SCIENTIFIC REPORTS

OPEN

Improving randomness characterization through Bayesian model selection

Received: 4 November 2016
Accepted: 7 April 2017
Published online: 08 June 2017

Rafael Díaz Hernández Rojas[1], Aldo Solís[2], Alí M. Angulo Martínez[2], Alfred B. U'Ren[2], Jorge G. Hirsch[2], Matteo Marsili[3] & Isaac Pérez Castillo[1,4]

Random number generation plays an essential role in technology with important applications in areas ranging from cryptography to Monte Carlo methods, and other probabilistic algorithms. All such applications require high-quality sources of random numbers, yet effective methods for assessing whether a source produce truly random sequences are still missing. Current methods either do not rely on a formal description of randomness (NIST test suite) on the one hand, or are inapplicable in principle (the characterization derived from the Algorithmic Theory of Information), on the other, for they require testing all the possible computer programs that could produce the sequence to be analysed. Here we present a rigorous method that overcomes these problems based on Bayesian model selection. We derive analytic expressions for a model's likelihood which is then used to compute its posterior distribution. Our method proves to be more rigorous than NIST's suite and Borel-Normality criterion and its implementation is straightforward. We applied our method to an experimental device based on the process of spontaneous parametric downconversion to confirm it behaves as a genuine quantum random number generator. As our approach relies on Bayesian inference our scheme transcends individual sequence analysis, leading to a characterization of the source itself.

Random numbers have acquired an essential role in our daily lives because of our close relationship with communication devices and technology. There are also numerous scientific techniques and applications that rely fundamentally on our ability for generating such numbers and typically pseudo-random number generators (pRNGs) suffice for those purposes. A new alternative has been proposed by exploiting the inherently probabilistic nature of quantum mechanical systems. These Quantum Random Number Generators (QRNGs) are in principle superior to their classical counterparts and recent experiments have shown ref. 1 that they can reach the same quality as commercial pRNGs. However, the natural question of how to assess whether a sequence is truly random is not yet fully established. Pragmatically, the NIST test suite[2] has become the standard method for analysing sequences coming from a RNG. The suite is based on testing certain features of random sequences that are hard to reproduce algorithmically, such as its power spectrum, longest string of consecutive 1's, and so on. Even though it constitutes an easily applicable procedure, recent findings show that its reliance on *P*-values is a drawback[3,4], while its lack of formality is a major disadvantage. On the other hand, although no definition of randomness is deemed absolute, a rigorous characterization is presented by the Algorithmic Theory of Information (ATI) but it is unfortunately inapplicable in real cases[5]. An alternative which overcomes both formal and applicability issues is the Borel-normality criterion[6] (BN). Intuitively, this approach works by successively compressing a given dataset, e.g. $\hat{s} = \{0101010010101010101011010\cdots\}$ of M bits, by taking strings of β consecutive bits and computing the frequency of occurrences $\gamma_i^{(\beta)}$ of each of those $i = 0, 1, \ldots, 2^{\beta} - 1$ possible strings. For example, $\beta = 1$ corresponds to looking for the frequencies of the strings $\{0, 1\}$ in the dataset $\hat{s}$, while $\beta = 2$ corresponds to analysing the frequencies of the strings $\{00, 01, 10, 11\}$, and so on. The whole sequence is said to be Borel-normal if the frequencies are bounded individually according to

[1]Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, Cd. Mx., C.P., 04510, Mexico. [2]Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Apdo. Postal 70-543, Cd. Mx., C.P., 04510, Mexico. [3]The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151, Trieste, Italy. [4]London Mathematical Laboratory, 14 Buckingham Street, London, WC2N 6DF, United Kingdom. Correspondence and requests for materials should be addressed to I.P. (email: isaacpc@fisica.unam.mx)

$$\left|\gamma_i^{(\beta)} - \frac{1}{2^\beta}\right| < \sqrt{\frac{\log_2 M}{M}}, \tag{1}$$

and with β an integer ranging from 1 to $\beta_{\max} = \log_2 \log_2 M$. It is important to mention that BN criterion is a (nearly) necessary condition for a sequence to be considered random[5]. Note that this test is restricted to a-single-sequence classification, so it cannot determine the random character of the generating *source*.

In the present work, we show that randomness characterization can also be addressed using a Bayesian inference approach for model selection[7], borrowing the compression scheme of BN. For simplicity, for a fixed β we denote each string with its decimal base representation $j \in \{0, 1, \ldots, 2^\beta - 1\} \equiv \Xi_\beta$. The first step consists in identifying the models which could have generated a compressed dataset $\hat{s}$. For instance if $\beta = 1$, we can describe it as M realizations of a Bernoulli process, leading to two possible models: with and without bias. Similarly, for $\beta = 2$, a model represents a way of constructing $\hat{s}$ with bias in some of the 2^2 possible strings. A simple combinatorial counting reveals that all the possible bias assignations correspond to all partitions of the four strings of Ξ_2.

Thus, in general, given the set Ξ_β, let $\mathcal{P}_{\Xi_\beta}$ denote the family of its $B_{2^\beta} = \sum_{K=1}^{2^\beta} \left\{ {2^\beta \atop K} \right\}$ possible partitions[8], with B_{2^β} the Bell's numbers and $\left\{ {2^\beta \atop K} \right\}$ the Stirling numbers of the second kind, which counts the different ways of grouping 2^β elements into K sets. Formally, $\alpha_\ell^{(K)} = \left\{\omega_\ell^{(1)}, \ldots, \omega_\ell^{(K)}\right\} \in \mathcal{P}_{\Xi_\beta}$ would refer to the ℓ-th partition into K subsets, but for notational simplicity we will omit henceforth the index ℓ. To each partition $\alpha^{(K)}$ there corresponds a unique model $\mathcal{M}_{\alpha^{(K)}}$ which assigns a probability p_j to string $j \in \Xi_\beta$ according to the following rule:

$$\mathcal{M}_{\alpha^{(K)}} = \left\{ p_j = \frac{\theta_r}{|\omega^{(r)}|}; \quad \forall r = 1, \ldots, K; \quad \forall j \in \omega^{(r)} \right\}. \tag{2}$$

This means that all strings contained in a given subset $\omega^{(r)}$ are deemed equiprobable within the specified model. Thus, keeping β fixed, the likelihood of observing the given dataset $\hat{s}$ in a model $\mathcal{M}_{\alpha^{(K)}}$ is:

$$P(\hat{s}|\mathcal{M}_{\alpha^{(K)}}, \{\theta_r\}_{r=1}^K) = \prod_{r=1}^K \left(\frac{\theta_r}{|\omega^{(r)}|} \right)^{k_{\omega^{(r)}}}, \tag{3}$$

where $k_j^{(\beta)}$ is the frequency of string $j \in \Xi_\beta$ and we have defined $k_{\omega^{(r)}} = \sum_{j \in \omega^{(r)}} k_j^{(\beta)}$ as the aggregate frequencies of the strings in the subset $\omega^{(r)}$. (For further use, we also introduce the relative aggregate frequencies $\gamma_{\omega^{(r)}} = \frac{\beta}{M} k_{\omega^{(r)}}$). From this perspective, only the model that is symmetric under any reordering of the possible strings is identified with a complete random source, because any other model entails biases assignations according to the strings' grouping represented by the corresponding partition. This symmetry only exists when the partition is the set Ξ_β itself, hence we denote $\mathcal{M}_{\alpha^{(1)}} = \mathcal{M}_{\text{sym}}$.

Consider now that when characterising randomness the only essential feature is whether bias for or against some strings is present, but the degree of bias is irrelevant. We can eliminate the dependence on the bias parameters by multiplying with a prior for $\{\theta_r\}_{r=1}^K$ and derive the so called *evidence* for a given model[9]. Following[10], we use the Jeffreys prior for it yields a model's probability distribution invariant under reparametrization and provides a measure of a model's complexity, thus giving a mathematical representation of Occam's Razor principle[10–12]. After integrating in the parameter space, we arrive at (see Supplementary Information (SI), Sec. 2)

$$P(\hat{s}|\mathcal{M}_{\alpha^{(K)}}) = \frac{\Gamma(\frac{K}{2})}{\Gamma^K(\frac{1}{2})} \prod_{r=1}^K \left(\frac{1}{|\omega^{(r)}|}\right)^{\frac{M}{\beta}\gamma_{\omega^{(r)}}} \frac{\prod_{r=1}^K \Gamma(\frac{1}{2} + \frac{M}{\beta}\gamma_{\omega^{(r)}})}{\Gamma(\frac{K}{2} + \frac{M}{\beta})}. \tag{4}$$

Eq. (4) is our main result, for it will let us perform the model selection straightforwardly. For $\mathcal{M}_{\text{sym}}$, its evidence is fairly intuitive:

$$P(\hat{s}|\mathcal{M}_{\text{sym}}) \equiv P(\hat{s}|\mathcal{M}_{\alpha^{(1)}}) = 2^{-M}. \tag{5}$$

Finally, we want to infer the model that best describes our source, *after* a dataset $\hat{s}$ is given. Using Bayes' theorem the posterior distribution $P(\mathcal{M}_{\alpha^{(K)}}|\hat{s})$ reads:

$$P(\mathcal{M}_{\alpha^{(K)}}|\hat{s}) = \frac{P(\hat{s}|\mathcal{M}_{\alpha^{(K)}})P_0(\mathcal{M}_{\alpha^{(K)}})}{\sum_\gamma P(\hat{s}\left|\mathcal{M}_\gamma\right.)P_0(\mathcal{M}_\gamma)}. \tag{6}$$

Henceforth we will consider a uniform prior over models (which is justified in SI), so the model's posterior is simply proportional to its evidence.

Suppose now we want to assess whether a source can be considered truly random. This is performed in two steps. As the first step, we need a model ranking procedure based on the posterior distribution. The second step consists in quantifying the goodness of our choice of model.

As a decision rule for the ranking process we use the Bayes Factor[13] perspective,

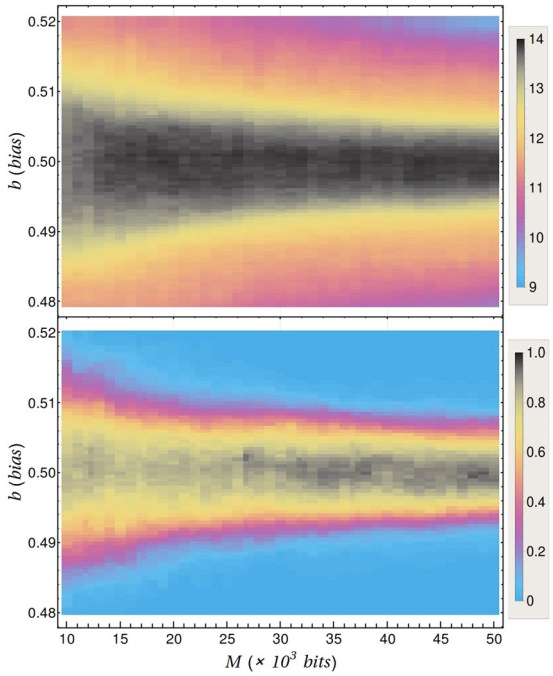

Figure 1. Phase diagram of Randomness Characterisation. Division of the parameter space into regions according to the likeliest model. The top figure corresponds to $\beta = 1$ in terms of the frequency γ_0 of the string 0 and the sample size M. The green curves corresponds to Borel's normality criterion, while the red curves are Borel-type bounds obtained by an approximation obtained from Eq. (4) (see Sec. 3 of SI). The bottom plot corresponds to $\beta = 2$ where each coloured area identifies the likeliest model in that region. Here we fixed the frequencies $\gamma_1 = 1/6$ and $\gamma_2 = 1/4$ and varied the frequency γ_0 of the string 00 and the sample size M.

$$\mathrm{BF}_{\alpha,\alpha'} = \frac{P(\mathcal{M}_\alpha|\hat{s})}{P(\mathcal{M}_{\alpha'}|\hat{s})} = \frac{P(\hat{s}|\mathcal{M}_\alpha)}{P(\hat{s}|\mathcal{M}_{\alpha'})}. \tag{7}$$

Thus, we will choose $\mathcal{M}_\alpha$ over $\mathcal{M}_{\alpha'}$ whenever $\mathrm{BF}_{\alpha,\alpha'} > 1$. It has been shown that $\mathrm{BF}_{\alpha,\alpha'}$ provides a measure of goodness of fit and $\lim_{M\to\infty} \mathrm{BF}_{\alpha,\alpha'} = \infty$ if $\mathcal{M}_\alpha$ is the true model[14].

To implement the second step, which is nothing more than a hypothesis testing problem, we have two alternatives: either we check whether $\log_{10} \mathrm{BF}_{\alpha,\alpha'} \geq 2$ which is considered decisive in favour of model $\mathcal{M}_\alpha$[13], or we compute the ratio between the posterior and the prior of a given model to assess how certain the posterior has become under the information provided by the dataset.

From a computational point of view notice that the evaluation of the posterior requires to being able to compute the normalization factor $\sum_\gamma P(\hat{s}|\mathcal{M}_\gamma)P_0(\mathcal{M}_\gamma)$ that appears in (6). When the number of models is very large we can choose either to work with a subspace of models or use the logarithm of the Bayes Factor, as in this case the normalisation factor cancels out.

It is clear that a full test of randomness requires different values of β to be used for the same dataset, while the strings should be short enough so that the M bits allow for each of the possible models to be sampled at least once. Thus, heuristically, $B_{2^{\beta_{\max}}} \sim M$ whence we can reproduce the BN limit[6], $\beta_{\max} \sim \log_2 \log_2 (M)$, after using an asymptotic expansion for the Bell number.

Note that by fixing β we have the set of parameters $(\{\gamma_j\}_{j=0}^{2^\beta - 1}, M)$, whose space can be divided into regions identifying the likeliest model according to Eq. (4). As illustrative cases, in Fig. 1 we show a phase-type diagram for $\beta = 1$ and $\beta = 2$ (upper and lower panel, respectively), where the orange-filled area delimits the parameters values that renders $\mathcal{M}_{\mathrm{sym}}$ the likeliest model. The top panel includes the bounds according to the BN criterion (green curves) given by Eq. (1), and shows that for any sequence length, M, our method allows for considerably smaller variations of γ_0. This is a significant improvement, since only necessary criteria exist for testing randomness. The lower panel depicts the analogous regions when $\beta = 2$, for which there are fifteen models (see a list in the SI) and we have fixed two frequencies: $\gamma_1 = 1/6$ and $\gamma_2 = 1/4$. The complete models distribution can be deduced from the structure of this graph, by distinguishing, *a posteriori*, the equiprobable strings for which the corresponding model is the likeliest. Thus more information than complete randomness classification can be readily obtained from our method.

Also in Fig. 1, the red curves of the $\beta = 1$ case are bounds obtained by comparing the likelihood of $\mathcal{M}_{\mathrm{sym}}$ with models involving partitions into $K = 2$ subsets. Agreement with the regions boundary is excellent. Our choice of $K = 2$ is justified as we would expect that models corresponding to partitions into two subsets to be the closest

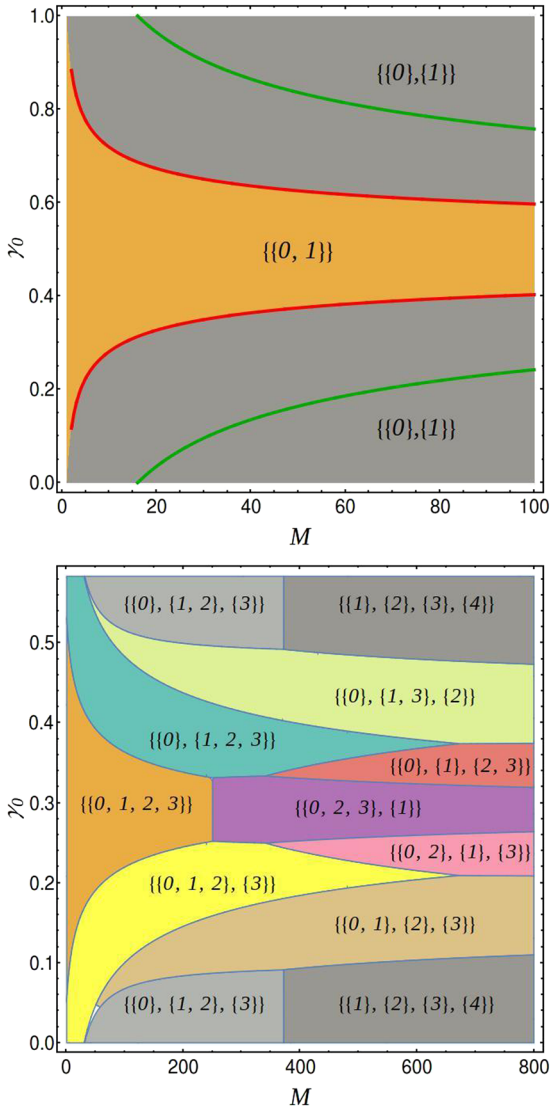

Figure 2. Comparison with NIST Suite test. Comparison of the bias allowed on a given sequence for it to be considered random using the NIST suite (upper panel) and our Bayesian method for randomness characterisation (lower panel).

ones to the model $\mathcal{M}_{\text{sym}}$. An explicit expression for these bounds is derived in SI, Sec. 3, and Extended Data Figs 2 and 3 depict that they also bound considerably well the region in which $\mathcal{M}_{\text{sym}}$ is the likeliest for $\beta = 2$.

For further benchmarking, we have compared our method against the NIST test suite[2]. The result is depicted in Fig. 2, as a function of the sequence length M and bias b employed to generate a 0. The upper panel on Fig. 2 shows the averaged number of tests passed when employing the NIST suite, while the lower one shows the frequency of $\mathcal{M}_{\text{sym}}$ being the likeliest, for $\beta = 1$, 2 and 3. We believe that our technique can contribute to test the quality of RNG in a more stringent form, since by applying a single test thrice (once for each value of β), we determined more precisely the random character of the sample of sequences.

As an application, we have tested our method in a bit sequence obtained experimentally from the differences in time detection in the process of spontaneous parametric down conversion (SPDC). Sequences generated via a SPDC photon-pair source have been shown to fulfil with ease the BN criterion, and to pass comfortably the NIST's suite[1]. In the SPDC process a laser pump beam illuminates a crystal with a $\chi^{(2)}$ nonlinearity, leading to the annihilation of pump photons and the emission of photon pairs, typically referred to as signal and idler[15]. Our experimental setup is shown in Extended Fig. 1 and we explain how to construct a 0 or 1 symbol from the detection signals in Section 1 of SI. We generated a 4×10^9 bits sequence, so $\beta_{\text{max}} \sim 4$. When $1 \leq \beta \leq 3$, we used all the possible models in the comparison, while, for computational ease, when $\beta = 4$, we restricted the model space to the 32, 768 models corresponding to $K = 1$ and $K = 2$ subsets (consider that $B_{2^4} = 10^{10}$). Our inference showed that $\mathcal{M}_{\text{sym}}$ was the likeliest model for every value of β.

As explained above, to achieve a full characterization of our QRNG as a random *source*, we need to go further from the model ranking based on the Bayes Factor and measure our certainty that $\mathcal{M}_{\text{sym}}$ is the true model gov-

β	$P(\mathcal{M}_{\text{sym}}\|\hat{s})$	$\log_{10}\text{BF}_{\text{sym},\alpha'}$
1	0.99993	4.15
2	0.99927	$\geq$3.55
3	0.95374	$\geq$1.84
4	0.31862	$\geq$3.16

Table 1. Posterior $P(\mathcal{M}_{\text{sym}}|\hat{s})$ calculated for a dataset of 4×10^9 bits.

erning the source. This (un)certainty quantification is the hallmark of Bayesian statistics, since $P(\mathcal{M}_{\text{sym}}|\hat{s})$ represents the probability that modelling our QRNG as a random source is correct. Computing this posterior distribution directly from Bayes' Theorem, Eq. 6, we arrive at the values shown in Table 1 for each β. The first three values are at least 0.95, but the corresponding to $\beta = 4$ is about 0.32, considerably smaller. However, this represents an improvement of order 10^4 when compared with the initial value for the prior, $P_0(\mathcal{M}_{\text{sym}}) = 1/32,768 \approx 3.1 \times 10^{-5}$. Alternatively, we computed $\log_{10}\text{BF}_{\text{sym},\alpha'}$ for each value of β. The values reported in Table 1 correspond to the comparison of $\mathcal{M}_{\text{sym}}$ and the second likeliest model, hence the inequality for $\beta > 2$. These two criteria combined lead us to conclude that there is decisive evidence for our hypothesis that $\mathcal{M}_{\text{sym}}$ is the underlying model driving our source, thus verifying that the photonic RNG is strictly random in the sense described in the article.

From a more general perspective, we propose that $P(\mathcal{M}_{\alpha^{(K)}}|\hat{s})$ quantifies our certainty on the hypothesis that a sequence $\hat{s}$ was generated using the biases on strings associated with $\alpha^{(K)}$. Because Bayesian methods entails a model's generalizability[9,10], the likeliest model provides a characterization of the source of $\hat{s}$. All partitions can be identified with standard computational packages, although it can be computationally demanding for sequences of $\sim 10^{10}$ bits. In any case, once a partition is given, its model's likelihood is easily found using Eq. (4). A simplified analysis can be performed with the BN-type bounds given in Section 3 of the SI, which also leads to more stringent criteria than other approaches.

References

1. Solis, A. *et al*. How random are random numbers generated using photons? *Physica Scripta* **90**, 074034 (2015).
2. Rukhin, A. *et al*. *Statistical test suite for random and pseudorandom number generators for cryptographic applications*, *nist special publication* (Citeseer, 2010).
3. Pareschi, F., Rovatti, R. & Setti, G. Second-level NIST randomness tests for improving test reliability. In *Circuits and Systems*, *2007. ISCAS 2007. IEEE International Symposium on*, 1437–1440 (IEEE, 2007).
4. Wasserstein, R. L. & Lazar, N. A. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 129–133 (2016).
5. Calude, C. S. *Information and Randomness*: *An Algorithmic Perspective*, 2nd edn. (Springer Publishing Company, Incorporated, 2010).
6. Calude, C. Borel normality and algorithmic randomness. In *Developments in Language Theory*, vol. 355, 113–129 (Citeseer, 1993).
7. Haimovici, A. & Marsili, M. Criticality of mostly informative samples: a Bayesian model selection approach. *Journal of Statistical Mechanics: Theory and Experiment* **2015**, P10013 (2015).
8. Pemmaraju, S. & Skiena, S. S. *Computational Discrete Mathematics*: *Combinatorics and Graph Theory with Mathematica* (Cambridge university press, 2003).
9. MacKay, D. J. Bayesian interpolation. *Neural computation* **4**, 415–447 (1992).
10. Myung, I. J., Balasubramanian, V. & Pitt, M. A. Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences* **97**, 11170–11175 (2000).
11. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural computation* **9**, 349–368 (1997).
12. Balasubramanian, V. A geometric formulation of Occam's razor for inference of parametric distributions. *arXiv preprint adap-org/9601001* (1996).
13. Robert, C. *The Bayesian choice*: *from decision-theoretic foundations to computational implementation* (Springer Science & Business Media, 2007).
14. Verdinelli, I. & Wasserman, L. *et al*. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *The Annals of Statistics* **26**, 1215–1241 (1998).
15. Burnham, D. C. & Weinberg, D. L. Observation of simultaneity in parametric production of optical photon pairs. *Physical Review Letters* **25**, 84 (1970).

Acknowledgements

I.P.C. and R.D.H.R. thank hospitality to the Abdus Salam ICTP. R.D.H.R. also thanks Susanne Still, Valerio Volpati, and Aaron King for helpful discussions regarding the choice of models priors. This work has received partial economical support from Consejo Nacional de Ciencia y Tecnología (Conacyt): SEP-Conacyt and RedTC-Conacyt, Mexico, PAPIIT-UNAM project IN109417, and PAPIIT-UNAM project IA103417. We also want to thank Mark Buchanan for his helpful feedback for writing the manuscript.

Author Contributions

I.P.C., R.D.H.R. and M.M. developed the Bayesian approach for the current application and derived the analytic expressions for the evidence of models. A.S., J.G.H., A.U., and A.M.A.M. furnished our work as a randomness characterization and provided the experimental datasets. The comparison with the NIST test suite and BN criterion was done by R.D.H.R. and A.S. All authors discussed the results and commented the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03185-y

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.