

# Randomness extraction from Bell violation with continuous parametric down conversion

Lijiong Shen,<sup>1,2</sup> Jianwei Lee,<sup>1</sup> Le Phuc Thinh,<sup>1</sup> Jean-Daniel Bancal,<sup>3</sup>  
Alessandro Cerè,<sup>1</sup> NIST team,<sup>4</sup> Valerio Scarani,<sup>1,2</sup> and Christian Kurtsiefer<sup>1,2</sup>

<sup>1</sup>Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543

<sup>2</sup>Department of Physics, National University of Singapore, 2 Science Drive 3, Singapore 117551

<sup>3</sup>Department of Physics, University of Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland

<sup>4</sup>NIST

(Dated: April 18, 2018)

We present a violation of the CHSH inequality without the fair sampling assumption with a continuously pumped photon pairs source combined with two high efficiency superconducting detectors. Due to the continuous nature of the source, the choice of the duration of each measurement round effectively controls the average number of photon pairs participating in the Bell test. We observe a maximum violation of  $S = 2.01602(32)$  with average number of pairs per round of  $\approx 0.32$ , compatible with our system overall detection efficiencies. Systems that violate a Bell inequality are guaranteed to generate private randomness, with the randomness extraction rate depending on the observed violation and on the repetition rate of the Bell test. For our realization, the optimal rate of randomness generation is a compromise between the observed violation and the duration of each measurement round, with the latter realistically limited by the detection time jitter. We calculate an asymptotic rate of  $\approx 1300$  random bits/s.

Based on a violation of a Bell inequality, quantum physics can provide randomness that can be certified to be private, i.e., uncorrelated to any outside process [1–3]. Initial experimental realizations of such sources of certified randomness are based on atomic or atomic-like systems, but exhibit extremely low generation rates, making them impractical for most applications [2, 4]. Advances in high efficiency infrared photon detectors [5, 6], combined with highly efficient photon pair sources, allowed experimental demonstrations of loophole free violation of the Bell inequality using photons [7, 8]. The random bit generation rate in these setups is on the order of tens per second [9, 10]. **NIST please check order of magnitude of rate per second**, mainly limited by the fixed repetition rate of the photon pair source in combination with the small observed violation of the Bell inequality.

In this work, we use a source of polarization entangled photon pairs operating in a continuous wave (CW) mode, and define measurement rounds by organizing the detection events in uniform time bins. The binning is set independently of the detection time, thus avoiding the coincidence loophole [11]. Superconducting detectors with a high detection efficiency allow us to close the detection loophole. We show how, for fixed overall detection efficiency and pair generation rate, the time bin duration determines the observed Bell violation. We then estimate the rate of random bits that can be extracted from the system and its dependence on time bin width.

*Theory.* – Bell tests are carried out in successions of rounds. In each round, each party chooses a measurement and records an outcome. The simplest meaningful scenario involves two parties, each of which can choose between two measurements with binary outcome. Alice and Bob’s measurements are labelled by  $x, y \in \{0, 1\}$ , respectively; their outcomes are labelled  $a, b \in \{+1, -1\}$ . As figure of merit we use the Clauser-Horne-Shimony-

Holt (CHSH) expression

$$S = E_{00} + E_{01} + E_{10} - E_{11}, \quad (1)$$

where the correlators are defined by

$$E_{xy} := \Pr(a = b|x, y) - \Pr(a \neq b|x, y). \quad (2)$$

As well known, if  $S > 2$ , the correlations cannot be due to pre-established agreement; and if they can’t be attributed to signaling either, the underlying process is necessarily random. This is not only a qualitative statement: the amount of extractable private randomness can be quantified. In the limit in which the statistics are collected from an arbitrarily large number of rounds, the number of random bits per round **is given by** [2]

$$r_\infty = 1 - h\left(\frac{1}{2} + \frac{1}{2}\sqrt{\frac{S^2}{4} - 1}\right), \quad (3)$$

where  $h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$  is the binary entropy function.


Besides the no-signaling assumption, *this certification of randomness is device-independent*: it relies on the value of  $S$  extracted from the observed statistics, but not on any characterisation of the degrees of freedom or of the devices used in the experiment. All that matters is that in every round both parties produce an outcome. In our case, we decide that, if a party’s detectors did not fire in a given round, that party will output +1 for that round. This convention allows us to use only one detector per party [12, 13]: in the rounds when the detector fires, the outcome will be -1; in the others, it will be +1.


While the certification is device-independent, the design of the experiment requires detailed knowledge and control of the physical degrees of freedom. Our experiment uses photons entangled in polarisation, produced by spontaneous parametric down-conversion (SPDC).

Let us first consider a simplified model, in which a pair of photons is created in each round. Eberhard [14] famously proved that, when the collection efficiencies  $\eta_A$  and  $\eta_B$  are not unity, higher values of  $S$  are obtained using non-maximally entangled pure states. So we aim at preparing

$$|\psi\rangle = \cos\theta|HV\rangle - e^{i\phi}\sin\theta|VH\rangle, \quad (4)$$

where  $H$  and  $V$  represent the horizontal and vertical polarization modes, respectively. The state and measurement that maximise  $S$  are a function of  $\eta_A$  and  $\eta_B$ . For  $\phi = 0$ , the optimal measurements correspond to linear polarisation directions, denoted  $\cos\alpha_x\hat{e}_H + \sin\alpha_x\hat{e}_V$  and  $\cos\beta_y\hat{e}_H + \sin\beta_y\hat{e}_V$ .

For a down-conversion source, the number of photons produced per round is not fixed. If the duration  $\tau$  of a round is much longer than the single-photon coherence time, and no multi-photon states are generated (a realistic assumption in a CW pumped scenario), the output of the source is accurately described by independent photon pairs, whose number  $v$  follows a Poissonian distribution  $P_\mu(v)$  of average pairs per round  $\mu$ . The main contribution to  $S > 2$  will come from the single-pair events; notice that  $P_\mu(1) \leq \frac{1}{e} \approx 0.37$  for a Poissonian distribution. So there is always a large fraction of other pair number events, and the observed value of  $S$  depends significantly on it [15]. For  $\mu \rightarrow 0$ , almost all rounds will give no detection, that is  $P(+1, +1|x, y) \approx 1$  which leads to  $S = 2$ . So, for  $\mu \ll 1$  we expect a violation  $S \approx P_\mu(1)S_{\text{qubits}} + (1 - P_\mu(1))2$ , where  $S_{\text{qubits}}$  is the value achievable with state (4). In the other limit,  $\mu \gg 1$ , almost all round will have a detection, that is  $P(-1, -1|x, y) \approx 1$  and again  $S = 2$ . Before this behavior kicks in, when more than one pair is frequently present we expect a drop in the value of  $S$ , since the detections may be triggered by independent pairs. An accurate modelling for any value of  $\mu$  is conceptually simple but notationally cumbersome.  leave it for Appendix A.

Photon pair sources based on pulsing quasi-CW sources with a fixed repetition rate control the value of  $\mu$  by limiting the pump power. With true CW pumping the average number of pairs per round is  $\mu = (\text{pair rate}) \cdot \tau$ , where  $\tau$  is the duration of . The resulting repetition rate of the experiment is  $1/\tau$ . In this work, we fix the pair rate, while  $\tau$  is a free parameter that can be optimized to extract the highest amount of randomness.

*Experimental setup.* – A sketch of the experimental setup is shown in Fig. 1. The source for entangled photon pairs is based on the coherent combination of two collinear type-II SPDC processes [16]. We pump a periodically poled potassium titanylphosphate crystal (PP-KTP,  $2 \times 1 \times 10 \text{ mm}^3$ ) from two opposite directions with light from the same laser diode (405 nm). Both pump beams have the same Gaussian waists of  $\approx 350 \mu\text{m}$  located within the crystal. Light at 810 nm from the two SPDC process is overlapped in a polarizing beam splitter (PBS), entangling the polarization modes, and collected

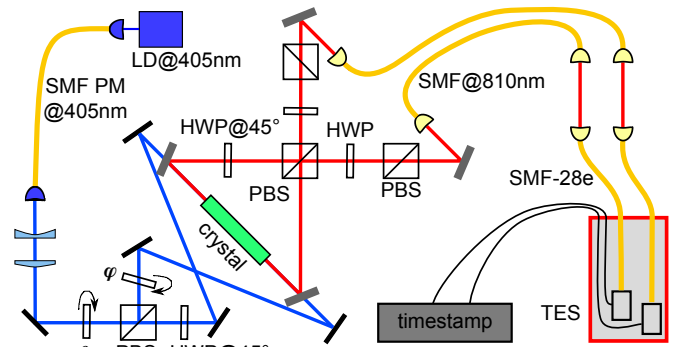


FIG. 1: Schematic of the experimental setup, including the source of the non-maximally entangled photon pairs. A PP-KTP crystal, cut and poled for type II spontaneous parametric down conversion from 405 nm to 810 nm, is placed at the waist of a Sagnac-style interferometer and pumped from both sides. Light at 810 nm from the two SPDC process is overlapped in a polarizing beam splitter (PBS), generating the non-maximally entangled state described by Eq. (4) when considering a single photon pair. A laser diode (LD) provides the continuous wave UV pump light. The combination of a half wave plate and polarization beam splitter (PBS) sets  $\theta$  by controlling the relative intensity of the two pump beams, while a thin glass plate controls their relative phase  $\phi$ . The pump beams enter the interferometer through dichroic mirrors. At each output of the PBS, the combination of a HWP and PBS projects the mode polarization before coupling into a fiber single mode for light at 810 nm (SMF@810). A free space link is used to transfer light from SMF@810 to single mode fibers designed for 1550 nm (SMF-28e). Eventually the light is detected with high efficiency superconducting Transition Edge Sensors (TES), and timestamped with a resolution of 2 ns.

into single mode fibers. When a single photon pair is generated, the resulting polarization state is given by Eq. (4), where  $\theta$  and  $\phi$  are determined by the relative intensity and phase of the two pump beams set by rotating a half wave plate before the first PBS, and the tilt of a glass plate in one of the pump arms.

The effective collection modes for the downconverted light, determined by the single mode optical fibers and incoupling optics was chosen to have a Gaussian beam waist of  $\approx 130 \mu\text{m}$  centered in the crystal in order to maximize collection efficiency [17, 18]. The combination of a zero-order half-wave plate and another PBS (extinction rate 1:1000 in transmission) sets the measurement bases for light entering the single mode fibers. All optical elements are anti-reflection coated for 810 nm. Light from each collection fiber is sent to superconducting transition edge sensor (TES) optimized for detection at 810 nm [5], which are kept at  $\approx 80 \text{ mK}$  within a cryostat. As the detectors show the highest efficiency when coupled to telecom fibers (SMF28+), the light collected in to single mode fibers from the parametric conversion source is transferred to these fibers via a free-space link. The TES output signal is translated into photodetection event arrival times using a constant fraction discrimina-

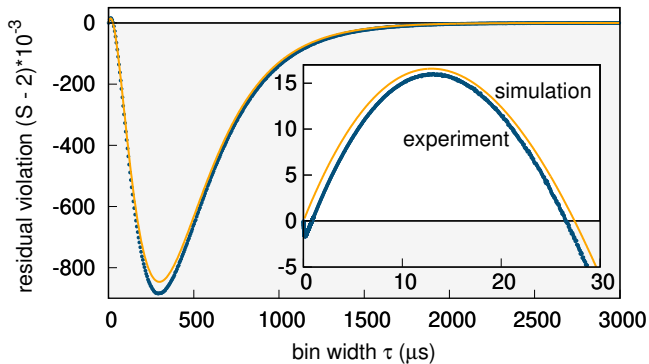


FIG. 2: Measured CHSH violation as function of bin width  $\tau$  (blue circles). A theoretical model (orange continuous line) is sketched in the main text and described in detail in Appendix A. Both the simulation and the experimental data show a violation for short  $\tau$  (zoom in inset). The uncertainty on the measured value, calculated assuming IID, corresponding to one standard deviation due to a Poissonian distribution of the events, is not visible in the graph. For  $\tau \lesssim 1 \mu\text{s}$  the detection jitter ( $\approx 170 \text{ ns}$ ) is comparable with the time bin, resulting in a loss of observable correlation and a fast drop of the value of  $S$ .

tor with an overall timing jitter  $\approx 170 \text{ ns}$ , and recorded with a resolution of  $2 \text{ ns}$ . Setting Alice's and Bob's analyzing waveplates in the natural basis of the combining PBS,  $HV$  and  $VH$ , we estimate heralded efficiencies of  $82.42 \pm 0.31 \%$  ( $HV$ ) and  $82.24 \pm 0.30 \%$  ( $VH$ ). We identified two main sources of uncorrelated detection events: intrinsic detector and background events at rates of  $6.7 \pm 0.58 \text{ s}^{-1}$  for Alice and  $11.9 \pm 0.77 \text{ s}^{-1}$  for Bob, respectively, and fluorescence caused by the UV pump in the PPKTP crystal [19], contributing  $0.135 \pm 0.08\%$  of the signal. With a total pump power at the crystal of  $5.8 \text{ mW}$  we estimate a pair generation rate  $\approx 24 \times 10^3 \text{ s}^{-1}$  (detected  $\approx 20 \times 10^3 \text{ s}^{-1}$ ), and dark count / background rates of  $45.7 \text{ s}^{-1}$  (Alice) and  $41.5 \text{ s}^{-1}$  (Bob).

*Violation.* – For the measured system efficiencies ( $\eta_A \approx 82.4\%$ ,  $\eta_B \approx 82.2\%$ ) and and rate of uncorrelated counts at each detector ( $45.7 \text{ s}^{-1}$  Alice,  $41.5 \text{ s}^{-1}$  Bob), a numerical optimisation gives the following values of the state and measurement parameters (see Appendix A for details):  $\theta = 25.9^\circ$ ,  $\alpha_0 = -7.2^\circ$ ,  $\alpha_1 = 28.7^\circ$ ,  $\beta_0 = 82.7^\circ$ , and  $\beta_1 = -61.5^\circ$ . These are close to optimal for all values of  $\mu$ , and the maximal violation is expected for  $\mu = 0.322$ .

We collected data for approximately 40 minutes, changing the measurement basis every 2 minutes. We periodically ensure that  $\phi \approx 0$  by rotating the phase plate until the visibility in the  $+45^\circ / -45^\circ$  basis is larger than 0.985.

In Fig. 2 we show the result of processing the time-stamped events for different bin widths  $\tau$ . The largest violation  $S = 2.01602(32)$  is observed for  $\tau = 13.150 \mu\text{s}$ , which, with the cited pair generation rate of  $24 \times 10^3 \text{ s}^{-1}$ , corresponds to  $\mu \approx 0.32$ . The uncertainty is calculated

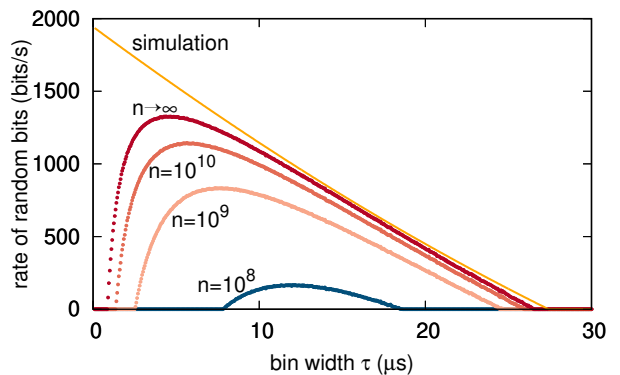


FIG. 3: Randomness generation rate  $r_n/\tau$  as a function of  $\tau$  for different block sizes  $n$ . The points are calculated via Eq. (5) for finite  $n$  (Eq. (3) for  $n \rightarrow \infty$ ) and the violation measured in the experiment, assuming  $\gamma = 0$  (no testing rounds) and  $\epsilon_c = \epsilon_s = 10^{-10}$ . The continuous line is the asymptotic rate Eq. (3) evaluated on the values of  $S$  of the simulation shown in Fig. 2, for the same security assumptions.

assuming that measurement results are independent and identically distributed (IID). The slight discrepancy between the experimental violation and the simulation is attributed to the non-ideal visibility of the state generated by the photon pair source. When  $\tau$  is comparable to the detection jitter, detection events due to a single pair may be assigned to different rounds, decreasing the correlations. This explains the drop of  $S$  below 2 (which our simulation does not capture because we have not included the jitter as a parameter).

*Randomness extraction.* – In order to turn the output data generated from our experiment into uniformly random bits, we need to employ a randomness expansion protocol [20]. Such a protocol consists of a pre-defined number of rounds  $n$ , forming a block. Each round is randomly assigned (with probability  $\gamma$  and  $1 - \gamma$ , respectively) to one of two tasks: testing the device for faults or eavesdropping attempts, or generating random bits. When the test rounds show a sufficient violation, one applies a quantum-proof randomness extractor to the block, obtaining  $m$  random bits. The performance of the extraction protocol is determined by completeness and soundness security parameters,  $\epsilon_c$  and  $\epsilon_s$ . To ensure the resulting string is uniform to within  $\approx 10^{-10}$ , we choose  $\epsilon_c = \epsilon_s = 10^{-10}$ . The extraction protocol we consider does not assume IID, is composable in the sense that it can be used as part of a larger protocol (such as quantum key distribution) without compromising its security, and it is secure against a quantum adversary holding quantum side information [20]. The details of the protocol execution and its security proof are given in Appendix B.

For a block consisting of  $n$  rounds, the number of random bits per round is given by

$$r_n = \eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}) - 4 \frac{\log n}{n} + 4 \frac{\log \epsilon_{\text{EX}}}{n} - \frac{10}{n}, \quad (5)$$

where the function  $\eta_{\text{opt}}$  depends on the block size  $n$ ,

detected violation  $S$ , and auxiliary security parameters  $\epsilon', \epsilon_{EA}, \epsilon_{EX}$ . The choice of these auxiliary security parameters is required to add up to the chosen level of completeness and soundness. In the limit  $n \rightarrow \infty$  one recovers the asymptotic value (3).

The extractable randomness rate  $r_n/\tau$  based on the observed  $S$  is presented in Fig. 3 for various block sizes  $n$ . For comparison, we also plot the asymptotic value  $r_\infty/\tau$  with  $S$  given by the simulation. The most obvious feature is that the highest randomness rate is not obtained at maximal violation of the inequality. There one gets highest randomness per round, but it turns out to be advantageous to sacrifice randomness per round in favor of a larger number of rounds per unit time. This optimization will be part of the calibration procedure for a random number generator with an active switch of measurement bases. As explained previously, the detection jitter affects the observable violation for  $\tau$  comparable to it. This causes the sharp drop for short time bins observed for the experimental data.

If we consider blocks of size  $n = 10^8$ , with  $\tau = 12.15 \mu\text{s}$  we calculate a rate of randomness of  $\approx 167$  bits/s, a rate comparable with recent reported results [20] with the advantage that the time required to acquire a single block is approximately 20 minutes. Increasing the block size by one order of magnitude,  $n = 10^9$ , with  $\tau = 7.35 \mu\text{s}$ , the rate of generated randomness increases to  $\approx 834$  bits/s, with the time required to acquire a single block of  $\approx 2$  hours, compatible with the experimental stability of our source. These numbers are not optimal; more sophisticated analysis demonstrated higher randomness extraction for the same detected violation [10, 21].

For fixed detector efficiencies, we expect the randomness rate to increase with higher photon pair generation rate, that is by increasing the pump power, and to be ultimately limited by the detection time jitter. Here, the use of efficient superconducting nanowire detectors will be a significant advantage.

*Conclusion.* – In conclusion, we experimentally observed a violation of CHSH inequality with a continuous wave photon entangled pair source without fair-sampling assumption combining a high collection efficiency source and high detection efficiency superconducting detectors, with the largest detected violation of  $S = 2.01602(32)$ .

The generation rate of all probabilistic sources of entangled photon pairs is limited by the probability of gen-

eration of multiple pairs per experimental round, according to Poissonian statistics. The flexible definition of an experimental round permitted by the CW nature of our setup allowed us to study the dependence of the observable violation as function of the average number of photon pairs per experimental round. This same flexibility can be exploited to reduce the time necessary to acquire sufficient statistics for this kind of experiments: an increase in the pair generation rate is accompanied by a reduction of the round duration. This approach shifts the experimental repetition rate limitation from the photon statistics to the other elements of the setup, e.g. detectors time response or active polarization basis switching speed.



The observation of a Bell violation also certifies the generation of randomness. We estimate the amount of randomness generated per round both in an asymptotic regime and for a finite number of experimental rounds, assuming a required level of uniformity of  $10^{-10}$ . When considering the largest attainable rate of random bit generation, the optimal round duration is the result of the trade-off between observed violation and number of rounds per unit time. While for an ideal realization the optimal round duration would be infinitesimally short, it is limited in our system by the detection jitter time. Our proof of principle demonstration can be extended into a complete, loophole-free random number source. This requires closing the locality and freedom-of-choice loopholes, with techniques not different from pulsed photonic sources, with the only addition of a periodic calibration necessary for determining the optimal time-bin.

## Acknowledgments

This research is supported by the Singapore Ministry of Education Academic Research Fund Tier 3 (Grant No. MOE2012-T3-1-009); by the National Research Fund and the Ministry of Education, Singapore, under the Research Centres of Excellence programme; by the Swiss National Science Foundation (SNSF), through the Grants PP00P2-150579 and PP00P2-179109; and by the Army Research Laboratory Center for Distributed Quantum Information via the project SciNet.

- 
- [1] R. Colbeck, *Quantum And Relativistic Protocols For Secure Multi-Party Computation*, Ph.D. thesis, University of Cambridge (2007).
  - [2] S. Pironio, A. Acín, S. Massar, A. B. de la Giroday, D. N. Matsukevich, P. Maunz, S. Olmschenk, D. Hayes, L. Luo, T. A. Manning, and C. R. Monroe, *Nature* **464**, 1021 (2010).
  - [3] A. Acín and L. Masanes, *Nature* **540**, 213 (2016).
  - [4] B. J. Hensen, H. Bernien, A. E. Dréau, A. Reiserer,

- N. Kalb, M. S. Blok, J. Ruitenbergh, R. F. L. Vermeulen, R. N. Schouten, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminau, and R. Hanson, *Nature* **526**, 682 (2015).
- [5] A. E. Lita, A. J. Miller, and S. W. Nam, *Opt. Express* **16**, 3032 (2008).
- [6] F. Marsili, V. B. Verma, J. A. Stern, S. Harrington, A. E. Lita, T. Gerrits, I. Vayshenker, B. Baek, M. D. Shaw,

- R. P. Mirin, and S. W. Nam, *Nature Photonics* **7**, 210 (2013).
- [7] M. Giustina, M. A. M. Versteegh, S. Wengerowsky, J. Handsteiner, A. Hochrainer, K. Phelan, F. Steinlechner, J. Kofler, J.-A. Larsson, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, J. Beyer, T. Gerrits, A. E. Lita, L. K. Shalm, S. W. Nam, T. Scheidl, R. Ursin, B. Wittmann, and A. Zeilinger, *Phys. Rev. Lett.* **115**, 250401 (2015).
- [8] L. K. Shalm, E. Meyer-Scott, B. G. Christensen, and P. Bierhorst, *Phys. Rev.* **115**, 250402 (2015).
- [9] Y. Liu, X. Yuan, M.-H. Li, W. Zhang, Q. Zhao, J. Zhong, Y. Cao, Y.-H. Li, L.-K. Chen, H. Li, T. Peng, Y.-A. Chen, C.-Z. Peng, S.-C. Shi, Z. Wang, L. You, X. Ma, J. Fan, Q. Zhang, and J.-W. Pan, arXiv (2017), 1709.06779 .
- [10] P. Bierhorst, E. Knill, S. Glancy, Y. Zhang, A. Mink, S. Jordan, A. Rommal, Y.-K. Liu, B. Christensen, S. W. Nam, M. J. Stevens, and L. K. Shalm, arXiv (2018), 1803.06219 
- [11] J.-A. Larsson and R. D. Gill, *EPL* **67**, 707 (2004).
- [12] M. Giustina, A. Mech, S. Ramelow, B. Wittmann, J. Kofler, J. Beyer, A. E. Lita, B. Calkins, T. Gerrits, S. W. Nam, R. Ursin, and A. Zeilinger, *Nature* **497**, 227 (2013).
- [13] J.-D. Bancal, L. Sheridan, and V. Scarani, *New J. Phys.* **16**, 033011 (2014).
- [14] P. H. Eberhard, *Phys. Rev. A* **47**, R747 (1993).
- [15] V. Caprara Vivoli, P. Sekatski, J.-D. Bancal, C. C. W. Lim, B. G. Christensen, A. Martin, R. T. Thew, H. Zbinden, N. Gisin, and N. Sangouard, *Phys. Rev. A* **91**, 012107 (2015).
- [16] M. Fiorentino, G. Messin, C. E. Kuklewicz, F. N. C. Wong, and J. H. Shapiro, *Phys. Rev. A* **69**, 041801 (2004).
- [17] R. S. Bennink, *Phys. Rev. A* **81**, 053805 (2010).
- [18] P. Ben Dixon, D. Rosenberg, V. Stelmakh, M. E. Grein, R. S. Bennink, E. A. Dauler, A. J. Kerman, R. J. Molnar, and F. N. C. Wong, *Phys. Rev. A* **90**, 043804 (2014).
- [19] S. M. Hegde, K. L. Schepler, R. D. Peterson, and D. E. Zelmon, in *Defense and Security Symposium*, edited by G. L. Wood and M. A. Dubinskii (SPIE, 2007) p. 65520V.
- [20] R. Arnon-Friedman, R. Renner, and T. Vidick, ArXiv e-prints (2016), arXiv:1607.01797 [quant-ph] 
- [21] E. Knill, Y. Zhang, and P. Bierhorst, ArXiv e-prints (2017), arXiv:1709.06159 [quant-ph] .
- [22] T. S. Hao and M. Hoshi, *IEEE Transactions on Information Theory* **43**, 599 (1997).
- [23] W. Mauerer, C. Portmann, and V. B. Scholz, ArXiv e-prints (2012), arXiv:1212.0520 [cs.IT] .

## Appendix A: Modelling the violation of CHSH by a Poissonian source of qubit pairs

The output of a CW-pumped SPDC process can be accurately described as the emission of independent pairs distributed according to Poissonian statistics of average  $\mu$ , if the time under consideration (in our case, the length of a round) is much longer than the single-photon coherence time. The goal of this appendix is to provide an estimate of the observed CHSH parameter  $S$  for such a source.

The pairs being independent, it helps to think in two steps. First, each pair is converted into classical information  $(\alpha, \beta) \in \{+, -\}$  with probability

$$P_Q(\alpha, \beta|x, y) = \text{Tr}(\rho \Pi_\alpha^x \otimes \Pi_\beta^y), \quad (\text{A1})$$

where the  $\Pi$ 's are measurement operators. If some of the events have  $\alpha = -$  ( $\beta = -$ ), Alice's (Bob's) detector may be triggered, leading to the observed outcome  $a = -1$  ( $b = -1$ ).

For the purpose of studying CHSH, it is sufficient to consider  $P(-1, +1|x, y)$  and  $P(+1, -1|x, y)$ , since

$$E_{xy} = 1 - P(-1, +1|x, y) - P(+1, -1|x, y). \quad (\text{A2})$$

With our convention of outcomes,  $P(-1, +1|x, y)$  is the probability associated with the case when Alice's detector clicks and Bob's does not. Thus, Bob's detector should not be triggered by any pair: each pair will contribute to  $P(-1, +1|x, y)$  with

$$D(\alpha) \equiv P_Q(\alpha, +|x, y) + (1 - \eta_B)P_Q(\alpha, -|x, y). \quad (\text{A3})$$

Now, let us look at the contribution by  $v$  pairs to  $P(-1, +1|x, y)$ . At least one of the  $\alpha$ 's must be  $-$  for the detector to be triggered; and multiple detections will also be treated as  $a = -1$ . Thus, a configuration in which exactly  $k$   $\alpha$ 's are  $-$  leads to  $a = -1$  with probability  $1 - (1 - \eta_A)^k$  (i.e. at least one  $\alpha = -$  must trigger a detection). Obviously there can be  $\binom{v}{k}$  such configurations, so the contribution of the  $v$  pair events to  $P(-1, +1|x, y)$  is

$$D_v = \sum_{k=1}^v \binom{v}{k} [1 - (1 - \eta_A)^k] D(-)^k D(+)^{v-k}. \quad (\text{A4})$$

Finally,

$$P(-1, +1|x, y) = \sum_{v=0}^{\infty} P_\mu(v) D_v. \quad (\text{A5})$$

The calculation of  $P(+1, -1|x, y)$  is identical, with  $D(\beta) \equiv P_Q(+, \beta|x, y) + (1 - \eta_A)P_Q(-, \beta|x, y)$  and  $\eta_B$  instead of  $\eta_A$  in (A4).

Because the quantum probabilities appear in such a convoluted way, the optimal parameters for both the state and the measurements are not the same as for the single-pair case. Upon inspection, however, the values are close, as expected from the fact that the violation is mostly contributed by the single-pair events.

The curves presented in Fig. 2 have been obtained with a slightly modified model that includes the effect of the background events. The quantum probabilities  $P_Q(\alpha, \beta|x, y)$  have been computed with  $\rho = |\psi\rangle\langle\psi|$  given in (4) and with projective measurements, with the values of the parameters given in the main text.

## Appendix B: Protocol and security proof

For completeness, we will present the protocol studied in [20] and give explicit constants in its security proof. It will be more convenient for us to switch to the notation  $a, b \in \{0, 1\}$  for the outcome labels (instead of  $a, b \in \{+1, -1\}$  of the main text) and use the language of nonlocal games with winning condition

$$w_{\text{CHSH}}(a, b, x, y) = \begin{cases} 1 & \text{if } a \oplus b = x \cdot y, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{B1})$$

The game winning probability is then  $w = 1/2 + S/8$  in terms of the CHSH value. The optimal classical winning strategy achieve a winning probability of 0.75, while the optimal quantum strategy achieves a winning probability of  $(2 + \sqrt{2})/4 \approx 0.85$ .

A randomness expansion protocol is a procedure that consumes  $r$ -bits of randomness and generates  $m$ -bits of almost uniform randomness. Formally, a  $(\epsilon_c, \epsilon_s)$ -secure  $r \rightarrow m$  randomness expansion protocol if given  $r$  uniformly random bits,

- (Soundness) For any implementation of the device it either aborts or returns an  $m$ -bit string  $Z \in \{0, 1\}^m$  with

$$(1 - \Pr[\text{abort}]) \|\rho_{ZRE} - \rho_{U_m} \otimes \rho_{U_r} \otimes \rho_E\|_1 \leq \epsilon_s,$$

where  $R$  is the input randomness register,  $E$  is the adversary system, and  $\rho_{U_m}, \rho_{U_r}$  are the prepared states on appropriate registers.

- (Completeness) There exists an honest implementation with  $\Pr[\text{abort}] \leq \epsilon_c$ .

We remark that this security definition is a composable definition assuming quantum adversary, but not composable assuming a no-signalling adversary [20].

For a concrete randomness expansion protocol, we present the protocol studied in [20]. The protocol takes parameters  $\gamma$  expected fraction (marginal probability) of test rounds,  $\omega_{\text{exp}}$  expected winning probability for an honest (perhaps noisy) implementation, and  $\delta_{\text{est}}$  width of the statistical confidence interval for the estimation test. In an execution, for every round  $i \in \{1, \dots, n\}$ :

- Bob chooses a random bit  $T_i \in \{0, 1\}$  such that  $\Pr(T_i = 1) = \gamma$  using the interval algorithm [22].
- If  $T_i = 0$  (randomness generation), Alice and Bob chooses deterministically  $(X_i, Y_i) = (0, 0)$ , otherwise  $T_i = 1$  (test round) they choose uniformly random inputs  $(X_i, Y_i)$ .
- Alice and Bob uses the physical devices with the said inputs  $(X_i, Y_i)$  and record their outputs  $(A_i, B_i)$ .
- If  $T_i = 1$ , they compute

$$C_i = w_{\text{CHSH}}(A_i, B_i, X_i, Y_i). \quad (\text{B2})$$

They abort the protocol if  $\sum_j C_j < (\omega_{\text{exp}}\gamma - \delta_{\text{est}})n$  where  $j$  is the index of test rounds, otherwise they return  $\text{Ext}(\mathbf{AB}, \mathbf{Z})$  where  $\text{Ext}$  is a randomness extractor and  $\mathbf{Z}$  is a uniformly random seed.

More precisely, we use a Trevisan extractor in [23] based on polynomial hashing with block weak design, because of its efficiency in terms of seed length. This is a function  $\text{Ext} : \{0, 1\}^{2n} \times \{0, 1\}^d \rightarrow \{0, 1\}^m$  such that if  $H_{\min}(\mathbf{AB}|E) \geq 4 \log \frac{1}{\epsilon_1} + 6 + m$  then

$$\frac{1}{2} \|\rho_{\text{Ext}(\mathbf{AB}, \mathbf{Z})ZE} - \rho_{U_m} \otimes \rho_{U_d} \otimes \rho_E\| \leq m\epsilon_1 \quad (\text{B3})$$

The seed length of this extractor is  $d = a(2\ell)^2$  where

$$a = \left\lceil \frac{\log(m - 2e) - \log(2\ell - 2e)}{\log(2e) - \log(2e - 1)} \right\rceil \quad (\text{B4})$$

$$\ell = \left\lceil \log 2n + 2 \log \frac{2}{\epsilon_1} \right\rceil \quad (\text{B5})$$

Now it can be shown that the entropy accumulation protocol gives the completeness and soundness of our randomness expansion protocol. However, let us mention how the input randomness affects the soundness and completeness of the final protocol.

In the protocol we assume access to a certain uniform randomness source, from which the random bits required in the protocol are generated: the  $T_i$ ,  $X_i$  and  $Y_i$ . In a certain rounds,  $X_i$  and  $Y_i$  are either deterministic or fully random bits and can be directly obtained from the source. On the other hand,  $T_i$  must be simulated from the uniform source (except when  $\gamma = 1/2$  which is not usually the case in practice). This can be done efficiently by the interval algorithm [22]: the expected number of random bits needed to generate one Bernoulli( $\gamma$ ) is at most  $h(\gamma) + 2$  and the maximum number of random bits needed is at most  $L_{\max} := \max\{\log \gamma^{-1}, \log(1 - \gamma)^{-1}\}$ . Then Lemma 16 of [20] gives us: let  $\gamma > 0$ , for any  $n$  there is an efficient procedure that given (at most)  $6h(\gamma)n$  uniformly random bits either it aborts with probability at most  $\epsilon_{\text{SA}} = \exp(-18h(\gamma)^3 n / L_{\max})$  or outputs  $n$  bits  $T_1, \dots, T_n$  whose distribution is within statistical distance at most  $\epsilon_{\text{SA}}$  of  $n$  i.i.d. Bernoulli( $\gamma$ ) random variables. This raises both the completeness and soundness parameter of the final protocol by  $\epsilon_{\text{SA}}$ .

In an honest implementation of the protocol, Alice and Bob execute the protocol with a device that performs i.i.d. measurements on a tensor product state resulting in an expected winning probability  $\omega_{\text{exp}}$ . Here Lemma 8 of [20] bounds the probability of aborting using Hoeffding's inequality. That is, the probability that our randomness expansion protocol aborts for an honest implementation is

$$\Pr[\text{abort}] \leq \exp(-2n\delta_{\text{est}}^2) =: \epsilon_{\text{est}}. \quad (\text{B6})$$

Therefore, the total completeness is bounded by  $\epsilon_{\text{SA}} + \epsilon_{\text{est}}$ . (Note that  $\epsilon_{\text{est}}$  is actually the completeness parameter of the entropy accumulation protocol in [20].)

For the soundness, Corollary 11 of [20] ensures that for any  $\epsilon_{\text{EA}}, \epsilon' \in (0, 1)$  either the protocol aborts with probability greater than  $1 - \epsilon_{\text{EA}}$  or

$$H_{\min}^{\epsilon'}(\mathbf{AB|XYTE})_{\rho|_{\text{pass}}} > n \cdot \eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}). \quad (\text{B7})$$

Together with our extractor, for all  $\epsilon_1 \in (0, 1)$ , if the length  $m$  of the final string satisfies

$$n \cdot \eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}) = 4 \log \frac{1}{\epsilon_1} + 6 + m \quad (\text{B8})$$

then we are guaranteed that

$$\frac{1}{2} \|\rho_{SRE} - \rho_{U_m} \otimes \rho_{U_R} \otimes \rho_E\| \leq \epsilon'/2 + m\epsilon_1. \quad (\text{B9})$$

Here  $\eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}})$  is given by the following equations: for  $h$  the binary entropy and  $\gamma, p(1) \in (0, 1]$

$$\eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}) = \max_{\frac{3}{4} < \frac{p_t(1)}{\gamma} < \frac{2+\sqrt{2}}{4}} \eta(\omega_{\text{exp}}\gamma - \delta_{\text{est}}, p_t, \epsilon', \epsilon_{\text{EA}}), \quad (\text{B10})$$

$$\eta(p, p_t, \epsilon', \epsilon_{\text{EA}}) = f_{\min}(p, p_t) - \frac{1}{\sqrt{n}} 2 \left( \log 13 + \frac{d}{dp(1)} g(p)|_{p_t} \right) \sqrt{1 - 2 \log(\epsilon' \epsilon_{\text{EA}})}, \quad (\text{B11})$$

$$f_{\min}(p, p_t) = \begin{cases} g(p) & \text{if } p(1) \leq p_t(1), \\ \frac{d}{dp(1)} g(p)|_{p_t} \cdot p(1) + \left( g(p_t) - \frac{d}{dp(1)} g(p)|_{p_t} \cdot p_t(1) \right) & \text{if } p(1) > p_t(1) \end{cases} \quad (\text{B12})$$

$$g(p) = \begin{cases} 1 - h \left( \frac{1}{2} + \frac{1}{2} \sqrt{16 \frac{p(1)}{\gamma} \left( \frac{p(1)}{\gamma} - 1 \right) + 3} \right) & \text{if } \frac{p(1)}{\gamma} \in \left[ 0, \frac{2+\sqrt{2}}{4} \right] \\ 1 & \text{if } \frac{p(1)}{\gamma} \in \left[ \frac{2+\sqrt{2}}{4}, 1 \right] \end{cases} \quad (\text{B13})$$

Combined with the input sampling soundness, the total soundness is bounded by  $\epsilon_{\text{SA}} + \epsilon_{\text{EA}} + \epsilon'/2 + m\epsilon_1$ .

Parameter	Meaning
$\epsilon_c$	completeness (bounding honest abort probability)
$\epsilon_s$	soundness (bounding randomness security)
$\epsilon_{\text{SA}}$	input sampling error tolerance
$\epsilon'$	smoothing parameter
$\epsilon_1$	1-bit extractor error tolerance
$\epsilon_{\text{EX}}$	randomness extractor error tolerance
$\epsilon_{\text{est}}$	Bell estimation error tolerance
$\epsilon_{\text{EA}}$	soundness of entropy accumulation protocol

TABLE I: Meaning of security parameters.

Finally, let us count the number of random bits consumed in the protocol. It consists of the randomness used to decide if a round is a test or generation round, the randomness used to pick the inputs in a test round, and the randomness used for the Trevisan extractor. Taking into account finite statistical fluctuations, we need at most  $6h(\gamma)n$  bits to choose between test and generation except with probability  $\epsilon_{\text{SA}}$ . This results in at most  $2\gamma n$  testing rounds except with probability  $\epsilon_{\text{SA}}$ , which equates to  $2 \times 2\gamma n$  random bits being consumed for generating the inputs for test rounds. The randomness for Trevisan extractor is  $d$  bits. (Practically, after the first run of the protocol, we can omit this amount because the extractor is

a strong extractor: we can reuse the seed for next run of the protocol). Summing these up, we have consumed at most  $6h(\gamma)n + 4\gamma n + d$  uniformly random bits with probability at least  $1 - 2\epsilon_{\text{SA}}$ .

In summary, for a device with  $\omega_{\text{exp}}$ , any choice of  $\gamma, \epsilon_1, \epsilon', \epsilon_{\text{EA}} \in (0, 1)$ , and  $n$  large enough, our protocol is an  $(\epsilon_{\text{SA}} + \epsilon_{\text{est}}, \epsilon_{\text{SA}} + \epsilon_{\text{EA}} + \epsilon'/2 + m\epsilon_1)$ -secure  $[6h(\gamma)n + 4\gamma n + d] \rightarrow m$  randomness expansion protocol. That is either our protocol abort with probability greater than  $1 - \epsilon_{\text{EA}}$ , or it produces a string of length  $m$  such that  $\frac{1}{2} \|\rho_{SRE} - \rho_{U_m} \otimes \rho_{U_R} \otimes \rho_E\| \leq \epsilon_{\text{SA}} + \epsilon'/2 + m\epsilon_1$ . The protocol consume at most  $6h(\gamma)n + 4\gamma n + d$  uniformly random bits with probability at least  $1 - 2\epsilon_{\text{SA}}$ .

### Appendix C: Input/Output randomness analysis

The previous Appendix gives a complete picture of the (one-shot) behavior of our randomness expansion protocol. For the purpose of this paper, it suffices to obtain rough estimates on the randomness output, but further optimization can be done.

For simplicity, we introduce some bounds on the resources. Since  $m \leq 2n$  we can let  $m\epsilon_1 \leq 2n\epsilon_1 =: \epsilon_{\text{EX}}$  which gives  $\epsilon_1 = \epsilon_{\text{EX}}/(2n)$ . Plugging this back in (B8) gives us the number of random bits one can extract,

$$m = n \cdot \eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}) - 4 \log n + 4 \log \epsilon_{\text{EX}} - 10, \quad (\text{C1})$$

for a given level of soundness  $\epsilon_{\text{SA}} + \epsilon_{\text{EA}} + \epsilon'/2 + \epsilon_{\text{EX}}$ . Moreover, the protocol consumes  $6h(\gamma)n + 4\gamma n + d$  bits of randomness with probability at least  $1 - \epsilon_{\text{SA}}$ , where


$$d = a(2\ell)^2 \text{ with } a \leq \frac{\log(2n - 2e) - \log(2\ell - 2e)}{\log(2e) - \log(2e - 1)} + 1$$

$$\text{and } \ell \leq 3 \log n + 6 - 2 \log \epsilon_{\text{EX}}. \quad (\text{C2})$$

This leads to an expansion of  $m - 6h(\gamma)n - 4\gamma n - d$ .

Hence, the output randomness rate per unit time is

$$r_n = \frac{1}{\tau} \left( \eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}) - 4 \frac{\log n}{n} + 4 \frac{\log \epsilon_{\text{EX}}}{n} - \frac{10}{n} \right), \quad (\text{C3})$$

and the net randomness rate per unit time 

$$r_n^{\text{net}} = \frac{1}{\tau} \left( \eta_{\text{opt}}(\epsilon', \epsilon_{\text{EA}}) - 4 \frac{\log n}{n} + 4 \frac{\log \epsilon_{\text{EX}}}{n} - \frac{10}{n} - 6h(\gamma) - 4\gamma - \frac{d}{n} \right). \quad (\text{C4})$$

These formulas are of course given for a protocol with completeness  $\epsilon_{\text{SA}} + \epsilon_{\text{est}}$  and soundness  $\epsilon_{\text{SA}} + \epsilon_{\text{EA}} + \epsilon'/2 + \epsilon_{\text{EX}}$ , where

$$\epsilon_{\text{SA}} = \exp(-18h(\gamma)^3 n / L_{\text{max}}) \quad (\text{C5})$$

$$L_{\text{max}} = \max\{\log \gamma^{-1}, \log(1 - \gamma)^{-1}\} \quad (\text{C6})$$

$$\epsilon_{\text{est}} = \exp(-2n\delta_{\text{est}}^2). \quad (\text{C7})$$

The asymptotic rate for the block size  $n \rightarrow \infty$  is given by taking the limit of block size  $n$

$$r_\infty = \frac{1}{\tau} \left[ 1 - h \left( \frac{1}{2} + \frac{1}{2} \sqrt{\frac{S^2}{4} - 1} \right) \right]. \quad (\text{C8})$$

For the net asymptotic rate we could also take the same limit, however we could obtain a better bound by the expected behavior of the interval algorithm. Since the expected number of random bits needed to generate  $T_1, \dots, T_n$  is  $nh(\gamma) + 2$  by [22], and  $\gamma n$  of which is expected to be test rounds each consuming 2 random bits, we have the asymptotic net rate

$$r_\infty^{\text{net}} = \frac{1}{\tau} \left[ 1 - h \left( \frac{1}{2} + \frac{1}{2} \sqrt{\frac{S^2}{4} - 1} \right) - h(\gamma) - 2\gamma \right]. \quad (\text{C9})$$

From an end-user perspective, one may argue that the only parameters of interest are the completeness and

soundness security parameters which will constrain the rest of protocol parameters— $\gamma, \delta_{\text{est}}, n, \epsilon$ 's—for a given objective such as maximizing randomness rate or net randomness rate. Therefore, we will fix  $\epsilon_c = 10^{-10}, \epsilon_s = 10^{-10}$  in illustrative plots which lead to the following constraints

$$\epsilon_{\text{SA}} + \epsilon_{\text{EA}} + \epsilon'/2 + \epsilon_{\text{EX}} = 10^{-10}, \quad (\text{C10})$$

$$\epsilon_{\text{SA}} + \epsilon_{\text{est}} = 10^{-10}. \quad (\text{C11})$$

One then maximizes randomness output or net randomness output given these constraints. This gives us the best (as measured by our objective function) protocol parameters within the relaxations made to obtain (C1) and (C2).

However, in the main text we take a simpler approach without optimizing over the variables  $\gamma, \delta_{\text{est}}, \epsilon$ 's. For each block size  $n$ , we compute  $\epsilon_{\text{SA}}$  as given by (C5) which then fixes  $\delta_{\text{est}}$  via  $\epsilon_{\text{est}} = 10^{-10} - \epsilon_{\text{SA}}$  and (C7). The remaining  $\epsilon$ 's which have weight  $10^{-10} - \epsilon_{\text{SA}}$  are chosen in a 1 : 2 : 1 ratio of  $\epsilon_{\text{EA}} : \epsilon' : \epsilon_{\text{EX}}$ , which is guaranteed to add up to the specified level of completeness and soundness. This approach is not far from optimal in the regime of large block size  $n$ . This results in the experimental points reported in Figure 3.