Quantum key distribution (QKD) promises information theoretic security that is guaranteed by the laws of quantum mechanics [1]. Its potential as a cryptographic primitive has stimulated significant developments in recent years and has resulted in pilot network field trials in several continents [2–8]. However, actual components in QKD implementations can deviate from their ideal behaviour, creating side-channels that might be exploited by an eavesdropper (Eve), thus threatening the theoretical security promised by QKD [9–14].

To mitigate this security risk, active or passive countermeasures can be considered. In an active approach, the legitimate QKD users (Alice and Bob) monitor in real time the device parameters that change under Eve’s attack [15, 16]. In a passive approach, they add extra guarding components to thwart Eve’s attempt [17, 18]. Using the Trojan-horse attack [19, 20] as an example, Alice and Bob can employ either a watchdog detector to actively detect Eve’s presence [21] or a combination of an optical filter, attenuator and isolator to passively prevent Eve’s light from reaching the encoding devices [17].

Single photon detectors are the most vulnerable component in a QKD system, especially when inappropriately operated [22, 23], due to their optical exposure to Eve through the quantum channel. Those based on semiconductor avalanche photodiodes (APDs) can be attacked using a strong optical signal in the so-called “blinding attack” [9, 10]. The severity of this attack has motivated a number of active countermeasures, including monitoring the APD current [22], varying [24, 25] or calibrating [16] in real time its detection efficiency, or using a watchdog detector [26]. Extraordinarily, detector side-channel free QKD [27, 28] closes all side channels in the measurement devices, but requires significant changes in users’ apparatus and does not offer an easy upgrade to existing systems.

Self-differencing (SD) InGaAs APDs are an important class of detectors for high bit rate QKD systems [29, 30]. They enable detection of extremely weak signals using a passive circuit for cancellation of the intense background capacitive response [31], thus supporting count rates in the GHz range [32], detection efficiency up to 55% [33], room-temperature operation [34] and superior resilience to background noise photons [35]. On the other hand, the self-cancellation nature of the SD circuit prevents the detection of consecutive identical signals [36]. This might be exploited by Eve to perform the blinding attack with a lower optical power [37], especially for certain settings of the detector [23].

Here, we propose taking preemptive action against blinding by placing an intensity modulator (IM) in front of the receiver’s measurement apparatus. The use of low extinction ratio modulation will not severely attenuate the incoming quantum signal, but the IM will create sufficient modulation in the detector’s photocurrent, which is detectable by SD circuitry and discrimination electronics. We experimentally demonstrate this method on an SD InGaAs APD.

This idea is similar in spirit to the random variation of an APD’s detection efficiency, suggested and partially implemented in [24], which was shown to be ineffective against a refined Eve’s attack [25]. However, our proposal contains some notable differences. An APD endowed with an SD circuit sets a more challenging target to the eavesdropper. In fact, Eve has to send sequential light pulses with identical intensity to cause blinding. Any small deviation from this condition is likely to cause a detection event with an associated 50% quantum bit error rate (QBER). At the same time, a long sequence of bright optical pulses generates a high photocurrent which is easily detectable [38]. When an IM

Quantum key distribution is rising as an important cryptographic primitive for protecting the communication infrastructure in the digital era. However, its implementation security is often weakened by components whose behavior deviates from what is expected. Here, we analyse the response of a self-differencing avalanche photodiode, a key enabler for high speed quantum key distribution, to intense light shone from a continuous-wave laser. Under incorrect settings, the cancellation entailed by the self-differencing circuitry can make the detector insensitive to single photons. However, we experimentally demonstrate that even in such cases intensity modulation can be used as an effective measure to restore the detector’s expected response to the input light.
with random modulation is added on top of this already compelling situation, the constraints for Eve become extremely stringent. In particular, we found no room for blinding in our experiment.

II. BLINDING ATTACKS AND COUNTERMEASURES

Single photon sensitivity of an APD relies on having an electrical excess bias above its breakdown voltage to enable avalanche multiplication of a single photo-carrier. In the blinding attacks, Eve erodes this excess by inducing an electrical current flowing through the biasing circuit [22] or heating up the device to raise its breakdown voltage [39]. These attacks are realised through injecting a strong laser signal into Bob’s module through the quantum channel and making both of his detectors blind, i.e., insensitive to single photons. At this point, Eve performs a modified intercept-resend attack to take control of Bob’s detectors. She measures the state prepared by Alice and re-sends a suitably prepared faked state encoded in a strong optical pulse. This will then trigger a detector count only when Bob chooses a measurement basis that is identical to Eve’s. In this attack, Eve can gain full information about the final key [40].

Consider an APD detector operated in Geiger mode with an excess voltage of $V^\text{ex}_{\text{apd}}$, as shown in Fig. 1(a). To completely erode this excess and blind the detector, Eve needs to create a photocurrent $I$ that can be approximated as

$$I = \frac{V^\text{ex}_{\text{apd}}}{R_{\text{bias}} + R_{\text{apd}} + R_s/2},$$

where $R_{\text{bias}}$, $R_{\text{apd}}$ and $R_s$ are the resistance values for the biasing resistor, the APD itself and the sensing resistor, respectively. In a usual setup, $R_{\text{bias}} = 0$ and the current is determined mainly by the value of $R_{\text{apd}}$. Its typical value is on the order of 1 mA, see Fig. 1(b). This large current, together with the gain modulation effect by the detector gating, has previously enabled gated-APDs to avoid the blinding attack when their discrimination levels are appropriately set [22].

Here we propose a different measure, schematically shown in Fig. 1(c). We insert an IM, driven by a quantum random number generator (QRNG), in front of the optical fibre input of the APD detector and an SD circuit after its electrical signal output. The SD circuit splits the APD output into two equal components, delays one of them, and then combines the two signals differentially, see Fig. 1(d). The positive peaks of the resulting photocurrent can then be detected by the discrimination electronics.

The IM acts as an optical shutter and stops any incoming light for a short duration at random times. Under normal conditions, i.e., in the absence of Eve, this would cause a decrease in the counts seen by Bob every time the IM is activated. Correspondingly, the resulting avalanche current would exhibit a waveform containing a positive current peak followed by a negative dip. On the contrary, if Eve sends her blinding pulses into Bob’s module, the IM’s activation would increase the counts seen by Bob, due to the SD effect, and the resulting photocurrent would exhibit a negative current dip followed by a positive peak. Because this outcome is distinctively different from that under normal conditions, it represents a clear signature of Eve’s presence.

Even without correlating Bob’s counts to the IM’s activation times, the presence of the IM and SD circuitry make it possible to restore the APD’s count rate and prevent its blinding. This is a simple consequence of the fact that, irrespective of the polarity of the photocurrent, the positive peak is always well above the detector’s discrimination level, for a detector that has been correctly set up. So, for simplicity, we decided to not take advantage in this work of the “fine-grained” signatures based on correlating the counts with the IM or based on the avalanche’s polarity and focus rather on the “coarse-grained” signature represented by the APD’s counts. The analysis of the whole statistics available to Bob is left for future studies and can only reinforce the results presented here.

III. EXPERIMENTAL SETUP

To investigate the efficacy of the proposed measure, we adopt the experimental setup shown in Fig. 2, which includes both an IM and an SD circuit. An InGaAs/InP APD is thermo-electrically cooled to $-30 \, ^\circ\text{C}$ and operated with a constant DC bias of 51.6 V and a 1 GHz square-wave signal with a peak-to-peak amplitude of 4.6 V. A telecom C-band continuous wave laser diode emitting polarised light is used to illuminate the APD. The APD resistance is measured to be $R_{\text{apd}} = 1 \, \text{k}\Omega$, and no
biasing resistor is used, $R_{\text{bias}} = 0$. We use a variable optical attenuator to provide a 120 dB intensity variation range and a LiNbO$_3$ intensity modulator driven by a pattern generator to modulate the optical power. A high-bandwidth oscilloscope (16 GHz) is used to analyze the SD output. We determine the appropriate discrimination level for the SD-APD to be 18 mV and measure a detection efficiency of 26% for pulsed light and a dark count rate of 23 kHz at this level. We note that under this operation condition it is not possible for Eve to blind the detector using continuous-wave illumination because the increasing APD capacitive response is sufficient to counter Eve’s blinding effort [41].

### IV. EFFECT OF THE INTENSITY MODULATION ON THE COUNT RATE

We first demonstrate an experimental condition under which the SD-APD can be blinded. By deliberately setting the discrimination level inappropriately high, at 26 mV, we measure the count rate of the SD-APD as a function of incident optical power measured directly after the IM. Here, the RF input to the IM is disabled and its DC bias is adjusted to have a maximum transmission. Figure 3 shows the count rate and photocurrent as a function of the incident optical power. The detector exhibits a blinding gap between 300 $\mu$W and 3 mW, within which the detector count rate falls to zero. Such a blinding gap enables Eve to blind the detector using continuous-wave illumination because the increasing APD capacitive response is sufficient to counter Eve’s blinding effort [41].

The photocurrent follows the count rate and grows linearly before the count rate saturation, and then grows sub-linearly (100 nW - 1 $\mu$W) before becoming quasi-linear with the incident optical power ($> 1$ $\mu$W). In the blinding gap, the photocurrent is measured to exceed 1 mA.

Such a large photocurrent offers an opportunity to close the blinding gap by modulating the intensity of the attacking signal. As illustrated earlier with Fig. 1(d), an intensity modulation creates a dip in the photocurrent. The SD circuit converts each dip to a positive peak which can trigger the detector discrimination circuit when there is sufficient modulation depth. A random pattern produces photocurrent dips at a rate that is 1/4 of the signal clock rate. For simplicity we simulate this rate in our experiment by applying an RF signal to the IM using a repetitive modulation pattern “0001”, which we label as “1/4” modulation. We set the RF amplitude to 4 V, achieving half-wave modulation and an intensity extinction ratio of 23 dB. This pattern carves a 1 ns hole for every 4 ns duration in the attacking light intensity.

Using the ill-set discrimination level of 26 mV, we measure the count rate versus the incident optical power with the result shown as black solid circles in Fig. 4(a). The intensity modulation causes distinctly different count rate behavior for higher optical power when compared with the case without intensity modulation (open squares). The count rate stays above 250 MHz from 100 nW to 7.5 mW, without any sign of it falling. Despite the high discrimination level, the IM successfully removes the former SD-APD’s blinding gap.

We attribute the closure of the blinding gap to the applied intensity modulation. To illustrate this, we compare two SD-APD output waveforms recorded under vastly different optical powers. In the single photon counting regime, the APD produces a positive, current spike and therefore its SD output becomes a positive spike followed by its negative copy 1 ns afterwards, see waveform 1 in Fig. 4 (b). With an optical power of 1 mW, the SD-APD output waveform reverses its polarity (waveform 2) because the intensity modulation carves a hole in the photocurrent, instead of a current spike for a single-photon induced avalanche. The signal level is about 9 times as strong as the single-photon induced avalanche, and can therefore overcome the detector discrimination level. This confirms the intuition given at the beginning of the paper. For the case of 1/4 modulation, the count rate saturates close to 750 MHz, which can be explained
by having two ripples after the main avalanche peak in Fig. 4(b) overcoming the discrimination threshold as well.

Figure 4(c) plots the signal level of the main positive peak induced by the IM as a function of incident optical power. Over the incident power spanning over 4 orders of magnitude between 0.7 $\mu$W and 7 mW, the IM induced signal has a significant margin to overcome the discrimination level, even though it was inappropriately set. At an optical power of 1 $\mu$W, where the count rate with no IM (open squares in Fig. 4(a)) begins to fall, the signal level in Fig. 4(c) is over 50 mV and continues to increase in amplitude to over 300 mV at an optical power of greater than 1 mW. Within the power range Eve needs for blinding, each intensity modulation is guaranteed to produce at least one detector count. This is in agreement with our experiment using sparser modulation patterns, as shown in Fig. 4(a). A sparser modulation results in a proportionally lower bottom-out count rate in the blinding gap.

The significant margin in the signal level shown in Fig. 4(c) offers room to relax the requirement on the IM’s modulation contrast, thus minimising the loss that the IM would introduce. We determine the lowest modulation contrast by measuring the count rate probability as a function of the RF signal amplitude applied to the IM with the DC set to maximum transmission. Here, the modulation frequency is 1/128 of 1 GHz and the incident optical power is 1 mW. As shown in Fig. 5, a modulation signal with amplitude 0.3 V can always induce at least one detector count. More counts are possible due to the ripples in the output waveform also overcoming the discrimination threshold. This RF level corresponds to an intensity contrast of 0.06 dB. The number of counts increases above unity at a modulation amplitude higher than 1.5 V because the amplitude of the signal ripple (see Fig. 4(b)) rises above the discrimination level.

**V. INTENSITY MODULATION TO PREVENT BLINDING**

In this section we discuss how the IM represents a potential countermeasure to the blinding attack. In doing so, we exclude side effects due to imperfect electronics or to a wrong setting of the parameters. We also neglect any artificial additions used to facilitate post-processing or software implementation, such as dead time, which may be exploited by Eve with a modified attack [13]. We assume that the IM is driven by a true random generator, so Eve cannot deterministically predict the modulation effect and prepare her blinding pulses accordingly. We also consider a sufficient modulation depth to ensure a strong signal difference between modulated and unmodulated gates, which guarantees the occurrence of a detector count, as we have shown. The synchronization of the IM pulses, as well as the differential delay of the SD circuit, must be carefully chosen such that the counts by IM activation fall within the acceptance time window of the QKD system. Finally, a spectral filter should be applied in the QKD receiver to limit Eve’s choice of wavelength.
and ensure that the modulation effect on Eve’s attacking signal takes place.

The main security observation is that each IM-induced electrical signal overcomes the discrimination level and deterministically generates at least one detector count. These counts have equal probabilities of contributing a correct or incorrect bit in the sifted key, thus generating an overall 50% QBER. Therefore we can choose the probability to activate the IM, $P_{IM}$, such that the resulting QBER in the presence of Eve exceeds the security tolerance of the protocol. If this happens, the protocol is aborted and no insecure key is distilled.

To decide the correct value for $P_{IM}$, let us consider the BB84 protocol [1], which features a security tolerance of 11%. Suppose that $P_{IM} = 25\%$. In this case, we have a guarantee that if Eve always launches her attack, Bob will see at least 1 click in 4 input pulses, which would cause a QBER, $Q_b$, equal to or larger than 12.5%. In this case the key rate is zero, $R(Q_b) = 0$. If Eve works in “burst-mode”, she cannot do better than in the previous scenario. On the N preparations effected by Alice, she could intercept-and-resend $N_b$ consecutive blinding signals in bursts and mount her attack only on these bursts, while blocking the remaining $N - N_b$ quantum signals. Even in this case, the IM would guarantee the final QBER to be above 12.5%, causing zero key rate. Moreover, at the beginning of the train of blinding pulses prepared by Eve, the SD effect would cause one additional count in Bob’s detectors, making this scenario more favourable to the legitimate users.

On the other hand, Eve could let a fraction of Alice’s signals pass undisturbed. In this case, Eve gets no information on the undisturbed signals, but the resulting QBER would be smaller than the protocol’s tolerance and the users would not abort the transmission. However this case is still secure due to the fact that the key rate is a convex function of the QBER (see e.g. [42]). Suppose that $C_b$ and $C_u$ ($Q_b$ and $Q_u$) are the count rates (QBERs) pertaining to blinding and undisturbed pulses, respectively. Then the average QBER seen by the users is $Q = C_b Q_b + C_u Q_u$. The convexity of the key rate and the fact that $R(Q_b) = 0$ imply that

$$R(Q) \leq R(Q_b) + R(Q_u) = R(Q_u),$$

where $R(Q_b)$ and $R(Q_u)$ are the key rates from separate blinding and undisturbed pulses, respectively. Eq. (2) shows that there is at least a fraction $R(Q_u)$ of secure bits in the users’ signals. This comes from the fact that $R(Q_u)$ is associated with the undisturbed pulses. In the real case, the users measure $Q$ and distill a secure fraction $R(Q)$. This, by virtue of Eq. (2), is a pessimistic estimate of the fraction of secure bits in the sample, hence the protocol is secure.

VI. SUMMARY AND DISCUSSION

In conclusion, we have devised and experimentally demonstrated a new technique to mitigate detector blinding. By using an intensity modulator and an SD circuit, we modulated the incoming light to create uneven avalanches for the case of strong input light. Significantly, whilst this protects the detector from Eve it only introduces a small intrinsic attenuation of Alice’s signal. We showed that a modulation depth of 0.06 dB is sufficient to prevent an SD detector from being blinded. In our experimental test, we adopted a continuous-wave laser. A pulsed laser would be no more effective at blinding an SD detector as it creates more intensity fluctuations, to which the detector is very sensitive. Although intensity modulation to prevent Eve’s faked-state attack has been previously addressed in the literature [25], this was concerned with controlling an already blinded detector. Our approach, on the contrary, includes a SD circuit to prevent blinding in the first place and thus eliminate the possibility of a faked-state attack at the root.

The proposed IM measure entails a considerable extrinsic loss penalty of around 2.5–5 dB, arising from imperfect intensity modulators based on LiNbO$_3$ and will therefore negatively impact the secure key rate. The loss associated with the product of the modulation rate of 1/4 and extinction ratio of 0.06 dB is comparatively negligible, hence the key rate in the presence of modulation would be 0.315 times the unmodulated key rate and the distance would be shortened by 25 km, assuming a maximum insertion loss of 5 dB. Although QKD systems typically have two or more detectors, placing an intensity modulator in front of only one would be sufficient to demonstrate the presence of Eve. We note that existing modulators were designed to achieve high modulation depth which requires a lengthy crystal waveguide for electro-optical interaction. With a reduced modulation depth, the insertion loss can be made significantly smaller, thereby alleviating the loss penalty. Although intensity modulators are typically polarization sensitive, Eve cannot mount an attack such that she simply sends light of a polarization which does not experience modulation. This is because the detectors in a QKD system always see a fixed polarization, whether in phase-encoded schemes such as [43], which contain an electronic polarization controller followed by a polarizing beamsplitter (PBS) or in polarization-encoded schemes, such as in [10], which also have PBSs before the detectors. The use of an IM also requires a random number generator (RNG). Since Bob typically already has an RNG for the purpose of active basis selection, he can use the same RNG operated at twice the clock rate for the IM, which would not open additional side-channels.
ACKNOWLEDGEMENTS

A. K.-S. gratefully acknowledges financial support from the EPSRC and Toshiba Research Europe Ltd.