A Statistical Look at Maps of the Discrete Logarithm

Joshua Holden and Nathan Lindle, Rose-Hulman Institute of Technology

Several algorithms in cryptography are based on the apparent difficulty of solving the discrete logarithm. It, like integer factorization, is attractive in cryptography because the inverse (modular exponentiation) is much easier to compute. The paper "Mapping the Discrete Logarithm" by Daniel Coulter and Joshua Holden takes a look at the functional graphs that can be generated using the function $x \mapsto g^x (\text{mod } p)$ where p is a prime number. It turns out the structure of the graph is largely determined by the interaction between g and $p-1$, and this interaction gives us an easy way to generate many binary functional graphs by choosing the correct values for g. In "Mapping the Discrete Logarithm" the authors extracted some statistics from graphs with p near 100,000. Their work showed evidence that binary functional graphs generated through modular exponentiation were very close to the theoretical values for random binary functional graphs.

This second look tells a slightly different story though. Using many of the same techniques as in the previous paper, we were able to generate values for the theoretical variance in several of the statistics which were measured. While the variance in the number of components and the number of cyclic nodes is similar to the theoretical variance for binary functional graphs, the variance in the average cycle length and the average tail length are far from what was expected. T-tests were also performed to determine if the theoretical and observed means were statistically similar. The results show that in some cases this is true, but in others the test shows that seemingly small differences could actually be quite significant. We also plan to study the differences in the variances more closely to determine their statistical significance.

Through some optimizations to the code used for the previous paper, as well as a conversion from C++ to C, we are now able to complete trials much more quickly. This has allowed us to test values of p near 250,000, and we hope to try higher numbers to produce even more convincing results. The following is a tabulation of the statistics we have gathered so far, along with the variation, theoretical variation, and the P-value obtained after running a two-tailed t-test on the observed statistic comparing it to the theoretical value.

	100043			100057		
	Predicted	Observed	P-value	Predicted	Observed	P-value
Components	6.392	6.364	0.006	6.392	6.389	0.920
Variance	5.158	5.098	NA	5.158	5.117	NA
Cyclic nodes	395.42	395.86	0.690	395.4	395.3	1
Variance	42543	42781	NA	42549	42227	NA
Avg cycle	198.21	198.22	1	198.22	198.32	1
Variance	27211	**20681**	NA	27214	**20393**	NA
Avg. Tail	197.21	196.77	0.548	197.23	197.18	1
Variance	27211	**7336**	NA	27215	**7363**	NA
Max cycle	247.49	247.30	NA	247.50	247.26	NA
Variance	NA	23985	NA	NA	23806	NA

Table 1: Observed and theoretical statistics for p=100043 and p=100057

	106261			250007		
	Predicted	Observed	P-value	Predicted	Observed	P-value
Components	6.422	6.370	0.022	6.850	6.859	0.230
Variance	5.188	5.176	NA	5.616	5.637	NA
Cyclic nodes	407.55	408.43	0.690	625.67	627.74	0.028
Variance	45200	44488	NA	106678	107922	NA
Avg. Cycle	204.28	206.61	0.162	313.33	313.84	0.548
Variance	28908	**22004**	NA	68181	**52192**	NA
Avg. Tail	203.279	201.644	0.368	312.34	313.19	0.272
Variance	28908	**7578**	NA	68181	**18536**	NA
Max cycle	255.06	256.99	NA	391.25	391.95	NA
Variance	NA	25629	NA	NA	60649	NA

Table 2: Observed and theoretical statistics for p=250007 and p=106261