Components and Cycles of a Random Function*

J. M. DeLaurentis
Sandia National Laboratories
Albuquerque, New Mexico 87185

Abstract

This investigation examines the average distribution of the components and cycles of a random function. Here we refer to the mappings from a finite set of, say, n elements into itself; denoted by Γ_n. Suppose the elements of Γ_n are assigned equal probability, i.e. $P(\gamma) = n^{-n}$, $\gamma \in \Gamma_n$. The directed graph that is naturally associated with γ consists of several components, each with a unique cycle. Define $X_n(s,t)(\gamma)$ as the number of components in γ containing at least the fraction s of the total number of nodes, with the size of each component's cycle not exceeding $tn^{1/2}$. We show that the expected value of $X_n(s,t)$ can be approximated by the double integral

$$EX_n(s,t) \approx \int_t^1 \int_0^s \frac{1}{\sqrt{2\pi}} \frac{\exp[-y^2/(2x)]}{\sqrt{x^3(1-x)}} \, dydx \quad .$$

The average number of components of a given size with cycles of a specified length approximately equals the volume under the graph of the integrand. This expression can be used to estimate the probability that a function has a component which contains a significant percentage of the total number of nodes and yet its cycle is relatively small.

Introduction

Random functions often arise as a model for the pseudo-random functions generated by a cryptosystem. Typically, the experiments performed on the latter are concerned with the

*This work performed at Sandia National Laboratories supported by the U.S. Dept. of Energy under contract No. DE-AC04-76DP00789.

C. Pomerance (Ed.): Advances in Cryptology - CRYPTO '87, LNCS 293, pp. 231-242, 1988.

size of its components and the length of the corresponding cycles. It is natural then to address similar problems for random mappings. This paper analyses the expected number components of a given size that contain cycles of a specified length. The asymptotic expression of this average value for random mappings provides some insight into the behavior of pseudo-random functions.

To better understand Hellman's time-memory cryptanalytic scheme [1], Hellman and Reyneri [2] estimated the expected size of the largest component of a random function. They compare this value with the average size of the largest component of the pseudo-random functions generated by the Data Encryption Standard (DES). In this case they considered mappings $f(\cdot)$ from the key-space into itself. Specifically, the functional value $f(k)$ was defined by applying the DES operation $S_k(\cdot)$ to a fixed plaintext block P_o and then reducing the 64-bit block $S_k(P_o)$ to 56-bits through a reduction operation $R(\cdot)$,

$$f(k) = R(S_k(P_o)) \ .$$

Choosing a different plaintext block defines a new function. They found their statistical tests to be in close agreement with the expected outcome for random maps.

More recently, at Crypto '86, G. J. Simmons presented a study by Quisquater [3] in which the cycling experiments involved DES functions similar to those described above. In his investigation Quisquater found a function with a relatively large component ($\approx$ 3% of the total number of nodes) that contained a relatively small cycle (cycle size $\approx 2^{16}$). As we will see, the probability of such an event for random functions is $\approx 10^{-3}$.

A variety of different functions have been introduced to analyze cryptosystems. To examine the closure properties of DES, Kaliski et al. [4] defined a set of mappings from the cipher-space into itself. In contrast to the preceding example, they applied a pseudo-random function $g(\cdot)$ to the cipher x to obtain a key $k = g(x)$. In turn, this key was used in the DES operation to produce

$$f(x) = S_k(x) = S_{g(x)}(x) \ .$$

Again, these studies detected no statistical anomalies. Assuming that random functions are a reasonable model for the maps in question, the methods developed in this paper are applicable.

Beyond these practical motivations the study of random functions is of some intrinsic combinatorial interest. Examples of probability distributions related to random functions are presented in [5], [6], [7], and [8]. A relationship between branching processes and random maps is discussed in [9] and [10]. For a survey of results, see [11].

In the following section we introduce the necessary definitions and notation. First we consider the average number of components of size k containing a cycle of length ℓ (a (k,ℓ)-component). Next we analyze the expected number of such components when k and ℓ are allowed to range over an entire region. Finally we estimate the probability of discovering a function that has a component which contains a significant percentage of the total number of nodes and such that its cycle is relatively small.

Components and Cycles

The set of mappings from an n element set into itself endowed with the uniform distribution is called the set of random functions or random mappings and is denoted $\Gamma_n, (P(\gamma)=n^{-n}, \gamma\in\Gamma_n)$. The directed graph naturally associated with each function γ is the graph with a vertex for each element of the domain and a directed edge from vertex i to vertex j if and only if $\gamma(i) = j$. The components of such a graph consist of a cycle with trees attached to its nodes (cyclic points). (In the following we sometimes write that γ has a certain property when in fact it is the associated graph that has this property.) We are interested in estimating the average number of components of a given size and with a specified cycle length.

First we define the random variable $Y_n(k,\ell)(\gamma)$ that counts the number of (k,ℓ)-components in γ (a (k,ℓ)-component has k nodes and ℓ cyclic points). Next we introduce

the function

$$f(x,y) = \frac{1}{\sqrt{2\pi}} \frac{\exp[-y^2/2x]}{\sqrt{x^3(1-x)}} .$$

The following lemma, whose proof is postponed until the end of this section, provides an estimate for the average value of $Y_n(k,\ell)$.

Lemma

For $Y_n(k,\ell)$ defined as above with $1 \le \ell \le k/2$ and $k \le n - n^{1/3}$ we have

$$E\, Y_n(k,\ell) = f(x_k, y_\ell) n^{-3/2} \, (1 + R_n(k,\ell)) , \tag{1}$$

where $x_k = k/n$, $y_\ell = \ell/\sqrt{n}$ and $R_n(k,\ell) = O(\ell^3/k^2 + k^{-1} + n^{-1/3})$. For $n - n^{1/3} < k \le n$ the left hand side is $O\left[f(x_{k-1}, y_\ell) n^{-3/2}\right]$.

The main features of this result are best explained by means of the graph of $f(x,y)$ (see figure 1).

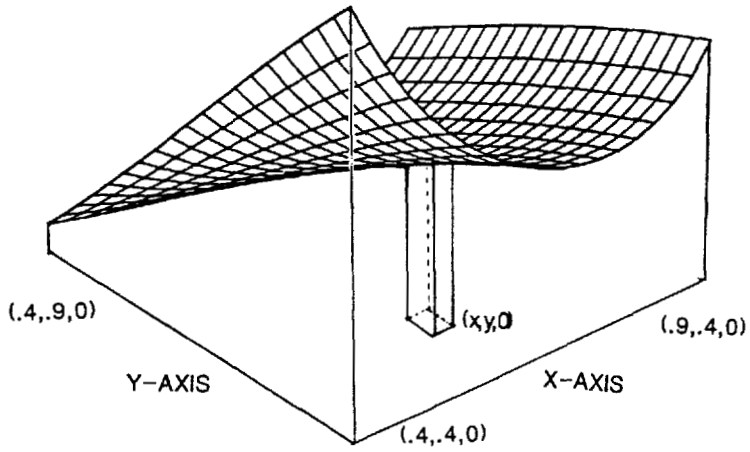


fig 1 GRAPH OF F(X, Y)

From the graph we see that the expected number of (k,ℓ)-components approximately equals the volume of the parallelepiped with height $f(x_k,y_\ell)$ whose basis is the rectangle of area $n^{-3/2}$ centered at (x_k,y_ℓ).

Except for k extremely close to 0 or n the sum over a range of these average values approximately equals the volume under the graph of f over a given region. For example, let $X_n(s,t)$ represent the number of (k,ℓ)-components with $sn \le k \le n$ $(0<s<1)$ and $1 \le \ell \le tn^{1/2}$, that is

$$X_n(s,t) = \sum_{sn\le k\le n} \sum_{1\le \ell\le tn^{1/2}} Y_n(k,\ell) \ . \tag{2}$$

Excluding the terms $n-n^{1/3} < k \le n$, the expected value of the expression on the right is asymptotically equivalent to the integral of $f(x,y)$ over $s \le x \le 1 - n^{-2/3}$, $0 \le y \le t$. The sum involving the excluded terms is measured by the integral of $\min\{t,1\}\left[x^3(1-x)\right]^{-1/2}$ over the interval $1 - n^{-2/3} \le x \le 1$, and the latter is $O(\min\{t,1\}n^{-1/3})$. It follows from the lemma that for s fixed, $0 < s < 1$, and n sufficiently large, we have:

Theorem 1

The average number of (k,ℓ)-components in the region $sn \le k \le n$, $1 \le \ell \le tn^{1/2}$ is approximately

$$EX_n(s,t) = \int_s^1 \int_0^t f(x,y)dydx\left[1+R_n(s,t)\right] \ , \tag{3}$$

where $R_n(s,t) = O(t^3s^{-2}n^{-1/2}+s^{-1}n^{-1}+n^{-1/3})$.

Note: The error made by replacing the sum with the integral has been included in the remainder term. Also, notice that the x-coordinate refers to the component's size and the y-coordinate indicates cycle length (see figure 2).

Although (3) is an asymptotic expression for an expected value, it can be used to estimate probabilities. As an example, we consider the probability of finding a function that has a component which contains a significant percentage of the total number of nodes

and yet its cycle is relatively small. That is, we want to estimate the probability that a random function has a (k,ℓ)-component in which k is a significant fraction of the total number of nodes and yet ℓ is relatively small.

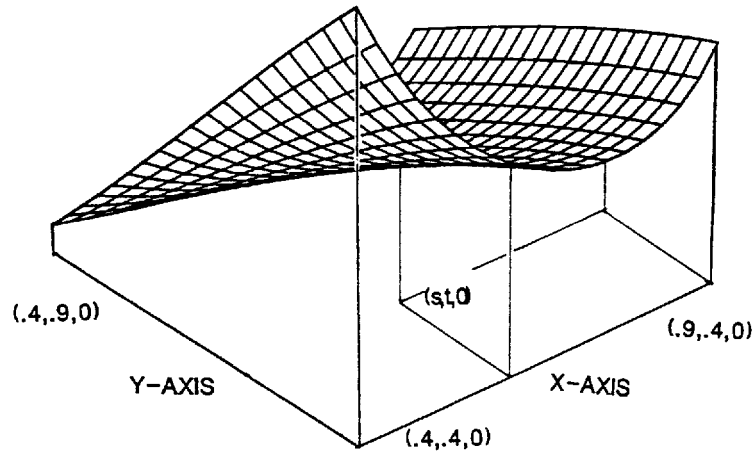


fig 2 GRAPH OF F (X, Y)

To make these terms precise, we consider a cycle of order $n^{1/2-\alpha}$, $0 < \alpha < 1/2$, as small (since the total number of cyclic points is on average $n^{1/2}$). A component is considered large if it contains at least the fraction s of the total number of nodes, $0 < s < 1$. The number of such components is given by $X_n(s,t)$ where $t = n^{-\alpha}$. Using this notation the problem is to estimate the probability that $X_n(s,t)$ is at least one.

The main idea is that if a function possesses a relatively large component with a small cycle, then it is most likely the only such component. It follows that

$$P(X_n(s,t) \geq 1) \approx EX_n(s,t) \ . \tag{4}$$

More precisely, by Bonferroni's inequality [12], we have

$$EX_n - E\left[X_n(X_n-1)\right] \leq P(X_n \geq 1) \leq EX_n .$$

It can be shown (see the appendix) that $E\left[X_n(X_n-1)\right]$ is $O(s^{-1}n^{-2\alpha})$. Thus we need only estimate the mean of $X_n(s,t)$.

Omitting, for the present, the error terms in Theorem 1, we obtain

$$EX_n(s,t) \approx (2\pi)^{-1/2} \int_s^1 x^{-3/2}(1-x)^{-1/2} \int_0^t e^{-y^2/2x} dy dx \ , \tag{5}$$

recall that $t = n^{-\alpha}$. For small t the innermost integrand approxmiately equals one. That is, replacing the exponential in (5) with $1 + O(y^2/s)$ yields

$$EX_n(s,t) \approx (2\pi)^{-1/2} n^{-\alpha} \int_s^1 x^{-3/2}(1-x)^{-1/2} dx \left[1+O(s^{-1}n^{-2\alpha})\right] \ . \tag{6}$$

Fortunately, the integral in (5) can be evaluated explicitly as

$$c(s) \equiv \left[\frac{2(1-s)}{\pi s}\right]^{1/2} = (2\pi)^{-1/2} \int_s^1 x^{-3/2}(1-x)^{-1/2} dx \ . \tag{7}$$

Using the estimate for the second factorial moment and including the remainder terms given in (3) leads to the conclusion

Theorem 2

The probability of finding at least one (k,ℓ)-component with $sn \le k \le n$, $1 \le \ell \le tn^{1/2}$ where $0 < s < 1$ and $t = n^{-\alpha}$ is given by

$$P(X_n(s,t) \ge 1) = c(s) n^{-\alpha} \{1+O[r_n(s)]\} \ , \tag{8}$$

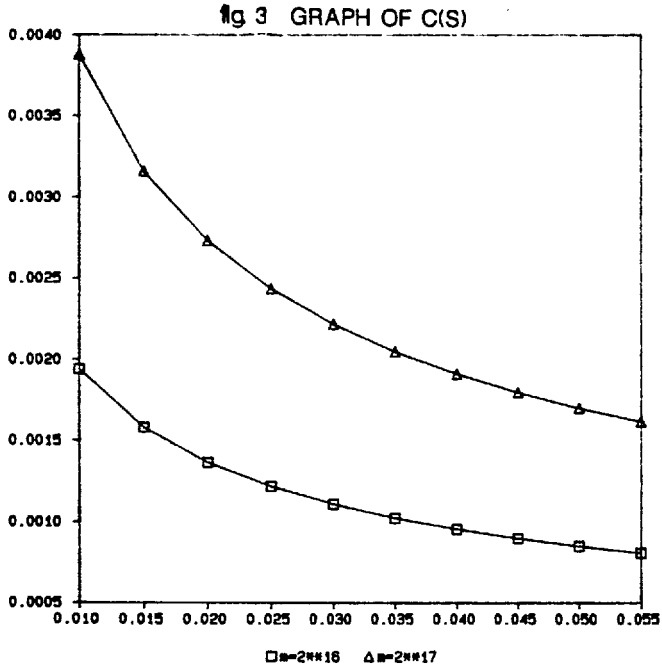
here $r_n(s) = s^{-2}n^{-1/2-3\alpha} + s^{-1/2}n^{-\alpha} + n^{-1/3}$.

Consider the example presented in the introduction [3]. In this case $n = 2^{56}$, $tn^{1/2} = 2^{16}$ (i.e., $t = n^{-3/14} = 2^{-12}$ and $\alpha = 3/14$), and $s = 0.03$. It follows that

$$P(X_n(s,t) \ge 1) \approx c(0.03)2^{-12} \approx 10^{-3} \ ,$$

with $r_n(0.03) \le 2\times10^{-3}$ (see figure 3). The key observation is that even if a DES cycling

experiment occasionally produces a relatively large component with a small cycle, this does not necessarily imply a statistical irregularity in DES.



Note: Here $m = tn^{1/2} = n^{1/2-\alpha}$ is the maximal cycle length.

We turn now to the proof of the lemma. Let $C(k,\ell)$ denote the number of connected mappings on k nodes with ℓ cyclic points. It is known [11] that

$$C(k,\ell) = (\ell-1)! \begin{bmatrix} k-1 \\ \ell-1 \end{bmatrix} k^{k-\ell} \quad . \tag{9}$$

Using Stirling's formula [12]

$$k! = \sqrt{2\pi k} \left(\frac{k}{e}\right)^k \left[1 + O(\frac{1}{k})\right] \quad , \tag{10}$$

we compute

$$C(k,\ell) = k^{k-1}e^{-\ell^2/2k}\,[1 + O(\ell^3/k^2)] \quad , \tag{11}$$

for $1 \le \ell \le k/2$.

The main idea in the proof involves writing $Y_n(k,\ell)$ as the sum of identically distributed random variables. We define $\epsilon_i = 1$ if the i-th node belongs to a (k,ℓ)-component, otherwise set $\epsilon_i = 0$.

It follows that

$$Y_n(k,\ell) = \frac{1}{k} \sum_{1\le i\le n} \epsilon_i \text{ and } E(Y_n(k,\ell)) = \frac{n}{k}\, P(\epsilon_i=1) \quad . \tag{12}$$

The probability that the i-th node belongs to a (k,ℓ)-component is given by

$$P(\epsilon_i=1) = \binom{n-1}{k-1} C(k,\ell)\ (n-k)^{n-k}/n^n \quad . \tag{13}$$

The first term is the number of ways to select the other members of the (k,ℓ)-component; the second term is the number of connected mappings consisting of k nodes and ℓ cyclic points; the third term is the number of functions on the n-k remaining elements; and the last term is the total number of mappings on an n element set. Combining (11) - (13) yields the expression

$$E\left[Y_n(k,\ell)\right] = \binom{n}{k}\ (n-k)^{n-k}k^{k-1}e^{-\ell^2/2k}n^{-n}[1+O(\ell^3/k^2)] \quad . \tag{14}$$

Applying Stirling's formula (10) to the first term in (14) and simplfying leads to the desired result

$$E\,[Y_n(k,\ell)] = f(x_k,y_\ell)n^{-3/2}[1+O(\ell^3/k^2+k^{-1}+n^{-1/3})] \quad , \tag{15}$$

for $1 \le \ell \le k/2$, $k \le n-n^{1/3}$, $x_k = k/n$ and $y_\ell = \ell/\sqrt{n}$. The proof of the last statement in the lemma is similar. The key step in these arguments is the introduction of the auxiliary random variables ϵ_i.

Summary

To better understand the structure of random functions we have examined the average distribution of its components and cycles. The relationship between a component's size and its cycle length is best illustrated by the graph of $f(x,y)$. The volume under the graph and over a specified region represents the expected number of components in a given range with cycle lengths belonging to a prescribed interval. In turn this mean value is used to estimate the probability of discovering a function containing a relatively large component with a small cycle.

Appendix

The derivation of the asymptotic expression for the second factorial moment $E[X_n(X_n-1)]$ is similar to the development for the mean of X_n. First, we may assume that $s \le 1/2$ since $X_n(s,t)(X_n(s,t)-1) = 0$ if $s > 1/2$. Expanding the product yields

$$X_n(s,t)(X_n(s,t)-1) = \sum_{\substack{sn\le k,k'\\ k'+k\le n}} \sum_{1\le \ell,\ell' \le tn^{1/2}} Y_n(k,\ell)[Y_n(k',\ell')-\delta] , \tag{16}$$

where $\delta = 1$ if $k = k'$ and $\ell = \ell'$; otherwise $\delta = 0$. So the problem is reduced to estimating the mean value of the terms in the sum.

As in the proof of the lemma, the main idea is to represent these terms as the sum of identically distributed random variables. Fix (k,ℓ) and (k',ℓ'); set $\epsilon_{ij} = 1$ if i,j belong to different (k,ℓ), (k',ℓ')-components, respectively; otherwise, let $\epsilon_{ij} = 0$. A

straightforward calculation shows that

$$Y_n(k,\ell)[Y_n(k',\ell')-\delta] = \frac{1}{kk'} \sum_{i\neq j} \epsilon_{ij} \quad . \tag{17}$$

The average value of the right-hand side of (17) is given by

$$\frac{n(n-1)}{kk'} P(\epsilon_{ij}=1) \tag{18}$$

$$= \frac{n!}{k!k'!(n-k-k')!} C(k,\ell)C(k',\ell')(n-k-k')^{n-k-k'} n^{-n} \quad .$$

Here we have used arguments similar to the ones employed in the derivation of (13). As before we apply Stirling's formula (10) and expression (11) to obtain the estimate

$$E\{Y_n(k,\ell)[Y_n(k',\ell')-\delta]\} = O[g(x_{k-1},x'_{k'-1})n^{-3}] \quad , \tag{19}$$

with

$$g(x,x') = x^{-3/2}x'^{-3/2}(1-x-x')^{-1/2}$$

and $x_k = k/n$, $x'_{k'} = k'/n$. Notice that the right-hand side of (19) does not depend on ℓ or ℓ'. Summing (19) over ℓ,ℓ' where $1 \leq \ell,\ \ell' \leq tn^{1/2} = n^{1/2-\alpha}$ leads to

$$\sum E\{Y_n(k,\ell)[Y_n(k',\ell')-\delta]\} = O\left[g(x_{k-1},x'_{k'-1})n^{-2-2\alpha}\right] \quad .$$

Finally, replacing the sum over k,k' where $sn \leq k',k,\ k'+k \leq n$ by the double integral of $g(x,x')$ over the region $s \leq x,x',\ x+x' \leq 1$, produces the desired conclusion.

References

1. M. E. Hellman, "A Cryptanalytic Time-Memory Trade-Off," IEEE Transactions on Information Theory, Vol. IT-26, No. 4, July 1980.

2. M. E. Hellman and J. M. Reyneri, "Drainage and the DES," Advances in Cryptology: Proceedings of Crypto '82, Plenum Press (New York, 1983).

3. J. J. Quisquater, "Some DES Cycling Results," Crypto '86.

4. B. S. Kaliski, R. L. Rivest, and A. T. Sherman, "Is DES a Pure Cipher," Advances in Cryptology: Proceedings of Crypto '85, Springer-Verlag (Berlin Heidelberg, 1986).

5. B. Harris, "Probability Distributions Related to Random Mappings," Annals of Mat. Statistics, 31 (1959), 1045-1062.

6. P. W. Purdom and J. H. Williams, "Cycle Length in a Random Function," Transactions of the American Mathematics Society, 133 (1968), 547-551.

7. P. G. Pittel, " On Distributions Related to Transitive Closures of Random Finite Mappings," Annals of Probability, Vol. II, No. 2 (1983), 428-441.

8. Yu. L. Povlov, "A Case of the Limit Distribution of the Maximum Size of a Tree in a Random Forest," Mat. Zametki, Vol. 25, No. 5 (1979), 751-760.

9. I. B. Kalugin, "Branching Processes and Random Mappings of Finite Sets," Mat. Zametki, Vol. 34, No. 5 (1983), 757-771.

10. I. B. Kalugin, "Characterization of Random Mappings," Mat. Zametki, Vol. 39, No. 3 (1986), 424-430.

11. J. W. Moon, Counting Labelled Trees: A Survey of Methods and Results, Canadian Mathematical Monographs, No. 1, 1970.

12. W. Feller, An Introduction to Probability Theory and Its Applications, Vol. I, John Wiley, New York (1968).