POLITECNICO DI TORINO
Repository ISTITUZIONALE

Second-level testing revisited and applications to NIST SP800-22

Original
Second-level testing revisited and applications to NIST SP800-22 / Pareschi, F., Rovatti, R., Setti, G.. - STAMPA. - (2007), pp. 627-630. (European Conference on Circuit Theory and Design 2007, ECCTD 2007 Seville, esp August 26-30, 2007) [10.1109/ECCTD.2007.4529674].

Availability:
This version is available at: 11583/2850185 since: 2020-10-27T22:52:04Z

Publisher:
IEEE Computer Society

Published
DOI:10.1109/ECCTD.2007.4529674

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2007 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

15 June 2026

This is the author's version of the article that has been presented at IEEE ECCTD2007
The editorial version of the paper is available at http://dx.doi.org/10.1109/ECCTD.2007.4529674

Second-level testing revisited and applications to NIST SP800-22

Fabio Pareschi*‡, Riccardo Rovatti†‡ and Gianluca Setti*‡
*ENDIF - University of Ferrara, via Saragat 1, 44100 Ferrara - ITALY
†DEIS - University of Bologna, viale risorgimento 2, 40136 Bologna - ITALY
‡ARCES - University of Bologna, via Toffano 2/2, 40125 Bologna - ITALY
Email: {fabio.pareschi,gianluca.setti}@unife.it, rrovatti@arces.unibo.it

***Abstract*— The use of second-level testing to reduce Type II errors in RNG validation was suggested from the very beginning though rarely employed in real-world cases. Yet, as security requirements become more critical and the availability of even faster RNG more commonplace, second-level testing will be key to distinguishing RNGs based on the quality of very large chunks of their output. This paper addresses some principles governing the proper design of second-level tests (i.e. how to divide available data into chunks and how to compute second-level p-values) as well as its implications on the design of the underlying basic tests.**

I. Introduction

Being able to test and validate a random number generator (RNG) is a complex and extremely important task, especially in applications like cryptography where RNGs represent a critical point for security [1]. Many tests have been developed in recent years for determining if a sequence can be considered randomly generated or not. In this paper we focus on the class of tests known as *statistical tests for randomness* [2], [3], which represent the most used and the most studied tests in literature.

Statistical tests are based on the standard statistical hypothesis testing approach. Given the hypothesis $\mathcal{H}_0$ that the sequence has been randomly generated, they compute a p-value, that is a probability measure indicating the strength of the evidence provided by the data against the hypothesis of a random sequence. More detailed, a test looks at some statistical features of the sequence, expresses them as a numerical quantity and compares it with the quantity expected if the sequence were coming from an ideal random number generator. The p-value p_v is computed in a way that $p_v = 1$ means that the observed statistical feature is exactly the expected one from a random sequence, while $p_v = 0$ means that the feature is completely different from what expected. Furthermore, the p-values coming from all possible sequences generated by a perfect RNG are uniformly distributed in the interval $[0, 1]$.

The interpretation of a test is the following: having chosen the number α, $\mathcal{H}_0$ is rejected (i.e. the test is considered *failed*) if $p_v < \alpha$, while $\mathcal{H}_0$ is accepted if $p_v \geq \alpha$. With this statistical approach, two errors can be committed:

- reject $\mathcal{H}_0$ when the sequence is generated by a perfect random generator *(Type I error)*
- accept $\mathcal{H}_0$ when the sequence is generated by a generator that is non random *(Type II error)*.

As far as the Type I error is concerned, its probability is α; for this reason, α is also called *level of significance*. However, the computation of the probability of a Type II error is not possible because it would require the characterization all possible non-random generators.

We can notice that if the significance level α is chosen too high, then the test may reject many sequences that were, in fact, random. On the other hand, if α is too low, then there is the danger that the test may accept plenty of sequences even though they were not randomly produced. A typical value is $\alpha = 0.01$ that is the value suggested by US National Institute of Standard and Technology (NIST) in its test suite [2].

In this paper we discuss the two concepts of reliability and accuracy of a random test. We define a test as *not reliable* when, due to errors or approximations in the p-value computation, the distribution of p-values for sequences generated by a perfect RNG is not uniform; in this way the probability of a Type I error can be very different from α (and also much higher). Instead, we refer to *accuracy* looking at Type II errors. Even if the natural definition of accuracy is linked to Type II error probability, we relate accuracy to the ability, given a non-random generator, of recognizing its sequences as non-random. This is a practical definition due to the impossibility of computing Type II error probability.

As reliability is a fundamental requirement for any statistical test, accuracy can be used to compare different tests: the higher the accuracy, the better the test.

A typical way to increase the accuracy of a statistical test is to consider a meta-test composed by a number T of different basic tests that look for different statistical features, and to apply them to the same sequence. In other words, the meta-test is now sensitive to deviation from ideality in all the statistical features observed by the basic tests.

In the following we consider another, complementary kind of meta-test that we address as *second-level* testing. Instead of considering several results from different tests over the same sequence, we take into account several results from the same test over different sequences. It has already been shown in [4] that this approach produces more accurate results; it also been shown that the number of basic tests involved in a second-level test must be limited in order to get reliable results. In this paper some different ways to aggregate basic test results into a single second-level test are analyzed, with the aim to identifying which method provides the highest accuracy. We also provide some considerations about the design of a basic test, to get the most reliable results from the second-level testing approach. In this paper we consider tests coming from the SP 800-22 test suite from NIST. The main reason we focus our attention on this suite is that, from an engineering point of view, it has several appealing properties. First, it is uniform: it is composed of several different tests, each of which are applied to the same sequence of n bits (the NIST suggests $n = 10^6$). Second, even if some flaws still exist [5], for all tests in the suite an exhaustive mathematical treatment is available and well documented.

Copyright (C) 2007 IEEE. Personal use is permitted.
For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org

This is the author's version of the article that has been presented at IEEE ECCTD2007
The editorial version of the paper is available at http://dx.doi.org/10.1109/ECCTD.2007.4529674

The paper is organized as follows. In section II, we analyze a real test as a case study with an aim to identifying any error in the computation of p_v. In section III, we describe some methods to aggregate several tests into a single second-level test, and make some considerations about accuracy and reliability of the approaches based on the results of the former section.

As a final remark, in this paper we address two pseudorandom generators. The first one is the 32 bits version of the KISS [6], which is a very simple but effective generator designed by G. Marsaglia; this generator will be used for checking the accuracy of the test, looking for a test capable of recognizing this generator as non random. The second one is used for checking the reliability, and it is the BBS generator. This is a computationally very heavy generator, but which is proven to be cryptographically secure (i.e. it passes al polynomial-time tests [7]). The code used for the testing comes directly from the NIST website[1] [8]; all additional mathematical code comes from [9].

II. Basic Test: a Case Study

The Binary Matrix Rank Test is included both in the NIST suite and in Marsaglia Die-Hard suite [3]. It works as follows.

Divide the input sequence $X_i = \{+1, -1\}$, $i = 0 \dots n-1$, into M contiguous non-overlapping strings of $P \cdot Q$ bits, with $n = M \cdot P \cdot Q$. With each string build a binary $P \times Q$ matrix and compute its rank r, $0 \leq r \leq \min(P, Q)$. Note that this has to be computed using the binary algebra based only on the symbols $\{-1, +1\}$. The probability that such a matrix has rank r is given by [2]

$$p_r = 2^{r(P+Q-r)-PQ} \prod_{i=0}^{r-1} \frac{\left(1-2^{i-P}\right)\left(1-2^{i-Q}\right)}{\left(1-2^{i-r}\right)} \qquad (1)$$

In the NIST version of this test, $P = Q = 32$; for these values we get

$$\begin{aligned} p_{32} &\simeq 0.289 \\ p_{31} &\simeq 0.578 \\ p_{30} &\simeq 0.128 \end{aligned} \qquad (2)$$

while all other probabilities are negligible. The rank of all M matrices is computed; the observed distribution is compared with the expected one by means of chi-square goodness-of-fit test, using $k = 3$ bins, one counting matrices with rank 32, one rank 31 and the last all other matrices. Given that the chi-square test is a statistical test, its output is a p-value; this is used directly as the output of the test.

Regrettably, this test (as any other basic test in the NIST suite) is not accurate enough for our purpose, as can be observed in the first column of Table I. Regarding its reliability, let us try to estimate the error in the p-value computation. Under the assumption of random input sequences, an error can arise due to approximations in the computation of the reference distribution, or approximations in the comparison of the observed distribution with the reference one.

First of all, Equation (1) comes in a closed form: this test has an exact mathematical background, and this is the main reason we use the rank test as a case study. So, any error in the test comes from the chi-square test employed. As proven by Pearson in [10], if the deviation of the observed frequencies from the expected values are normal (here, the deviation of the observed ratio of the rank of the matrices from (2)), the chi-square test is also exact. The problem is that this hypothesis is not usually satisfied: the deviations are normal when considering a very large number M of matrices (this result is given by the central limit theorem), and the speed of convergence is given by the Berry and Esseén inequality [11]. Regrettably, a formulation for the error that this approximation introduces in the chi-square test has not yet been found.

[1] At the time of this paper the latest version available is 1.8, March 2005.

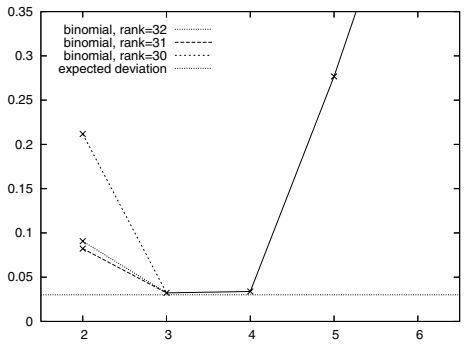

Fig. 1. Comparison between expected deviation from uniformity and measured deviation in the distribution of $N = 10,000$ p-values generated by the modified rank test with different numbers k of bins.

To get at least an intuitive idea if the introduced error is negligible or not, we have performed a simulation using the BBS generator. We have considered the distribution of $N = 10,000$ p-values from a rank test, looking for deviation from the uniformity in the distribution of the p-values in 10 intervals. Ideally, we have an expected deviation of $\sigma = 0.03$; a comparison with this value is shown in Figure 1.

In addition to the standard rank test we have also considered a modified version, in which the chi-square test used for testing the distribution of the observed rank is done in a number of bins ranging from $k = 6$ to $k = 2$. We notice that when the number of bins is too high, the observed deviation is significantly higher than the expected, proving a not-negligible error in the p-value computation. As the number of bins used decreases to $k = 4$ or to $k = 3$, the observed deviation decreases to the expected value. So, it seems that, the lower k is, the lower the error introduced.

However, when the number of bins is reduced to $k = 2$ (in this case the chi-square test degenerates into a binomial test, like the Frequency Test already analyzed in [4]) the test is again not reliable. This goes against the common intuition, because one would think that, with the binomial test being the simplest test, with the simplest hypotheses, this test would be the most reliable. One could argue from this observation that the chi-square test produces incorrect results. However, at this point it is not possible to prove this.

Let us consider the binomial case in the modified rank test. In this the basic event is that the matrix has rank r, and since all matrices are independent, it is regulated by the stochastic process

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q \end{cases} \qquad (3)$$

where $p = p_r$; $q = 1 - p_r$ is also used for simplicity of notation. We have considered three variants, where the event X_i corresponds to have a matrix with rank $r = 32$, $r = 31$ and $r = 30$.

The error introduced in the p-value computation by this normal approximation can be now bounded by the Berry

Copyright (C) 2007 IEEE. Personal use is permitted.
For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org

This is the author's version of the article that has been presented at IEEE ECCTD2007
The editorial version of the paper is available at http://dx.doi.org/10.1109/ECCTD.2007.4529674

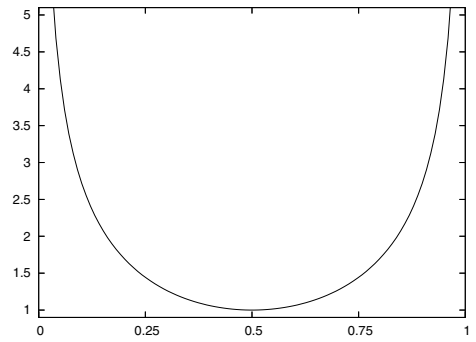

Fig. 2. Value of β in relation to the value of p in equation (4)

and Esseén inequality. Actually, this inequality needs a zero-average random variable $Z_i = X_i - p$. The random variable $s = \sum_{i=0}^{M-1} Z_i/\sigma_i\sqrt{M}$ has a cumulative distribution F_s and has a standard normal distribution Φ as a limit distribution for large M. The error of this approximation is bound by the Berry and Esseén inequality

$$\sup_x |F_s(x) - \Phi(x)| = C\frac{E\left[|Z_i|^3\right]}{\sigma_i^3\sqrt{M}}$$

with σ_i^2 and $E\left[|Z_i|^3\right]$ the second and third order moment of all Z_i, and $C \simeq 0.8$.

Since the p-value of a binomial test is computed as $p_v = 2 - 2\Phi(|s|)$, the maximum error in the p-value computation ε is twice the above error.

We can also compute

$$\beta = \frac{E\left[|Z_i|^3\right]}{\sigma^3} = \frac{p^2+q^2}{\sqrt{pq}} \tag{4}$$

Figures 2 shows the value of β in relation to the probability p: the nearer p is to 0.5, the lower the introduced error. This is confirmed by Figure 1: the case $p = p_{31}$ presents the lowest deviation from uniformity, while the case $p = p_{30}$ presents the highest deviation.

Note that, with the parameters used the maximum p-value error in all three cases is at least one order of magnitude smaller than α, so the test itself can be considered reliable. However in the next section we analyze how this error propagates in a second level test, making it sometimes not reliable.

III. Second-level Testing Approach

We define a *second-level* test as a test involving a number N of results from the same basic test applied to a number N of non-overlapping generated sequences. The basic idea underlying this approach is that, while it is very easy to design a pseudorandom generator capable of generating sequences that always pass a basic test, at the same time an ideal RNG has a certain probability to fail the same test. If all generated sequences pass a basic test, the generator being tested cannot be considered random.

Second-level tests are already present in the NIST document [2, Section 4]. However, up to now they have been employed rarely in real cases, due to the high number of bits (that is $n \cdot N$ instead of n) and the high computational power required. The NIST addresses two second level tests:

- if a generator is ideal, a generated sequence has a probability equal to $1-\alpha$ to pass a basic test with the above described methodology. Furthermore, since all generated sequences are non-overlapping, their associated p-values are independent. The event of passing a basic test is a binary random variable as (3) with $p = 1-\alpha$. The ratio of sequences passing a basic test $s = \sum_{i=0}^{N-1} X_i/N$ is binomial; for large N it can be assumed normal with the mean value $\mu = p$ and variance $\sigma^2 = pq/N$. NIST suggests to apply a 3σ criterion, and consider this second level passed if the ratio of sequences passing the basic test pertains in $[\mu - 3\sigma, \mu + 3\sigma]$.
- for an ideal generator, p-values are uniformly distributed in $[0, 1]$; the observed distribution is compared with the uniform distribution with a goodness-of-fit test, which NIST suggests to be the Pearson chi-square test. Since we are comparing two continuous distributions, a discretization is necessary; let us use k here for indicating the number of intervals (i.e. the bins) in which $[0, 1]$ is divided for the comparison; the value suggested by NIST is $k = 10$. This test is again a statistical test, and gives a p-value (in this case, a second-level p-value). However in this case NIST suggests to consider a level of significance $\alpha = 0.0001$.

Let us briefly comment on these approaches. The first one comes in the form of an on/off test; however, it is indeed a statistical test, with a probability of Type I error that is about 1% (i.e. the probability that a normal variable is out of the 3σ interval). We suggest to use it in its p-value form: a second-level p-value can be computed from s as

$$p_v = 2 - 2\Phi\left(\frac{|s-\mu|}{\sigma}\right) = \text{erfc}\left(\frac{|s-\mu|}{\sqrt{2}\sigma}\right) \tag{5}$$

To compare these two different approaches, we have used them to test the KISS generator. Results from a basic test and these two different second-level tests for different values of N are shown in Table I. In the example, for all tests we have considered a level of significance $\alpha = 0.01$, thus making possible a direct comparison between all the different methodologies. All p-values smaller than this level of significance are stressed in bold.

We notice that a second-level test may be accurate enough to recognize the KISS as a non random generator. Accordingly, the example confirms the two following intuitive ideas. (a) The 3σ test looks only at the distribution of the p-values in a very small interval, and this is less accurate than checking for the distribution of p-values in the whole $[0, 1]$. In fact, the 3σ test, like a basic test, always fails in identifying the KISS as a non random generator. (b) In chi-square approach, the larger N is, the more accurate the test. Roughly speaking, if N is small, deviations from uniformity in the distribution of p-values can be hidden by statistical errors in the distribution. For example, in the distribution of N uniformly distributed objects in k bins, the ratio of p-values in a bin has $\sigma^2 = (k-1)/Nk^2$. As N grows, σ decreases, and small deviations from uniformity can be more easily observed.

Note that results from the secure BBS generator (not reported here) do not present any failure, so reliability is assumed. However, let us make a more detailed analysis on the reliability of a chi-square based second-level test.

Assuming a maximum error ε in the computation of a p-value in a basic test, a p-value that should belong to a bin can be found in the nearby one only if the distance from the

Copyright (C) 2007 IEEE. Personal use is permitted.
For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org

This is the author's version of the article that has been presented at IEEE ECCTD2007
The editorial version of the paper is available at http://dx.doi.org/10.1109/ECCTD.2007.4529674

SP800-22 test	basic test	2nd level test on N=1,000 p-values		2nd level test on N=5,000 p-values		2nd level test on N=10,000 p-values	
		$\pm 3\sigma$	chi-square	$\pm 3\sigma$	chi-square	$\pm 3\sigma$	chi-square
Frequency	0.713479	0.340356	0.126946	0.319767	**0.005546**	0.421380	**0.000220**
Block Frequency	0.129962	0.112037	0.035124	0.088082	0.717492	0.840693	**0.003014**
Cumulative Sums	0.833869	0.203628	0.730660	0.117942	**0.001288**	0.687673	**0.000196**
Runs	0.768154	0.203628	0.475509	0.033006	0.318464	0.011985	0.118078
Longest Run of Ones	0.736930	0.203628	0.862417	0.477289	0.263704	0.687673	0.161804
Matrix Rank	0.224896	0.525010	0.869123	0.064640	0.697236	0.027030	0.277894
Spectral (DFT)	0.060580	0.750621	0.666313	0.010515	**0.000027**	0.011985	0.015227
NOT Matching	0.085400	0.750621	0.602290	0.776205	0.627274	0.763025	0.341393
OT Matching	0.105840	0.525010	0.026010	0.046605	0.606187	0.015861	**0.006872**
Universal	0.711080	0.056530	0.172742	1.000000	0.172263	0.027030	0.061237
Approx. Entropy	0.029426	0.750621	0.293897	0.477289	0.757106	0.546494	0.195730
Random Excursion	0.692131	0.197527	0.304499	0.274259	0.773039	0.027706	0.219808
Random Exc. Variant	0.280164	0.743512	0.716543	0.226180	0.175516	0.346918	0.466767
Serial	0.870041	0.056530	0.812929	0.886974	0.494979	0.421380	0.797496
Linear Complexity	0.998535	0.750621	0.841301	0.088082	0.932385	0.481728	0.781680

TABLE I
RESULTS OF RANDOMNESS TEST FOR THE KISS GENERATOR.

endpoint of the bin is less than ε. Since we are using the interval $[0, 1]$, ε is also the ratio of p-values that can be found in the wrong bin. This is independent of the number of bins. Since all bins (but the first and last), have two neighbors, the maximum propagated deviation $\Delta = 2\varepsilon$; for a binomial test

$$\Delta = 4\beta \frac{C}{\sqrt{M}}$$

Having this bound, we can express a reliability condition for a second-level test employing a chi-square test. If we look at the distribution of p-values in bins, the expected statistical deviation of the ratio a p-values found in a bin is $\sigma = \sqrt{(k-1)/Nk^2}$. If there is an error in the p-value computation, we are expecting that this error propagates into an additional deviation. If this propagated deviation is small with respect to the expected statistical deviation, the second-level test is reliable.

In all the three cases of the previous section, a second level test would not be reliable; in fact, if we impose $\Delta \leq \sigma$, we get

$$M \geq \frac{16NC^2k^2}{k-1}\beta^2 \tag{6}$$

With the parameters used, and considering $\beta = 1$, we get about $M \geq 10^6$; in the test of the example we have $M \simeq 1,000$.

To confirm these results, we have repeated the modified rank test with $M = 10^6$, increasing n to $n = 256 \cdot 10^6$ and reducing $P = Q = 8$. Results of the simulations (not reported here) confirm that with the original parameters, a second-level testing on the binomial rank test is not reliable; with these new parameters the test is now reliable.

Note that Equation (6) links reliability of the test to the parameter k used in the chi-square. One can think to substitute the chi-square test with other goodness of fit tests, such as the *Kolmogorov-Smirnov* (KS) test, which do net require any parameters. A closed form for the reliability of a second-level KS test is currently under investigation by authors.

IV. CONCLUSION

We have shown that the second-level approach in testing a RNG provides the best results in terms of accuracy. However, it may presents some flaws due to reliability problems. In this paper, starting from a case study that is the Binary Matrix Rank Test, we have found some guidelines to minimize the error on the p-value computation and ensure the reliability of the second level test employing a chi-square test:

- the basic test used should be based on a binary random event; in this case we are able to express the maximum error in the p-value computation;
- this random variable should describe two events with a probability as near as possible to 0.5; this probability value (as in Figure 2) introduces the minimum error in the p-value computation;
- the number of random binary events observed must satisfy condition (6). This condition can also be used, based on the number of acquired bits from the RNG being tested, to determine the ratio between n (the bits used in the basic test) and N (the number of basic test performed) when trying to get the maximum accuracy from the test.

ACKNOWLEDGMENT

This work has been supported by MIUR under the FIRB framework.

REFERENCES

[1] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1996.
[2] National Institute of Standard and Technology, "A statistical test suite for random and pseudorandom number generators for cryptographic applications", Special Publication 800-22, May 15, 2001. Available at http://csrc.nist.gov/rng/SP800-22b.pdf
[3] G. Marsaglia, "The diehard test suite", 2003. Available at http://www.csis.hku.hk/~diehard/
[4] F. Pareschi, R. Rovatti, and G. Setti "Second Level NIST Randomness Test for Improving Test Reliability", in *Proceedings of ISCAS2007*, pp. 1437–1440. New Orleans (USA), May 27–30, 2007.
[5] K. Hamano, "The Distribution of the Spectrum for the Discrete Fourier Transform Test included in SP800-22", in *IEICE Transactions on Fundamentals*, vol. E88, no. 1, pp. 67–73, January 2005.
[6] G. Marsaglia, and A. Zaman, "The KISS generator", Technical Report, Department of Statistics, University of Florida, 1993.
[7] L. Blum, M. Blum, and M. Shub, "A Simple Unpredictable Pseudo-Random Number Generator", in *SIAM Journal on Computing*, vol. 15, pp. 364–383, May 1986.
[8] National Institute of Standard and Technology, "Random Number Generation and Testing". Available at http://csrc.nist.gov/rng/
[9] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press, 1992. Available at http://www.nrbook.com/a/bookcpdf.php
[10] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to hove arisen from random sampling", in *Philosophical Magazine*, no. 50, pp. 157–172, 1900.
[11] A. N. Shiryaev, and R. P. Boas "Probability" (Graduate Texts in Mathematics), Springer-Verlag, 1995.

Copyright (C) 2007 IEEE. Personal use is permitted.
For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org