Cornell University
Department of Computer Science

Ronan Le Bras
Ph.D. Candidate, Computer Science
Cornell University

Computer Science Department
344 Gates Hall, Ithaca NY 14850
(607) 333-8758

lebras@cs.cornell.edu
www.cs.cornell.edu/~lebras

Research Statement

Overview

The field of computer science is undergoing a fundamental change. The big data revolution calls for improved computational methods, beyond the exponential growth of storage or computational capacity. At the same time, pressing issues in sustainability, including complex environmental, social and economic aspects, are also reshaping the field of computer science, as they often raise new challenges and warrant fundamentally innovative techniques.

In this context, the goal of my research is to advance computational methods in reasoning, inference, machine learning and human computation to process and interpret large and increasingly complex real-world datasets, with a focus on the emerging field of computational sustainability. In my dissertation, I develop techniques for large-scale combinatorial optimization, leverage human insights through alternative representations and visualizations, and exploit hidden structures in the data for accelerated and improved solution techniques. This work is motivated by a series of applications, especially in sustainability-related areas, and has led to a series of scientific discoveries in graph theory, experimental design, combinatorics, and discrepancy theory as well as in materials science, conservation biology and ecology.

Dissertation Research

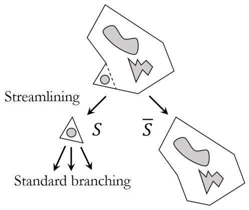
A key insight of my research is that most problems, from theoretical problems in combinatorics to real-world applications, comprise structural properties not directly captured by the problem definition. The research question becomes how to uncover these hidden structures, characterize and exploit them in order to boost and advance the state-of-the-art optimization techniques.

1 Leveraging Human Insights into Problem Structure for Scientific Discovery

Can we exploit modern automated reasoning tools for scientific discovery? Can we use human insights to uncover and exploit hidden structure in combinatorial satisfaction and optimization problems?

In this area, this research led to results that appear in the proceedings of AAAI 2012 [10], CP 2014

[12], HCOMP 2014 [13], and IJCAI 2013 [11, 5].

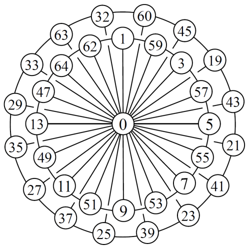


Streamlining

Efficiently solving complex decision and optimization problems may require exploiting intrinsic additional properties of the problems. Automated reasoning techniques derive additional constraints that are implied by the model, i.e. that must be satisfied in all solutions of the problem. In this work, I exploit properties that may not necessarily be derived from a given constraint model. This includes structural properties such as symmetries or regularities that only arise in a subspace of the solution space. This also encompasses hidden structures such as backdoors [5] (i.e. sets of variables that, once instantiated, simplify the remaining problem to a tractable class), which capture the practical complexity of a problem instance. I exploit these structural properties through streamlining [10], a strong branching mechanism that evaluates and propagates a set of constraints corresponding to these properties, and proceeds to search for solutions with these properties. This branching mechanism intentionally discards entire subspaces of the search space in order to focus on a highly structured subspace, thereby boosting constraint reasoning.

I propose to leverage and further extend the concept of streamlined combinatorial search, coupling it with a human computation component, in a complementary, iterative approach. The human computation component is used to identify possible patterns in solutions and suggest insightful regularities [10] or potential backdoors [5]. In practice, through alternative representations and visualizations, the data reveals these properties by exhibiting visual structural patterns, and these observed properties are evaluated through streamlining in order to dramatically speed up the search.

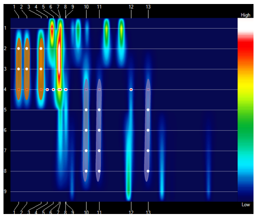
Application to Discovery in Finite Mathematics The study of challenging problems in combinatorics and finite algebra has given rise to significant progress in the area of search, constraint satisfaction, and automated reasoning. In turn, it has led to the discovery of interesting discrete structures with intricate mathematical properties. While some of those results have resolved open questions and conjectures, a key shortcoming is that they generally fail to provide further mathematical insights, from which one could derive more general observations.



Graceful double-wheel graph

Contrastingly, the proposed approach specifically exploits insights about the combinatorial objects. The human computation component allows to conjecture properties about the solutions of the problem. As these properties generalize to larger problem sizes, they are combined to dramatically boost reasoning techniques and obtain solutions of sizes out of reach of traditional techniques [12]. These properties also constitute the building blocks for efficient, constructive procedures for generating classes of complex combinatorial objects [10, 11]. This approach led to a series of results on open problems in finite mathematics, and in particular in experimental design [10], graph theory [11], and discrepancy theory [12]. For example, spatially-balanced designs, used in agronomy to evaluate various soil treatments, were not known, yet conjectured, to exist for 36 treatments or more. I prove that such a structure exists for most sizes, and provide efficient constructions to generate such designs. Similarly, double-wheel graphs were conjectured to be graceful, an important property in graph theory. Yet previous approaches could generate graceful double-wheel graphs only up to size 24. This work formally proves the gracefulness property of any such graph of arbitrary size.

Extension to Crowdsourcing Pattern Decomposition in Big Data I investigate how the identification of structural properties can be translated into abstract pattern visualization tasks, with no information about the original combinatorial optimization problem in question, in a way that allows for crowdsourcing. The results show how human computation and crowdsourcing insights can be key to identifying backdoor variables in combinatorial optimization problems, dramatically speeding up the performance of combinatorial solvers [5]. The approach leverages the complementary strength of human input, providing global insights into the problem structure, and the power of combinatorial solvers to exploit complex local constraints.



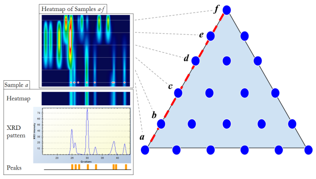
Human task for pattern identification

The framework proposed in [13] decomposes large, complex tasks into units suitable for human feedback, aggregates all user input, and extracts partial solutions to feed a combinatorial solver. This approach provides feedback on the human input, automatically corrects it, and overall leads to significant performance gains. We also show how incorporating expert knowledge into factor-based pattern decomposition techniques lead to significantly more accurate interpretations of the data [3].

2 Computational Sustainability Applications

Pressing issues in sustainability urge us to find new strategies to better manage and allocate our resources. Some of the most challenging problems in sustainability require innovative computational methods to help balance environmental, economic, and societal needs for a sustainable future. I study computational aspects of high-dimensional decision problems and large-scale data analysis under noise conditions and uncertainty in sustainability-related applications. I investigate how to leverage human insights and domain-specific knowledge about these applications. The close collaboration with experts in the fields of materials science, agronomy, conservation biology, and resource economics has made this interdisciplinary research possible. Results in this area appear in the proceedings of AAAI 2013-2014-2015 [8, 1, 5, 6, 3], CP 2011 [7], HCOMP 2014 [13], IJCAI 2013 [5] and SAT 2012-2013 [2, 4].

Combinatorial Materials Discovery for Sustainable Energy Accelerating the discovery and deployment cycle of new advanced materials is instrumental to achieving sustainable, clean energy, as underlined by the White House Materials Genome Initiative. For example, hydrogen fuel cells and solar fuel cells are among the most promising enabling technologies for electric cars and renewable energy storage. Yet, their constituent materials, used for their catalytic or light-absorbency properties, currently limit the full potential and widespread use of these technologies. To find more effective alternatives, materials scientists search for candidate materials in a high-throughput regime, able of outputting millions of materials a day, each of them associated with complex characterization information. The success of this approach relies, however, on efficient, robust and scalable automated analysis techniques capable of identifying the next-generation materials.



Materials discovery

In close collaboration with materials scientists at Cornell and Caltech, we introduce a novel approach that analyzes a library of sample materials using X-ray diffraction data and their composition data, and identifies the key underlying crystal structures. The novelty lies in the integration of complex a priori scientific domain knowledge into state-of-the-art combinatorial optimization techniques [7, 2, 4, 3] and how experts in materials science as well as complete novices may provide valuable input to this solution approach [5, 6]. To date, this approach provides the **most accurate, physically-meaningful** and **reliable interpretation** of the materials data. This breakthrough has been recognized by the scientific community and I have had the honor of being selected for an **invited talk** at the Materials Research Society meeting in Spring 2016.

Design of Scientific Experiments for Sustainable Fertilizer Practices In order to compare various soil treatments (e.g. fertilizers), agronomic laboratories must devise experimental designs that avoid the introduction of biases in the assessment of these treatments. Fertilizers have both a high socio-economic and high environmental impact: better experiment designs directly translate into a better assessment of their quality and their impact on the environment.

In that respect, so-called spatially-balanced designs are formally proved to minimize the potential spatial bias, but such designs were only known to exist for up to 35 treatments. Applying the approach described previously, we provide the first polynomial-time constructive procedure for spatially-balanced designs, proving that they exist for an infinite family of sizes [10]. This construction is now used in a **commercial tool** specialized in agronomic designs. Note that, in addition to field experiments, these spatially-balanced designs have **applications** in greenhouse trials, chemical analyses involving multi-well titer plates and genomics research involving microarray slides.

1	2	3	4	5	6
2	4	6	5	3	1
3	6	4	1	2	5
4	5	1	3	6	2
5	3	2	6	1	4
6	1	5	2	4	3

Spatially-balanced experimental design

Wildlife Conservation & Ecological Monitoring Biodiversity underpins all ecosystems and the goods and services they provide. Yet, many species are threatened by habitat loss and fragmentation due to human land use and climate change. Designing robust, cost-efficient and scalable conservation strategies is an important and challenging computational task. Together with conservation biology experts, we address the problem of designing conservation strategies for landscape connectivity for endangered species. We apply our approach to solve a **large-scale real-world network design** problem for wolverine and lynx populations in Western Montana [8, 1]. The novelty lies in the **robustness** of the solutions to unexpected natural or anthropological disturbances, as well as the ability to handle **multiple species** in a single conservation plan.

Wolverines & Lynx Conservation in Montana

In collaboration with resource economists and ecologists, we use **crowd-sourcing** to assess the **vegetation conditions** in East Africa and improve rangeland and forage maps. We provide local herders with smartphones for them to submit vegetation images and surveys. In addition, we use the crowd (through Amazon Mechanical Turk) to validate the herders' submissions and provide feedback in near real time. The results show the crowd has a substantial positive impact on the **participation rate** of the herders as well as on the **data quality** of their submissions.

Research Plans

In my future work, I will continue developing foundational methods in automated reasoning, inference, learning and human computation. I will keep exploring sustainability-related topics as I aspire to be a major contributor to the emerging field of computational sustainability. Incorporating computational thinking into sustainability not only provides new insights into sustainability issues, but also raises new challenges and entails new methodologies in computer science. Therefore, it benefits the scientific community while having an important and relevant impact on the society at large.

In this interdisciplinary research, the collaboration with experts in other application domains is key and I intend to maintain and nurture an important network of collaborators.

Combinatorial Materials Discovery Combinatorial materials science is an extraordinarily promising avenue of research. Techniques in materials science have only recently allowed to efficiently collect massive amounts of data. As such, combinatorial materials discovery is still greatly underexploited and offers many opportunities for computer scientists, with tremendous potential impact, as was computational biology in its early days. This is a very rich area, with many facets that are still to be explored. For example, I intend to exploit the existing libraries of known or theoretical materials characterization to allow a more thorough and meaningful interpretation of the data. In addition, data analysis techniques should interact with the data collection itself, in an active learning or optimal learning setting, not only to collect data more efficiently but also to suggest promising combinations of elements and experiments to conduct. Another possible important extension is to combine multiple data sources (e.g., about various material properties and characterization) to augment the interpretation of the data and transfer knowledge between different analysis tasks.

Integrating model-driven and data-driven reasoning My work combines deductive, model-driven reasoning (automated solvers deriving properties from the problem statement) and inductive, data-driven reasoning (humans deriving properties from the data). I think a tighter integration of these two methodologies will enable the next generation of AI, and human computation has an important role to play in both methodologies. Indeed, deductive reasoning benefits from injecting refined prior human knowledge into the model, and inductive reasoning through human insights provides valuable input about instance-specific knowledge and features. This structure offers great opportunities, as it frames many exciting research directions involving knowledge extraction from digital content. In this context, new advances such as deep learning constitute a promising research direction, as it complements the proposed approach by learning complex, high-level features and suggesting new representations of the data.

Data science and Human-Computer Interaction Important applications and extensions of my work span research areas beyond Artificial Intelligence and Machine Learning. The inductive reasoning described above can be viewed as statistical inference through crowdsourcing, where human cognition plays the role of statistical tests. This opens up perspectives as how to evaluate the significance of the discovered patterns and which human tasks should be performed. Related to *Human-Computer Interaction*, I plan to explore new user interfaces to enable human insights and communicate them to the solvers, as well as other settings such as game-based incentive schemes. For example, in an application like *Materials Discovery*, I plan to develop other crowdsourcing approaches through alternative visualizations and the development of new user tasks and experiences. The richness of this application makes it particularly suitable to be framed as a multi-faceted citizen science project such as *Zooniverse* or a game with a purpose such as *FoldIt*.

Overall, I believe that this long-term research agenda offers many opportunities for impactful contributions to the field of computer science, the scientific community, and society at large.

References

[1] Dilkina, B., Gomes, C. P., Lai, K., **Le Bras**, R., McKelvey, K. S., Sabharwal, A., Schwartz, M. K., Suter, J. and Xue, Y. [2013], Large conservation landscape - synthetic and real-world datasets, *in* 'the 16th Conference on Artificial Intelligence', AAAI'13.

[2] Ermon, S., **Le Bras**, R., Gomes, C. P., Selman, B. and van Dover, R. B. [2012], Smt-aided combinatorial materials discovery, *in* 'the 15th International Conference on Theory and Applications of Satisfiability Testing', SAT'12.

[3] Ermon, S., **Le Bras**, R., Suram, S. K., Gregoire, J. M., Gomes, C. P., Selman, B. and van Dover, R. B. [2015], Pattern decomposition with complex combinatorial constraints: Application to materials discovery, *in* 'the 29th Conference on Artificial Intelligence', AAAI'15.

[4] Finger, M., **Le Bras**, R., Gomes, C. P. and Selman, B. [2013], Solutions for hard and soft constraints using optimized probabilistic satisfiability, *in* 'the 16th International Conference on Theory and Applications of Satisfiability Testing', SAT'13.

[5] **Le Bras**, R., Bernstein, R., Gomes, C. P. and Selman, B. [2013], Crowdsourcing backdoor identification for combinatorial optimization, *in* 'the 23rd International Joint Conference on Artificial Intelligence', IJCAI'13.

[6] **Le Bras**, R., Bernstein, R., Gregoire, J. M., Suram, S. K., Gomes, C. P., Selman, B. and van Dover, R. B. [2014], A computational challenge problem in materials discovery: Synthetic problem generator and real-world datasets, *in* 'the 28th Conference on Artificial Intelligence', AAAI'14.

[7] **Le Bras**, R., Damoulas, T., Gregoire, J. M., Sabharwal, A., Gomes, C. P. and van Dover, R. B. [2011], Constraint reasoning and kernel clustering for pattern decomposition with scaling, *in* 'the 17th International Conference on Principles and Practice of Constraint Programming', CP'11.

[8] **Le Bras**, R., Dilkina, B., Xue, Y., Gomes, C. P., McKelvey, K. S., Montgomery, C. and Schwartz, M. K. [2013], Robust network design for multispecies conservation, *in* 'the 16th Conference on Artificial Intelligence', AAAI'13.

[9] **Le Bras**, R., Ermon, S., Damoulas, T., Bernstein, R., Gomes, C., Selman, B. and van Dover, R. B. [2012], Materials discovery: New opportunities at the intersection of constraint reasoning and learning, *in* 'International Conference on Computational Sustainability', CompSust'12.

[10] **Le Bras**, R., Gomes, C. P. and Selman, B. [2012], From streamlined combinatorial search to efficient constructive procedures, *in* 'the 15th Conference on Artificial Intelligence', AAAI'12.

[11] **Le Bras**, R., Gomes, C. P. and Selman, B. [2013], Double-wheel graphs are graceful, *in* 'the 23rd International Joint Conference on Artificial Intelligence', IJCAI'13.

[12] **Le Bras**, R., Gomes, C. P. and Selman, B. [2014], On the erdos discrepancy problem, *in* 'the 20th International Conference on Principles and Practice of Constraint Programming', CP'14.

[13] **Le Bras**, R., Xue, Y., Bernstein, R., Gomes, C. P. and Selman, B. [2014], A human computation framework for boosting combinatorial solvers, *in* 'the 2nd AAAI Conference on Human Computation and Crowdsourcing', HCOMP'14.

[14] Zou, T., **Le Bras**, R., Salles, M., Demers, A. and Gehrke, J. [2015], 'Cloudia: a deployment advisor for public clouds', *The VLDB Journal, *Special Issue on the Best Papers of VLDB 2013** .

[15] Zou, T., **Le Bras**, R., Salles, M. V., Demers, A. and Gehrke, J. [2013], Cloudia: a deployment advisor for public clouds, *in* 'the 39th International Conference on Very Large Data Bases', VLDB'13.