Rapport de soutenance de thèse

Nom et prénom : Mr ZENIL CHAVEZ HECTOR

Titre de la thèse :
UNE APPROCHE EXPÉRIMENTALE À LA THÉORIE ALGORITHMIQUE DE LA COMPLEXITÉ

Discipline : INFORMATIQUE

A travers un exposé passionnant et clairement illustré. Mr Zenil a soutenu sa thèse sur une approche expérimentale à la théorie algorithmique de la complexité.

Il a su mettre en valeur un travail basé sur des concepts difficiles qu'il a présenté de manière pédagogique.

Le jury s'associe à l'éloge des rapporteurs sur la qualité et la diversité des travaux qui ont été présentés.

Les réponses apportées aux questions incisives du jury ont été pleinement convaincantes et montrent l'étendue des connaissances du candidat aussi bien au niveau pratique que théorique.

Le jury a également apprécié que le candidat de langue natale espagnole, présente ses transparents en anglais avec un discours en français de grande qualité.

Pour toutes ces raisons, le jury décerne à Mr Zenil le titre de docteur de l'Université Lille1 avec la mention très honorable.

Mention :

□ **Honorable** ☒ **Très Honorable**

L'USTL n'attribue plus de mention "Très Honorable avec les Félicitations" (par décision du Conseil Scientifique du 15 juin 2007)

Report on the Ph.D. Thesis submitted by Hector Zenil to the University of Lille: "Une approche expérimentale à la théorie algorithmique de la complexité"

The thesis submitted by the candidate does not consist of a single piece of work, but instead of a substantial body of published or publicly presented works that are collected together as individual chapters in this Ph.D. thesis. The whole is definitely greater than the sum of its parts, for although it may not be so obvious from the individual journal or conference papers that comprise the chapters of this thesis, the candidate has accumulated a body of work with a definite and novel point of view.

I should also comment on the excellent conceptual overview of the theory of algorithmic complexity at the beginning of the thesis, which it was a pleasure to read, and in which the candidate clearly exhibits mastery of his field.

Another important aspect of the candidate's work is that he has internalized the "experimental mathematics" proposals by Stephen Wolfram, Jonathan Borwein, Thomas Tymoczko and others regarding the utility of adopting an empirical approach even when working with the Platonic world of ideas. Instead of attempting to justify this approach, as happens in the writings of others, Zenil wastes no time on methodological considerations and immediately uses this "quasi-empirical" approach to good effect in his work. He thus represents a new kind of theoretician, one that did not exist a short time ago.

The fundamental intellectual tension in Zenil's work is the apparent contradiction between the fact that algorithmic complexity is extremely, dare I say violently, uncomputable, but nevertheless often is irresistible to apply in practical applications. So Zenil has, on the one hand, meditated deeply on the question of obtaining reasonable approximations to algorithmic complexity, and he simultaneously exhibits considerable dexterity and imagination in finding practical applications that may profitably be analyzed using the fore-mentioned reasonable approximations.

Since this *Rapporteur* is a theoretician with little interest in practical applications, let me comment on the considerable interest of Zenil's theoretical efforts. The theory of algorithmic complexity is of course now widely accepted, but was initially rejected by many because of the fact that algorithmic complexity depends on the choice of universal Turing machine and short binary sequences cannot be usefully discussed from the perspective of algorithmic complexity. Nevertheless Zenil has discovered, employing his empirical, experimental approach, the fact that most reasonable choices of formalisms for describing short sequences of bits give consistent measures of algorithmic complexity! So the dreaded theoretical hole in the foundations of algorithmic complexity turns out, in practice, not to be as serious as was previously assumed.

This and related considerations permit Zenil to feel free to use practical approximations to algorithmic complexity in his applied work, which he does with great skill, representing what one might term a new kind of "practical theoretician/experimentalist," and constituting a marked turn in the field of algorithmic complexity from deep theory to practical applications.

I therefore recommend that we welcome the candidate into the academic community by accepting this Ph.D. thesis.

In fact, Zenil has already been acting as a member of the academic community: He co-organized the October 31, 2008 through November 2 University of Indiana at Bloomington meeting on "What is Computation? How Does Nature Compute?," which attracted international attention. He also edited and is editing two important collections of papers in the field of theoretical computer science, collections which have attracted a long list of distinguished contributors: Zenil's *Randomness Through Computation*, World Scientific, already published in 2011, and his forthcoming book *Computation in Nature & The Nature of Computation*, World Scientific, to appear.

So it is a pleasure to finally be able to officially welcome Zenil into the international academic community; he has already been acting as a full member of that community for some time now.

Gregory Chaitin
University of Buenos Aires

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Cristian S. Calude, MAE
Chair Professor and CDMTCS Director
Department of Computer Science
Private Bag 92019
Auckland, New Zealand
www.cs.auckland.ac.nz/~cristian/

May 19, 2011

Report on the PhD Thesis

Une Approche Expérimentale à la Théorie Algorithmique de la Complexité

presented to the University Lille 1, Sciences and Technology by Zenil Chavez Hector

The *Thesis* contains an *Introduction* (dedicated to the main theoretical tools from computability theory, algorithmic complexity, and algorithmic probability) and three parts: *Foundations, Applications* and *Reflexions.*

The *Foundations* studies experimentally the plain complexity of short strings. First, the strings are generated with 4096 Turing machines: all 2-state, 2-symbol Turing machines and a statistical sample of 3-state, 2-symbol Turing machines, later extended to 4-state Turing machines for which the halting problem is solved thanks to various computations of the busy beaver problem. An approximation method of the algorithmic complexity of these small strings is proposed based on their frequencies and the coding theorem applied to a very large sample of Turing machines. A strong experimental confirmation of a theoretical result saying that most programs that halt actually stop in a very short time was obtained.

The second *Applications* starts with a compression-based investigation of dynamical properties of cellular automata. This approach classifies cellular automata into clusters according to their heuristic behaviour, which seems related to Wolframs main classes of cellular automata. A compression-based method to estimate a characteristic exponent for detecting phase transitions and measuring the resiliency or sensitivity of a system to its initial conditions is developed leading to a conjecture regarding the capability of a system to reach computational universality. Evidence for the validity of Wolframs Principle of Computational Equivalence was obtained. This part continues with the use of Bennetts logical depth to experimentally assess and quantify the information content of an image. The last chapter in this part is dedicated to Sloane's database of more than 150000 numerical sequences of integers. Many interesting questions can be asked bout this amazing database. For example, which numbers do not appear in Sloane's database? In August 2008 the smallest absent number was 8795; in February 2009 the smallest absent number was 11630. This time instability suggested the need to study the distribution of numbers rather than of their mere presence or absence. To this aim the author has studied the function giving the number of occurrences of of an integer k, $N(k)$, in the database. P. Guglielmetti observed the cloud-shaped formation of points determined by the function N: there are two parts separated by a clear zone, as if the numbers sorted themselves into two categories, the more interesting above the clear zone, and the less interesting below. This phenomenon, called by the authors Sloane's gap, is experimentally investigated.

The last part *Reflexions* starts with a test to check the claim that the world is algorithmic. A statistical comparison of the frequency distributions of data collected from various physical sources (from repositories of images, computer stored data, computer programs and DNA sequences) and of data algorithmically generated (by running abstract computing devices such as Turing machines, cellular automata and Post Tag systems). Statistical correlations have been found. This study is followed by an algorithmic information-theoretic model of the behaviour of financial markets and an experimental investigation of the trade-offs between program-size and time computational complexity for small Turing machines which reveal "phase-transitions" in the halting probability distribution.

The Thesis contains interesting experimental results which contribute to the theoretical understanding of a couple of challenging questions. The author has shown scientific maturity in various directions, including the choice of the material assimilated, of questions studied and tools used.

The Thesis can be defended on 21 June 2011 and, based on it, I recommend that Zenil Chavez Hector is awarded the title of PhD in Computer Science.

C. Calude

Cristian S. Calude

UMR 7089

LABORATOIRE
D'INFORMATIQUE
ALGORITHMIQUE :
FONDEMENTS
ET
APPLICATIONS

université PARIS DIDEROT PARIS 7

Report on Hector ZENIL's Thèse de doctorat
"Une approche expérimentale à la théorie algorithmique de la complexité"
by Serge GRIGORIEFF, emeritus professor, Université Paris 7 Denis Diderot

Background of the thesis.

- *Algorithmic information theory.* AIT introduces fundamental concepts to get a quantitative measure of information contents and randomness. Program-size complexity is the length of shortest programs to get a wanted output whereas Bennett's complexity is the execution time of a shortest program. They are viewed as measuring the information contents and the logical depth or physical complexity of objects. Various notions of programs give variants of these complexities.

- *The Coding Theorem.* The most important variants are those associated to prefix-free programs, i.e. to programming languages where no program is a prefix of another one. The reason lies in the so-called "Coding Theorem". Technically, this result asserts the equality, up to a multiplicative constant, of the quantities $2^{-K(x)}$ and $m(x) = \Sigma\{2^{-length(p)} \mid U(p){\downarrow}= x\}$ where $K(x)$ is the program-size complexity of the string x and U is a universal partial computable prefix-free function which maps finite binary words (seen as programs) into finite binary words (the outputs). The significance of this theorem lies in the fact that $m(x)$ can be seen as the probability (better called semi-measure since the sum of all $m(x)$'s is less than 1) that a universal prefix-free machine U outputs x. The intuitive meaning of this theorem is that, in an algorithmic world where everything is created by a program, objects of the same size are not uniformly distributed but are distributed according to their program size complexity. In other words, for binary words of length n, the probability of a word x is not 2^{-n} but is of the order of $2^{-K(x)}$.

- *Approximating AIT complexities by compressors.* As happens with most objects in computability theory, program-size and Bennett complexities are non computable functions. For applications they have to be approximated. The classical way to do so is to consider compressors and to approximate $K(x)$ with the size of the compressed file obtained from a given file x (considered as a binary word). Also, Bennet's complexity of x is approximated by the decompression time of this compressed file. This method of approximation via compression has been extensively used in the works by Paul Vitanyi, Cilibrasi and al.

Contents of the thesis.

- *Using AIT to revisit Wolfram classification of the dynamical properties of cellular automata.* Long ago, using extensive experiments, Wolfram classified the behaviour of the 256 two-state cellular automata in four groups. This classification has been the starting point of numerous subsequent works. In the paper "Compression-based investigations of the dynamical properties of cellular automata and other systems" (published in the Journal Complex Systems, 2010, cf. Chapter 4 of the thesis), Hector Zenil looks at the program size complexity of the successive configurations when the initial one has only one non quiescent cell. He approximates that complexity by the size of the compressed file. It turns out that the the different shapes of the function *time* ↦ *compressed length* and the clusters of compressed lengths are in close agreement with Wolfram classification. A meticulous analysis of phase transitions leads the author to a very interesting conjecture relating universality of a cellular automaton to some phase transition coefficient.

• *Using AIT to classify the dynamical properties of small Turing machines.* In the same paper, the same analysis done for cellular automata is done with small Turing machines having 3 states and 2 symbols. It involves delicate and heavy computations.

• *Image classification using Bennett complexity.* A limitation of program size complexity is that it does not discriminate random objects and richly structured ones: all have high complexity. Nevertheless, Bennett complexity does discriminate such objects. The execution time of best programs for random objects is much longer than that for richly structured ones. Zenil applies these ideas to classify images (cf. Chapter 5 and the associated paper). He uses approximations by compression: length of the compressed file and time to decompress. This chapter explains very carefully all the problems involved to measure a decompression time. In particular, he analyses and explains the solutions brought to the following difficulty: an execution time does depend of the current state of the operating system. The extent of work which was necessary to significant and fair computational experiments appears to be quite impressive.
The classification of images is illustrated by a panel of very different kinds of images. The result is fascinating and plainly corroborates the roles of program size and Bennett complexity as concerns randomness and rich structure.

• *AIT and Sloane's gap.* Sloane's encyclopedia of numeric sequences registers more than 150 000 numerical sequences. Using this encyclopedia, to any number one can attach the number of sequences in which it occurs. The resulting (discrete) map (with values in logarithmic scale) has a very special shape, concentrating on a horn-like area. Is there a mathematical reason for such a remarkable distribution? This is an open question. Zenil looks at it with AIT and argues that social contingencies may enter the question.

• *A new method to approximate AIT complexities of short strings.* Hector Zenil adds to the compression method a very original one (cf. Chapters 2 and 3 and the associated published paper by Zenil and Delahaye). Consider Turing machines starting with an initial blank tape and a given number n of states and consider the ratio $D_n(x)$ between the number of n-states TMs which halt and output x and the total number of such n-states TMs. It happens that the halting problem, which is unsolvable in general, is solvable for such n-tapes TMs in case $n \leq 4$. This allows to exactly know the map D_n. It turns out that 2-states (resp. 3-states, resp. 4-states) TMs output only 22 (resp. 128, resp. 1824) strings, all with length ≤ 4 (resp. 7, resp. 16). Each D_n orders the strings in its range: say x appears more frequently than y if $D_n(x) > D_n(y)$. A close examination of the values of the $D_n(x)$'s shows a remarkable agreement between D_2, D_3 and D_4 as concerns the way they order strings. Also, simple

strings such as 0000 are ranked above more complex ones of the same size.
Now, Zenil takes a very bold step: consider that $D_n(x)$ is an approximation of the semi-measure $m(x)$ (cf. supra) and, using the coding theorem, consider that $K(x)$ is approximated by $-\log D_n(x)$.
For obvious computational limitations, this approach only works for small strings. But this is not a limitation at all. On the contrary, it brings a possibility to approximate K in a case where compression fails: compressing a short file with gzip makes no sense.

- *AIT complexities in the micro-cosmos of small Turing machines.* Chapter 9 (and the associated published paper) develops the above analysis of small Turing machines and looks at the execution times. Some years ago Cristian Calude proved that "most programs stop quickly or never halt". Zenil obtains a spectacular confirmation of this result on small TMs. Moreover, best runtimes are obtained from simplest machines: as Zenil states, *"when computing a particular function, slow-down is more likely than speed-up if the TMs have access to more resources to perform their computations"*. Also, the runtimes probability distributions show unexpected phase transitions. This chapter opens a new area of theoretical research in computability theory: what Zenil observed by very clever and sophisticated experimentation is most probably the trace of still unnoticed theorems about models of computations.

- *AIT approach to the behaviour of financial market.* This subject has been the source of plenty of mathematical work based on analysis and probability theory. Zenil brings AIT into the subject by analyzing the program size complexity of price fluctuations (cf. Chapter 8 and the paper published in "Journal of Economic Surveys, 2011". The main idea is that explained supra: the distribution of data is not uniform when the source is of algorithmic nature. The author confronts the classical approaches to that of AIT and discusses what AIT can bring to the subject. This is a provocative first step into a subject which clearly needs some more mathematical control.

- *On the algorithmic nature of the world.* This is the title of a paper by J.P. Delahaye & H. Zenil (cf. Chapter 7) which raises a popular philosophico-scientific question which is the backbone of AIT. Though to such a question no definitive answer is given (the author is a reasonable person...), it is tackled with much talent using AIT and his very fine experiments involving impressive computational work.

This thesis is a very original one. It brings much substance to the applicability of Algorithmic Information Theory and opens a bunch of theoretical questions about computational models. The main ideas are clearly exposed and the amount of work behind the obtained experimental results is really impressive. The obtained results are fascinating and have been the source of nine papers, six being published, three being submitted. I have the highest opinion of this work and recommend that the thesis be accepted.

Paris, May 21st 2011