Description of Data

****

This is a description of the data that were taken September 18 and 19th, 2015. The information in this document is likely incomplete. Please contact us at belltest@nist.gov with any questions that arise. We intend to update this file in response to questions and comments from our readers.

****

This version of the description was last updated on December 23, 2015.

The first numbers in the file name are the time at which the run started in Mountain Daylight Time, HH_MM At the end of the file name, there is either the string 'alice' or 'bob' identifying where the file was recorded.

All raw files have same file format. Each file is a timetagged sequence of records. Each record consists of three 8-byte unsigned integers. The first unsigned integer is the channel number that the timetagged event occurred in. The second channel number is the 8-byte integer raw timetag. Each timetag bin corresponds to 78.125 ps (a 12.8 GHz clock). The third 8-byte integer represents how many times the computer transferred data from the timetagger buffer in the computer to disk. This should be equivalent to the number of seconds the data run was for.

The timetaggers are locked to the same 10 MHz external reference, which originates in the source room and is distributed over optical fiber to both alice and bob via a pulsed laser and a photodiode receiver. The absolute timestamps at alice and bob are different because they were started at different times; but they are counting at the same rate and are synchronized by the 10MHz external reference.

For both parties, the channel assignments are identical:

Using the labels on the front panel:

channel 1 = detection events
channel 3 = The RNG output was for the 0 setting
channel 5 = The RNG output was for the 1 setting
channel 6 = Pulse-per-second GPS signal, at 1 Hz. This signal has ~10 ns jitter, but can have 100 ns jumps
channel 7 = The arrival of a sync pulse to turn on the Pockels cell (the sync signal is the output of the interface circuit that samples the random number generators). This signal is the time a random number is pulled from the QRNG, delayed by 18 feet of coax cable.

Channel assignments on the front panel and in the raw data **ARE DIFFERENT.** In the stored data, the channel numbers start counting at 0.

In the file,

0: Detector click
2: RNG output 0
4: RNG output 1
5: GPS PPS
6: Sync

Details for each data run by original filename:

September 18, 2015

00_44_CH_pockel_100kHz.run3:
Data file after initial attempt to get the proper delays in the experiment. This timing correction was not done properly.

01_11_CH_pockel_100kHz.run4.afterTimingfix:
Data file is the first data set after we corrected the timing. This data set has a large bias on one of the RNG setting choices. Stopping criterion was approximately 15 GB.

02_54_CH_pockel_100kHz.run4.afterTimingfix2:
Data file is after we corrected the bias in the RNG. Near end of file, laser loses mode locking. Stopping criterion was approximately 15 GB.

03_31_CH_pockel_100kHz.run4.afterTimingFix2_training:
Training data for analysis. This dataset is bad because the laser is not mode locked at all during the file.

03_43_CH_pockel_100kHz.run4.afterTimingfix2_afterfixingModeLocking:
Continuation of above file, after modelocking was fixed. Fridge warms up near end of data run. Combined stopping with above data file was approximately 15 GB.

September 19, 2015

17_04_CH_pockel_100kHz.run.completeblind:
Complete blind data set. Stopping criterion of 1 hour. This is not in the repository and we have not yet analyzed it. This is being saved for later analysis for potentially more sophisticated hypothesis tests. We have no idea about the quality of this data.

19_44_CH_pockel_100kHz.run.nolightconeshift:
Another data set for optimal configuration. Stopping criterion was 30 minutes.

21_15_CH_pockel_100kHz.run.200nsadditiondelay_lightconeshift:
We delayed the Pockels cell turn-on times by an additional 200 ns at both Alice and Bob. Stopping criterion of 30 minutes.

22_20_CH_pockel_100kHz.run.200nsreduceddelay_lightconeshift:
We changed the Pockels cell turn-on times by removing 200 ns of delay from the optimal timing configuration. Stopping criterion of 30 min.

23_55_CH_pockel_100kHz.run.ClassicalRNGXOR:
We also XOR'd in classical random numbers. 30 min stopping criterion

00_25_CH_pockel_100kHz.run.ClassicalRNGXOR_2:
Tried realigning and doing an additional run. 30 min stopping criterion

02_31_CH_pockel_100kHz.run.ClassicalRNGXOR_3:
Realigned again and cleaned up Alice's computer. 30 min stopping criterion

03_12_CH_pockel_100kHz.run.Blind_2:
Another blind data run with no monitoring. After data were collected, we saw the fridge was warm. 30 min stopping criterion. This is not in the repository. This is being saved for later analysis for potentially more sophisticated hypothesis tests. We have no idea about the quality of this data.

Description of Folders:

compressed:

In this folder, the raw data from alice and bob has been converted it into a new file. The 8-byte unsigned integer that describes the channel has been compressed to a 1-byte unsigned integer. The 8-byte integer describing the computer transfer number has been compressed to a 2-byte unsigned integer. The data has also been 'zip'ed to keep the file size under 2 GB. For some files, the zip archive has been split into multiple files that are each less than 2 GB. Using the proper unzip command, the split zip files can be combined into one large data file. On Windows machines, use the 7-zip utility to unzip the split files.

There is a folder call anomoly_examples. This is data that has been extracted from the raw timetagger data for various people.

processed_compressed

This folder contains processed data by Sae Woo's programs to convert the time stamp files to a list of events. The data was processed using the 'compressed' data files, NOT the original raw files.

processed_compressed/cw45:

This is a subfolder of processed_compressed. The file is a time-ordered record of events. Each event is stored as 5 bytes. Byte #1 is the settings used: the two least significant bits (bits 0 and 1) contain Alice's settings; bits 2 and 3 describe Bob's settings. In other words:

bit #0: Alice RNG setting 0 (this is “1” if Alice’s RNG setting was 0; otherwise it is “0”)

bit #1: Alice RNG setting 1 (this is “1” if Alice’s RNG setting was 1; otherwise it is “0”)

bit #2: Bob RNG setting 0

bit #3: Bob RNG setting 1

On extremely rare occasions, one will observe events where both RNG settings fire at either Alice or Bob; or neither fires.

Bytes #2 and #3 are an unsigned 2-byte integer that represents which laser cycle during the Pockels cell window may have seen a click in Alice. There are 16 possible times while the Pockels cell is on that Alice could have recorded a click. The slot/laser cycle number is represented by the bit position in the 16-bit (2-byte) unsigned integer (i.e. slot one corresponds to bit 0, slot 16 corresponds to bit 15).

Bytes #4 and #5 are an unsigned 2-byte integer for Bob's clicks, with the same formatting.

Metadata about how events were defined is not contained in these files... This is a serious drawback about using these files for analysis. We will upload the yaml file that we used for input into the program used to create these files. However, we would discourage the use of these files because the full analysis chain is not documented. The folder name "cw45" was used to describe the timing window at Alice and bob used to detemine if a click occurred. radius 4 at alice, and radius 5 at bob... However, the time delay from the 'sync' pulse is not recorded in the file or folder names. It has to be extracted from the yaml file used by the analysis program to creates these files. The program used to create these files was: build_file_txt.py . It is in the 'bell_analysis_code' bitbucket repository

processed_compressed/hdf5:

These are files created by processing the 'compressed' files (NOT the raw data files) into a form that can be used later to generate 'event-by-event' files for input to calculate the CH inequality, J, p-values, etc. It is an attempt to convert the raw data into a form that can more easily be used to calculate parameters that we would like to extract from the experiment. The files are in hdf5 file format and contain some metadata about how the data were processed from raw data into this form. These files are also divided into two folders that represent the dates that the raw data were taken. The file names themselves start with the raw data filename that they are derived from. However, data from Alice and Bob are merged into one file, so there is no 'alice' or 'bob' string. There is a '.build' extension.

Things stored in this hdf5 file: settings for alice and bob, clicks at alice and bob described by unsigned 2-byte integers using the format described above; a list of "anomalous" syncs that were observed by Alice and Bob; how much of each file was initially skipped to determine the clock misalignment between Alice and Bob; how many syncs / events are in each processed data block; metadata about the relative delay from the sync used to define a click in alice/bob and its radius. There is also raw data about the delay for each click in alice's and bob's detectors relative to the sync so new "click" definitions can be calculated without having to go back to the raw data files. The sync number for each detector click in alice and bob is also recorded so new 'click' definitions can be calculated.

rawdata_from_servers:

This is a “zipped” copy of the raw data from the servers. The actual raw data files are sometimes greater than 2 GB. This causes a storage problem on some systems. Therefore, the datasets were compressed and split into files that were less than 2 GB. To reconstruct the full data file, a program that can rebuild the full file from potentially two or more zip files must be used. This folder includes not only the raw data, but also the data used to help us determine clock offsets right before taking data. This data, which has the extension "T1.dat" or "T2.dat," can be used to check/train analysis programs.

Peter’s Data Calculations:

This is a folder contains processed data and spreadsheets that Peter Bierhorst used to determine an exact p-value. The files are in text format. The file format depends upon what Peter and Sae Woo have agreed upon. It may not be the same for each file.

The “rate estimates file” contains the calculations that were used to determine the Nx cutpoints, following the method described in the Supplementary Material of our paper. The rough rate estimates for a data run were made by looking at the observed rates of previous data sets and looking at the first small portion of data for the run (that was discarded).

The folders “First three runs” and “3 classical XORS” contain the Excel files that were used to compute the p-values in Table S-I. These two folders contain six files overall. Each one of these files contains five tabs. Each tab corresponds to a pulse grouping, so the tab labelled 456 contains the data for the three-pulse grouping. (There is a one-off mismatch in indexing compared to Table S-I so this corresponds to pulse group 5-7 in the table.) Note that the largest pulse grouping was not reported in the table because the degree of confidence in spacelike separation for this pulse grouping was not high. In these flies, each tab contains 16 columns corresponding to the 16 possible outcomes, in the order ++ab, +0ab, 0+ ab, 00ab, followed by the same outcomes for ab’, then a’b, then a’b’. Due to a programming error, the 00 column entries may be incorrect; however, this bug was never fixed in these files because the 00 entries do not affect the calculation of the p-value. Column R contains a running tally of the relevant T trials (trials whose outcome appears in the Bell inequality being tested). The first row contains the sum of the data from the first 85% of the run (75% for one of the six files), and remaining rows record data in increments of 1/50 of a second starting from this point and going forward. This fine-graining allows one to find the Nx cutpoint in column R. This location is marked with the word “cut point” in column Q, and the p-values are calculated automatically to the right. Some other calculations may appear, such as settings frequencies, other Eberhard statistics, No-signalling statistics -- these represent artifacts of exploratory calculations and should not be taken to represent meaningful figures (the formulas may contain errors, and counting of 00 trials may cause problems); this caveat also applies to everything at the bottom of the columns in these files, which should also be ignored.

The folder “diagnostic statistics” contains the data files used to calculate the statistics in S-IV-C and Table S-III. The first row contains the 16 entries corresponding to the total outcome data from the entire run, following the same order explained in the last paragraph. (Note that the statistics here are calculated for the entire run, and not at the Nx cutpoint.) The files also contain one tab for each pulse grouping as before. The various statistical calculations are annotated.

Code:

This is a folder contains folders that contain the code used to take the data and code used to process the raw data into data that was used by Peter and others to compute exact p values.