Final progress report for ONR grant N000140910466

David I. Spivak

August 24, 2010

This report will give a broad overview of the progress I made toward fulfilling the goals of ONR grant N000140910466 ("Geometric Networks: a higher-dimensional approach to networks and databases") during its operative period, the 18 months between January 2009 and July 2010. As this was only half the tenure of the grant (I left the University of Oregon for MIT where a new incarnation of the grant is in operation, N000141010841), I was not able to achieve all its goals during the operative period. However, I made a good start on these goals, as I will detail below.

1   Goals of the project

The overarching goal of this project was to find a mathematical model for communication – communication between people, computers, businesses, or neurons in the brain. Given a network of individuals, each with their own way of storing and processing information, how might one model the ability to transfer information between them?

Within this larger goal of modeling information and communication, I aimed to find a mathematical formulation of networks, of databases, and of learning. These are related to the above idea as follows. The interactions I hoped to study take place on a network. Each node (person, computer, neuron) in that network represents an individual, and each individual has a database of "stored knowledge." Nodes speak to each other across lines of communication and try to access and update each others' databases; i.e. learn from and teach each other. Through this process each node learns something, as does the network as a whole. In what ways can all this be made precise?

To find a precise way to speak about information, one must find a precise and tangible setting in which information is surely found. I began with databases – it is clear to anyone looking at a database that there is information in it. Moreover, databases are completely tangible (from a mathematical point of view) in that they are man-made – nothing in their structure is hidden. So I began by understanding the dynamics of databases and related systems: how can a database be precisely described, and how can data transfer from one database to another be precisely described?

As mentioned above, each database represents the "internal state" or "worldview" of some individual in a network. I was interested to understand how the nodes might interact and learn from one another, and how the network itself might change as a result.

In summary, the goals of this project were:

- to find a mathematical model for information and communication, grounded in the reality of databases;

- to find a mathematical model of networks so that communication in a network takes the form of data migration in databases; and

- to look into how a network could "learn" from these interactions; i.e. to make mathematical statements about how communication in a network changes the network or the nodes in it.

2   Modeling networks

Currently, networks are usually modeled as graphs: a bunch of nodes and a bunch of arrows connecting them – the nodes correspond to individuals and the arrows correspond to lines of communication. But in the real world not every communication is sent from one person to one other person. Consider the difference between, on the one hand, three people talking all-together at a table, and on the other hand, the same three people whispering to each other or talking on telephones. The dynamics in a 3-way conversation are very different than the dynamics in three 2-way conversations.

I thus proposed a new mathematical formulation of networks that takes into account the possibility of many individuals engaged in a single conversation. This formulation uses geometric objects called simplicial sets, each of which can be drawn as a union of edges, triangles, tetrahedra, and so on. For example three people talking together would be modeled by a solid triangle, whereas three people whispering to one another would be modeled by just the edges in the perimeter of that triangle – three 2-way conversations.

The network of all human interactions would be a very large, constantly-changing shape, and to what extent we could get a handle on that shape is unknown. But still, we could make guesses about it and begin to experiment with how it changes in time and how it reflects realities about human relationships. The network of all neuronal connections is similarly difficult, but again perhaps something worthwhile could be said within this new model that is invisible in the classical, graph-theoretic model.

Also related to networks, I helped to organize a conference called *Agent-based complex systems* (https://www.ipam.ucla.edu/schedule.aspx?pc=onr2009) at the Institute for Pure and Applied Mathematics (at UCLA), as well as giving a talk there (entitled "Metric realization of fuzzy simplicial sets"). In this talk I hoped to show how one could model not just which nodes in a network (simplicial set)

are in communication, but how strong or weak (how fuzzy) their communication is. A distinction of this sort is surely part of any realistic model of communication.

3   Modeling databases

A database is a system of tables connected together in well-defined ways. It turns out that a database can be modeled by a simplicial set, just like networks can. This coincidence was surprising when I noticed it, but it suggests that information systems have a certain "fractal" nature, which was welcome. I developed a theory of *simplicial databases* which are more dynamic than normal databases, as well as having a straightforward mathematical description.

Each table in a database has some number of columns. If it has one column, imagine it as a point; if it has two, imagine it as a line segment; if it has three, imagine it as a solid triangle; if it has four, imagine it as a solid 4-pointed pyramid; and so on. In short, each table looks like a triangular tile, and these tiles can be glued together if some columns from one table match with some columns of another table. This is how we get a simplicial set – a union of tiles. Each tile, representing a table in a database, contains a set of facts.

We can imagine ourselves going to the computer with a question in mind. We think of what we want to know and what we already know; we ask: how can we get from here to there? So we go to the computer, and through a bit of investigation we begin to find tiles that represent facts which can be put together to obtain the answer. We attach these tiles in the appropriate way, draw a curve through the shape from the "what we know" section to the "what we want to know" section, and voila! The computer supplies the answer. Amazingly, this possibility naturally falls out of the rigorous category theory which defines simplicial databases. Although reminiscent of movies like "Minority Report," the ability to build detailed queries from a collection of relevant information-tiles is provably possible.

My paper on simplicial databases http://math.mit.edu/ dspivak/cs/SD.pdf was noticed by various researchers, and after a few months I was contacted by Dr. David Balaban, Vice President for Informatics at Amgen Inc. He and some members of his team, as well as Dr. Alan Brown from Microsoft Research, met with me in San Francisco (February 2010) to discuss my work. The meeting was successful in that they came away with a good understanding of the ideas and presented me with some concrete questions to work on. We continue to work together to this day.

4   Modeling learning

Given a network of communicating entities, I had hoped to understand how this network as a whole could process information. Unfortunately, I did not have time to really look into this question. I did however consider how two

individuals in a network may learn from one another. Myself and Dr. Mathieu Anel in Montreal, QC collaborated to develop a "communication protocol" for entities with different worldviews.

The basic idea is that two people each have their own worldview (database) and share a "common ground" worldview which may be incomplete. When one node speaks to another, the speaking node uses ideas from the common ground to develop new ideas within the listening node. The listening node parses the incoming information and does three things:

- it interprets some of the new ideas as already part of its worldview and updates the common ground worldview accordingly;

- it learns from some of the new ideas and updates its worldview as well as the common ground worldview accordingly; and

- it rejects some of the information as either incompatible with its worldview or as simply non-sensical.

This parsing is reported to the speaker node, although it does not have access to how the information is interpreted or learned. The two may then reverse roles, or perhaps the speaker will continue to speak.

The point is that all of this has been done within the rigors of category theory. For example, if all the " worldviews" above really are databases, then the protocol will result in actual updates of these databases. A paper for this communication protocol is still in preparation.

5   Conclusions

Aside from the above tangible outcomes of this grant, I made a lot of progress in simply understanding what needs to be done. I spent time learning about formal grammars, ontologies, programming languages, topos theory, informatics, basic neuroscience, and much more. I also ran two seminars, one called "Mathematical methods in computer science" in the computer science department at the University of Oregon, and one called "Informatics" in the mathematics department there. These enabled me to see things from new angles and hence strengthen my intuition about the subject.

I am now (Summer 2010) a postdoctoral associate at MIT, and I have a new grant from the ONR entitled "Categorical information theory." I am confident that the work I did under the Oregon grant will enable me to make fast progress these next three years.