Detection loophole attacks on semi-device-independent quantum and classical protocols

Michele Dall'Arno[a]

ICFO-Institut de Ciencies Fotoniques, Mediterranean Technology Park
Castelldefels (Barcelona), 08860, Spain

Graduate School of Information Science, Nagoya University
Chikusa-ku, Nagoya, 464-8601, Japan

Centre for Quantum Technologies, National University of Singapore
3 Science Drive 2, 117543 Singapore, Singapore

Elsa Passaro

ICFO-Institut de Ciencies Fotoniques, Mediterranean Technology Park
Castelldefels (Barcelona), 08860, Spain

Rodrigo Gallego

ICFO-Institut de Ciencies Fotoniques, Mediterranean Technology Park
Castelldefels (Barcelona), 08860, Spain

Dahlem Center for Complex Quantum Systems, Freie Universität Berlin,
14195 Berlin, Germany

Marcin Pawłowski

Institute of Theoretical Physics and Astrophysics, University of Gdansk,
80-952 Gdansk, Poland

Department of Mathematics, University of Bristol
Bristol BS8 1TW, United Kingdom

Antonio Acín

ICFO-Institut de Ciencies Fotoniques, Mediterranean Technology Park
Castelldefels (Barcelona), 08860, Spain

ICREA-Institucio Catalana de Recerca i Estudis Avançats
Lluis Companys 23, 08010 Barcelona, Spain

Semi-device-independent quantum protocols realize information tasks – e.g. secure key distribution, random access coding, and randomness generation – in a scenario where no assumption on the internal working of the devices used in the protocol is made, except their dimension. These protocols offer two main advantages: first, their implementation is often less demanding than fully-device-independent protocols. Second, they are more secure than their device-dependent counterparts. Their classical analogous is represented by random access codes, which provide a general framework for describing one-sided classical communication tasks. We discuss conditions under which detection inefficiencies can be exploited by a malicious provider to fake the performance of semi-device-independent quantum and classical protocols – and how to prevent it.

Keywords: semi-device-independent protocols, random access codes, detection loophole

[a]cqtmda@nus.edu.sg

1 Introduction

In the last decades, the distinguishing properties of quantum theory have been exploited to accomplish tasks which are unfeasible in classical theory [1]. For example, protocols were proposed for secure quantum key distribution (QKD) [2, 3], quantum teleportation [4, 5], and quantum randomness generation (QRG) [6, 7]. With no exceptions, the first protocols to be proposed were device dependent, namely their success critically relies on the agreement between the description of the setup and its implementation. Since this hypothesis is never exactly fulfilled, in experimental implementations a plethora of related problems arises. In fact, the unavoidable mismatch between theoretical requirements of security proofs and experimental implementations has recently been exploited to hack quantum cryptography systems [8, 9, 10, 11].

Subsequently, fully-device-independent protocols were proposed, in a scenario where devices are completely uncharacterized. Celebrated examples of device-independent protocols include schemes for randomness generation [12, 13, 14] or secure quantum key distribution [15, 16, 17]. While first security proofs of these protocols were partial and required some extra assumptions, such as, for instance, to assume that devices were memoryless, general security proofs have recently been obtained in [18, 19]. Unfortunately, these security proofs turn out to be quite demanding in terms of noise (in fact, the proof in [19] does not tolerate any noise). Moreover, from an implementation point of view, device-independent protocols are quite challenging, as they require a detection-loophole-free Bell inequality violation between two distant parties.

Recently, intermediate solutions between device-dependent and fully device-independent protocols have been proposed. For example, in the so-called measurement-device-independent (MDI) QKD [20, 21], the measurement device is left uncharacterized. While offering a weaker form of security than their device-independent counterparts, the implementation of MDI protocols is much easier as they can tolerate higher losses and their performance is comparable with that of standard schemes. Security of measurement-device-independent QKD protocols under detection loophole hacks is discussed in Ref. [21], and experimental implementations are reported in Refs. [22, 23].

Another example are the so-called semi-device-independent quantum protocols, where only the dimension of the exchanged system is assumed while the devices are uncharacterized. At the price of upper bounding the dimension of the system, secure QKD is possible [24] in a measure and prepare scheme, and semi-device-independent protocols for QRG are also known [25]. The importance of semi-device independent protocols relies on the fact that they are secure against a wide class of attacks, where the eavesdropper can alter the characteristics of systems and measurements, but can not change their dimensionality. For example, in the attacks proposed in Refs. [8, 9, 11], by flashing bright light into the devices, an eavesdropper gets control of detection efficiency for each measurement, but there is no way for him to make the devices encode the information in a system of a higher dimensionality.

Semi-device-independent protocols are based on the quantum certification provided by dimension witnesses for a fixed dimension [26]. While the recently developed formalism of dimension witness [26, 27, 28] allows to lower bound the dimension of a system in a device-independent fashion, a device-independent upper bound is prohibited by fundamental arguments. Indeed, for any given system, no experiment is guaranteed to exploit all the available

degrees of freedom. Then, only some a priori knowledge of the underlying physical model can provide a (device dependent) upper bound on the dimension of the system. On the other hand, it is worth to stress that in semi-device-independent protocols no additional assumption on the model needs to be done. For example, for information encoded in the polarization of a photon, only the unsurprising assumption that the system is two dimensional is required, while no assumption is required on what the polarization directions are.

The classical analogous of semi-device-independent quantum protocols – namely, the case in which the exchanged system is classical – is known as the problem of random access codes (RACs) [29]. In the context of RACs, the aim of two distant parties is to optimally perform some one-sided communication task under a constraint on the amount of classical information exchanged.

Despite their security, real world implementations of semi-device-independent (quantum or classical) protocols are subject to detection loophole (DL) attacks [a]– as happens for any fully-device-independent protocol. In this attack, a malicious provider exploits non-ideal detection efficiencies to skew the statistics of the experiment and ultimately faking its result. The aim of this work is to provide conditions under which DL attacks are harmless in faking the result of a semi-device-independent (quantum or classical) protocol.

The paper is organized as follows. In Sec. 2 we introduce DL attacks and present our main results. In Sec. 3 we derive conditions under which DL attacks on semi-device-independent quantum protocol are harmless, in the general framework where only the statistics of the protocol is taken into account. In Sec. 4 we address the problem of the certification of semi-device-independent classical protocols, in the framework of RACs. Finally we summarize our results and delineate some further developments in Sec. 5.

2 Detection loophole attack

We start by presenting the general structure of the semi-device-independent (quantum and classical) protocols considered in this work. The existing quantum protocols for QKD [24] and QRG [25], as well as classical RACs, are examples of this structure. We consider protocols in which two distant parties, Alice and Bob, have access to uncorrelated random number generators.[b] For each round, we denote by j (i) the random variable generated by Alice's (Bob's) generator and with q_j (p_i) its probability distribution. As said, these probability distributions are independent. Random variables j and i represent the strategy that Alice and Bob apply, respectively. This scheme is depicted in Fig. 1.

In each run, Alice and Bob get classical inputs a and b respectively. Alice sends a message A – which may be classical or quantum – to Bob, who then returns a classical value B. Finally they collect the statistics of several runs (the asymptotic case is always considered), obtaining

[a] The problem of DL attack on semi-device-independent protocols shares analogies with that of the robustness to loss of device-independent dimension witness (DIDWs) [26, 27, 28, 30], addressed in [31]. Nevertheless, while in the former the task is to exploit non-ideal detection efficiencies to fake the result of a protocol producing an input/output statistics which would be forbidden in the absence of DL, in the latter the task is to devise conditions under which dimension witnessing is indeed possible even in the presence of loss.

[b] Notice that the assumption of uncorrelation is fulfilled by a broad class of protocols. Indeed, in any semi-device-independent setup one necessarily has to assume that the devices are shielded – namely they can not communicate except through message A. Then to have shared (classical or quantum) randomness one is forced to introduce a trusted third party random generator, or to allow for infinite local memory on each device storing previously distributed randomness.

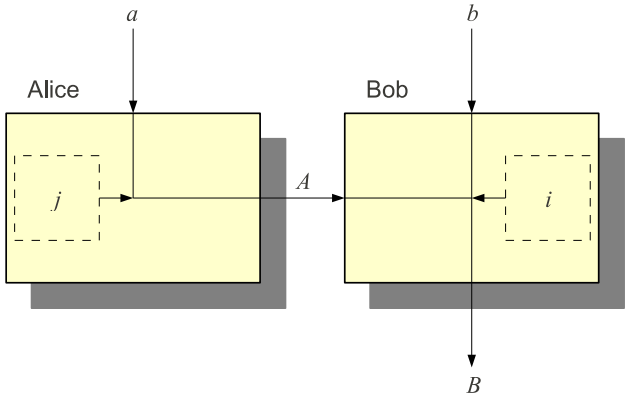

Fig. 1. Scheme of a generic semi-device-independent (quantum or classical) protocol. Two distant parties Alice and Bob are provided a black box each (bold-line boxes in the Figure). Alice's and Bob's boxes receive classical input a and b respectively. Each box is allowed to use a classical random generator (dashed-line boxes), which outcome – j for Alice's box and i for Bob's – is not accessible to the parties but can influence the outcome of the box. Alice's box sends a (quantum or classical) message A to Bob, that finally outputs classical message B.

the conditional probability distribution $P(B|a,b)$ of outcome B given inputs a and b, namely

$$P(B|a,b) := \sum_{i,j,A} p_i q_j P_i(B|A,b) P_j(A|a). \tag{1}$$

It is important to stress that – as Eq. (1) clearly shows – access is granted only to the inputs a, b and the output B, while no knowledge of the internal behavior of the black boxes (including the random variables i, j) and of the message A is provided. The goal is to exploit the correlations between the two parties, encapsulated by $P(B|a,b)$ to solve an information task, e.g. to distribute a secure key or generate random numbers.

When studying DL attacks, we assume that for each round of the experiment Alice or Bob can claim that their "detector did not click", and in this case this round of the experiment is discarded from the statistics. In general, Alices box can decide whether to send state A to Bob after receiving her input a and random variable j, while Bob's box after receiving his input b, the message A and random variable i. Thus, the detection efficiencies, i.e. the probabilities that the detector clicks, are denoted with $\eta_j(a)$ for Alice and $\eta_i(A,b)$ for Bob. Notice that these probabilities cannot be estimated as they depend on the variables i and j internal to the devices. The conditional probability distribution of outcome B given inputs a and b in the presence of a DL attack is given by

$$P_{DL}(B|a,b) := \frac{\sum_{i,j,A} p_i q_j \eta_i(A,b) \eta_j(a) P_i(B|A,b) P_j(A|a)}{\sum_{i,j,A} p_i q_j \eta_i(A,b) \eta_j(a) P_j(A|a)}. \tag{2}$$

We use the suffix DL whenever a distribution is obtained resorting to DL attack. Without loss of generality, we are assuming that for every input a, b there is a non-zero probability of click, namely denominator in Eq. (2) is strictly larger than 0 for any a and b. Indeed, if this is not the case, namely if there exist a and b such that the corresponding probability of click

is zero, one can discard their occurrences from the statistics without affecting the success of the protocol.

Notice that whether Alice uses DL is not relevant[c], since any settings she can prepare with DL can also clearly be achieved without resorting to it, so Eq. (2) can be simplified as

$$P_{DL}(B|a,b) = \frac{\sum_{i,A} p_i \eta_i(A,b) P_i(B|A,b) P(A|a)}{\sum_{i,A} p_i \eta_i(A,b) P(A|a)}, \tag{3}$$

where $P(A|a) := \sum_j q_j \eta_j(a) P_j(A|a) / \sum_j q_j \eta_j(a)$.

Independently of the task to be realized, all the known examples of semi-device-independent quantum protocols are based on the quantum certification provided by dimension witnesses [26] or, in other words, on the fact that, for a fixed dimension of the exchanged system A, there are quantum distributions $P(B|a,b)$ that cannot be attained when system A is classical - a system is classical if the states in which it can be prepared are pairwise commuting. This quantum certification plays here the same role as Bell violations for fully device-independent protocols. Our purpose is then to understand how a DL attack can mimic correlations that are intrinsically quantum exploiting the losses in the implementation. That is, rather than analyzing the effect of losses for a given quantum protocol, we study situations in which the observed correlations are useless for any quantum protocol. This is analogous to what is done when studying the detection loophole for Bell inequalities.

On the other hand, for classical protocols such a general approach is obviously not possible. However, when addressing the problem of classical RACS, one is usually interested in maximizing some figure of merit related to the particular communication task, such as the worst case or average probability of correct detection. In this work we will focus on the former - being the latter related by Yao's principle [29, 32] – and we will devise conditions under which it can not be improved resorting to DL attack.

3 Certification of semi-device-independent quantum protocols

In this Section, we focus on semi-device-independent quantum protocols, namely where the exchanged system A is quantum. As mentioned, the success of semi-device-independent quantum protocols depends on the generated statistics. Usually, for a given protocol, a large enough value of a particular function of such statistics ensures the success of the protocol. For instance, the protocol in [24] is secure only when it is assumed that the dimension of the measured systems is two and a large value of a dimension witness is observed. Yet, in general, a necessary condition for the successful performance of any protocol is the ability to discriminate whether the source is intrinsically quantum or it can be described as a classical distribution, building only on the knowledge of the conditional probability distribution $P(B|a,b)$. That is, it is necessary to certify that the observed correlations cannot be explained classically and, therefore, are potentially useful for quantum protocols without classical analogue. The advantage of this approach is that it allows one to evaluate necessary conditions

[c] This may be no more true if other constraints are introduced, since in this case the sets of distributions $P(A|a)$ and $P_{DL}(A|a)$ attainable by Alice can be different. For example, suppose that Alice is computationally constrained to prepare message A in time polynomial in the size of a. On the one hand, without resorting to DL it is impossible to obtain the distribution $P(A|a) = \delta_{A,f(a)}$, with $f(a)$ some NP-hard function (as long as we assume that $P \neq NP$). On the other hand, exploiting DL Alice can randomly choose A and check in polynomial time whether $A = f(a)$, clicking only in this case.

for security irrespectively of the particular protocol considered. Indeed, finding a DL attack able to fake an intrinsically quantum distribution by exploiting detection inefficiencies makes the observed correlations useless for any protocol. In this Section we provide conditions under which DL attack can by no means recast a classical $P(B|a,b)$ into an intrinsically quantum $P_{DL}(B|a,b)$ thus faking the result of the protocol.

We say that a conditional probability distribution $P(B|a,b)$ of outcome B given inputs a on Alice's side and b on Bob's side admits a classical (quantum) d-dimensional model if it can be written as

$$P(B|a,b) = \sum_{A,i} p_i P_i(B|A,b) P(A|a),$$

where

$$\sum_A P(A|a) = 1, \ \forall a, \quad \sum_B P_i(B|A,b) = 1, \ \forall A,b,i, \tag{4}$$
$$P(A|a) \geq 0 \ \forall A,a, \quad P_i(B|A,b) \geq 0 \ \forall B,A,b,i$$

for some probability p_i and where A is a classical (quantum) d-dimensional system. Given some correlations with losses, we say that DL attacks are harmless whenever there is no classical attack faking the correlations.

The probability of click on Bob's side given he received message A from Alice and input b is given by

$$Q(B \neq NC|A,b) := \sum_i p_i \eta_i(A,b),$$

where we denoted with NC the no-click event. The following Lemma shows that whenever

$$Q(B{\neq}NC|A,b) = Q(B{\neq}NC|b). \tag{5}$$

then, DL attacks are harmless.

Lemma 1 *If $Q(B{\neq}NC|A,b) = Q(B{\neq}NC|b)$ for any A,b, then if $P_{DL}(B|a,b)$ does not admit a d-dimensional classical (quantum) model then also $P(B|a,b)$ does not admit a d-dimensional classical (quantum) model.*

Proof: First we show that under the hypothesis $Q(B{\neq}NC|A,b) = Q(B{\neq}NC|b)$, if $P(B|a,b)$ admits a d-dimensional classical (quantum) model then also $P_{DL}(B|a,b)$ admits a d-dimensional classical (quantum) model.

Let then us assume that $P(B|A,b)$ admits a classical (quantum) model, namely it can be written as

$$P(B|a,b) = \sum_{A,i} p_i P_i(B|A,b) P(A|a),$$

with $P_i(B|A,b), P(A|a)$ satisfying Eq. (4) and for some probability p_i. Then by definition

$$P_{DL}(B|a,b) = \frac{\sum_{i,A} p_i \eta_i(A,b) P_i(B|A,b) P(A|a)}{\sum_{i,A} p_i \eta_i(A,b) P(A|a)}.$$

Upon introducing the hypothesis $Q(B{\neq}NC|A,b) = Q(B{\neq}NC|b)$ one has

$$P_{DL}(B|a,b) = \frac{\sum_{i,A} p_i \eta_i(A,b) P_i(B|A,b) P(A|a)}{Q(B{\neq}NC|b)}.$$

Upon setting

$$P_{DL}(B|A,b) = \frac{\sum_{i} p_i \eta_i(A,b) P_i(B|A,b)}{Q(B{\neq}NC|b)},$$

one clearly has $\sum_B P_{DL}(B|A,b) = 1$ and $P_{DL}(B|A,b) \geq 0$ for any B, A, b. Then $P_{DL}(B|a,b)$ admits the d-dimensional classical (quantum) model

$$P_{DL}(B|a,b) = \sum_A P_{DL}(B|A,b) P(A|a).$$

Then, whenever $P_{DL}(B|a,b)$ does not admit a d-dimensional classical (quantum) model, also $P(B|a,b)$ does not. □

At this point it is convenient to discuss condition (5) in relation with the so-called *fair sampling assumption*. The latter states that the set of events in which the detectors clicked is a randomly chosen sample from the total set of events that one would have obtained with perfect detectors. That is,

$$P_{DL}(B|a,b) = P(B|a,b). \tag{6}$$

One can clearly see by using Eq. (3), that (5) does not necessarily imply the fair sampling assumption. Indeed, in order to fulfill the fair sampling assumption for every choice of $\{p_i, P(A|a), P_i(B|A,b)\}$, one needs that $\eta_i(A,b) = \eta(b)$. On the other hand, in order to fulfill (5) for every function p_i suffices that $\eta_i(A,b) = \eta_i(b)$. In this sense, Lemma 1 generalizes the fair sampling assumption, providing strictly weaker hypothesis under which DL-attacks are harmless.

Nonetheless, the fair sampling assumption and our slightly more general condition (5) have in common that they refer to properties of the internal working of the devices. In particular, condition $\eta_i(A,b) = \eta_i(b)$ – or the more constraining fair sampling assumption – cannot be verified solely from the statistics, since the message A sent by Alice is not directly accessible to the parties.

Next Proposition provides a much stronger condition for DL attacks to be harmless, as it is stated only in terms of the probability $Q(B{\neq}NC|a,b)$ of click given inputs a on Alice's side and b on Bob's side, namely

$$Q(B{\neq}NC|a,b) := \sum_{i,A} p_i \eta_i(A,b) P(A|a). \tag{7}$$

Notice that this probability is accessible to the parties, being a function of the inputs a, b which are in turn accessible. The following proposition shows that whenever statistics of bidimensional systems fulfill

$$Q(B{\neq}NC|a,b) = Q(B{\neq}NC|b) \tag{8}$$

DL-attacks are harmless.

Proposition 1 *If $Q(B{\neq}NC|a,b) = Q(B{\neq}NC|b)$ for any a, b, then if $P_{DL}(B|a,b)$ does not admit a 2-dimensional classical model then also $P(B|a,b)$ does not admit a 2-dimensional classical model.*

Proof: By hypothesis, for any input a_0, a_1 on Alice's side one has

$$\sum_A Q(B{\neq}NC|A,b)\,[P(A|a{=}a_0) - P(A|a{=}a_1)] = 0,$$

where the sum is over $A = 0, 1$.

Rearranging explicitly the terms in previous Equation and using the fact $P(A = 1|a) = 1 - P(A = 0|a)$ for any a, one obtains that either

$$P(A{=}0|a{=}a_0) = P(A{=}0|a{=}a_1),$$

for any a_0, a_1, namely the message A sent by Alice is independent on her input a, or

$$Q(B{\neq}NC|A{=}0,b) = Q(B{\neq}NC|A{=}1,b),$$

for any b, namely the detection probability on Bob's side is independent on the message A received from Alice.

In the former case $P(B|a,b)$ clearly admits a classical local model, namely one in which no message is sent from Alice to Bob, and the same holds true for $P_{DL}(B|a,b)$ due to Eq. (3). In the latter case the hypothesis of Lemma 1 is satisfied, and thus the statement is proven. □

As said, contrary to Lemma 1, this result is much stronger, as it is proven under an assumption that can be verified only from the observed statistics. The price to pay is that it only holds for systems of dimension two. Condition (8) is, therefore, highly inequivalent to (5) or the fair sampling assumption. In fact, the attack presented in [8, 9] fulfills condition (8), however clearly violates the fair sampling assumption (but also violates the assumption on the dimension).

4 Certification of semi-device-independent classical protocols

In this Section, we focus on semi-device-independent classical protocols, namely where the exchanged system A is classical. We devise functions of the input/output statistics that can not be altered by DL attacks. Thus, any certification for semi-device-independent classical protocols building only on the value of these functions will be immune to DL attacks. Again, the main advantage is that, as above, these functions can be verified only from the observed statistics.

A semi-device-independent classical protocol can be viewed as a random access code [33, 34] (RAC), and in the following it will be convenient to work in the framework of RACs. In this framework, the aim of the two distant parties Alice and Bob is to optimally perform some communication task by means of one-sided communication of classical information. RACs are usually denoted with the notation $n \to m$. Here n is the number of input bits of Alice, namely the dimension of input a is $\dim(a) = 2^n$, while m is the number of bits sent by Alice, namely the dimension of message A is $\dim(A) = 2^m$ (see Fig. 1).

In this scenario, the relevant figures of merit usually considered are the worst case or the average success probability to have that $B = f(a,b)$ for a specific Boolean function

$f(a,b) \in \{0,1\}$. Here we will focus on the former, being the latter related through Yao's principle [32]. The worst case probability of success P^{wc} is defined as

$$P^{wc} := \min_{a,b} P(B{=}f(a,b)|a,b).$$

The probability that $B = f(a,b)$ with the DL attack is given by

$$P_{DL}(B{=}f(a,b)|a,b) = \frac{\sum_{i,A} w_i(A,a,b) P_i(B{=}f(a,b)|A,b)}{\sum_{i,A} w_i(A,a,b)}, \tag{9}$$

where $w_i(A,a,b) = p_i \eta_i(A,b) P(A|a)$ and the worst case probability that $B = f(a,b)$ is given by

$$P^{wc}_{DL} := \min_{a,b} P_{DL}(B{=}f(a,b)|a,b).$$

The following Proposition provides conditions under which the worst case success probability of a RAC can not be increased resorting to DL exploit. When these hypotheses are satisfied, a protocol relying on the worst case success probability may not be affected by DL attack.

Proposition 2 *Given a RAC, if the worst case success probability without resorting to DL attack is $P^{wc}{=}1/2$, then the worst case probability of success resorting to DL attack is $P^{wc}_{DL}{=}1/2$.*

Proof: The proof proceeds by absurd assuming $P^{wc} = 1/2$ and $P^{wc}_{DL} > 1/2$.

Equation (9) is the weighted sum over indices i and A of the numbers $P_i(B = f(a,b)|A,b)$ with weights $w_i(A,a,b)/\sum_{i,A} w_i(A,a,b)$ and therefore is upper bounded by

$$P_{DL}(B{=}f(a,b)|a,b) \leq \max_{A,i} \{P_i(B{=}f(a,b)|A,b)\},$$

and one has

$$P^{wc}_{DL} \leq \min_{a,b} \max_{A,i} \{P_i(B{=}f(a,b)|A,b)\}.$$

Since we are assuming $P^{wc}_{DL} > 1/2$ there exists a strategy i_0 of Bob and a message A_0 of Alice such that for all a,b one has $P_{i_0}(B = f(a,b)|A_0,b) > 1/2$. Then Bob can exploit a new strategy where he applies strategy i_0 whenever he gets A_0 and returns a random number otherwise, for which the probability $\tilde{P}(B{=}f(a,b)|a,b)$ of $B{=}f(a,b)$ given inputs a and b is given by

$$\tilde{P}(B{=}f(a,b)|a,b) = \left[P_{i_0}(B{=}f(a,b)|A_0,b) - \frac{1}{2}\right] P(A_0|a) + \frac{1}{2}.$$

This new strategy does not resort to DL and since $P_{i_0}(B = f(a,b)|A_0,b) > 1/2$ it has the worst case success probability greater than

$$\tilde{P}^{wc} = \min_{a,b} \{\tilde{P}(B{=}f(a,b)|a,b)\} > \frac{1}{2}$$

which contradicts the assumptions.□

The following Corollary shows that for any $n \to 1$ RAC, DL attacks are harmless.

Corollary 1 *For any $n \to 1$ RAC the worst case success probability resorting to DL attack is P_{DL}^{wc}=1/2.*

Proof: In [35] it was shown that for any $n \to 1$ RAC the hypothesis of Proposition 2 are fulfilled, namely $P^{wc} = 1/2$, so the statement follows.□

One may ask whether it is possible to relax the hypothesis of Proposition 2. We provide here an example of RAC with worst case success probability larger than 1/2, and show that this probability can be increased using DL attack. Consider the $3 \to \log 6$ RAC. Alice is given three independent bits a_0, a_1, a_2, namely $a = a_0 \otimes a_1 \otimes a_2$, and she can send to Bob a 6-dimensional message or, equivalently, one bit A_0 and one trit A_1, namely $A = A_0 \otimes A_1$. Bob's input is the trit $b = 0, 1, 2$ and the function to be computed is $f(a, b) = a_b$. Here we show that the worst case success probability P^{wc} without resorting to DL of $3 \to \log 6$ RAC is $P^{wc} < 0.981$, while there exists a DL attack such that the worst case success probability is $P_{DL}^{wc} = 1$.

First, we prove that for the $3 \to \log 6$ RAC one has $P^{wc} < 0.981$. An explicit upper bound for the worst case quantum success probability – which is clearly at least as large as the classical one P^{wc} – was derived in [36] in the context of quantum finite automata, namely

$$(1 - h(P^{wc}))n \leq m,$$

where $h(.)$ is the Shannon binary entropy function. Setting $n = 3$ and $m = \log 6$ we get $P^{wc} < 0.981$.

Now we provide a protocol using DL which achieves the worst case success probability $P_{DL}^{wc} = 1$. The idea is to use part of the communicated message to distribute randomness. Alice can choose the trit A_1 at random and encode $A_0 = a_{A_1}$, in other words she sends one of her bits randomly to Bob but also sends him information regarding which bit it is. If $b = A_1$ then Bob returns $B = A_0$ which is equal to a_b. If $b \neq A_1$ his detector does not click. The detection efficiency of Bob's device with this protocol is given by $\eta_i(A, b) = \delta_{b,A_1}$, and the worst case success probability is given by $P_{DL}^{wc} = 1$.

5 Conclusion

In this work we addressed the problem of how non-ideal detection efficiencies can be exploited to fake the result of semi-device-independent quantum and classical protocols through DL attacks.

For quantum protocols, we discussed general conditions under which DL attacks are harmless in terms of the detection probability. First, we proved in Lemma 1 that a sufficient condition is that the measurements can be modeled as ideal ones affected by non ideal detection efficiency - in terms of positive operator-valued measures (POVMs) [1], this means that the element corresponding to no-click event is proportional to the identity. Perhaps surprisingly, the "quantumness" of the generated statistics does not depend on the detection efficiency, as far as this is strictly larger than zero, thus making our certification strategies extremely robust. Unfortunately, this condition is not practical, as it requires some additional assumption on the

model underlying the measurements. Then, we derived in Proposition 1 a sufficient condition which is practical, namely it can be verified only from the knowledge of the input/output statistics. Also in this case the certification is extremely robust, as it can be obtained for any non-null value of the detection efficiency. Notice that, after the submission of this work, an analogous result for device independent dimension witnessing with uncorrelated devices was reported in Ref. [30]. There, indeed, it was shown that a particular dimension witness can certify the quantumness of a qubit for any non-null value of the detection efficiency.

For classical protocols, we provided conditions under which DL attacks can not increase the worst case success probability of a RAC. Our main results can be used as a guideline to devise quantum and classical protocols resistant to DL attacks, being thus of relevance for applications such as QKD, QRG, and RAC. Our approach for quantum protocols is very general, providing necessary conditions for the quantum certification of the devices. A natural follow-up question is to understand how DL attacks apply to specific examples of semi-device-independent quantum and classical protocols.

Some of the presented results – namely Proposition 1 and Corollary 1 - hold in the hypothesis that the message A sent by Alice is 2-dimensional. For the classical case, we showed through the example $3 \to \log 6$ RAC that this assumption can not be relaxed trivially. Thus, it remains as an open problem how to devise more general conditions under which DL attacks are harmless.

Acknowledgements

We are grateful to Nicolas Brunner for very useful discussions and suggestions. M. D. thanks Anne Gstottner and the Human Resources staff at ICFO for their invaluable support. This work was funded by the Spanish project FIS2010-14830 and Generalitat de Catalunya, UK EPSRC, the European PERCENT ERC Starting Grant and Q-Essence Project, the JSPS (Japan Society for the Promotion of Science) Grant-in-Aid for JSPS Fellows No. 24-0219, the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), NCN grant no. 2013/08/M/ST2/00626, the Ministry of Education and the Ministry of Manpower (Singapore).

References

1. I. L. Chuang and M. A. Nielsen (2000) *Quantum computation and quantum information*, Cambridge University Press.
2. C. H. Bennett and G. Brassard (1984) *Quantum cryptography: public key distribution and coin tossing*, Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing, pp. 175-179.
3. A. K. Ekert (1991), *Quantum cryptography based on Bell's theorem*, Phys. Rev. Lett., 67, pp. 661-663.
4. C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters (1993), *Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels*, Phys. Rev. Lett., 70, pp. 1895-1899.
5. D. Bouwmeester, J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger (1997), *Experimental quantum teleportation*, Nature, 390, pp. 575-579.
6. J. G. Rarity, P. C. M. Owens, and P. R. Tapster (1994), *Quantum random-number generation and key sharing*, J. Mod. Optics, 41, pp. 2435-2444.

7. M. Fiorentino, C. Santori, S. M. Spillane, R. G. Beausoleil, W. J. Munro (2007), *Secure self-calibrating quantum random-bit generator*, Phys. Rev. A, 75, 032334.
8. L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov (2010), *Hacking commercial quantum cryptography systems by tailored bright illumination*, Nature Phot., 4, pp. 686-689.
9. I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Skaar, C. Kurtsiefer, V. Makarov (2011), *Full-field implementation of a perfect eavesdropper on a quantum cryptography system*, Nature Commun., 2, 349.
10. F. Xu, B. Qi, and H. Lo (2010), *Experimental demonstration of phase-remapping attack in a practical quantum key distribution system*, New J. Phys. **12**, 113026.
11. N. Jain, C. Wittmann, L. Lydersen, C. Wiechers, D. Elser, C. Marquardt, V. Makarov, and G. Leuchs (2011), *Device Calibration Impacts Security of Quantum Key Distribution*, Phys. Rev. Lett. 107, 110501.
12. S. Pironio, A. Acín, S. Massar, A. Boyer de la Giroday, D. N. Matsukevich, P. Maunz, S. Olmschenk, D. Hayes, L. Luo, T. A. Manning, and C. Monroe (2010), *Random numbers certified by Bell's theorem*, Nature, 464, pp. 1021-1024.
13. R. Colbeck (2007), *Quantum and relativistic protocols for secure multi-party computation*, PhD dissertation, Univ. Cambridge.
14. R. Colbeck and A. Kent (2011), *Private randomness expansion with untrusted devices*, J. Phys. A, 44, 095305.
15. A. Acín, N. Brunner, N. Gisin, S. Massar, S. Pironio, V. Scarani (2007), *Device-independent security of quantum cryptography against collective attacks*, Phys. Rev. Lett., 98, 230501.
16. S. Pironio, A. Acín, N. Brunner, N. Gisin, S. Massar, and V. Scarani (2009), *Device-independent quantum key distribution secure against collective attacks*, New J. Phys., 11, 045021.
17. J. Barrett, L. Hardy and A. Kent (2005), *No Signalling Quantum Key Distribution* Phys. Rev. Lett., 95, 010503.
18. U. Vazirani and T. Vidick, *Fully device independent quantum key distribution*, arXiv:1210.1810.
19. B. W. Reichardt, F. Unger, and U. Vazirani, *A classical leash for a quantum system: Command of quantum systems via rigidity of CHSH games*, arXiv:1209.0448.
20. S. L. Braunstein and S. Pirandola (2012), *Side-Channel-Free Quantum Key Distribution*, Phys. Rev. Lett. 108, 130502.
21. H. Lo, M. Curty, and B. Qi (2012), *Measurement-Device-Independent Quantum Key Distribution*, Phys. Rev. Lett. 108, 130503.
22. Z. Tang, Z. Liao, F. Xu, B. Qi, L. Qian, and H.-K. Lo (2014), *Experimental Demonstration of Polarization Encoding Measurement-Device-Independent Quantum Key Distribution*, Phys. Rev. Lett. 112, 190503.
23. Y. Liu, T.-Y. Chen, L.-J. Wang, H. Liang, G.-L. Shentu, J. Wang, K. Cui, H.-L. Yin, N.-L. Liu, L. Li, X. Ma, J. S. Pelc, M. M. Fejer, C.-Z. Peng, Q. Zhang, and Jian-Wei Pan (2013), *Experimental Measurement-Device-Independent Quantum Key Distribution*, Phys. Rev. Lett. 111, 130502.
24. M. Pawłowski and N. Brunner (2011), *Semi-device-independent security of one-way quantum key distribution*, Phys. Rev. A, 84, 010302.
25. H.-W. Li, M. Pawłowski, Z.-Q. Yin, G.-C. Guo, and Z.-F. Han (2012), *Semi-device-independent randomness certification using $n \to 1$ quantum random access codes*, Phys. Rev. A, 85, 052308.
26. R. Gallego, N. Brunner, C. Hadley, and A. Acín (2010), *Device-independent tests of classical and quantum dimensions*, Phys. Rev. Lett., 105, 230501.
27. M. Hendrych, R. Gallego, M. Mičuda, N. Brunner, A. Acín, and J. P. Torres (2012), *Experimental estimation of the dimension of classical and quantum systems*, Nature Phys., 8, pp. 588-591.
28. J. Ahrens, P. Badziąg, A. Cabello, and M. Bourennane (2012), *Experimental device-independent tests of classical and quantum dimensions*, Nature Phys., 8, pp. 592-595.
29. A. Ambainis, D. Leung, L. Mancinska, and M. Ozols (2008), arXiv:0810.2937.
30. J. Bowles, M. T. Quintino, and N. Brunner (2013), *Certifying the dimension of classical and quantum systems in a prepare-and-measure scenario with independent devices*, Phys. Rev. Lett.

112, 140407 (2014).
31. M. Dall'Arno, E. Passaro, R. Gallego, and A. Acín (2012), *Robustness of device-independent dimension witnesses*, Phys. Rev. A, 86, 042312.
32. A. Yao (1977), *Probabilistic computations: Toward a unified measure of complexity*, Proceedings of the 18th IEEE Symposium on Foundations of Computer Science (FOCS), pp. 222-227.
33. A. Ambainis, A. Nayak, A. Ta-shma, and U. Vazirani (2002), *Dense quantum coding and quantum finite automata*, J. ACM, 49, pp. 496-511.
34. M. Hayashi, K. Iwama, H. Nishimura, R. Raymond, S. Yamashita (2006), *(4,1)-quantum random access coding does not exist – one qubit is not enough to recover one of the four bits*, New J. Phys, 8, 129.
35. A. Ambainis, A. Nayak, A. Ta-Shma, and U. Vazirani (1999), *Dense quantum coding and a lower bound for 1-way quantum automata*, Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC'99), pp. 376-383.
36. A. Nayak (1999), *Optimal lower bounds for quantum automata and random access codes*, Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science (FOCS'99), pp. 369-376.