From Practice to Theory: The "Bright Illumination" Attack on Quantum Key Distribution Systems

Rotem Liss(✉) and Tal Mor

Computer Science Department, Technion – Israel Institute of Technology,
Technion City, Haifa 3200003, Israel
{rotemliss,talmo}@cs.technion.ac.il

Abstract. The "Bright Illumination" attack [Lydersen et al., Nat. Photon. 4, 686–689 (2010)] is a practical attack, fully implementable against quantum key distribution systems. In contrast to almost all developments in quantum information processing (for example, Shor's factorization algorithm, quantum teleportation, Bennett-Brassard (BB84) quantum key distribution, the "Photon-Number Splitting" attack, and many other examples), for which theory has been proposed decades before a proper implementation, the "Bright Illumination" attack preceded any sign or hint of a theoretical prediction. Here we explain how the "Reversed-Space" methodology of attacks, complementary to the notion of "quantum side-channel attacks" (which is analogous to a similar term in "classical"—namely, non-quantum—computer security), has missed the opportunity of predicting the "Bright Illumination" attack.

Keywords: Quantum Cryptography · QKD · Security · Reversed-Space Attacks · Bright Illumination · Practice · Theory · Side-Channel Attacks

1 Introduction

In the area of quantum information processing, theory usually precedes experiment. For example, the BB84 protocol for quantum key distribution (QKD) was suggested in 1984 [2], five years before it was implemented [1], and it still cannot be implemented in a perfectly secure way even today [16,25]. The "Photon-Number Splitting" attack was suggested in 2000 [5,6], but it is not implementable today. Quantum computing was suggested in the 1980s (see, e.g., [7]), but no useful and universal quantum computer (with a large number of clean qubits) has been implemented until today [22]. The same applies to Shor's factorization algorithm [27,28], to quantum teleportation [3] (at least to some extent; see also [21]), and to many other examples.

In contrast to the above examples, the "Bright Illumination" attack against practical QKD systems was presented and fully implemented in 2010 [18], *prior* to any theoretical prediction of the possibility of such an attack. Here we ask the question: could the "Bright Illumination" attack have been theoretically predicted?

The work of T.M. and R.L. was partly supported by the Israeli MOD Research and Technology Unit, and by the Gerald Schwartz & Heather Reisman Foundation.

C. Martín-Vide et al. (Eds.): TPNC 2020, LNCS 12494, pp. 82–94, 2020.
https://doi.org/10.1007/978-3-030-63000-3_7

Quantum key distribution (QKD) makes it possible for two parties, Alice and Bob, to agree on a shared secret key. This task, that is impossible for two parties using only classical communication, is made possible by taking advantage of quantum phenomena: Alice and Bob use an insecure quantum channel and an authenticated (unjammable) classical channel. The resulting key is secure even against an adversary (Eve) that may use the most general attacks allowed by quantum physics, and remains secret indefinitely, even if Eve has unlimited computing power.

For example, in the BB84 QKD protocol [2], Alice sends to Bob N qubits, each of them randomly chosen from the set of quantum states $\{|0\rangle, |1\rangle, |+\rangle \triangleq \frac{|0\rangle+|1\rangle}{\sqrt{2}}, |-\rangle \triangleq \frac{|0\rangle-|1\rangle}{\sqrt{2}}\}$, and Bob measures each of them either in the computational basis $\{|0\rangle, |1\rangle\}$ or in the Hadamard basis $\{|+\rangle, |-\rangle\}$, chosen randomly. Thereafter, Alice and Bob post-process the results by using the classical channel. If Alice and Bob use matching bases, they share a bit (unless there is some noise or eavesdropping); if they use mismatching bases, Bob's results are random. Alice and Bob reveal their basis choices, and discard the bits for which they used mismatching bases. After that, they publicly reveal a random subset of their bits in order to estimate the error rate (then discarding those exposed test bits), aborting the protocol if the error rate is too high; and they perform error correction and privacy amplification processes, obtaining a completely secret, identical final key.

The security promises of QKD are true in theory, but its *practical* security is far from being guaranteed. The practical implementations of QKD use realistic photons; therefore, they deviate from the theoretical protocols, which use ideal qubits. These deviations make possible various attacks [16,24], related to the idea of "side-channel attacks" in classical (i.e., non-quantum) computer security.

For example, the "Photon-Number Splitting" (PNS) attack [5,6] (see Subsect. 2.1) takes advantage of some specific imperfections: while the quantum state sent by Alice *should* be encoded in a single photon, Eve exploits the fact that in most implementations, Alice sometimes sends to Bob more than one photon (e.g., two photons). The PNS attack was found using more realistic notations—the Fock space notations; the main insight of [5,6] is that using proper notations is vital, both when theoretically searching for possible loopholes and attacks against QKD, and when attempting to prove its security.

The "Bright Illumination" practical attack [18] uses a weakness of Bob's measurement devices that allows Eve to "blind" them and fully control Bob's measurement results. Eve can then get full information on the secret key, without inducing any error.

In Sect. 2 we explain experimental QKD systems and their weaknesses: we introduce the Fock space notations, the "Photon-Number Splitting" (PNS) attack, and two imperfections of Bob's detection process. In Sects. 3 and 4 we describe the practical "Bright Illumination" attack and the "Reversed-Space" methodology of attacks [4,10,11], respectively, and in Sect. 6 we bring together all the above notions for explaining the theory underlying the "Bright Illumination" attack. As an important side issue, in Sect. 5 we describe the notion of "quantum side-channel attacks", partially related to all the above. We conclude that while the "Bright Illumination" attack is not a "side-channel" attack, it can be modeled as a "Reversed-Space" attack [11]:

this attack and similar attacks could and should have been proposed or anticipated by theoreticians.

2 Experimental QKD, Imperfections, and the Fock Space Notations

The BB84 protocol may be experimentally implemented in a "polarization-based" implementation, that we can model as follows: Alice's quantum states, that are sent to Bob, are single photons whose polarizations encode the quantum states. The four possible states to be sent by Alice are $|0\rangle$, $|1\rangle$, $|+\rangle$, and $|-\rangle$, where $|0\rangle = |\leftrightarrow\rangle$ (a single photon in the horizontal polarization) and $|1\rangle = |\updownarrow\rangle$ (a single photon in the vertical polarization). The states $|+\rangle = |\swarrow\!\!\!\!\!\nearrow\rangle$ and $|-\rangle = |\nwarrow\!\!\!\!\!\searrow\rangle$ correspond to orthogonal diagonal polarizations.

For measuring the incoming photons, Bob uses a polarizing beam splitter (PBS) and two detectors. Bob actively configures the PBS for choosing his random measurement basis (the computational basis $\{|0\rangle, |1\rangle\}$ or the Hadamard basis $\{|+\rangle, |-\rangle\}$). If the PBS is configured for measurement in the computational basis, it sends any *horizontally* polarized photon to one arm and sends any *vertically* polarized photon to the other arm. In the end of each arm, a detector is placed, which clicks whenever it detects a photon. Therefore, the detector in the first arm clicks *only* if the $|0\rangle$ qubit state is detected, and the detector in the second arm clicks *only* if the $|1\rangle$ qubit state is detected. A *diagonally* polarized photon (i.e., $|+\rangle = |\swarrow\!\!\!\!\!\nearrow\rangle$ or $|-\rangle = |\nwarrow\!\!\!\!\!\searrow\rangle$) would cause exactly one of the detectors (uniformly random) to click. Similarly, if the PBS is configured for measurement in the Hadamard basis, it distinguishes $|+\rangle$ from $|-\rangle$. This implementation may be slow, because Bob needs to randomly choose a basis for each arriving photon.

A more practical—yet imperfect—variant of this implementation uses a "passive" basis choice (e.g., [15]). This variant uses one polarization-independent beam splitter, two PBSs, and *four* detectors. In this variant, the polarization-independent beam splitter randomly sends each photon to one arm or to another. A photon going to the first arm is then measured (as described above) in the computational basis, while a photon going to the second arm is measured (as described above) in the Hadamard basis. This "passive" variant is exposed to various attacks; see Sect. 4.

2.1 The Fock Space Notations and the "Photon-Number Splitting" (PNS) Attack

We use the Fock space notations for describing practical QKD systems:

- In the simplest case, there are $k \geq 0$ photons, and all these photons belong to *one* photonic mode. The Fock state $|k\rangle$ represents k photons in this single mode: for example, $|0\rangle$ is the vacuum state, representing no photons in that mode; $|1\rangle$ represents one photon in that mode; $|2\rangle$ represents two photons in that mode; and so on.
- For describing several different *pulses* of photons (for example, photons traveling on different arms or at different time bins, or any other *external* degree of freedom), we need several photonic *modes*. For example, if we assume a single photon in two

pulses (and, thus, in two modes), we can describe a qubit[1]: for the computational basis $\{|\mathrm{o}\rangle, |\imath\rangle\}$ of a single qubit, we write $|\mathrm{o}\rangle = |0\rangle \otimes |1\rangle \equiv |0\rangle|1\rangle$ and $|\imath\rangle = |1\rangle \otimes |0\rangle \equiv |1\rangle|0\rangle$. (Those two modes are mathematically described using a tensor product, but we omit the $\otimes$ sign for brevity.) A superposition, too, describes a single photon in those two pulses: for example, the Hadamard basis states are $|\pm\rangle = \frac{|0\rangle|1\rangle \pm |1\rangle|0\rangle}{\sqrt{2}}$.

- More generally, if we have $k = k_1 + k_0$ photons in two different pulses (two modes), where k_1 photons are in one pulse and k_0 photons are in the other pulse, we write $|k_1\rangle|k_0\rangle$. Subscripts are added for specifying the types of pulses—for example, $|k_1\rangle_{t_1}|k_0\rangle_{t_0}$ for the two time bins t_1, t_0, or $|k_1\rangle_A|k_0\rangle_B$ for the two arms A, B.
- For describing more than two pulses (namely, more than two modes), we use generalized notations: for example, $k = k_2 + k_1 + k_0$ photons in three time bins are denoted $|k_2\rangle_{t_2}|k_1\rangle_{t_1}|k_0\rangle_{t_0}$. In particular, the vacuum state (absence of photons) is denoted $|0\rangle$ for one mode, $|0\rangle|0\rangle$ for two modes, $|0\rangle|0\rangle|0\rangle$ for three modes, and so on.

The above notations assume the photon *polarizations* (which are an *internal* degree of freedom) to be identical for all k photons. However, a single photon in a single pulse generally has two orthogonal polarizations: horizontal $\leftrightarrow$ and vertical $\updownarrow$. For each pulse, the two polarizations are described as two modes; therefore, m pulses mean $2m$ modes.

In this paper, we denote *polarization* modes of $k = k_1 + k_0$ photons by $|k_1, k_0\rangle$ (without any subscript), and denote only *pulse* modes by $|k_1\rangle|k_0\rangle$ (always with subscripts). Thus:

- For a *single* pulse, the two *polarization* modes describe a qubit if there is exactly *one photon* in the pulse. The computational basis states are $|\mathrm{o}\rangle = |0, 1\rangle$ (representing *one* photon in the horizontal polarization mode and *zero* photons in the vertical polarization mode) and $|\imath\rangle = |1, 0\rangle$ (where the single photon is in the vertical mode).
- Similarly to the above, we can also describe: (a) superpositions; (b) the state $|k_1, k_0\rangle$ of $k = k_1 + k_0$ photons in those two polarization modes (k_1 photons in the vertical mode and k_0 photons in the horizontal mode); and (c) the vacuum state $|0, 0\rangle$.

We have seen that the Fock space notations extend *much* beyond the ideal single-qubit world, which is represented by the two-dimensional space $\mathrm{Span}\{|\mathrm{o}\rangle, |\imath\rangle\}$. Ideally, in BB84, Alice should send a qubit in this two-dimensional space; however, in practice, Alice sometimes sends states in a higher-dimensional Fock space.

The “Photon-Number Splitting” (PNS) attack [5,6] (which showed all QKD experiments done until around 2000 to be insecure) is based on analyzing the *six*-dimensional Hilbert space $\mathrm{Span}\{|0,0\rangle, |0,1\rangle, |1,0\rangle, |0,2\rangle, |2,0\rangle, |1,1\rangle\}$, which represents all typical pulses with two polarizations if we can neglect the case of three or more photons—namely, if we assume $k_1 + k_0 \leq 2$. The PNS attack is based on three observations [5,6]: (a) Alice sometimes sends *two-photon pulses* in one of the four allowed polarizations; (b) Eve can, in principle, distinguish a two-photon pulse from a single-photon pulse without influencing the polarizations; and (c) Eve can, in principle, split such a two-photon pulse into two pulses, each containing a single photon, without influencing the polarizations. Thus, Eve can “steal” a single photon from each such two-photon pulse (without influencing the other photon), save it, and, after learning the basis, get full

[1] The notations $|\mathrm{o}\rangle, |\imath\rangle, |\pm\rangle$ are used for the standard qubit (in a two-dimensional Hilbert space).

information about this pulse without being noticed. This attack could have been detrimental to the security of QKD, but counter-measures [13,17,23,30] have been found later.

2.2 Imperfections of Bob's Detectors

Two important examples of imperfections (see [10]) are highly relevant to various "Reversed-Space" attacks. As we show in this paper, those two imperfections must be *combined* for understanding the "Bright Illumination" attack.

Imperfection 1: Our realistic assumption, which is true for standard detectors in QKD implementations, is that Bob's detectors cannot *count* the number of photons in a pulse. Thus, they cannot distinguish *all* Fock states $|k\rangle$ from one another, but can only distinguish the Fock state $|0\rangle$ (a lack of photons) from the Fock states $\{|k\rangle : k \geq 1\}$. Namely, standard detectors can only decide whether the mode is empty ($k = 0$) or has at least one photon ($k > 0$). In contrast, we assume that Eve can (in principle) do anything allowed by the laws of quantum physics; in particular, Eve may have such "photon counters".

In particular, let us assume that there are *two* pulses, each of them consisting of a single mode. Bob cannot know whether a pulse contains one photon or two photons; therefore, he cannot distinguish between $|1\rangle|0\rangle$ and $|2\rangle|0\rangle$ (and, similarly, he cannot distinguish between $|0\rangle|1\rangle$ and $|0\rangle|2\rangle$). For example, assume that Alice sends the $|1\rangle|0\rangle$ state (a qubit) to Bob, and Eve replaces Alice's state by $|2\rangle|0\rangle$ and sends it to Bob instead (or, similarly, assume that Eve replaces $|0\rangle|1\rangle$ by $|0\rangle|2\rangle$). In this case, Bob cannot notice the change, and no error can occur; still, Bob got a state he had not expected to get. It may be possible for Eve to take advantage of this fact in a fully-designed attack.

Imperfection 2: Our realistic assumption is that Bob cannot know exactly *when* the photon he measured arrived. For example (in a polarization-based implementation):

- Alice's ideal qubit arrives at time t (states denoted $|0,1\rangle_t|0,0\rangle_{t+\delta}$, $|1,0\rangle_t|0,0\rangle_{t+\delta}$).
- Eve's photon may arrive at time $t + \delta$ (states denoted $|0,0\rangle_t|0,1\rangle_{t+\delta}$, $|0,0\rangle_t|1,0\rangle_{t+\delta}$).

Again, Eve may take advantage of this fact in a fully-designed attack.

Similar imperfections can be found if Bob cannot know exactly what the *wavelength* of the photon is, or *where* the photon arrives.

The conceptual difference between the two imperfections is in whether Bob can (ideally) avoid measuring the extra states sent by Eve, or not:

- In Imperfection 1, Eve may send more than one photon, and Bob must measure the state (while he cannot count the number of photons using current technology).
- In Imperfection 2, Eve sends states in two separate subsystems. Bob can, in principle, ignore the "wrong" subsystem in case he knows for sure it has not been sent by Alice.

3 The “Bright Illumination” Attack

The “Bright Illumination” blinding attack [18] works against QKD systems that use Avalanche Photodiodes (APDs) as Bob’s detectors. As an example, we describe below the implementation of this attack against a system implementing the BB84 protocol in a polarization-based scheme, but it is important to note that the attack can be adapted to most QKD protocols and implementations that use APDs [18].

The APDs can be operated in two “modes of operation”: the “linear mode” that detects only a light beam above a specific power threshold, and “Geiger mode” that detects even a single photon (but cannot count the number of photons). In this attack, the adversary Eve sends a continuous strong light beam towards Bob’s detectors, causing them to operate *only* in the linear mode (thus “blinding” the detectors).

After Bob’s detectors have been blinded (and in parallel to sending the continuous strong beam, making sure they are kept blind), Eve performs a “measure-resend” attack: she detects the qubit (single photon) sent by Alice, measures it in one of the two bases (exactly as Bob would do), and sends to Bob a *strong* light beam depending on the state she measured, a little above the power threshold of the detectors. For example, if Eve measures the state $|1,0\rangle$, she sends to Bob the state $|k,0\rangle$ for $k \gg 1$. Now, if Bob chooses the same basis as Eve, he will measure the same result as Eve; and if Bob chooses a different basis, he will measure nothing, because the strong light beam will get split between the two detectors. This means that Bob will always either measure the same result as Eve or lose the bit.

In the end, Bob and Eve have exactly the same information, so Eve can copy Bob’s classical post-processing and get the same final key as Alice and Bob do. Moreover, Eve’s attack causes no detectable disturbance, because Bob does not know that his detectors have operated in the wrong mode of operation; the only effect is a loss rate of 50% (that is not problematic: the loss rate for the single photons sent by Alice is usually much higher, so Eve can cause Bob to get the same loss rate he expects to get).

This attack was both developed and experimentally demonstrated against commercial QKD systems by [18]. See [18] for more details and for diagrams.

4 “Reversed-Space” Attacks

The “Reversed-Space” methodology, described in [8,10,11], is a theoretical framework of attacks exploiting the imperfections of Bob. This methodology is a special case (easier to analyze) of the more general methodology of “Quantum Space” attacks [8,9], that exploits the imperfections of *both* Alice and Bob; the “Reversed-Space” methodology assumes Alice to be ideal and only exploits Bob’s imperfections [4,8,10,11]. (Another special case of a “Quantum Space” attack is the PNS attack [5,6] described above.)

In the ideal QKD protocol, Bob expects to get from Alice a state in the Hilbert space $\mathscr{H}^{\mathrm{A}}$; however, in the “Reversed-Space” attack, Bob gets from Eve an unexpected state, residing in a larger Hilbert space called the “space of the protocol” and denoted by $\mathscr{H}^{\mathrm{P}}$. In principle, Eve could have used a huge space $\mathscr{H}'$ such that $\mathscr{H}^{\mathrm{A}} \subseteq \mathscr{H}^{\mathrm{P}} \subseteq \mathscr{H}'$: the huge Hilbert space $\mathscr{H}'$ consists of *all* the quantum states that Eve *can possibly* send to Bob, but it is too large, and most of it is irrelevant.

Because “Reversed-Space” attacks assume a “perfect Alice” (sending prefect qubits), it is usually easy to find the *relevant* subspace $\mathscr{H}^{\mathrm{P}}$, as we demonstrate by three examples below; $\mathscr{H}^{\mathrm{P}}$ is only enlarged (relative to the ideal space $\mathscr{H}^{\mathrm{A}}$) by Bob’s imperfections. Therefore, $\mathscr{H}^{\mathrm{P}}$ is the space that includes all the states that may be useful for Eve to send to Bob. The space $\mathscr{H}^{\mathrm{P}}$ is defined by taking all the possible measurement results of Bob and reversing them in time; more precisely, it is the span of all the states in $\mathscr{H}^{\mathrm{A}}$ *and* all the states that Eve can send to Bob so that he gets the measurement results she desires.

Whether Bob is aware of it or not, his experimental setting treats not only the states in $\mathscr{H}^{\mathrm{A}}$, but all the possible inputs in the “space of the protocol” $\mathscr{H}^{\mathrm{P}}$. Bob then classifies them into three classes: (1) valid states from Alice, (2) losses, and (3) invalid states. *Valid states* are always treated in conventional security analysis: a random subset is compared with Alice for estimating the error rate, and then the final key is obtained using the error correction and privacy amplification processes. *Losses* are expected, and they are not counted as noise. *Invalid states* are usually counted as errors (noise), but they do not appear in ideal analyses of ideal protocols. We note that loss rate and error rate are computed separately: the error rate must be small (e.g., around 10%) for the protocol not to be aborted by Alice and Bob, while the loss rate can be much higher (even higher than 99%). Any “Reversed-Space” attack takes advantage of the possibility that Bob treats some states in $\mathscr{H}^{\mathrm{P}}$ in the wrong way, because he does not expect to get those states.

Eve’s attack is called “Reversed-Space” because Eve can devise her attack by looking at Bob’s possible measurement results: Eve finds a measurement result she wants to be obtained by Bob (because he interprets it in a way desired by her) and reverses the measurement result in time for finding the state in $\mathscr{H}^{\mathrm{P}}$ she should send to Bob. In particular, if Bob applies the unitary operation $\mathscr{U}_{\mathrm{B}}$ on his state prior to his measurement, Eve should apply the inverted operation $\mathscr{U}_{\mathrm{B}}^{-1} = \mathscr{U}_{\mathrm{B}}^{\dagger}$ to each state corresponding to each possible measurement outcome of Bob.

We present three examples of “Reversed-Space” attacks. For simplicity, we only consider BB84 implemented in a polarization-based scheme (as described in Sect. 2), but the attacks may be generalized to other implementations, too. We emphasize that all three examples have been chosen to satisfy two conditions, also satisfied by the “Bright Illumination” attack: (a) Eve performs a “measure-resend” attack in a basis she chooses randomly, and (b) it is possible for Eve to get full information without inducing noise.

Example 1 (a special case of the “Trojan Pony” attack [12]): This example exploits Imperfection 1 and assumes Bob uses an “active” basis choice (see Sect. 2 for both).

In this attack, Eve performs a “measure-resend” attack—namely, she measures each qubit state sent from Alice to Bob in a random basis, and resends “it” towards Bob. However, instead of resending it as a single photon, she resends a huge number of photons towards Bob: she sends many *identical* photons, all with the same polarization as the state she measured ($|0\rangle$, $|1\rangle$, $|+\rangle$, or $|-\rangle$). If Bob chooses the same basis as Eve, he will get the same result as her, because Imperfection 1 causes his system to treat the incoming states $|0,k\rangle$ and $|k,0\rangle$ (for any $k \geq 1$) as if they were $|0,1\rangle$ and $|1,0\rangle$, respectively; but if he chooses a different basis from Eve, both of his detectors will (almost surely) click. If Bob decides to treat this *invalid* event (a two-detector click) as an "error", the error rate will be around 50%, so Alice and Bob will abort the protocol; but if he naively decides to treat this event as a "loss", Eve can get full information without inducing errors.

Alice sends an ideal qubit (a single photon), while Eve may send any number of photons. Therefore, using the Fock space notations, $\mathscr{H}^{\mathrm{A}} = \mathscr{H}_2 \triangleq \mathrm{Span}\{|0,1\rangle, |1,0\rangle\}$ and $\mathscr{H}^{\mathrm{P}} = \mathrm{Span}\{|m_1,m_0\rangle : m_1,m_0 \geq 0\}$.

Example 2 (a special case of the "Faked States" attack [8,19,20]): This attack exploits Imperfection 2 (Sect. 2). We assume that Bob has four detectors (namely, that he uses the "passive" basis choice variant of the polarization-based encoding; see Sect. 2), and that his detectors have different (but overlapping) *time gates* during which they are sensitive: given the three different times $t_0 < t_{1/2} < t_1$, the detectors for the computational basis are sensitive only to pulses sent at t_0 or $t_{1/2}$ (or in between), and the detectors for the Hadamard basis are sensitive only to pulses sent at $t_{1/2}$ or t_1 (or in between). Alice normally sends her pulses at $t_{1/2}$ (when both detectors are sensitive), but Eve may send her pulses at t_0, $t_{1/2}$, or t_1.

Eve performs a "measure-resend" attack by measuring Alice's state in a random basis, and resending it towards Bob as follows: if Eve measures in the computational basis, she resends the state at time t_0; and if she measures in the Hadamard basis, she resends the state at time t_1. Therefore, Bob gets the same result as Eve if he measures in the same basis as hers, but he gets a loss otherwise (because Bob's detectors for the other basis are not sensitive at that timing). This means that Eve gets full information without inducing any error.

Using the same notations as in Imperfection 2, the state $|m_1,m_0\rangle_{t_0}|n_1,n_0\rangle_{t_{1/2}}|o_1,o_0\rangle_{t_1}$ consists of the Fock states $|m_1,m_0\rangle$ sent at time t_0, $|n_1,n_0\rangle$ sent at time $t_{1/2}$, and $|o_1,o_0\rangle$ sent at time t_1. Alice sends an ideal qubit (a single photon at time $t_{1/2}$), while Eve may send a single photon at any of the times t_0, $t_{1/2}$, or t_1, or a superposition.

Therefore, $\mathscr{H}^{\mathrm{A}} = \mathscr{H}_2 \triangleq \mathrm{Span}\{|0,0\rangle_{t_0}|0,1\rangle_{t_{1/2}}|0,0\rangle_{t_1}\ ,\ |0,0\rangle_{t_0}|1,0\rangle_{t_{1/2}}|0,0\rangle_{t_1}\}$ and $\mathscr{H}^{\mathrm{P}} = \mathrm{Span}\{|0,1\rangle_{t_0}|0,0\rangle_{t_{1/2}}|0,0\rangle_{t_1}\ ,\ |1,0\rangle_{t_0}|0,0\rangle_{t_{1/2}}|0,0\rangle_{t_1}\ ,\ |0,0\rangle_{t_0}|0,1\rangle_{t_{1/2}}|0,0\rangle_{t_1}\ ,\ |0,0\rangle_{t_0}|1,0\rangle_{t_{1/2}}|0,0\rangle_{t_1}\ ,\ |0,0\rangle_{t_0}|0,0\rangle_{t_{1/2}}|0,1\rangle_{t_1}\ ,\ |0,0\rangle_{t_0}|0,0\rangle_{t_{1/2}}|1,0\rangle_{t_1}\}$.

Example 3 (the "Fixed Apparatus" attack [4]) can be applied by Eve if Bob uses a "passive" basis choice (Sect. 2). In this attack, Eve sends to Bob an unexpected state, and this state "forces" Bob to obtain the basis Eve wants. This attack makes it possible for Eve to force Bob choose the same basis as her (and, therefore, get the same outcome as her), thus stealing the whole key without inducing any errors or losses. The attack is only possible if Eve has a one-time access to Bob's laboratory, because it requires Eve to first compromise Bob's device (otherwise, she cannot send him that unexpected state).

Assume that Bob uses a polarization-independent beam splitter that splits the incoming beam into two different output arms (as described in Sect. 2). This beam splitter has two input arms: a *regular arm*, through which the standard incoming beam comes, and a *blocked arm*, where the incoming state is always assumed to be the zero-photon beam $|0,0\rangle$ (the vacuum state of two polarizations). If Eve can drill a small hole in Bob's device, exactly where the blocked arm gets its input from, then she can send a beam to the blocked arm and not only to the standard arm. It is proved [4] that Eve can then cause the beam splitter to choose an output arm to her desire, instead of choosing a "random" arm. The state $|m_1,m_0\rangle_r|n_1,n_0\rangle_b$ consists of the Fock state $|m_1,m_0\rangle$ sent through the *regular arm* of the beam splitter and the Fock state $|n_1,n_0\rangle$ sent through the *blocked arm*. Alice sends an ideal qubit (a single photon through the regular arm), while Eve may send a single photon through any of the two arms or a superposition. Therefore, $\mathscr{H}^{\mathrm{A}} = \mathscr{H}_2 \triangleq \mathrm{Span}\{|0,1\rangle_r|0,0\rangle_b \ , \ |1,0\rangle_r|0,0\rangle_b\}$ and $\mathscr{H}^{\mathrm{P}} = \mathrm{Span}\{|0,1\rangle_r|0,0\rangle_b \ , \ |1,0\rangle_r|0,0\rangle_b \ , \ |0,0\rangle_r|0,1\rangle_b \ , \ |0,0\rangle_r|1,0\rangle_b\}$.

5 Quantum Side-Channel Attacks

Shamir's "Quantum Side-Channel Attack" on Polarization-Based QKD: The following attack was proposed by Adi Shamir in a meeting with one of the authors (T.M.) around 1996–1999 [26], and it may have never been published (but see similar attacks below). Shamir's attack only applies to QKD implementations that use "*active*" basis choice (as opposed to the "passive" basis choice, which leads to the "Fixed Apparatus" attack described in Example 3 of Sect. 4). The attack is related to Imperfection 2 described in Sect. 2: Bob's apparatus must be fully or partially ready to receive Alice's photon before it arrives. For example, if the photon is supposed to arrive at time t, then Bob's setup is already partially ready at time $t-\delta$; in particular, Bob decides the *basis choice* and configures the polarizing beam splitter accordingly before time $t-\delta$. The attack also assumes that the detectors themselves are still inactive (blocked) at time $t-\delta$, and are activated just before time t. Therefore, at time $t-\delta$, the polarizing beam splitter is already configured to match the required basis (the computational basis or the Hadamard basis), while the detectors are still blocked.

Eve's attack is sending a strong pulse at time $t-\delta$, that hits the polarizing beam splitter (but not the blocked detectors) and gets reflected back to Eve, containing full or partial information on the direction of the polarizing beam splitter—and, thus, on the basis choice. Assuming Eve gets the information on Bob's basis choice *before* she receives Alice's pulse, Eve could employ the following full attack: Eve measures the photon coming from Alice *in the same basis chosen by Bob*, learns the qubit's value, and resends to Bob the resulting state (in the same basis), obtaining full information while inducing no errors and no losses.

One can suggest two ways to possibly prevent the attack: (a) opening the detection window (activating the detectors) *shortly* after the polarizing beam splitter is configured according to the basis choice (if the time difference is sufficiently short, Eve cannot find Bob's basis choice on time for employing the full attack); or (b) blocking access to the polarizing beam splitter until the detectors are activated (although this solution may be hard to implement).

As we explain in Sect. 6, the "Bright Illumination" attack could have been predicted by adding Imperfection 1 described in Sect. 2 (namely, detection of multi-photon pulses) to the above idea of a strong pulse sent at time $t-\delta$ towards Bob (i.e., Imperfection 2, as already discussed here) and using the Fock space notations.

"Conventional Optical Eavesdropping" and "Quantum Side-Channel Attacks": Other attacks, similar to Shamir's attack, have been independently developed—for example, the "Large Pulse" attack [29], which attacks both Alice's and Bob's set-ups. As written in [29]: "This [large pulse] attack is one of the possible methods of conventional optical eavesdropping, a new strategy of eavesdropping on quantum cryptosystems, which eliminates the need of immediate interaction with transmitted quantum states. It allows the eavesdropper to avoid inducing transmission errors that disclose her presence to the legal users."

Instead of restricting ourselves to "conventional optical eavesdropping on quantum cryptosystems", we make use of a different sentence from [29]—"eavesdropping on quantum cryptosystems which eliminates the need of immediate interaction with transmitted quantum states"—and we define "quantum side-channel attacks" as follows:

> A *quantum side-channel attack* is any eavesdropping strategy which eliminates the need of any immediate interaction with the transmitted quantum states.

According to the above definition, both Shamir's attack and the "Large Pulse" attack are "quantum side-channel attacks", because they attack the devices and not the quantum states themselves. On the other hand, the "Reversed-Space" attacks and the "Quantum Space" attacks (see Sect. 4) can be fully described using a proper description of the QKD protocol, which uses the Fock space notations; therefore, they should *not* be considered as "quantum side-channel attacks". In fact, we can say they are *complementary* to "quantum side-channel attacks", and we name them "*state*-channel attacks".

In a classical communication world, the notion of "side-channel attacks" makes use of any information leaked by the *physical* execution of the algorithm (see, for example, [14]). Accordingly, other researchers (e.g., [24]) have chosen to adopt a wide definition of "quantum side-channels", which also includes the "Photon-Number Splitting" attack and many other practical attacks. However, we prefer to take a narrower view of "quantum side-channel attacks", as explained above.

6 From Practice to Theory: The Possibility of Predicting the "Bright Illumination" Attack

The "Bright Illumination" attack could have been predicted, because it simply combines Imperfections 1 and 2 that were described in Sect. 2: namely, detecting many photons at time $t-\delta$, while the single "information" photon should have arrived at time t. In some sense, it seems to merge a "Reversed-Space" attack and a "quantum side-channel attack", because it attacks both the transmitted quantum states and the detectors themselves. However, because Bob's detectors are fully exposed to Eve at both times t and $t-\delta$ (unlike the "Large Pulse" attack [29], where the detectors are not exposed at time $t-\delta$), we see the "Bright Illumination" attack as a special (and fascinating) case of "Reversed-Space" attack, and not as a "quantum side-channel attack".

The "Bright Illumination" attack is made possible by a *lack of information* on the "space of the protocol" $\mathscr{H}^{\mathrm{P}}$: Eve sends many photons (as in Imperfection 1) at time $t-\delta$ (as in Imperfection 2), and Bob does not notice her disruption because he cannot *count* the number of photons and cannot *block* the detectors at time $t-\delta$.

For preventing all the possible attacks and proving full security, it must be known how Bob's detectors treat *any* number k of photons sent to him by Eve, and it must also be known how Bob's detectors treat multiple pulses. In particular, a detector definitely cannot operate properly in the hypothetical scenario where an infinite number of photons (with infinite energy) arrives as its input. A potentially secure system must have an estimated threshold N, such that if $k \lesssim N$ photons arrive, they are correctly measured by the detectors (treated as one photon), and if $k \gtrsim N$ photons arrive, the measurement result is clearly invalid and is known to Bob (for example, smoke comes out of the detectors, or the detectors are burned). N is estimated, so there is a small unknown range near it.

Prior to the "Bright Illumination" attack, it seems that no systematic effort has been invested in finding or approximating the threshold N and characterizing the detectors' behavior on *all* possible inputs (any number of photons k). A proper "Reversed-Space" analysis would have suggested that experimentalists *must* check what N is and fully analyze the behavior of Bob's detectors on each quantum state; such an analysis would then have found the "space of the protocol" $\mathscr{H}^{\mathrm{P}}$ which is available for Eve's attack.

A careful "Reversed-Space" analysis—if it had been carried out—would then have found that instead of *one* estimated threshold N (with some small unknown range around it), there are *two* estimated thresholds N_1, N_2, such that $N_1 < N_2$, with a some small unknown range around each of them, and a *large* difference between them. Therefore, there are three main ranges of the numbers of photons k: (a) for $k \lesssim N_1$ photons, Bob's detectors work well (and click if at least one photon arrives); (b) for $N_1 \lesssim k \lesssim N_2$ photons, it would have become *known* that some strange phenomena happen—for example, that Bob's detectors switch to the "linear mode"; and (c) for $k \gtrsim N_2$ photons, Bob's detectors malfunction (e.g., the detectors are burned).

Thus, surprisingly, even if the experimentalist had not known about the two modes of operation ("Geiger mode" and the "linear mode") existing for each detector, he or she could still have discovered the two different thresholds N_1, N_2 and then investigated the detectors' behavior in the middle range $N_1 \lesssim k \lesssim N_2$. This would have allowed him or her to discover the "linear mode" and realize that there is also a need to check *multiple* pulses for finding the correct "space of the protocol" and for analyzing the security against "Reversed-Space" attacks. Namely, the "Reversed-Space" approach makes it possible to discover attacks even if the detectors are treated as *a black box* whose internal behavior is unknown. By theoretically trying to prove security against any theoretical "Reversed-Space" attack, it would have been possible to find the practical "Bright Illumination" attack; it would have even been possible to study the operation of a "*black-box*" detector and discover, for example, that it has a "linear mode" of operation (even if this mode of operation had not been already known for realistic detectors).

7 Conclusion

We have seen a rare example (in quantum information processing) where experiment preceded theory. We can see now that this experimental attack could have been theoretically predicted: for a system to be secure, Bob must be sure that Eve cannot attack by sending an unexpected number of photons, and he must know what happens to his detectors for any number of photons. Otherwise—Eve can attack; and we could have known that this may be possible. We have also defined the general notion of “quantum side-channel attacks”, distinguishing “state-channel attacks” (including “Reversed-Space” and “Quantum Space” attacks) that interact with the transmitted (prepared or measured) quantum states, from “quantum side-channel attacks” that *do not interact* with the transmitted quantum states.

References

1. Bennett, C.H., Bessette, F., Brassard, G., Salvail, L., Smolin, J.: Experimental quantum cryptography. J. Cryptol. **5**(1), 3–28 (1992). https://doi.org/10.1007/BF00191318
2. Bennett, C.H., Brassard, G.: Quantum cryptography: public key distribution and coin tossing. In: International Conference on Computers, Systems & Signal Processing, pp. 175–179 (1984)
3. Bennett, C.H., Brassard, G., Crépeau, C., Jozsa, R., Peres, A., Wootters, W.K.: Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. Phys. Rev. Lett. **70**, 1895–1899 (1993). https://doi.org/10.1103/PhysRevLett.70.1895
4. Boyer, M., Gelles, R., Mor, T.: Attacks on fixed-apparatus quantum-key-distribution schemes. Phys. Rev. A **90**, 012329 (2014). https://doi.org/10.1103/PhysRevA.90.012329
5. Brassard, G., Lütkenhaus, N., Mor, T., Sanders, B.C.: Limitations on practical quantum cryptography. Phys. Rev. Lett. **85**, 1330–1333 (2000). https://doi.org/10.1103/PhysRevLett.85.1330
6. Brassard, G., Lütkenhaus, N., Mor, T., Sanders, B.C.: Security aspects of practical quantum cryptography. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 289–299. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45539-6_20
7. Deutsch, D.: Quantum theory, the Church-Turing principle and the universal quantum computer. P. Roy. Soc. Lond. A Mat. **400**, 97–117 (1985). https://doi.org/10.1098/rspa.1985.0070
8. Gelles, R.: On the security of theoretical and realistic quantum key distribution schemes. Master’s thesis, Technion - Israel Institute of Technology, Haifa, Israel (2008)
9. Gelles, R., Mor, T.: Quantum-space attacks. arXiv preprint arXiv:0711.3019 (2007)
10. Gelles, R., Mor, T.: Reversed space attacks. arXiv preprint arXiv:1110.6573 (2011)
11. Gelles, R., Mor, T.: On the security of interferometric quantum key distribution. In: Dediu, A.-H., Martín-Vide, C., Truthe, B. (eds.) TPNC 2012. LNCS, vol. 7505, pp. 133–146. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33860-1_12
12. Gottesman, D., Lo, H.K., Lütkenhaus, N., Preskill, J.: Security of quantum key distribution with imperfect devices. Quantum Inf. Comput. **4**, 325–360 (2004)
13. Hwang, W.Y.: Quantum key distribution with high loss: toward global secure communication. Phys. Rev. Lett. **91**, 057901 (2003). https://doi.org/10.1103/PhysRevLett.91.057901
14. Köpf, B., Basin, D.: An information-theoretic model for adaptive side-channel attacks. In: Proceedings of the 14th ACM Conference on Computer and Communications Security, CCS 2007, pp. 286–296 (2007)
15. Kurtsiefer, C., et al.: A step towards global key distribution. Nature **419**, 450 (2002). https://doi.org/10.1038/419450a
16. Lo, H.K., Curty, M., Tamaki, K.: Secure quantum key distribution. Nat. Photon. **8**, 595–604 (2014). https://doi.org/10.1038/nphoton.2014.14910.1038/nphoton.2014.149
17. Lo, H.K., Ma, X., Chen, K.: Decoy state quantum key distribution. Phys. Rev. Lett. **94**, 230504 (2005). https://doi.org/10.1103/PhysRevLett.94.230504

18. Lydersen, L., Wiechers, C., Wittmann, C., Elser, D., Skaar, J., Makarov, V.: Hacking commercial quantum cryptography systems by tailored bright illumination. Nat. Photon. **4**, 686–689 (2010). https://doi.org/10.1038/nphoton.2010.214
19. Makarov, V., Anisimov, A., Skaar, J.: Effects of detector efficiency mismatch on security of quantum cryptosystems. Phys. Rev. A **74**, 022313 (2006). https://doi.org/10.1103/PhysRevA.74.022313
20. Makarov, V., Hjelme, D.R.: Faked states attack on quantum cryptosystems. J. Mod. Opt. **52**, 691–705 (2005). https://doi.org/10.1080/09500340410001730986
21. Pfaff, W., et al.: Unconditional quantum teleportation between distant solid-state quantum bits. Science **345**, 532–535 (2014). https://doi.org/10.1126/science.1253512
22. Preskill, J.: Quantum computing in the NISQ era and beyond. Quantum **2**, 79 (2018)
23. Scarani, V., Acín, A., Ribordy, G., Gisin, N.: Quantum cryptography protocols robust against photon number splitting attacks for weak laser pulse implementations. Phys. Rev. Lett. **92**, 057901 (2004). https://doi.org/10.1103/PhysRevLett.92.057901
24. Scarani, V., Bechmann-Pasquinucci, H., Cerf, N.J., Dušek, M., Lütkenhaus, N., Peev, M.: The security of practical quantum key distribution. Rev. Mod. Phys. **81**, 1301–1350 (2009). https://doi.org/10.1103/RevModPhys.81.1301
25. Scarani, V., Kurtsiefer, C.: The black paper of quantum cryptography: real implementation problems. Theor. Comput. Sci. **560**, 27–32 (2014). https://doi.org/10.1016/j.tcs.2014.09.01510.1016/j.tcs.2014.09.015
26. Shamir, A.: Personal communication (around 1996–1999)
27. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings 35th Annual Symposium on Foundations of Computer Science, pp. 124–134 (1994)
28. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Rev. **41**, 303–332 (1999). https://doi.org/10.1137/S0036144598347011
29. Vakhitov, A., Makarov, V., Hjelme, D.R.: Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography. J. Mod. Opt. **48**, 2023–2038 (2001). https://doi.org/10.1080/09500340108240904
30. Wang, X.B.: Beating the photon-number-splitting attack in practical quantum cryptography. Phys. Rev. Lett. **94**, 230503 (2005). https://doi.org/10.1103/PhysRevLett.94.230503