Hamming Distance and Data Compression of 1-D CA

Raied Salman

School of Computer Information Systems
Stratford University
11104 West Broad Street
Glen Allen, VA 23060
rsalman@stratford.edu

ABSTRACT :

In this paper an application of von Neumann correction technique to the output string of some chaotic rules of 1-D Cellular Automata that are unsuitable for cryptographic pseudo random number generation due to their non uniform distribution of the binary elements is presented. The one dimensional (1-D) Cellular Automata (CA) Rule space will be classified by the time run of Hamming Distance (HD). This has the advantage of determining the rules that have short cycle lengths and therefore deemed to be unsuitable for cryptographic pseudo random number generation. The data collected from evolution of chaotic rules that have long cycles are subjected to the original von Neumann density correction scheme as well as a new generalized scheme presented in this paper and tested for statistical testing fitness using Diehard battery of tests. Results show that significant improvement in the statistical tests are obtained when the output of a balanced chaotic rule are mutually exclusive ORed with the output of unbalanced chaotic rule that have undergone von Neumann density correction.

KEYWORDS :

Component; Formatting; Cellular Automata, Hamming Distance, Pseudo Random Number Generator (PRNG)

1. INTRODUCTION

Random numbers are in large demand for such wide spectrum of applications such as cryptography, Mont Carlo simulation, VLSI testing, etc. Pure random numbers are only obtainable from natural sources are not suitable because they are not reproducible. Hence, pseudo random number generation (PRNG) has been established as the best alternative. These PRNs can be produced by mathematical formulae where repeatability is their stagnant problem albeit they produce good statistical properties. Recently, however, cellular automata (CA) have been attempted and proved to be quite viable with the added advantage of ease of hardware implementation and running. One significant advantage of some of the chaotic rules of the CA is that it can produce outputs that are amenable to mathematical representation and therefore hard to

Sundarapandian et al. (Eds) : CSE, CICS, DBDM, AIFL, SCOM - 2013
pp. 01–10, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3301

reproduce by the adversary, a necessary condition for cryptographic applications. The problem is to find the suitable rule or rules out of a large size of rule space. Researchers have long sought to classify CA rules [1-3]. A seminal and widely referenced attempt is that due to [4]. Wolfram's classification scheme was influential, and thorough. The extensive computer simulation carried out by Wolfram has relied heavily on the inferences drawn from phenomenological study of the space-time diagrams of the evolution of all the $2^{2^{2r+1}}$ rules, where r is the radius of the neighborhood of the center cell that is being updated in discrete time steps running under Galois Field $GF(2)$ [2,5,6]. Although $r=1$ was mostly adopted in order to make the rule space practically realizable with the availability of the computational powers of the existing computers, larger values of r nevertheless have also been attempted mostly with genetic algorithms [7]. Some prominent researchers have introduced ad hoc parameters in their attempts to classify the rule space [8,9]. Unfortunately, none of these methods have culminated in a well-defined classification of the CA rule space. For a binary one dimensional (1-D) CA and a neighborhood of radius r the rule space is $2^{2^{2r+1}}$. Even for an elementary 1-D CA where $r=1$ the rule space is reduced to $2^8 = 256$ and still making an exhaustive search a difficult and time consuming process. For a mere 1-bit larger neighborhood radius $r=2$ produce humongous rule space of 2^{32} rendering any linear search scheme prohibitively and computationally unfeasible. One useful and statistically dependable approach is cross correlation between two delayed versions of the evolution runs of the CA. This research presents a new approach that can partially resolve the search problem by attempting to use the Hamming Distance (HD) between consecutive configurations in the time evolution of the CA and observing the cyclic behavior of this metric. This approach can in a straight forward manner show that rules that result in a cyclic HD are actually cyclic and therefore can be decided to be unsuitable for PRN generation. Since this operation does not require large amount of data, the search process can be finished in a relatively very short time. It has been observed that the HD approach can discover Wolfram’s category IV (the so called complex rules) much faster than expected. In fact the difference between category II and category IV almost diminish. Both of these categories as well as category I are unsuitable for PRN generation.

2. PRELIMINARIES

This paper deals with a homogeneous lattice of one dimensional cellular automata 1-D CA. The present state of any cell at time t is denoted by $\sigma(t,l)$ where $l \in L$ is the spatial index of a lattice length of L bits.

The CA can evolve using a single rule or can use multiple rules in either or both the space and time dimension. When more than one rule is used it is usually referred to as Hybrid CA. In this paper a single rule will be used and the CA will be referred to as a uniform 1-D CA. In order to limit the size of the lattice *cyclic boundary* conditions will be applied. This means the end cells will wrap around the lattice. If the rules deal with the center cell and the two nearest neighbors such that the radius from the center cell to the neighboring left and right cells is $r=1$ the CA is usually referred to as Elementary CA (ECA). Therefore the rule acting on cell $\sigma(t,1)$ will consider the left neighboring cell $\sigma(t,L)$ and the right neighboring cell $\sigma(t,2)$ as depicted in Figure 1. Similarly the rule will act on the right most cell $\sigma(t,L)$ such that the left neighboring cell will be $\sigma(t,L-1)$ and the right neighboring cell will be $\sigma(t,1)$. The center cell at an arbitrary location l and time t will be denoted by $\sigma(t,l)$ and the left neighboring cell as

$\sigma(t,l-1)$ whereas the right neighboring cell will be $\sigma(t,l+1)$. The initial configuration will thus be denoted by

$$\Gamma_0 = \{\sigma(0,1), \sigma(0,2), ..., \sigma(0,l-1), \sigma(0,l), \sigma(0,l+1), \\ ..., \sigma(0,L-1), \sigma(0,L)\} \qquad (1)$$

while an arbitrary configuration will be

$$\Gamma_t = \{\sigma(t,1), \sigma(t,2), ..., \sigma(t,l-1), \sigma(t,l), \sigma(t,l+1), \\ ..., \sigma(t,L-1), \sigma(t,L)\} \qquad (2)$$

where $t \in T$ and T is the total evolution time. The rule R_n where n is the rule number according to the numbering scheme adopted by [4], is a mapping $\mathrm{R}_n : \{0,1\} \rightarrow \{0,1\}^3$ and the next state of the cell under this rule can be represented by

$$\sigma(t+1,l) : f(\sigma(t,l-1), \sigma(t,l), \sigma(t,l+1)) \qquad (3)$$

The Hamming distance measures the distance between two binary strings by counting the number of different bits and can be defined by

$$HD(t) \; @ \, \Gamma_t \oplus \Gamma_{t+1} = \sum_{l=1}^{L} (\sigma(t,l) \oplus \sigma(t+1,l)) \qquad (4)$$

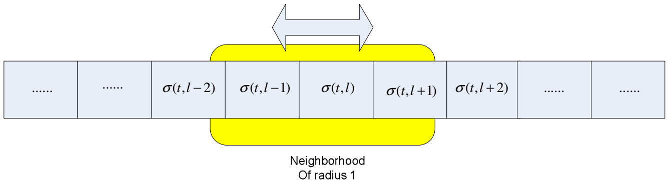


Figure 1. Local Rule Representation

3. Space Rule Classification

When applying the HD on an arbitrary time-space set of data two results can be extracted. One is the transient from the initial configuration until the start of a cycle if that cycle exists within the time evolution of the data set. The second is the length of the cycle if the cycle is captured during the time evolution. For example the variation in the for a rule that belongs to category I according to Wolfram's [4] typical classification is a very short transient that terminates very sharply to an 0. The small transient length seems to be a typical feature of category I rules, as shown in Figure 2 for Rule 255 in both cases of random initial seed or an active center cell and the rest of the cells are inactive. The difference in the first is of course due to the initial seed . Category II rules, represented by Rule 1, Figure 3, exhibit a relatively longer transient but again stabilizes at a constant which has different values depending on the initial seed . Category IV rules, represented by Rule 35, Figure 4, again seem to exhibit similar behavior. The transient length is again different depending on the initial seed while the asymptotically stabilizes to a constant value. This behavior is also recurrent with category III rules, Figure 5, albeit on a larger scale but the main thing is that in this case it is not clear whether the is indicative of the cycle length or whether it is a symptom of some hidden but repetitive behavior that cannot be captured from the space-time diagram. It is a worthwhile topic for further investigation and research. This process is simple and fast since it requires a relatively very short evolution time to produce results that may prove to be significant in the testing of PRNs. It can be conjectured that the may be able to be used as a fast and efficient tool for testing PRNs for suitability in cryptographic applications. Based on the data it can also be concluded that category III rules are the best suited for PRNs.

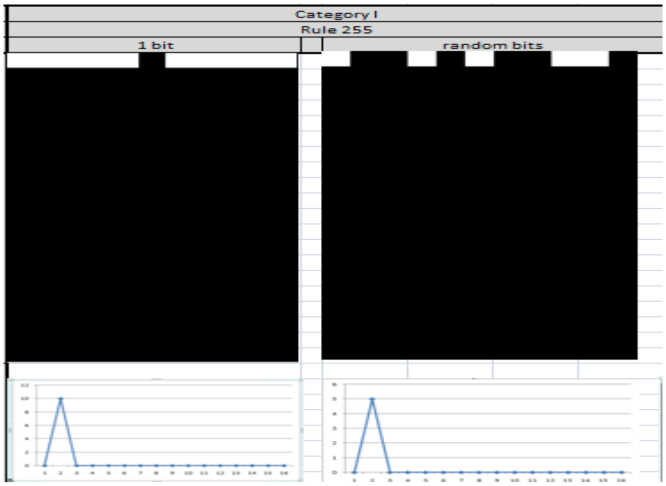


Figure 2. Time-Space and *HD* plots for Rule 255

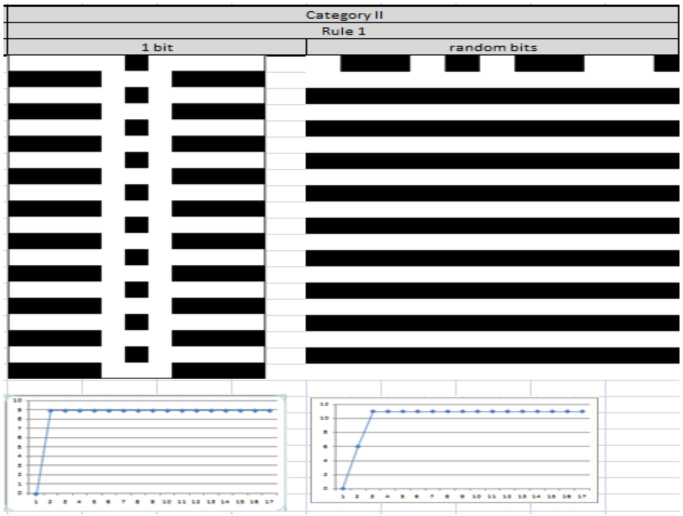


Figure 3. Time-Space and *HD* plots for Rule 1

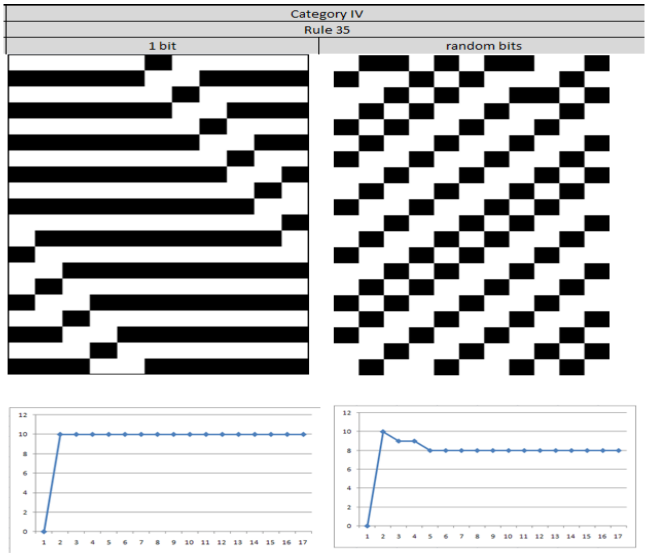


Figure 4. Time-Space and *HD* plots for Rule 35

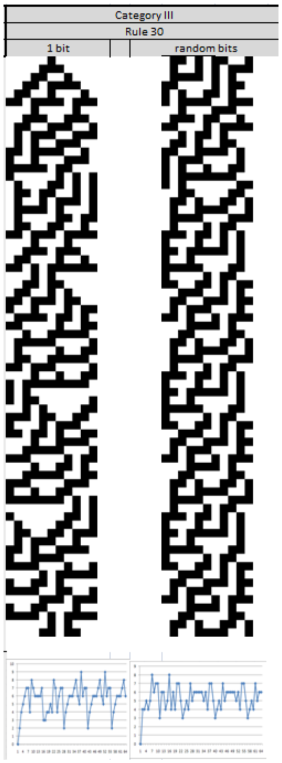


Figure 5. Time-Space and *HD* plots for Rule 255

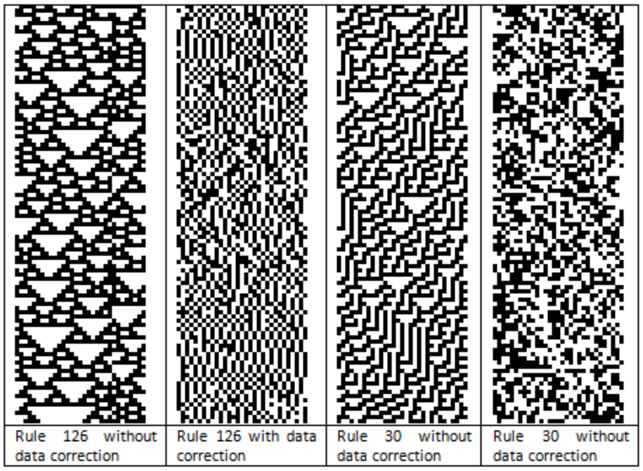
4. CA Density Correction And Data Compression

The rule space classification usually does not touch upon the density of the CA evolution. Such metric is an essential criterion for suitability to generate cryptographically strong PRNs. Since the rules that may be suitable for PRN generation is restricted to category III it can be seen that some of the rules in this category do not produce uniform density. The density must be uniform such that the number of one's should be equal or differ by at most one bit from the number of zero's in the data according to Golomb's randomness postulate number 1 [10]. Such a requirement isolates a number of rules in category III that can possibly be considered as candidates for PRNs. For example Rule 22 and Rule 126 both cannot produce the 0.5 uniform density but they are still chaotic and belong to category III. The performance of such rules when tested using Diehard is consequently very poor. If the density of these rules can be corrected then these rules can be reconsidered for PRN generation and the repertoire of rules available for PRN generation can be widened. Luckily there exists a very effective and yet very simple approach that is originally attributed to von Neumann. The method effectively compresses the data according to the steps depicted in Table 1.

Table 1 von Neumann correction Scheme

Original Data	Resultant Data
01	0
10	1
11	delete
00	delete

As an example, a 1-D CA of lattice length $L = 31$ bit was run for an evolution time of $T =$ 2,645,000 time steps under Rule 126 produced a density of one's equal to 0.527746. When von Neumann reduction scheme described in Table 1 was applied on the same data the density was corrected to 0.5. In addition this density correction is usually accompanied with two important features in as far as PRN generation is concerned. The first is that the resultant data is now extremely hard to reproduce, a fundamental and necessary requirement for cryptographically strong PRNs. This is clearly due to the loss of information from both rules in the correction process. Therefore the process can be considered as an irreversible process. The second is a byproduct which is an improvement in the statistical properties of the rule. For this particular example the data was tested for statistical strength by the Diehard battery of tests and passed two tests but another test was also passed when the density was corrected. A more significant example is Rule 30 under the same parameters passed 51 tests whereas the number of passes jumped to 129 when the data was run after the application of von Neumann correction scheme. It is very clear from the time-space diagrams depicted in Figure 6 that Rule 126 and Rule 30 have undergone significant randomization which were reflected the time-space diagrams as well as in the number of passes for both rules but it was more pronounced with Rule 30 as mentioned above.



Rule 126 without data correction

Rule 126 with data correction

Rule 30 without data correction

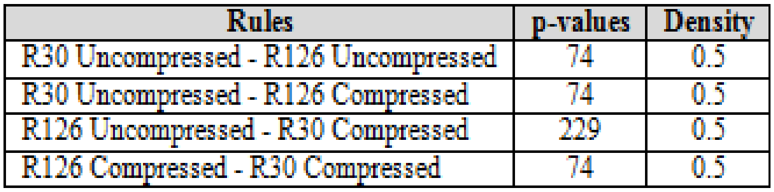
Rule 30 without data correction

Figure 6. Space-Time diagrams for Rules 30 and 126

When the two rules in their uncompressed and compressed forms were linearly mixed with a mutual exclusion operation as depicted in Figure 7 some astonishingly remarkable results were produced as shown in Table 2. The three combinations R30 uncompressed with R126 uncompressed, R30 uncompressed with R126 compressed, R30 compressed with R126 uncompressed, produced identical results when tested with the Diehard test suite and the density was also maintained at the favorable 0.5 level. The combination of R30 compressed with R126 uncompressed Figure 8, produced the best results and passed all the 229 Diehard tests and of course maintained the same ideal density of 0.5. It is generally accepted that passing all the Diehard tests is a strong indication that the PRN generator is suitable for cryptographic applications. This is in addition to the above stated hardness in reproducing the sequence generated. Further research and more details are deemed necessary in order to validate the initial findings in this paper. It can also be conjectured that the other chaotic rules can produce the same results.

Table 3 shows the variation in the Diehard test results for all the runs for Rules 30 and 126 as well as their mixtures. It can be seen that the Overlapping Sums test number 15 and the GCD test number 2 were the most difficult to pass except for the PRN8 (The combination of R30 compressed with R126 uncompressed) case.

Table 2. p-values and Density of Rules 30 and 126 mixtures



Rules	p-values	Density
R30 Uncompressed - R126 Uncompressed	74	0.5
R30 Uncompressed - R126 Compressed	74	0.5
R126 Uncompressed - R30 Compressed	229	0.5
R126 Compressed - R30 Compressed	74	0.5

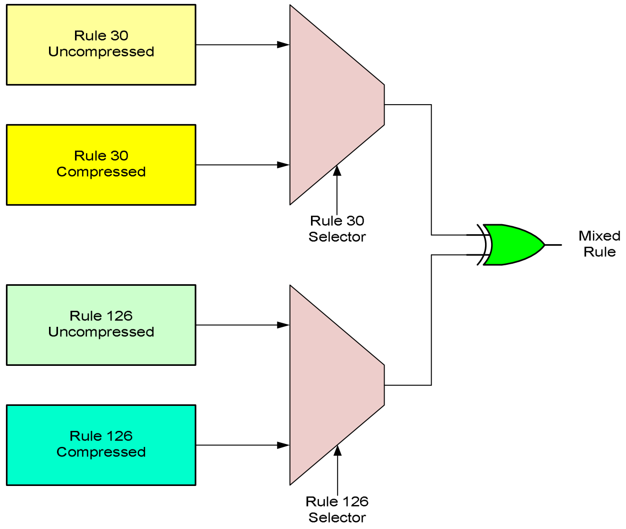
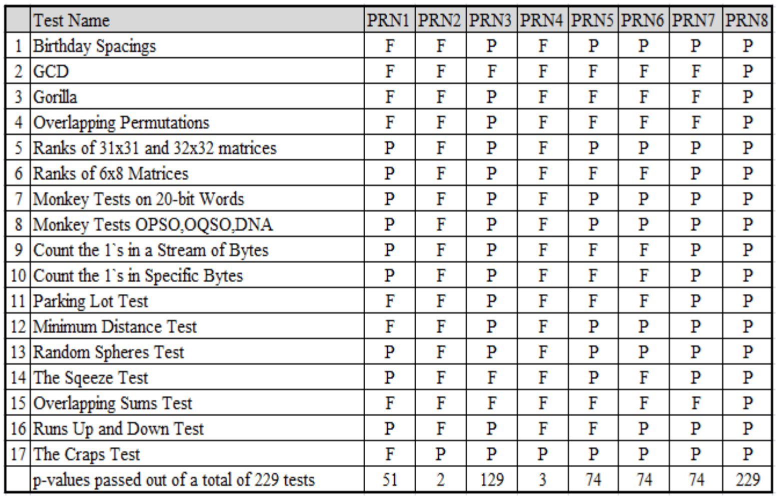
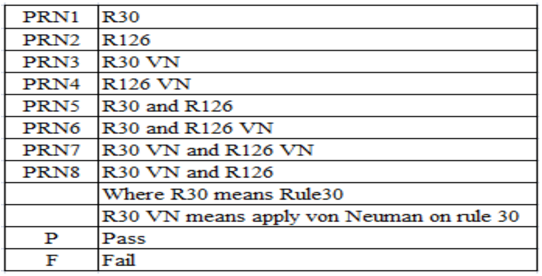


Figure 7. Rules 30 and 126 mixing scheme

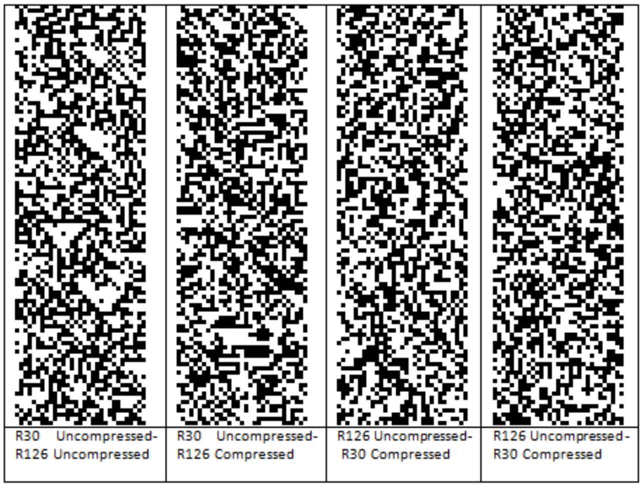
Table 3. Diehard Results for Rules 30 and 126



	Test Name	PRN1	PRN2	PRN3	PRN4	PRN5	PRN6	PRN7	PRN8
1	Birthday Spacings	F	F	P	F	P	P	P	P
2	GCD	F	F	F	F	F	F	F	P
3	Gorilla	F	F	P	F	F	F	F	P
4	Overlapping Permutations	F	F	P	F	F	F	F	P
5	Ranks of 31x31 and 32x32 matrices	P	F	P	F	P	P	P	P
6	Ranks of 6x8 Matrices	P	F	P	F	F	F	P	P
7	Monkey Tests on 20-bit Words	P	F	P	F	P	P	P	P
8	Monkey Tests OPSO,OQSO,DNA	P	F	P	F	P	P	P	P
9	Count the 1`s in a Stream of Bytes	P	F	P	F	F	F	P	P
10	Count the 1`s in Specific Bytes	P	F	P	F	F	F	P	P
11	Parking Lot Test	F	F	P	F	F	F	P	P
12	Minimum Distance Test	F	F	P	F	P	P	P	P
13	Random Spheres Test	P	F	P	F	P	P	P	P
14	The Sqeeze Test	P	F	F	F	P	F	P	P
15	Overlapping Sums Test	F	F	F	F	F	F	F	P
16	Runs Up and Down Test	P	F	P	F	F	P	P	P
17	The Craps Test	F	P	P	P	P	P	P	P
	p-values passed out of a total of 229 tests	51	2	129	3	74	74	74	229



PRN1	R30
PRN2	R126
PRN3	R30 VN
PRN4	R126 VN
PRN5	R30 and R126
PRN6	R30 and R126 VN
PRN7	R30 VN and R126 VN
PRN8	R30 VN and R126
	Where R30 means Rule30
	R30 VN means apply von Neuman on rule 30
P	Pass
F	Fail



R30 Uncompressed-R126 Uncompressed

R30 Uncompressed-R126 Compressed

R126 Uncompressed-R30 Compressed

R126 Uncompressed-R30 Compressed

Figure 8. Space-Time diagrams for Rules 30 and 126

5. CONCLUSIONS

In this paper the Hamming Distance was revisited and applied to the 1-D CA. The original motivation was the classification of the rule space of the CA. This has been achieved in a very simple and yet effective approach. The results show a well defined behavior of the chaotic rules of category III as compared to the behavior of the rules of the other three categories. The oscillations of the hamming distance in the transient stage are indicative of the chaotic nature of the rule. In other words, the high value of the hamming distance in the transient stage is actually indicative of rules Category I or II. The hamming distance values during the oscillation period do not vary very much as is the case during the transient stage. It can be concluded that Category III rules are the best rules suited for PRN generation. The behavior of category I rules seem to be very clear and their time evolution reach a hamming distance equal to 0 after one or two time steps only depending on the initial seed. Category II and IV Rules seem to behave in a similar manner. They both reach a constant hamming distance after a very short transient cycle with a slight difference in the values of the hamming distance during the transient cycle but the asymptotic behavior is the same. Therefore, the new categorization of the rule space is that they are indeed of three distinct types, Category I, Category II and IV combined, and the third is Category III. This seems to agree with the findings of some past researchers that argued strongly against the separate categorization of Category IV. The finding in this paper can reduce the rule search significantly. The correlation technique that is usually used in the analysis of pseudo random number generation can indicate the amount of correlation between two delayed versions of the data as well as the distance between the cycles if the cycles exist. In this paper the hamming distance is used as an alternative. The advantage of the Hamming Distance approach as compared with the Correlation approach is that the hamming distance can show the transient stage (the number of time steps to finish the transient orbit or system time constant) as well as showing the cycles with clear repetition a feature that the correlation technique is unable to produce. In addition the hamming distance can arrive at the results in a very short time while the cross correlation technique requires the full length of the data and much more computational effort.

It is also clear from the results in this paper and the findings of previous research that not all rules of Category III are suitable for PRN generation. One stagnant problem with the rules that are deemed unsuitable is attributed to the non-uniform density output of some of these rules, such as Rule 126. The application of von Neumann reduction scheme proved to be beneficial. The density has been corrected to the desirable value of 0.5. However, a byproduct to this was the improvement in the randomization as depicted in the images produced which was also validated in the increase of test passes. A more significant improvement was in the number of tests passed by Rule 30 that jumped from 51 prior to the application of the reduction scheme to 129 after the application of the scheme. Another remarkable result was achieved when the two types of rules R126 and R30 were linearly mixed together. When a reduced output of R30 was mutually exclusive ORed with the output of unreduced output of R126, the output data has passed all the 229 Diehard tests. A result that is extremely difficult to achieve by other PRN sources. This result may require further effort to validate the findings in this paper as well show that the approach is equally applicable to the other chaotic rules.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.