Technical Proposal

BAA number: #09-001

Title of Proposal: Geometric Networks: A higher-dimensional approach to networks and databases.

Identity of prime Offeror: David Spivak

Technical contact:
David Spivak
317 Fenton Hall
University of Oregon
Eugene, OR 94703
dspivak@uoregon.edu
541-346-5629

Administrative/business contact:
Chingling Reed
218 Fenton Hall
University of Oregon
Eugene, OR 97403

Duration of Effort: 3 years: 2009/01/01 to 2011/12/31.

Contents

Part I
Statement of Work

1 Introduction

The importance of networks in the modern era cannot be overstated. The goal of this project is to study networks and related topics, such as databases, information transfer, and learning.

Roughly, a network is a group of individual objects, called nodes, which are connected with other nodes in various ways, and which are able to communicate through these connections. The organization of a network can change in time, and often this change is dictated by the information that flows through the network.

We aim to study networks by constructing a suitable mathematical model. After some preliminary research, it has become clear that a category-theoretic construction of databases is needed before we can give a rigorous definition for networks. We already have produced such a construction, and will continue to refine it, as needed. Once networks have been defined, we will attempt to interpret learning within our model; that is, we will examine how information is transferred in a network, and how that information can lead to changes in the shape and structure of the network.

We believe that the models that are currently used to study networks and databases are inadequate for many reasons. Most notably, they utilize graphs, which are 1-dimensional objects, when in fact higher-order relationships can only be studied by higher-dimensional graphs, called simplicial sets. In the following sections, we discuss networks, databases, and finally learning. In each section we begin by conveying the inadequacies of the current models. We then give a proposed solution to these problems, including specifics where available. Finally, we lay out some questions and directions for future research. In the last section of the Statement of Work, we summarize by stating the scope, objectives, and technical approach of the proposal.

2 Networks

Networks are the main object of study in the proposed project. Our goal is to find a model that can be used to study a wide variety of networks, including social networks of individuals, economic networks, and the network of neurons that constitutes the brain. In each of these examples, the network is a construct in which individual nodes are connected and use these connections to transfer information. In fact, we propose that the primary function of a network is to process information, i.e. to learn.

In our study of networks, we will attempt to understand their structure, how information is transferred via that structure, and how this information transfer leads to learning. Examples of networks, such as those listed above, are important for deciding which mathematical structures are more appropriate and which are less appropriate for modeling networks.

See a previously written grant proposal about networks, written for a lay audience, at

`http://www.uoregon.edu/~dspivak/lokey.pdf`

2.1 Inadequacies of the current model

Networks are currently modeled as graphs, where the edges represent lines of communication. Sometimes the edges of these graphs are annotated with numbers, which represent the speed of the connection. To explain why this approach is inadequate, we will show that it is not appropriate for modeling information transfer.

The first inadequacy comes from the fact that graphs are 1-dimensional objects. We propose that the correct model is a higher-dimensional object called a simplicial set. Simplicial sets have not only vertices and edges (called 0-simplices and 1-simplices), but also filled-in triangles (called 2-simplices), filled in tetrahedra (3-simplices) and so on. A higher-dimensional simplex can be used to represent a group of nodes that is capable of interacting as a unit.

For example, consider the case of a dinner party. When a person speaks, he often speaks to the whole group in a way that he believes the whole group will understand. The communication is delivered to the group as a whole. Now consider the case in which the same people are present, and each person can speak to every other person in the room, but this time only whispering is allowed. That is, each time a person speaks, only one other person can hear the message. The difference between these two situations is tremendous, both socially and information-theoretically. One should view it as analogous to the difference between a filled-in simplex and its underlying 1-dimensional graph. It is clear that the 1-dimensional graph model is a gross reduction of the higher-dimensional model.

The second inadequacy of the current model is that it does not address the fact that different parts of a network communicate in different ways. Each node has a language or vocabulary that it understands, as well as a repository of information represented in that language. For example, each person has her own way of speaking and her own memory of experiences and knowledge. Often overlooked is the fact that higher-dimensional simplices are equipped with the same structure. That is, each simplex should come equipped with a vocabulary that it "understands" in some sense, as well as some information in that language. Let us expand on this idea for a moment.

We are suggesting a model of networks in which nodes, edges, and higher simplices each come equipped with a database of information. The database attached to an edge of the simplicial set corresponds to the language and the information which is common to and assumed by the two vertices of the edge. Similarly, for a higher simplex, which represents a group of nodes acting as a unit, there is a database of shared language and information used by that unit.

This is most obvious in the case of social networks. For example, when siblings converse together, although ostensibly speaking in English, it may be the case that other English speakers can not understand what is being said, both because the language is specialized and because the source of shared information is not accessible to outsiders. Similarly, when mathematicians converse together, they convey information in their own language with their own knowledge base. The same goes for myriad other social cliques. When a member of such a clique speaks with someone outside the clique, they use different language and refer to a different (possibly shallower) set of shared data.

The current model of networks does not address this issue at all. Annotating each edge with a number which refers to how quickly information is transferred along that edge may be seen as a primitive approach to this issue. Our point, however, is that speed of transfer is almost irrelevant when compared to language and shared knowledge. A person can convey certain information to his sister thousands of times faster than he could convey it to a stranger, not because he is speaking faster, but because he is speaking in a more meaningful language with access to a shared past. Current models convey the quantity of information which can be transferred along a channel, whereas we believe that the quality (shared language, etc.) of the channel is much more relevant.

Finally, the definition of network should be fractal and hierarchical. One often has a network of networks of networks. In the human brain, for example, research has shown that higher order structures query lower order structures, looking for specific information (see Hawkins, J. On Intelligence.) The inner workings of these lower-order structures are not accessible by the higher order structures. For example, the higher visual cortices may be interested in a particular feature in the visual field and direct lower cortices to ascertain data about that feature. Mathematically, we want a way to define a network whose nodes are lower-order networks of a given type, and whose higher simplices constitute certain correspondences between these lower networks. While often discussed in science, such considerations have not been approached mathematically (as far as we know).

2.2 Proposed model

We propose to model networks using simplicial sets. The vertices represent individuals, and the higher-dimensional simplices represent communication channels. The topology of a simplicial set is quite different from the topology of its underlying graph; for example, the higher homotopy and homology groups vanish for graphs, whereas they do not vanish for arbitrary simplicial sets. We propose to study how the topology and geometry of a network influences the way in which information is transferred.

Under our model, the simplicial set X corresponding to a network will be endowed with a sheaf $\mathcal{O}_X$ that represents the language and available information held by each simplex in the network. Although we have not worked out the specifics yet, the idea is that each node has its own information repository, as does each edge and each higher simplex. These information repositories are in fact databases, which is why databases are important for studying networks (see Section **??**).

We present the following definition to give a flavor of what we have in mind for defining networks; we are not intending to imply that it is the best model, only a prototype. The reader with a background in algebraic geometry will notice a strong similarity to the definition of ringed space. The similarity comes from the fact that networks, like ringed spaces, are spaces equipped with extra structure.

Definition 2.1 *Let $\mathcal{D}$ denote a category of databases. A* network on $\mathcal{D}$ *is a pair $(X, \mathcal{O}_X)$, where X is a simplicial set and $\mathcal{O}_X\colon (\Delta \downarrow X)^{op} \to \mathcal{D}$ is a presheaf of databases on the category of simplices of X. A morphism of networks $(X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ consists of a pair $(f, f^\sharp)$, where $f\colon X \to Y$ is a morphism of simplicial sets and $f^\sharp\colon f^*\mathcal{O}_Y \to \mathcal{O}_X$ is a morphism of sheaves on X.*

2.3 Questions and future research

The model we have presented is far from being in its final form. For example, the sheaf $\mathcal{O}_X$ may not be a sheaf of databases but a sheaf in some related category (for example the category whose objects are databases but whose morphisms are retractions of databases). It is also unclear how information is transferred from a node to a group of nodes in models of this sort. Different models for the theory of networks will use possibly different sheaf-theoretic constructions (direct images, inverse images, extensions by zero, etc.) to transport data from the database over one node to the databases over other nodes. To find the formulation which actually models information transfer in a network will require the consideration of many examples.

Another question relates to how queries are processed up and down a hierarchical network. What is the protocol of interaction between a higher network and its subordinates?

Finally, there are questions of "distance." In the model we proposed, there is no notion of distance between nodes – either they are connected by a simplex or they are not. However, in social networks and the neural network of the brain, some connections are stronger than others. Humans are "more connected" to people they trust, certain internet connections are faster than others, and connections between neurons are strengthened by use.

There may be a variety of good ways to encode distance in our definition of networks, though perhaps what we have is already enough. That is, perhaps what is often perceived as proximity or distance between nodes is in fact just a function of the shared language and information between those nodes.

One other idea we have in mind, however, is to encode the strength of connections using fuzzy simplicial sets. The topos of fuzzy objects in a category has been defined (see Barr, Michael. "Fuzzy set theory and topos theory." Canad. Math. Bull. 29 (1986), no. 4, 501–508.) We are currently working on adapting this to the category of simplicial sets and perhaps finding a "metric realization functor" which can functorially convert a fuzzy simplicial set into a metric space (in a way that is compatible with the usual geometric realization functor on simplicial sets).

We will have more questions about networks in Section **??**.

3 Databases

The modern world would be impossible without databases. Thus, it is imperative that we have a good model with which to understand them. When databases get large, querying them becomes increasingly costly, and finding the correct organization of the database becomes increasingly important. Database integration and lossless joins are examples of issues which plague the models which are currently in use.

In this section we will explain the inadequacies of the relational model of databases, propose a solution (which is currently being prepared for publication), and then discuss areas of future research. We refer the reader to our survey paper about this topic, called "Geometric Databases," at

`http://www.uoregon.edu/~dspivak/atmcs-gd.pdf`

and to a slide presentation which is more rudimentary, but which provides a few more examples (including an example of a lossless join) at

`http://www.uoregon.edu/~dspivak/colloquium.pdf`

3.1 Inadequacies of the relational model

The relational model is the most commonly used model for databases, and it is founded on the theory of mathematical logic, more specifically on the theory of relations. A relation on types A, B, and C is a subset $R \subset A \times B \times C$. A database is usually defined as a set of relations, though this is clearly not an adequate definition because it does not mention the ways in which various relations are related to one another. The usual formulation leaves open the question: "what is a morphism of databases." That is, what is the fundamental structure that should be preserved when we transform one database into another. Finding the correct definition for a morphism of databases is the first step in approaching problems of database integration.

A major problem in the theory of relational databases is that of achieving lossless joins. For example, suppose one wants to join the following two tables:

Title	LastName
Dr.	Marx
Mr.	Marx

FirstName	LastName
Karl	Marx
Groucho	Marx

The outcome will be the following table:

Title	FirstName	LastName
Dr.	Karl	Marx
Dr.	Groucho	Marx
Mr.	Karl	Marx
Mr.	Groucho	Marx

This table has four entries, two of which are “accurate,” in that they describe real entities, and two of which are not. This occurs because the database cannot distinguish between the two instances of the last name Marx. Although it is difficult to make it clear in this short SOW, this reflects a deficiency of the relational model (see the sources listed above).

Relations tend to confuse entities with their attributes: if two entities have the same attributes then, to the relation, they are the same. The inflexibility which causes this problem is that relations are subsets such as $R \subset A \times B \times C$, which we can view as injective functions $R \hookrightarrow A \times B \times C$. We will propose in Section **??** that much is gained by allowing all functions, not just injective ones. Two different entities (elements of the domain set R) may have the same attributes (images in $A \times B \times C$), if we allow non-injective functions. This eliminates the confusion between an entity and its attributes. A (possibly non-injective) function, e.g. $f\colon R \to A \times B \times C$, is called a *table* rather than a relation.

In fact, there is a split along these lines between relational databases in theory and relational databases in practice. In practice, people use SQL to work with databases. Although ostensibly implementing the relational model, in fact SQL uses tables instead of honest relations. That is, SQL allows multiple rows with the same data, a situation which does not at all fit in with the purely relational model.

Another split between theory and practice comes from the use of nulls. Nulls, or locations in the table for which the desired data is not known, are used in SQL but are totally illegal in relational database theory. The split between theory and practice is a good sign that there is something wrong with the theory.

One of the big problems is to find good ways of organizing schemata. People use entity-relationship diagrams to map out the structure of a database. These are ornate pictures, possibly with ovals and squares, crows feet, etc. It would be hard to pin down exactly what an ER-diagram is, and thus hard to know when one can transform a database whose schema is based on ER-diagram A into a database whose schema is based on ER-diagram B. Without knowing the category of which these diagrams are objects, this process will always be difficult and ad-hoc.

One problem from our vantage point is that, as for networks, databases are often modeled 1-dimensionally, with entities as nodes and relationships as edges. Unfortunately, higher-order relationships do not find a place in this model. The dynamics of a single interaction of three entities are inherently different than the dynamics of three interactions between pairs of entities, and this difference cannot be captured in the 1-dimensional approach.

3.2 Proposed model

We proposed a model for databases in the survey paper mentioned above, entitled "Geometric databases." In that paper, we first relax the relation condition, thus allowing lossless joins and duplication of records. Nulls are also appear naturally in this theory. Second, we base the model on a foundation of simplicial sets, which serve as schema. These higher-dimensional graphs can be viewed as something like ER-diagrams. The morphisms between schema are easy to define. We do not reproduce the definitions here, because they take some space to set up, but we will try to make the idea more clear presently.

We view a database as a "bundle of data" over a schema. The schema is a type of space (i.e. a simplicial set), and we imagine that the data is "sitting over" that space as sections of the data bundle. For example, one can imagine space-time as 4-dimensional space, together with a projection down to the time line. This is a rank 3 bundle $p\colon \mathbb{R}^4 \to \mathbb{R}$. A global section $s\colon \mathbb{R} \to \mathbb{R}^4$ of this bundle is a choice, for every $t \in \mathbb{R}$, of a point $s_t \in \mathbb{R}^3$. A local section is a section over some open set $(a, b) \subset \mathbb{R}$, which amounts to a point in space for every time $t \in (a, b)$.

To relate this example back to databases, imagine we are tracking the location of objects in space. For every time interval (a, b), there are a certain number of these objects which exist during that interval. The "tracks" of these objects in 3-dimensional space are precisely sections of the bundle. This is a database whose schema is the time line, and whose data over an interval is the collection of tracks in $\mathbb{R}^3$, or sections of the data bundle.

In reality, we would only be able to store a finite number of locations for our objects. Thus, our schema would not be the real line, but instead perhaps the union of 1-simplices $A_t = \Delta^1$, for $t \in \mathbb{Z}$, where the left endpoint of A_{t+1} is identified with the right endpoint of A_t. Sections of the data sheaf over a vertex in the schema would be the locations of our objects at that time, and sections over the 1-simplices would allow us to connect the location of each object at time t to that object at time $t + 1$. If an object were "destroyed" or "created" during that time, there would be no section over the associated 1-simplex, but there would be a section over the first or second endpoint of the simplex (respectively).

In general, we may have schema with much more interesting shapes than simply long intervals. As anyone who has seen an ER-diagram knows, there can be many interrelationships between entities. Further, the simplicial model allows relationships of a significantly more general nature because of the existence of higher simplices to model higher-order relationships. Homological or other topological methods may be used to find interesting aspects of the data.

3.3 Questions and future research

Since the first incarnation of the category of geometric databases is already complete, our first goal is to test it out in practice. For this, we will need to work with a computer scientist to code up the ideas. We then need to test the model out in practice and see if anything needs to be modified.

There are many theoretical questions remaining as well. First, instead of a *set* of data types (as we used for the theory of geometric databases), it may be useful to consider a *category* of data types. This would be a monoidal category equipped with a monoidal functor to **Sets** (corresponding to the fact that if A, B, and C are data types, then $A \times B \times C$ is a data type). Any monoidal category has an underlying simplicial set of objects. We are interested in generalizing the work on geometric databases by adding the extra functionality afforded by working directly with a category of data types instead of its underlying simplicial set.

Query optimization is a major issue in the theory of databases. As suggested by Gunnar Carlsson, the shape of the schema is flexible in our model, and this may be useful for finding solutions to query optimization problems. That is, a certain schema shape may be more useful at one time than at another. This is reminiscent of "DNA folding" in biology, in which the DNA (a kind of data structure) folds up in different ways at various stages of development. In so doing, data that is more relevant to the task at hand is made more readily available. The structure and shape of the schema in a geometric database can be changed as well, depending on the current use.

4 Learning

Admittedly, we are not well informed about the current research in learning, as this is a vast and rapidly growing subject. In this section we simply present some ideas for how our theory of networks may be of use in modeling learning.

When a phenomenon occurs, either from within the network or from without, various nodes may perceive it and respond by communicating with their neighbors about what has occurred. When nodes agree on what is occurring, this agreement fosters "friendship," i.e. a stronger connection. Neuroscientists paraphrase this idea by the mantra "cells that fire together, wire together."

This idea fits very well within our theory of networks (see Section **??**). Each node of a network has its own database from which it can communicate to neighbors. The connections themselves are annotated with databases as well, which consists of words which all participating nodes *ostensibly* understand. However, there is no a priori reason to believe that one node's understanding of a common word is the same as another node's understanding of that word. That is, although there is a morphism from the database over a simplex in the network to the database over each subsimplex, there is no reason to believe that this morphism is sensible. The network is always trying to improve inter-nodal communication; in some sense this is the whole purpose of the given network.

Phenomena are precisely what make these improvements possible. As mentioned above, when a phenomenon occurs, nodes communicate their "impressions" of that phenomenon to neighboring nodes. When it is known that these communications are all referring to the same event, the communication becomes grounded by that event. This provides the network with a way of connecting together the languages used by different nodes. In other words, if two people respond to a phenomenon in similar ways, then the two people gain a new piece of shared experience and a new assurance that there is agreement in their languages. The bond between them is strengthened, as mathematically the database over their common edge is augmented. Similarly, when two nodes disagree about a phenomenon, it calls into question the agreements they have, and may diminish the database of shared language and experience between them.

Cultures, subcultures, and individuals all have databases of shared information and language. The given social unit (considered here as a network) strives to improve the quality of its connections by aiming towards a more coherent use of language among nodes as well as a larger body of shared knowledge and experience.

To reiterate, we propose that each phenomenon leads to learning in a network by providing an event about which nodes can communicate and agree or disagree. The structural changes may involve forming or severing connections (introduction or removal of simplices), or more commonly a change in the sheaf of databases lying over the network. Unexpected agreements or disagreements among the messages expressed by distinct subsimplices of a given simplex cause changes in the databases over that connecting simplex.

4.1 Issues and questions

We certainly have more questions than answers at this point in time. A good first step in our study of learning, networks, and communication, will be for us to visit scholars who do research these areas (such as Jeff Hawkins in northern California, Mark C. Taylor at Columbia, and various linguists at MIT). We now present a few questions which we propose to study.

Since the set of occurring phenomena is uncountably infinite whereas the set of expressible sentences is finite, communicating requires a drastic slicing of reality into pieces. The central problem in communication seems to be determining sameness versus distinction. The network constantly evolves to find more agreement on the classification of phenomena into communicable chunks, and boundary issues are one big difficulty. We imagine the use of "partitions of unity," which are (well-known mathematical) ways of smoothly partitioning a space into pieces, without cutting it. With a partition of unity, an object would not be exclusively named either A or B, but instead some combination of the two. This idea seems promising and must be explored.

The other difficulty in learning comes from unsound data in the network. It is somehow possible for nodes to misinterpret or misclassify phenomena. When this occurs, false information enters the network and is communicated between nodes. This is bad for the network, because it leads to miscommunications. It will be interesting to see how false information can be found and eliminated.

The mathematical question which we most need to solve is: what kind of structure is a learning network. Perhaps it is a kind of fiber bundle over the real line, where each fiber is a network, and in which the transition functions on overlaps of open sets are certain endomorphisms of these networks. This vision is far from rigorous and far from tested. This will certainly be a focus of our research in the coming years.

Finally, an emergence of unexpected order in the sheaf on a network would constitute a kind of evolutionary step for the network. These "insights" lead to large-scale organizational changes in the network. It is completely unclear how this works, and it may be outside the scope of what is possible to understand in just a few years of research, but it is a fascinating question.

5 Scope, objectives, and technical approach

To summarize, we propose to study networks, databases, and learning. Most basic among these are databases. We will continue to explore higher-dimensional databases. Our objective is to find a model that is both rigorous and flexible, and which models exactly what humans wish to use databases for, i.e. to classify the phenomena that they observe in a way that will be easily accessible later. We will do this by implementing Geometric databases and exploiting their advantages over the relational model, while working to correct any deficiencies we find.

We will then shift our focus to networks. Our objective is to understand how information is contained, transferred within, and input to and output from a network. A satisfactory theory should be broad enough to encompass neural networks, social networks, economic networks, ecological networks – any network which can be said to process information about its environment.

To understand how networks behave with regards to information, we will be interested in the resulting structural changes to those networks, changes which we call learning. Our objective is to produce a category-theoretic and/or topological model for the types of structural changes that occur in a network, as information is processed.

The technical approach we will take for studying databases is to attempt to implement them. For this we may wish to hire a graduate student in computer science. We will search for capacities in which our formulation is an improvement on the current designs, and capacities in which it can be improved.

To approach networks and learning, we will begin by gathering ideas. For this, we will need to visit experts in various related fields, such as network theory, neuroscience, complexity theory, linguistics, epistemology, and communication theory. All the while, we will be attempting to find a categorical model for networks and communication. Once a satisfactory model has been produced, we may use established mathematical methods to study the behavior of the model.

Part II
Future Naval Relevance

Modern day organizations have a pressing need for good database management systems, and the navy is no exception. It is our view that the entire underlying model must be restructured in order to achieve significant improvements in quality. The category-theoretic approach is not only rigorous, it is also flexible. If models for DBMS are expressed categorically, a user can determine using established mathematical techniques not only the similarities and differences between the distinct models, he or she should be able to convert databases from one model to another using functors. This flexibility to change will be important as the models evolve.

A good theory of networks and learning will also be relevant to the navy for the foreseeable future. Networks are ubiquitous and constitute a new paradigm under which to study many aspects of our modern world. Whether studying computer networks, networks of individuals, or the network that constitutes the navy itself, a good understanding of how networks behave and process information will be very useful.

Part III
Project Schedule and Milestones

We will begin by hiring a graduate student in computer science and attempting to implement geometric databases. We should be able to accomplish this by Summer, 2009. We will continue to consider the questions outlined in Section **??**, especially the use of monoidal categories of data types.

We will then begin an earnest study of networks and learning. We plan to travel extensively at this point, in order to gather information about various kinds of networks and how they behave. After travel, we will work to find a mathematical (topological and category-theoretic) model for what we have learned. We hope to have a rudimentary model of networks on the PI's website by January, 2010. We will then study this model using established mathematical methods. We will also discuss our model with experts in various areas of academia (computer science, neuroscience, economics, sociology, etc.) to determine the extent to which their objects of study fit into our models. If successful, this will foster cross-disciplinary research, which would be a very important milestone and measure of success.

Part IV
Assertion of Data Rights

All material produced under this research, including software, publications, reports, and data, will be public domain unless otherwise explicitly specified to the PI by the ONR.

Part V
Deliverables

The deliverables consist of theorems and algorithms associated with the principal foci of the project. Theoretical results will be reported in conferences, journals, and technical reports that will be made available on the PIs website. All associated software will be distributed as per the data rights approved by ONR.

Part VI
Other Agencies

Not applicable: the proposal has not been submitted to any other agencies.