Reduciing Uncertainty in Consecutive Prime Residue Classes with Reconstructability Analysis

Shawn Marincas
Systems Science PhD Student
Portland State University

December 11, 2020

Abstract

Consecutive prime numbers have been shown to exhibit bias in their reduced residue classes using simple statistics[7]. Here we uncover the same bias using reconstructability analysis, an information-theoretic machine-learning methodology[11], which captures the bias as "uncertainty reduction" in a probabilistic graphical model of the prime residue class data. Deeper exploration of the data using this method uncovered further gains in uncertainty reduction from more complex models using residue classes of multiple prime modulo.

Introduction

The 2016 paper by Lemke Oliver and Soundararajan[7] found statistically significant bias within pairs of consecutive prime number residue classes. The residue class (mod q), r, of $n \pmod q$ is the set of remainders, or residues, left over after dividing an integer, n, by another integer, q. Lemke Oliver and Soundararajan found was that consecutive prime numbers would tend not to have the same residue class for the the same values of q.

$\pi(x_0; 5, (A, B))$	A_1	A_2	A_3	A_4	$p(B)$
B_1	0.0462	0.0637	0.0601	0.0799	**0.2499**
B_2	0.0750	0.0444	0.0704	0.0601	**0.2499**
B_3	0.0743	0.0676	0.0444	0.0637	**0.25**
B_4	0.0544	0.0743	0.0750	0.0462	**0.2499**
$p(A)$	**0.2499**	**0.2499**	**0.25**	**0.2499**	**$\approx$ 1.0**

Table 1: Probabilities of consecutive prime residue class (mod 5) tuples in the first 10^8 consecutive primes.

Table 1 shows this bias most clearly along the diagonal representing the probabilities of consecutive primes sharing the same residue (mod 5). Stated information-theoretically, we would say that knowledge of a prime's residue (mod 5) captures information, or reduces our uncertainty, of the residue (mod 5) of the next prime.

$$H(x) = -\sum_{j}^{n} p_j log_2 p_j \tag{1}$$

To quantify uncertainty reduction we find the entropy of a particular model by applying Shannon's entropy[10] equation (1) to the contingency table built for a particular model from the data. This entropy value is compared that of a reference model to determine by how much the uncertainty is reduced by the particular model over the reference model. If we are using a uniform reference model then this measurement tells us how much better than a random number generator this model would be at reconstructing the data being analyzed.

Methodology - Reconstructability Analysis

Reconstructability analysis (RA) is based on the work of Ashby[1] and was developed in the systems community (Klir[4], Conant[2], Krippendorf [6], and others). A general overview is provided by Zwick[11], as well as a number of examples of RA applications in the fields of healthcare with Froemke[3], cellular automata with Shu[12], and genetics with Kramer[5].

Reconstructability analysis can be used to find and fit models in categorical data using information-theoretic methods, as described above, or using set-theoretical methods not explored in this paper. Information-theoretic RA is implemented by the open source OCCAM software[8], which is also provided as a web service application by the Systems Science department at Portland State University[9].

OCCAM searches a lattice of structures representing the possible RA models in a hierarchy, from the independence model at the bottom to the "saturated" model at the top. The saturated model includes every variable in the data and represents the maximum possible complexity and minimum possible uncertainty.

If we compile a dataset of consecutive prime residues, we should be able to find a reduction in uncertainty for the model outlined by Lemke Oliver and Soundararajan using this method.

Data - Consecutive Prime Residue Classes

Our data consists of the residue classes $\pmod{\{3, 4, 5, 6, 7, 8, 9, 11, 13, 16\}}$ of the first 10^8 consecutive primes, with exception to the first 10 for convenience. The variables for each residue class are in Table 2.

Variable	Cardinality	Description
A	2	Last prime residue (mod 3)
B	2	Last prime residue (mod 4)
C	4	Last prime residue (mod 5)
D	2	Last prime residue (mod 6)
E	6	Last prime residue (mod 7)
F	4	Last prime residue (mod 8)
G	6	Last prime residue (mod 9)
H	10	Last prime residue (mod 11)
I	12	Last prime residue (mod 13)
J	8	Last prime residue (mod 16)
Z_A	2	Current Prime residue (mod 3)
Z_B	2	Current Prime residue (mod 4)
Z_C	4	Current Prime residue (mod 5)
Z_D	2	Current Prime residue (mod 6)
Z_E	6	Current Prime residue (mod 7)
Z_F	4	Current Prime residue (mod 8)
Z_G	6	Current Prime residue (mod 9)
Z_H	10	Current Prime residue (mod 11)
Z_I	12	Current Prime residue (mod 13)
Z_J	8	Current Prime residue (mod 16)

Table 2: Residue class data variables, with cardinalities.

Table 3 shows a sample of this dataset with variables defined in Table 2. While the dataset contains all the IVs (independent variables) and DVs (dependent variables), each analysis ignores all but one of the DVs.

An OCCAM Search analyzes models of IVs and the DV, looking for models which maximize uncertainty reduction in the DV. After a model is found, an OCCAM Fit will build joint probability tables for the model components, from which a decision tree can be built.

From the Lemke Oliver & Soundararajan paper, we should expect an OCCAM Search to find uncertainty reduction in models including the current and previous prime of the same modulo, e.g. AZ_A, BZ_B, etc. Doing a Fit on these models should generate a contingency table that aproximates the same results in the LO&S paper.

Results - Uncertainty Reduction

When searching loop-less models a clear tendency was found for models combining the residues of prime moduli. The best models for many of the DVs, and especially the prime moduli DVs, was $ACEHI$ which is the model using all of the prime moduli in the data.

A	B	C	D	E	F	G	H	I	J	Z_A	Z_B	Z_C	Z_D	Z_E	Z_F	Z_G	Z_H	Z_I	Z_J
2	3	2	5	5	7	2	3	8	15	2	1	3	5	4	5	8	9	1	5
2	1	3	5	4	5	8	9	1	5	2	3	4	5	3	3	5	4	7	11
2	3	4	5	3	3	5	4	7	11	1	1	1	1	5	5	7	6	9	13
1	1	1	1	5	5	7	6	9	13	1	3	2	1	4	3	4	1	2	3
1	3	2	1	4	3	4	1	2	3	2	3	1	5	1	7	8	5	6	7
2	3	1	5	1	7	8	5	6	7	1	1	3	1	3	1	1	7	8	9
1	1	3	1	3	1	1	7	8	9	1	3	4	1	2	7	7	2	1	15
1	3	4	1	2	7	7	2	1	15	2	3	3	5	6	3	2	6	5	3
2	3	3	5	6	3	2	6	5	3	2	1	4	5	5	1	8	1	11	9

Table 3: A sample of prime residue class data, IVs on the left are residue classes of P_{N-1}, DVs on the right are the same residue classes of P_N.

ID	Model	Level	ΔDF	α	Information	%ΔH(DV)	ΔBIC	Inc. α	Prog.	%C(Data)	%cover
47	ABCDEFGHIJZc	10	13271037	0	1	7.7409	-222999094.7	1	45	39.6263	3.125
46	IV:ACDEFGHIJZc	9	6635517	0	1	7.7409	-100768300.2	1	42	39.6263	6.25
45	IV:BCDEFGHIJZc	9	6635517	0	1	7.7409	-100768300.2	1	43	39.6263	6.25
44	IV:ABCDEGHIJZc	9	3317757	0	1	7.7409	-39652902.99	1	39	39.6263	12.5
43	IV:BCDEFGHIZc	8	829437	0	0.99094402	7.6708	5989281.623	1	32	39.562	25
42	IV:ACDEFGHIZc	8	829437	0	0.99094402	7.6708	5989281.623	1	36	39.562	25
41	IV:ABCDEFGHIZc	9	1658877	0	0.99094402	7.6708	-9289567.689	1	42	39.562	12.5
40	IV:ABCDEFHIJZc	9	2211837	0	0.98793839	7.6476	-19539975.5	1	39	39.5401	6.25
39	IV:ABCDEHIJZc	8	552957	0	0.98793839	7.6476	11017723.12	1	29	39.5401	25
38	IV:ABCDEGHIZc	8	414717	0	0.98642803	7.6359	13531782.11	1	37	39.5295	25
37	IV:ABCEGHIZc	7	207357	0	0.98642803	7.6359	17351494.44	1	28	39.5295	50
36	IV:ACDEFHIZc	7	138237	0	0.98497423	7.6246	18593529.74	1	31	39.5189	50
35	IV:ACEFHIZc	6	69117	0	0.98497423	7.6246	19866767.18	1	25	39.5189	100
34	IV:ABCDEFHIZc	8	276477	0	0.98497423	7.6246	16047054.85	1	32	39.5189	25
33	IV:ABCEFHIZc	7	138237	0	0.98497423	7.6246	18593529.74	1	35	39.5189	50
32	IV:BCDEFHIZc	7	138237	0	0.98497423	7.6246	18593529.74	1	31	39.5189	50
31	IV:CDEFHIZc	6	69117	0	0.98497423	7.6246	19866767.18	1	24	39.5189	100
30*	IV:CEGHIZc	5	51837	0	0.98424046	7.6189	20169328.06	0	23	39.5134	100
29	IV:ABCDEHIZc	7	69117	0	0.98351178	7.6133	19835379.44	1	26	39.5082	50
28	IV:ABCEHIZc	6	34557	0	0.98351178	7.6133	20471998.16	0	22	39.5082	100
27	IV:BCDEHIZc	6	34557	0	0.98351178	7.6133	20471998.16	0	21	39.5082	100
26	IV:ACDEHIZc	6	34557	0	0.98279785	7.6078	20456675.42	1	24	39.5029	50
25*	IV:ACEHIZc	5	17277	0	0.98279785	7.6078	20774984.78	0	17	39.5029	100
24*	IV:CDEHIZc	5	17277	0	0.98279785	7.6078	20774984.78	0	18	39.5029	100
23*	IV:CEGHZc	4	4317	0	0.77355558	5.988	16522867.42	0	16	37.889	100
22	IV:ABCEHZc	5	2877	0	0.77349646	5.9876	16548124.28	1	19	37.8883	100
21	IV:BCDEHZc	5	2877	0	0.77349646	5.9876	16548124.28	1	20	37.8883	100
20*	IV:CDEHZc	4	1437	0	0.773444	5.9872	16573524.21	0	12	37.8881	100
19*	IV:ACEHZc	4	1437	0	0.773444	5.9872	16573524.21	0	13	37.8881	100
18*	IV:CDEIZc	4	1725	0	0.74289985	5.7507	15912667.01	0	14	37.583	100
17*	IV:ACEIZc	4	1725	0	0.74289985	5.7507	15912667.01	0	15	37.583	100
16*	IV:CEGZc	3	429	0	0.55980145	4.3334	12006801.98	0	10	36.0308	100
15*	IV:ACEZc	3	141	0	0.55979122	4.3333	12011887.64	0	9	36.03	100
14*	IV:CDEZc	3	141	0	0.55979122	4.3333	12011887.64	0	8	36.03	100
13*	IV:ACHZc	3	237	0	0.49048893	3.7968	10522723.14	0	7	34.8074	100
12*	IV:CDHZc	3	237	0	0.49048893	3.7968	10522723.14	0	7	34.8074	100
11*	IV:CEZc	2	69	0	0.32081986	2.4834	6884305.592	0	5	32.6711	100
10*	IV:CGZc	2	69	0	0.31871201	2.4671	6839065.995	0	4	32.9222	100
9*	IV:ACZc	2	21	0	0.31871053	2.4671	6839918.451	0	3	32.9204	100
8*	IV:CDZc	2	21	0	0.31871053	2.4671	6839918.451	0	2	32.9204	100
7*	IV:CHZc	2	117	0	0.28328243	2.1929	6077776.633	0	6	32.2527	100
6*	IV:CZc	1	9	0	0.17264152	1.3364	3705142.314	0	1	30.4308	100
5*	IV:EZc	1	15	0	0.01267883	0.0981	271842.2082	0	1	26.9375	100
4*	IV:GZc	1	15	0	0.01003313	0.0777	215059.1452	0	1	26.3393	100
3*	IV:AZc	1	3	0	0.01003277	0.0777	215272.5338	0	1	26.3393	100
2*	IV:DZc	1	3	0	0.01003277	0.0777	215272.5338	0	1	26.3393	100
1*	IV:Zc	0	0	1	0	0	0	0	0	25.0004	100

Table 4: OCCAM Search of loopless models for directed residue class (mod 5), Z_C

		observed data				calculated model								
C	freq	Zc=1	Zc=2	Zc=3	Zc=4	Zc=1	Zc=2	Zc=3	Zc=4	rule	#correct	%correct	p(rule)	p(margin)
1	24999432	18.493	30.019	29.718	21.77	18.493	30.019	29.718	21.77	2	7504611	30.019	0	0
2	25000399	25.496	17.757	27.02	29.727	25.496	17.757	27.02	29.727	4	7431869	29.727	0	0
3	25000130	24.044	28.175	17.77	30.011	24.044	28.175	17.77	30.011	4	7502895	30.011	0	0
4	25000024	31.966	24.051	25.492	18.492	31.966	24.051	25.492	18.492	1	7991430	31.966	0	0
	99999985	24.999	25	25	25	24.999	25	25	25	2	30430805	30.431		

Table 5: OCCAM Fit results for the CZ_C model, with $N = 10^8$ the calculated model matches the observed almost exactly and future Fit tables will show only the caclulated model.

DV	Model	Level	ΔDF	Alpha	Information	ΔH(DV)	ΔBIC	Inc. Alpha	Progenitor	%C(Data)	% Cover
Zc	IV:CZc	1	9	0	0.17264152	1.3364	3705142.314	0	1	30.4308	100
Zc	IV:CEZc	2	69	0	0.32081986	2.4834	6884305.592	0	5	32.6711	100
Zc	IV:ACEZc	3	141	0	0.55979122	4.3333	12011887.64	0	9	36.03	100
Zc	IV:ACEHZc	4	1437	0	0.773444	5.9872	16573524.21	0	13	37.8881	100
Zc	IV:ACEHIZc	5	17277	0	0.98279785	7.6078	20774984.78	0	17	39.5029	100

Table 6: Best model per level of a loopless OCCAM directed search for p_n (mod 5), Z_C, up to the best overall model at level 5.

Looking at the single variable models at Level 1 of the OCCAM Search results in Table 4, we see that the model CZc does indeed reduce the uncertainty, and does it better than any other model at the same level. Doing a Fit on this model shows the joint probability which demonstrates the matching bias along the diaganol against repeating residue classes.

Looking again at the Search results in Table 4 we see that this model is far from the best at reducing uncertainty. If we look at the models in Level 2 we see that every model of 3 variables improves on the uncertainty reduction for 2 variables, with the best model (CEZ_C) improving significantly over the single variable CZ_C model.

Each additional level provides similar improvements, with the best models building on the one below by adding an additional prime residue class variable, Table 6 and Figure 1.

Searching without the restriction to loopless models produces a similar increase in uncertainty reduction as new variables are added up to the top of the search at Level 10, as in Figure 2. However, the %ΔH(DV) for “loopy” models does not reach the maximum seen in the loopless model search.

If we look at uncertainty reduction over the prime DVs in 3 we also see a trend of increasing uncertainty reduction, with the best overall reduction being the model $ACEHIZi$ where we use the previous prime residue classes p_{n-1} (mod $q \in \{3, 5, 7, 11, 13\}$) to predict the current prime residue class, p_n (mod 13).

As we increase the number of prime IVs and the size of the prime DV in our model, we also see a steady increase in uncertainty reduction. If we do a Fit on the best model found across all DVs, $ACEHIZi$, we get a table of 5760 rows representing all the possible combinations of the residue classes (mod $q \in \{3, 5, 6, 11, 13\}$), see Table 7.

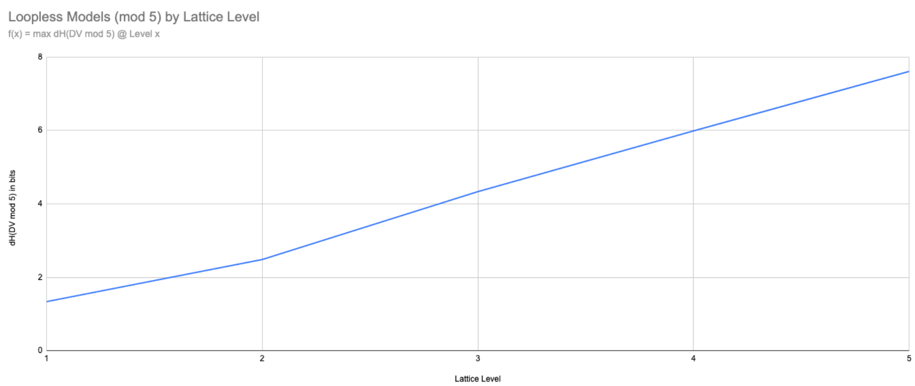

Figure 1: Best loopless models of uncertainty reduction in $p_n \pmod 5$, Z_C, proceeding up the lattice levels.

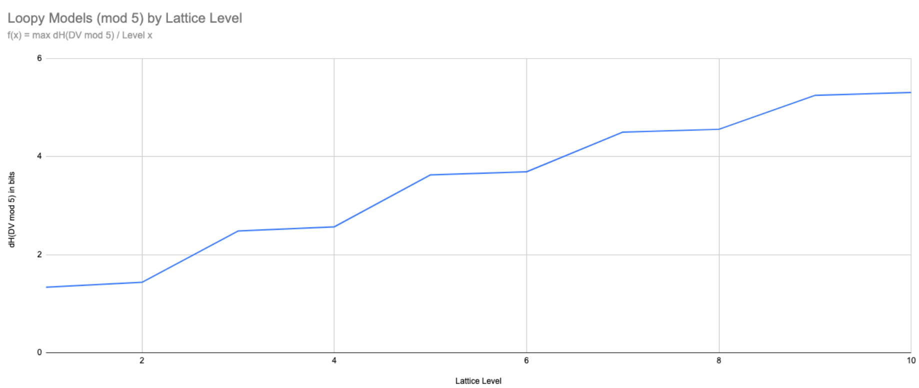

Figure 2: Best overall models of uncertainty reduction in $p_n \pmod 5$, Z_C, proceeding up the lattice levels.

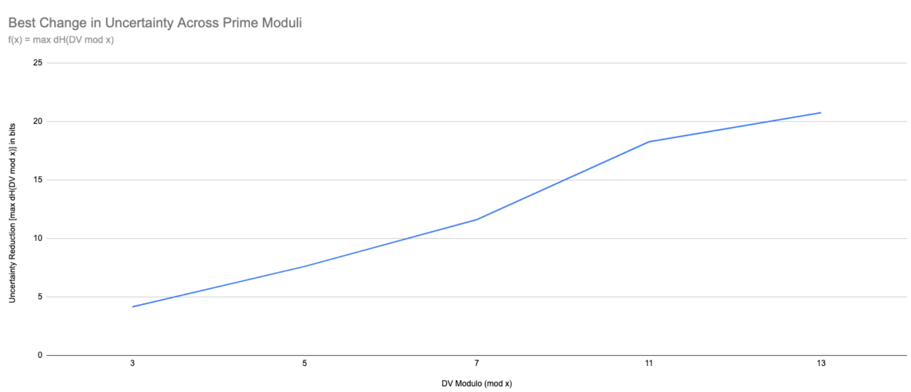

Figure 3: Best loopless models of uncertainty reduction across DVs.

A	C	E	H	I	freq	Zi=1	Zi=2	Zi=3	Zi=4	Zi=5	Zi=6	Zi=7	Zi=8	Zi=9	Zi=10	Zi=11	Zi=12	rule	#correct	%correct
1	4	4	4	1	17333	0.9	0.121	17.608	5.792	36.958	1.402	2.285	0.9	10.737	22.518	0.548	0.231	5	6406	36.958
1	3	4	5	1	17328	0.381	4.097	1.876	18.813	36.559	15.178	1.46	3.295	9.447	2.383	5.996	0.514	5	6335	36.559
1	4	4	5	1	17380	0.69	5.903	2.532	4.114	36.525	20.086	2.192	0.04	10.115	17.319	0.316	0.167	5	6348	36.525
1	2	4	5	1	17397	0.023	4.667	1.77	22.349	36.322	1.293	0.362	2.316	9.622	14.411	5.834	1.029	5	6319	36.322
1	1	5	5	1	17378	2.825	9.472	0.075	5.933	0.472	20.474	0	4.35	1.847	17.706	35.925	0.921	11	6243	35.925
1	2	4	6	1	17372	0.035	4.358	1.997	3.454	35.787	1.658	19.382	3.189	8.41	14.742	6.061	0.927	5	6217	35.787
1	4	4	10	1	17304	0.029	5.727	12.962	4.537	34.316	19.626	2.335	0.89	1.456	17.539	0.341	0.243	5	5938	34.316
1	4	4	2	1	17485	0.532	4.633	12.742	0.063	34.058	19.234	2.162	0.698	7.429	17.947	0.32	0.183	5	5955	34.058
...					...															
					99999985	8.333	8.334	8.333	8.333	8.333	8.334	8.333	8.334	8.335	8.333	8.333	8.333	2	27479970	27.480

Table 7: Sample of joint probability table for the OCCAM Fit of model $ACEHIZi$ with p(margin) and p(rule) columns excluded as they were 0 for all rows.

What is interesting is that we have many combinations with extremely low frequencies, including one in this sample that has 0 occurrences where $p_n \equiv 7 \pmod{13}$ when $p_{n-1} \equiv (1, 1, 5, 5, 1) \pmod{(3, 5, 7, 11, 13)}$, despite the $A_1C_1E_5H_5I_1$ state having a frequency of 17378 in the first 10^8 primes. There are many of these 0 states for each of the Z_I states, and it would be interesting to see how many of these cases would remain unseen in larger prime tuples.

Discussion

The results show that consecutive prime number residue classes have structure, such that a relatively simple model of five residue classes $p_{n-1} \pmod{\{3, 5, 7, 11, 13\}}$ can capture information about the next prime number residue class $p_n \pmod{13}$ by a significant margin in the first 10^8 consecutive primes.

If we look at the best model, $ACEHIZi$, each of the 5760 combinations for possible residues $p_{n-1} \pmod{q \in \{3, 5, 7, 11, 13\}}$ could be represented by each of the 5760 reduced residue classes mod 15015. 15015 is $\frac{1}{2}$ of 30030, which is $p_{\#7}$, the 7th primorial number, the product of the first 7 primes, $1 * 2 * 3 * 5 * 7 * 11 * 13$. Since residues $\pmod 2$ provide no useful information (all primes are $\equiv 1 \pmod 2$) we should be able to see equivalent uncertainty reduction using a single variable of of $p_N \pmod{\frac{1}{2}p_{\#q}}$, though this hypothesis remains to be tested.

Additional research might seek to test these models against sets of larger prime numbers, possibly looking to quantify the rate at which uncertainty reduction is reduced or the theoretical limits of uncertainty reduction possible given increasing sets of prime number residue classes and unlimited computational resource.

Conclusion

We have demonstrated how reconstructability analysis can be applied to an exploration of number theory via modular arithmetic and prime numbers. Other applications of machine learning to number theory are likely worth pursuing.

References

[1] W Ross Ashby. *Constraint analysis of many-dimensional relations.* Tech. rep. Illinois University Urbana Electrical Engineering Research Lab, 1964.

[2] Roger C Conant. "Set-theoretic structure modelling". In: *International Journal Of General System* 7.1 (1981), pp. 93–107.

[3] Cecily Froemke and Martin Zwick. "Predicting risk of adverse outcomes in knee replacement surgery with reconstructability analysis". In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI).* 2017 IEEE Symposium Series on Computational Intelligence (SSCI). Honolulu, HI: IEEE, Nov. 2017, pp. 1–6. ISBN: 978-1-5386-2726-6. DOI: 10.1109/SSCI.2017.8280870. URL: http://ieeexplore.ieee.org/document/8280870/ (visited on 10/21/2019).

[4] George J. Klir. "Identification of generative structures in empirical data†". In: *International Journal of General Systems* 3.2 (1976), pp. 89–104. DOI: 10.1080/03081077608934746. eprint: https://doi.org/10.1080/03081077608934746. URL: https://doi.org/10.1080/03081077608934746.

[5] Patricia Kramer et al. "Reconstructability analysis of genetic loci associated with Alzheimer disease". In: *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems.* 2012 Joint 6th Intl. Conference on Soft Computing and Intelligent Systems (SCIS) and 13th Intl. Symposium on Advanced Intelligent Systems (ISIS). Kobe, Japan: IEEE, Nov. 2012, pp. 2104–2110. ISBN: 978-1-4673-2743-5. DOI: 10.1109/SCIS-ISIS.2012.6505196. URL: http://ieeexplore.ieee.org/document/6505196/ (visited on 10/21/2019).

[6] Klaus Krippendorff. "An algorithm for identifying structural models of multi-var1ate data". In: *International Journal Of General System* 7.1 (1981), pp. 63–79.

[7] Robert J. Lemke Oliver and Kannan Soundararajan. "Unexpected biases in the distribution of consecutive primes". In: *Proceedings of the National Academy of Sciences* 113.31 (July 2016), E4446–E4454. ISSN: 1091-6490. DOI: 10.1073/pnas.1605366113. URL: http://dx.doi.org/10.1073/pnas.1605366113.

[8] Portland State University OCCAM Project Team. *OCCAM Open Source Project.* Mar. 8, 2019. URL: https://github.com/occam-ra/occam.

[9] PSU Systems Science Department. *OCCAM Web Interface.* Version 3.4.0. Apr. 19, 2019. URL: http://dmit.sysc.pdx.edu/weboccam.cgi.

[10] Claude Elwood Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/document/6773024/.

[11] Martin Zwick. "An overview of reconstructability analysis". In: *Kybernetes* 33.5/6 (2004), pp. 877–905.

[12] Martin Zwick and Hui Shu. "Set-Theoretic Reconstructability of Elementary Cellular Automata". In: *Advances in Systems Science and Applications* (I 1995), p. 6. ISSN: 1078-6236.