nm pulsed laser of width 50ns, at a repetition rate of 1MHz, for generating this displaced signal. Displacing the signal is achieved as follows. The VBS, with splitting ratio $\approx 99.9\%$, splits the pulse from the circulator into two. Less intense signal pulse in clockwise direction goes under Gaussian modulation by amplitude modulator (AM1) and phase modulator (PM1) and further heavily attenuated by isolator (connected in reverse to achieve an attenuation higher than 30dB). High intense pulse travels along anti-clockwise directions, referred as pump pulse, meets the signal pulse at VBS and displaces it²⁴. The amplitude modulator AM2 controls the intensity of the pump and thence the amount of displacement Δ . A PIN diode attached to the VBS helps to monitor the stability of displacement operation. The Sagnac configuration helps to lock the relative phase of pump pulse and signal pulse to zero. Finally, the circulator directs the displaced signal towards the polarization beam splitter (PBS) that polarization multiplexes the local oscillator and displaced signal to the output fibre channel.

Setup for incoherent laser pulse injection. In this version of attack, Eve sends external laser pulse of 20ns width, along with signal pulse in the same polarization but at different wavelength (1550.12 nm). The signal laser and incoherent laser pulses are synchronized with proper delay. At Bob station, he performs the same

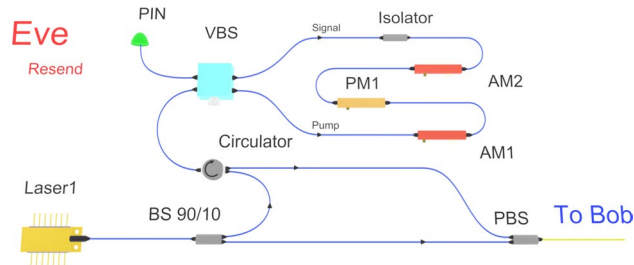


Figure 7. Experimental setup for generating displaced coherent state. AM: Amplitude Modulator, PM: Phase Modulator, BS: BeamSplitter. In the Sagnac loop, Gaussian modulated signals are prepared using the AM and PM modulators and are then displaced at the Variable Beam Splitter (VBS), based on a coherent interference between pump. Displaced signals is then sent to Bob along with local oscillator.

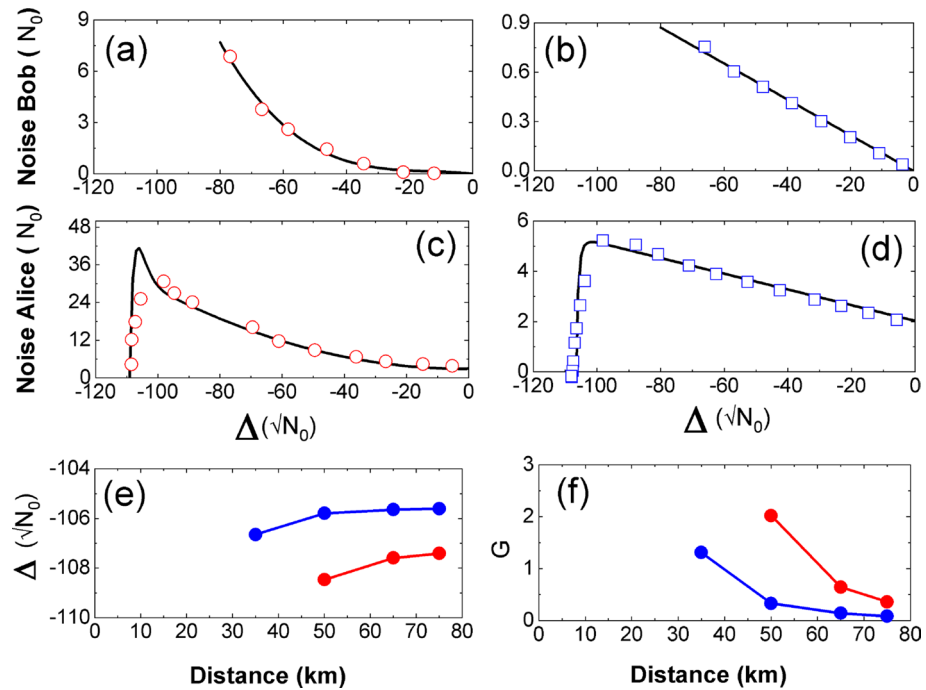


Figure 8. Excess noise due to displacement (a)–(d). Red circles and blue squares represents noise from coherent displacement and incoherent light, respectively. Black lines are theoretical fit with respective noise model. (a) and (b) show excess noise at Bob induced by Δ . Noise from coherent displacement shows quadratic behaviour while incoherent light adds noise from its own shot noise which is linear. Noise at Alice is shown in (c) and (d). (e)–(f) Optimal values of Δ and G at a different distances are calculated based on noise model from (a) and (b). Red and blue dots represent coherent and incoherent attack, respectively.

homodyne measurement as in the coherent attack strategy, where incoherent laser pulse is polarisation demultiplexed along with signal. It exploits two features of the homodyne setup: imbalance of the homodyne experienced by the light through signal port and also wavelength dependent splitting ratio of the beam splitter^{19,21,34}. By taking into account the wavelength dependent effect and the attenuator value adjusted for the local oscillator, the effective transmittance applied to the incoherent laser is approximately $T_{bs} \approx 49\%$, while the transmittance applied to local oscillator is about $T_{lo} \approx 50\% \pm 0.05\%$. Varying intensity of the incoherent light shifts the mean of the homodyne output towards the saturation limit α_1 and as a result affects the output variance.

Optimizing displacement Δ and gain G . In order to evaluate optimal values of Δ and G for successful attack, it is essential to characterize the noises associated with displacement as well as incoherent laser pulse. In the absence of resent signal, displacement pump/incoherent light is sent to Bob and amount of noise recorded for various values of Δ . This helps to model the excess noise at Bob, shown in Fig. 8a, b, and it is taken into account during optimization of Δ and G . The Fig. 8c, d show excess noise at Alice with respect to Δ . Value of detection limit α_1 is calibrated as $106\sqrt{N_0}$ ($-2.5V$ expressed in shot noise unit) for the optimization. For each transmission distance and for respective optimal V_A , Δ and G are calculated such that excess noise falls below the null key threshold. The optimal values are those that correspond to a maximum key rate, with $T_{sat} = T$, shown

in Fig. 8e, f. It can be seen that at a transmission distance shorter than 50 km and 35 km, respectively for coherent and incoherent attack strategy, no values of G and Δ are able to meet attack success conditions. The difference in feasible distances could have an impact on the attack potential as it affects the window of opportunity (WoO) rating factor (it is easier to meet attack conditions at shorter distance than longer distances). In the present case, the minimal distance to launch the attack is respectively 50 km for incoherent attack and 35 km for coherent attack. This difference is in line with the fact that have attributed a larger WoO rating to the coherent attack. In the incoherent attack strategy, the average power of the incoherent light required to reach the detection limit $\alpha_2 = 106\sqrt{N_0}$ is observed as 5.55 μW ³⁵.

Received: 18 December 2020; Accepted: 26 March 2021

Published online: 05 May 2021

References

- Bennett, C. H. & Brassard, G. Quantum cryptography: public key distribution and coin tossing. In *Proceedings IEEE International Conference on Computers, Systems and Signal Proceedings* 175–179 (1984).
- Gisin, N., Ribordy, G., Tittel, W. & Zbinden, H. Quantum cryptography. *Rev. Mod. Phys.* **74**, 145–195 (2002).
- Diamanti, E., Lo, H.-K., Qi, B. & Yuan, Z. Practical challenges in quantum key distribution. *npj Quant. Inf.* **2**, 1–12 (2016).
- Pirandola, S. *et al.* Advances in quantum cryptography. *Adv. Opt. Photon.* **12**, 1012–1236 (2020).
- Xu, F., Ma, X., Zhang, Q., Lo, H.-K. & Pan, J.-W. Secure quantum key distribution with realistic devices. *Rev. Mod. Phys.* **92**, 025002 (2020).
- Sasaki, M. Quantum key distribution and its applications. *IEEE Secur. Priv.* **16**, 42–48 (2018).
- Lewis, A. & Travagnin, M. A secure quantum communications infrastructure for Europe. *JRC Technical Papers, JRC116937* (2019).
- Renner, R. Security of quantum key distribution. *Int. J. Quant. Inf.* **6**, 1–127 (2008).
- Scarani, V. *et al.* The security of practical quantum key distribution. *Rev. Mod. Phys.* **81**, 1301 (2009).
- Lo, H.-K., Curty, M. & Tamaki, K. Secure quantum key distribution. *Nat. Photon.* **8**, 595–604 (2014).
- Lydersen, L. *et al.* Hacking commercial quantum cryptography systems by tailored bright illumination. *Nat. Photon.* **4**, 686–689 (2010).
- Gerhardt, I. *et al.* Full-field implementation of a perfect eavesdropper on a quantum cryptography system. *Nat. Commun.* **2**, 349 (2011).
- Brassard, G., Lütkenhaus, N., Mor, T. & Sanders, B. C. Limitations on practical quantum cryptography. *Phys. Rev. Lett.* **85**, 1330 (2000).
- Lvovsky, A. I., Sanders, B. C. & Tittel, W. Optical quantum memory. *Nat. Photon.* **3**, 706–714 (2009).
- Herrmann, D. S. *Using the Common Criteria for IT Security Evaluation* (CRC Press, 2002).
- Grosshans, F. & Grangier, P. Continuous variable quantum cryptography using coherent states. *Phys. Rev. Lett.* **88**, 057902 (2002).
- Grosshans, F. *et al.* Quantum key distribution using Gaussian-modulated coherent states. *Nature* **421**, 238–241 (2003).
- Häselser, H., Moroder, T. & Lütkenhaus, N. Testing quantum devices: practical entanglement verification in bipartite optical systems. *Phys. Rev. A* **77**, 032303 (2008).
- Huang, J.-Z. *et al.* Quantum hacking of a continuous-variable quantum-key-distribution system using a wavelength attack. *Phys. Rev. A* **87**, 062329 (2013).
- Ma, X.-C., Sun, S.-H., Jiang, M.-S. & Liang, L.-M. Local oscillator fluctuation opens a loophole for Eve in practical continuous-variable quantum-key-distribution systems. *Phys. Rev. A* **88**, 022339 (2013).
- Huang, J.-Z. *et al.* Quantum hacking on quantum key distribution using homodyne detection. *Phys. Rev. A* **89**, 032304 (2014).
- Qin, H., Kumar, R. & Alléaume, R. Quantum hacking: saturation attack on practical continuous-variable quantum key distribution. *Phys. Rev. A* **94**, 012325 (2016).
- Qin, H., Kumar, R., Makarov, V. & Alléaume, R. Homodyne-detector-blinding attack in continuous-variable quantum key distribution. *Phys. Rev. A* **98**, 012312 (2018).
- Paris, M. G. Displacement operator by beam splitter. *Phys. Lett. A* **217**, 78–80 (1996).
- Common Methodology for Information Technology Security Evaluation, Version 3.1, Revision 5. <https://www.commoncriteriaportal.org/files/ccfiles/CEMV3.1R5.pdf> (2017).
- Lodewyck, J. *et al.* Experimental implementation of non-Gaussian attacks on a continuous-variable quantum-key-distribution system. *Phys. Rev. Lett.* **98**, 030503 (2007).
- Kunz-Jacques, S. & Jouguet, P. Robust shot-noise measurement for continuous-variable quantum key distribution. *Phys. Rev. A* **91**, 022307 (2015).
- Gerhardt, I. *et al.* Full-field implementation of a perfect eavesdropper on a quantum cryptography system. *Nat. Commun.* **2**, 1–6 (2011).
- ISO/IEC JTC1 SC27 WG3, Security requirements, test and evaluation methods for quantum key distribution. <https://www.iso.org/standard/77097.html>.
- ETSI QKD-ISG DGS/QKD-016-PP, QKD Common Criteria Protection Profile for QKD. https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=58004.
- Application of Attack Potential to Smartcards and Similar Services. <https://www.sogis.eu/documents/cc/domains/sc/JIL-Application-of-Attack-Potential-to-Smartcards-v3-0.pdf> (2019).
- Chi, Y.-M. *et al.* A balanced homodyne detector for high-rate Gaussian-modulated coherent-state quantum key distribution. *New J. Phys.* **13**, 013003 (2011).
- Jouguet, P., Kunz-Jacques, S., Leverrier, A., Grangier, P. & Diamanti, E. Experimental demonstration of long-distance continuous-variable quantum key distribution. *Nat. Photon.* **7**, 378–381 (2013).
- Ma, X.-C., Sun, S.-H., Jiang, M.-S. & Liang, L.-M. Wavelength attack on practical continuous-variable quantum-key-distribution system with a heterodyne protocol. *Phys. Rev. A* **87**, 052309 (2013).
- Qin, H., Kumar, R. & Alléaume, R. Saturation attack on continuous-variable quantum key distribution system. In *Proc. SPIE 8899, Emerging Technologies in Security and Defence; and Quantum Security II; and Unmanned Sensor Systems X, 88990N*, vol. 8899, 88990N–88990N–7 (2013).

Acknowledgements

This research was supported by the European Union's Quantum Technology Flagship, through the Horizon 2020 research and innovation programme, under Grant agreements CIVIQ No 820466 and OPENQKD No

857156. H.Q is sponsored by Shanghai Pujiang Program. R. K. acknowledges support from EPSRC Grant no EP/T001011/1 (the EPSRC Quantum Communications Hub).

Author contributions

R.K. performed the experimental realization of the attacks, F.M. performed the attack rating analysis. H.Q. and R.A. developed the theoretical models of the attacks. R.A. supervised the Project. All authors contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-87574-4>.

Correspondence and requests for materials should be addressed to R.A.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021