2014 PROGRESS REPORT FOR ONR GRANT N000141310260

DAVID I. SPIVAK

1. Introduction

This report, submitted to Predrag Neskovic at the Office of Naval Research, summarizes my past-year's progress toward the goals of ONR grant N000141310260. Broadly speaking, my objective is to understand the fundamental nature of information and communication, and to express it mathematically. Claud Shannon's *Information theory* deliberately avoids the question of meaning; in my research, on the other hand, meaning is of primary importance.

In order for data to be meaningful, it must carry some structure. The relatively young mathematical subject of *Category theory* is the study of structure, and I have identified it as a particularly relevant field for describing the nature of information. The idea is that there exist many different notions of "informativity", i.e., many types of information-bearing structure, including databases, ontologies, computer programs, neural networks, etc. Each such notion includes an infinite array of conforming structures (e.g., an infinite variety of different database schemas are possible), and each such structure is instantiated by an infinite array of conforming data (e.g., there is an infinite variety of instances, or states, for a given database). This layering of abstraction is a major strength of category theory. A closely-related strength is the ability to translate between different paradigms (e.g., from one database to another, or from ontologies to databases), using functors.

In the following sections, I will summarize the events of the past year. These include some collaborations, outreach, transitions, and new hires, as well as multiple advances in research.

Contents

2. Collaborations, Outreach, Transitions, and Recent hires

My research focuses on how category theory can be applied to real-world problems. As such, it is important that the ideas get out of academia. I make an effort to talk with people in industry, as well as to disseminate my research and the categorical informatics philosophy in various ways.

In this section I will briefly mention some collaborations and outreach that have begun or continued this past year. There have also been some transitions of my research, which has enabled me to hire three postdocs.

2.1. **Collaborations.** As mentioned, my work is oriented toward applications. While connections between category theory and quantum physics is well-known,[1] my focus tends toward finding a rigorous formulation of ordinary (meaningful) information and communication. This work has sparked the interest of several pharmaceutical and biotech companies, as well as NIST.

This year I spent two weeks working with Peter Gates from Jansson (also known as J&J), who has been inspirational to my work, and who has been equally grateful for a mathematical formulation of much of what he has observed in his 30 years of experience in informatics.

I have also collaborated with Sub Eswaran and Al Jones at NIST. Our work has garnered enough excitement around NIST that they have hired an NRC postdoc, Spencer Breiner, who will help them bring category theory to NIST applications. Both NIST and Jansson are looking to ameliorate schema matching problems, in which different databases need to be unified or simply to exchange information.

Amgen's informatics team invited me to explain my work on the self-similar nature of wiring diagrams, e.g., of open dynamical systems, which they use to understand *compartmental models* of the body. I've been working with this team for several years now. A new collaboration along similar lines has begun with Eric Neumann at Foundation Medicine, although nothing concrete has emerged.

2.2. **Outreach.** My book, *Category Theory for the Sciences* was published this year by MIT Press. This book disseminates a good amount of my research, as well as promotes the categorical informatics idea. A free version is available through MIT Press.

I was invited to contribute a book chapter on mathematical modeling to a book called *Categories for the Working Philosopher*, edited by Elaine Landry. The article can be found at `http://arxiv.org/abs/1409.6067`.

Another interesting development was being invited by Piet Hut at the Institute for Advanced Study to give an interdisciplinary talk to the institute. He also invited me to Tokyo, Japan for a conference on Modeling the Origins of Life. I consider this outreach, because it is only tangential to my research, but I did begin a collaboration with systems biologist Aviv Bergman there.

2.3. **Transitions.** The research sponsored by ONR has led to several transitions of various sorts. My work on wiring diagrams and communication led to a new grant through the AFOSR, called *Categorical approach to agent interaction*, which began in December 2013.

The dissemination of my work, e.g., with the book, allowed Kevin Schweiker from Honeywell to cold-call me and suggest a collaboration. Together, we submitted a proposal for a NASA grant, called *Category-theoretic Approaches for the Analysis of Distributed Systems*, which is set to begin this month.

A talk I gave this year at Oracle also has generated interest. They may fund some work on their Knowledge Intensive Database System.

2.4. **Recent hires.** The new grants I received this year allowed me to hire several postdocs. In February I hired Ryan Wisnesky, who recently graduated with a PhD from the Computer Science department at Harvard University. He's been implementing my database ideas in an open source application. Ryan will likely start a company in summer 2015, based on these ideas.

In July I hired two more postdocs. The first, Marco Perez, graduated with a PhD from the Math department at UQÁM, advised by André Joyal (the renowned category theorist). The second, Patrick Schultz, graduated with a PhD from the Math department at the University of Oregon.

[1] See work by Bob Coeck, John Baez, and Samson Abramsky.

3. Advances in the functorial query language (FQL)

3.1. **Main problem and its importance.** The purpose of a category-theoretic approach to information management is two-fold: first, to bring categorical machinery to bear on existing problems in information management, and second, to suggest new possibilities for information management that may not be apparent with conventional approaches.

3.2. **Main contributions and their importance.** In many ways our work is a continuation of an existing line of work by Rosebrugh et all from ∼10 years ago. That line of work, while promising, was unable to overcome two challenges. First, because category theory characterizes objects up to isomorphism, it is mathematically non-trivial to use category theory to describe objects such as databases which must be characterized up to equality. We have developed several solutions to this problem (the "attribute problem") and are currently evaluating them against each other. Second, to interoperate with existing databases requires a way to represent existing databases using category theory and vice versa.

We have developed a query language, FQL, proved that it can express the negation-free fragment of the relational algebra, and proved that much of FQL can be implemented using the negation-free fragment of the relational algebra extended with a unique ID generator. Moreover, we have found an equivalence between categorical "lifting problems" and traditional database constraints known as "embedded dependencies" (EDs).

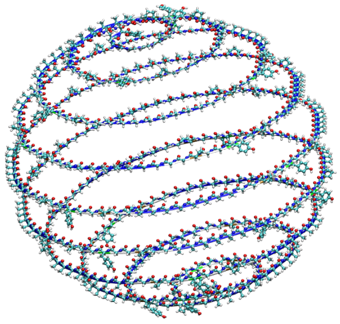
There are additional features of FQL and the categorical semantics of databases that are new and important. For example, the classical treatment of nulls is fairly ad-hoc; we get it for free (i.e., without making an arbitrary choice) from the category theory. We also have a more principled approach to data migration. using this formulation, the usually distinct notions of data migration, query, and update are all subsumed in the categorical notion of polynomial functors.

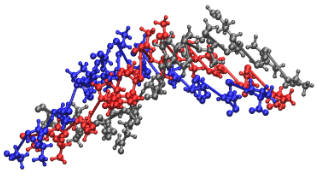
3.3. **Limitations of existing approaches.** The continued expansion of so-called "no-sql" systems over the past 15 years suggests that conventional approaches to information management are no longer sufficient at least for some types of problems (e.g., information integration), and category theory provides a mathematically rigorous way to pose and solve some of these problems. For example, current approaches to migrating data from one schema to another use sets of logical formulas (EDs) to express constraints that are required to hold between the input and output data. These formulas are low-level and much research over the past decade has been dedicated to finding higher-level abstractions for specifying data migrations. Category theory suggests a particularly powerful and concise way to specify integrations: as universal objects. For example, an output instance could be characterized as a push out or pullback of another instance. Existing relational tools lack this expressive power and initial tests with J&J and NIST suggests diagrams are useful in practice.

3.4. **Plan.** There are several technical problems that remain open. One is to find good solution to the attribute problem. Another is to use what we have learned about lifting problems and constraints to show that the traditional and widely used algorithm known as "the chase" is an instance of a categorical "small object argument". Third we need to validate our experimental software by performing an end-to-end data migration of a real-world dataset (obtained from an industrial partner such as J&J or NIST).

There are also theoretical advances which need to be made. One is to get a solid understanding of aggregation functions, such as sums and averages. These are not straightforward, but we are exploring some very interesting possibilities. Another is the "view update" problem. We hope to tackle this in the next few years.

4. Operadic approach to materials design





4.1. **Main problem and its importance.** Despite the increasing interest to use hierarchically structured protein materials as low-cost, environmentally benign, and yet high-performance materials, the interplay between nano-scale structure and global properties is not well understood. Materials Engineers need tools to explore the design-space for sequence, structure, and functionality of organic materials. In particular, these tools should allow for the creation of arbitrary *material architectures*, i.e., assemblies of basic building blocks into complex hierarchical structure.

4.2. **Main contributions and their importance.** A graduate student in Mechanical Engineering at MIT (Tristan Giesa), a freshman at Harvard University (Ravi Jagadeeson), and I developed a Python library, called *Matriarch* that allows the creation of complex material architectures. Our program is based on the category-theoretic notion of operads.

The idea is that material building blocks can be put together using various building instructions, and the result will be a new building block. Thus our operad $\mathcal{O}$ has building blocks as objects and building instructions as morphisms. Instructions include *attach*, which attaches two building blocks together (and forms the appropriate bonds); *twist*, which re-shapes a given building block so that it follows an arbitrary curve in $\mathbb{R}^3$; and *overlay*, which overlays different building blocks in the same region of space.

Our program outputs a Protein Data Bank (PDB) file for any given material architecture. This corresponds to an algebra $P\colon \mathcal{O} \to \mathbf{Set}$. The pictures at the top of Section 4 are (molecular dynamics) representations of the PDB outputs of Matriarch. In the first, we simply *twist* a chain, which is made using a series of *attach* commands applied to amino acids, into a spherical spiral. In the second, we *overlay* three alpha helices (obtained using twists of chains) into a collagen triple helix.

4.3. **Limitations of existing approaches.** The Matriarch software certainly shares features of other molecular design software packages (such as NanoEngineer, Material Studio, Charmm, and Sybil). The main advantage of Matriarch is the high complexity of possible conformations that can be created, from triple helices to multiple layers of strands, any geometric shape obtained by overlaying curves is possible. Furthermore, because Matriarch is an open source and freely available python library, it is possible to directly integrate the material architecture creation into the runtime of molecular dynamics simulations (such as LAMMPS).

4.4. **Plan.** We plan to publish our work in a materials science journal. We also will write up a User's Guide for the program, which we will release online. Finally, we will write a mathematics supplement, which will provide all the details of the operad $\mathcal{O}$ and its algebra $P\colon \mathcal{O} \to \mathbf{Set}$.

5. Other category-theoretic advances

In this brief section, I will discuss other advances that came out of my ONR research this year.

I wrote a paper with Jason Gross and Adam Chlipala, in which we present a fairly comprehensive Coq (proof assistant) library for doing category theory, especially tuned toward categorical databases. I also wrote a paper with Henrik Forssell and Håkon Gylterud, which gives a type-theoretic account of some of my early work sponsored by ONR, namely simplicial databases.

Finally Dylan Rupel and I discovered a strong connection between traced symmetric monoidal categories (TSMCs), which are often taken as the category theoretic model of processes with feedback, and our operad of wiring diagrams. We believe that there is a tight relationship between enriched TSMC's and lax functors out of the category of 1-cobordisms. Later, joined by Patrick Schultz, we found a substantial generalization of this fact. This is a result in pure category theory, but it stemmed from our work on wiring diagrams.

In passing I should note that I have done more work on wiring diagrams in the past year (see my ONR Progress Report from 2013), including giving a categorical account of how to interconnect dynamical systems. However, I am not detailing this work here, because it has transitioned to another grant (AFOSR), as mentioned in Section 2.

6. Papers and presentations

Below I will list some publications and presentations with which I have been involved, in connection with the ONR grant.

6.1. Books and book chapters.

- Spivak, D.I. (2014) *Category Theory for the Sciences.* Cambridge: MIT Press. 496 pages.
- Spivak, D.I. (2014) "Categories as mathematical models". To appear as a chapter in *Categories for the Working Philosopher.* Available online http://arxiv.org/abs/1409.6067

6.2. Papers.

- Gross, J.; Chlipala, A.; Spivak, D.I. (2014) "Experience Implementing a Performant Category theory Library in Coq". *5th conference on interactive theorem proving (ITP'14).* ePrint available: http://arxiv.org/abs/1401.7694
- Spivak, D.I.. (2014) "Database queries and constraints via lifting problems." *Mathematical structures in computer science.* ePrint available: http://arxiv.org/abs/1202.2591
- Vagner, D.; Spivak, D.I.; Lerman, E. (2014) "Algebras of Open Dynamical Systems on the Operad of Wiring Diagrams". Available online http://arxiv.org/abs/1408.1598
- Forssell, H.; Gylterud, H.K.; Spivak, D.I. (2014) "Type theoretical databases". Available online http://arxiv.org/abs/1406.6268
- Spivak, D.I.; Wisnesky, R. (2013) "A Functorial Query Language". *Data-Centric Programming workshop (DCP2014).* Available online: http://research.microsoft.com/en-us/events/dcp2014/wisnesky.pdf

6.3. Invited Presentations.

This year I spoke in the following seminars:

MIT (Programming languages seminar) 2014/04/15;
IAS (Bar talk) 2014/03/20;
PARC 2014/03/03;
Amgen 2014/03/04;
Oracle 2014/02/28;
UIUC (Topology seminar) 2014/02/25;
Harvard (PL seminar) 2014/02/19;
Carnegie Mellon U. (POP seminar) 2014/01/23.