

UNCLASSIFIED

TECHNICAL REPORT
SUMMARIZING AN APPROACH TO THE
DARPA SIMPLEX PROJECT

Conglomeration of Heterogeneous Content using Local Topology (CHCLT)

American University

Prof. Michael ROBINSON
Mathematics and Statistics
American University
4400 Massachusetts Ave NW
Washington, DC 20016
(202)885-3681
michaelr@american.edu

Dr. Cliff JOSLYN

Chief Scientist

Pacific Northwest National Laboratory
cliff.joslyn@pnnl.gov

Dr. Emilie HOGAN

Scientist

Pacific Northwest National Laboratory
emilie.hogan@pnnl.gov

Mr. Chris CAPRARO
Senior Systems Engineer
SRC, Inc.
ccapraro@srcinc.com

March 11, 2015



UNCLASSIFIED

CONTENTS

1. Executive Summary	2
2. Goals and Impact	4
3. Technical Plan	5
3.1. Elements of a Sheaf-Theoretical Knowledge Representation	6
3.2. Notional use case 1: situational awareness of a public event	7
3.3. Base space construction: relationships between analytics	9
3.4. Stalks: encoding knowledge as algebraic types	11
3.5. Restrictions: applying domain-specific models	13
3.6. Analytical methods	17
3.7. Knowledge updates	22
3.8. Scalability, modularity, and extensibility	23
3.9. Notional use case 2: multi-sensor experiment analysis	24
3.10. Interaction with teams focused on related areas	28
3.11. Technical risks and challenges	28
4. Capabilities	29
References	30

1. EXECUTIVE SUMMARY

American University, led by Prof. Michael Robinson (PI), in collaboration with the Pacific Northwest National Laboratory (PNNL) and SRC, Inc., propose to develop a unified framework for global representation and modeling of hybrid information systems, integrating both quantitative and qualitative information. Our approach uses the mathematical theory of *sheaves*, which is rooted in topology, combinatorics, and category theory. A sheaf is not just a mathematical representational structure that provides a common ground among heterogeneous data sources, it is *the canonical such structure*: any principled specification of the interaction between heterogeneous data sources will provably recapitulate some portion of sheaf theory [17]. The theory of sheaves is as foundational to heterogeneous data integration as function spaces are to the study of differential equations or as model theory is to logic. In contrast with previous approaches that leverage graph and network theory, sheaves allow *simple, functorial interfaces* that support natural multiscale reasoning about their structure.

SIMPLEX aims to automate inference, reasoning, hypothesis generation, and steering for context-aware data analysis and causal inference over heterogeneous data. Our project aims to provide a mathematical foundation for ALL such data fusion problems requiring local to global inference, independent of their source, nature, or data type. Thus we are aiming at the core aspect of heterogeneous integration which has stymied information systems development for years, and which SIMPLEX is taking central aim at, namely the ability to address both traditional quantitative sensors (such as radar and sonar) and semantic qualitative information sources (such as symbolic metadata or the outputs of natural language processing (NLP) analytics feeds) within the same modeling framework. Our approach aims precisely at this point by directly supporting the automatic construction and verification of global models from knowledge of local interactions. In this way, we will provide the theoretical foundation necessary for the accomplishment of the full range of SIMPLEX goals.

To demonstrate our sheaf-theoretical approach, we will explicate two notional use cases:

Situational Awareness of a Public Event: This example has been previously initiated for pedagogical purposes to explicate the way mathematical sheaves can encode complex interactions among heterogeneous information sources, including integrating semantic and quantitative data. Situational awareness of a public event in a US city is imagined through the integration of municipal police sources and public traditional and social media.

Multi-Sensor Experiment Analysis: Representation and integration of radar and image sensor data together with their semantic metadata annotations, specifically:

- (1) Measurements of moving targets from a heterogeneous collection of imaging and tracking sensors previously flown in support of Air Force Research Laboratory (AFRL) research activities; and
- (2) Semantic meta-data on these sources describing the types of targets and their behaviors.

These will be used to demonstrate that sheaves can be used to analyze dissimilar data types (radar, optical, tracking, and semantic) and to support automated inferences about target behavior and intent.

Sheaves have been studied extensively since their discovery in the 1940s, but have not seen substantial application outside of mathematics until recent work by Prof. Robinson and his collaborators. There are many important mathematical questions about sheaves that have only recently been posed, and nearly all remain unanswered. In addition to developing new implementations of sheaf-theoretic algorithms for knowledge representation and analysis, we will address the most relevant of these unanswered questions by developing new mathematical results. Because of the foundational nature of sheaves, answering these questions is critical for *any* work (theoretical or not) that attempts to represent the relationships between arbitrary sources of knowledge.

We propose to develop a knowledge representation pipeline that will ingest structured data, extracted facts from structured documents, and quantitative measurements (such as images or timeseries), encode these data elements and their underlying phenomenological models as a sheaf data structure, and then provide an interface to interrogate this data structure using queries providing mathematical guarantees. This interface will support the following two classes of queries: (1) detection of potential feedback loops (or higher dimensional analogs) that can lead to pitfalls in reasoning, and (2) “retasking” queries that suggest the data types that need to be acquired next to reduce the likelihood of these pitfalls. By working at a sufficient level of abstraction, this retasking query provides the capability to iteratively update the structure of a knowledge network, so that the types and sources of data can be incrementally changed without extensive recomputation. This addresses a major limitation of existing approaches – adding new data types is *ad hoc* and requires substantial manual intervention – but sheaf theory organizes the changes that need to be made and isolates their impact.

Central to the specification of a sheaf is a careful expression of the relations between different data types – this requires the insertion of domain knowledge appropriate to the represented data types. Our team is already domain experts in semantic (PNNL) and radar and imagery (SRC) data feeds. We are therefore poised to develop efficient and correct datafication methods that encode our proposed measurement data feeds into sheaves, and then to analyze the sheaf invariants that result from our processing. Because sheaves are agnostic to the data sources they represent, it is likely that they can generalize existing methods involving Bayesian reasoning, network analysis, and other popular data analysis methods.

Based on Prof. Robinson’s success in sonar and radar signal processing using sheaves [29], we are certain that sheaf theory is the correct platform for combining heterogeneous quantitative data sources. In support of a previous DARPA/I2O seedling, Prof. Robinson, Dr. Cliff Joslyn (PNNL), and Dr. Emilie Hogan (PNNL) successfully used sheaf theory to develop a use case involving a highly heterogeneous set of *semantic* and quantitative data sources[20]. This use case was intended for pedagogy; the proposed effort will load semantic and quantitative data into a sheaf for the first time ever, and at a realistic scale. By teaming with SRC, Inc., we will obtain domain knowledge concerning synthetic aperture radar, trackers, and optical sensors along with

datasets that feature these modalities. Thereby we have developed a program plan that will effectively and efficiently meet the objectives of Phases I, II, and III and transition to future efforts.

2. GOALS AND IMPACT

There is increasing concern within the data processing community about “swimming in sensors and drowning in data,” because current fusion approaches only address localized objects. This refrain is repeated throughout many scientific disciplines, because there are few treatments of unifying models for complex phenomena and it is difficult to infer these models from heterogeneous data.

Because sheaves are the correct mathematical framework for performing abstract local to global inference, they could also be the correct tool for making inferences about unified models in many scientific domains. The only place domain-specific knowledge is needed in CHCLT is in the construction of the sheaf itself through the specification of pairwise interactions between known sources. The existing mathematical theory of sheaves provides domain-agnostic tools for integrating and interrogating the resulting global knowledge. These pairwise linking methods have been developed to date only for certain data types. Work has been initiated on extending to new data types, in particular those appropriate for semantic data. This will be a substantial focus for this project.

Sheaf theory represents locally valid datasets in which the datatypes and contexts vary, and therefore provides a common, canonical language for heterogenous datasets. Other methods typically aggregate information either exclusively globally (on the level of whole semantic ontologies) or exclusively locally (through various Kalman filtering approaches). This limits the kind of inferences that can be made by these approaches. For instance, the data association problem frustrates local approaches (such as extended Kalman filters) and remains essentially unsolved [6]. The analysis of sheaves avoids both of these extremes by specifying information where it occurs – locally at each data source – and then uses global relationships to constrain how this information is interpreted.

The foundational and canonical nature of sheaves means that existing approaches that address aspects of the knowledge integration problem space already illuminate some portion of sheaf theory without exploiting its full potential. In contrast to the generality that is naturally present in sheaves, existing approaches to combining heterogeneous quantitative and qualitative data tend to reflect specific domain knowledge on small collections of data sources. Even in the most limited setting of pairs quantitative data sources, considerable effort has been expended in developing models of their joint behavior. Our approach leverages these pairwise models into models of multiple-source interaction. Additionally, where joint models are not yet available, sheaf theory provides a context for understanding the properties that such a model must have.

Bayesian or network theory approaches seem to be natural candidates for performing knowledge integration. However, these models tend to rely on the homogeneity of information sources in order to obtain strong theoretical results. Sheaf theory extends the reach of these methods by explaining that the most robust aspects of networks tend to be topological in nature. For

example, one of the strengths of Bayesian methods is that strong convergence guarantees are available. However, when applied to data sources arranged in a feedback loop, Bayesian updates can converge to the wrong distribution! The fact that this is possible is witnessed by the presence of a topological feature – a loop – in the relationships between sources. Sheaf theory provides the means for detecting topological features such as loops (and more subtle features as well), and therefore is able to identify relationships between information sources that present hazards to Bayesian reasoning.

In addition to reports and scientific articles, our primary deliverable is a sheaf-based pipeline for knowledge ingestion and representation, along with a standardized query format for knowledge interrogation. We expect that due to its generality, both the pipeline and its internal libraries will be of general interest to the scientific community. The time is ripe for such library to gain considerable attention. It is helpful to draw an analogy from the recent lesson of *computational homology*, which was previously funded by the DARPA/DSO Topological Data Analysis project. Because of their wide applicability to many scientific domains without any changes, computational homology libraries are now the basis of several start-up companies and are of considerable academic interest. Indeed, the Institute for Mathematics and its Applications (University of Minnesota) declared 2013-2014 a thematic year¹ in “Scientific and Engineering Applications of Algebraic Topology” – essentially due to the use of computational homology in numerous applications. However, computational homology is limited in that it only works with homogeneous, quantitative datasets. Since sheaves permit the use of heterogeneous data, we expect that our libraries will garner *even wider* appeal to a broader collection of researchers.

3. TECHNICAL PLAN

This project aims to develop a robust processing methodology that leverages the discovery of

- (1) Local topological invariants that detect when spurious phenomena can become dominant through feedback loops among information sources, or when unexpected phenomena can occur without being noticed by other methods and
- (2) Strategies for designing optimal incremental experiments or data collections to supplant existing knowledge, specifically to remedy the above.

This approach relies heavily on the innovative use of mathematical abstractions called *sheaves* to represent heterogeneous data and will develop new ways to exploit this representation to

- (1) Certify that collections or experiments are able to detect phenomena of interest,
- (2) Suggest new experiments that identify and collapse unwanted information feedback loops, and
- (3) Aggregate data from multiple sources across spatial, temporal, and complexity scales.

Sheaf theory is extremely well-established in the mathematical literature [10, 18, 1], because *sheaves are the canonical mathematical representation for aggregating data from multiple incomplete sources of information*. Because of its axiomatic formulation, any other representation must recapitulate some aspect of sheaf theory. However, the existing literature on sheaves deals

¹<http://www.ima.umn.edu/programs/annual/>

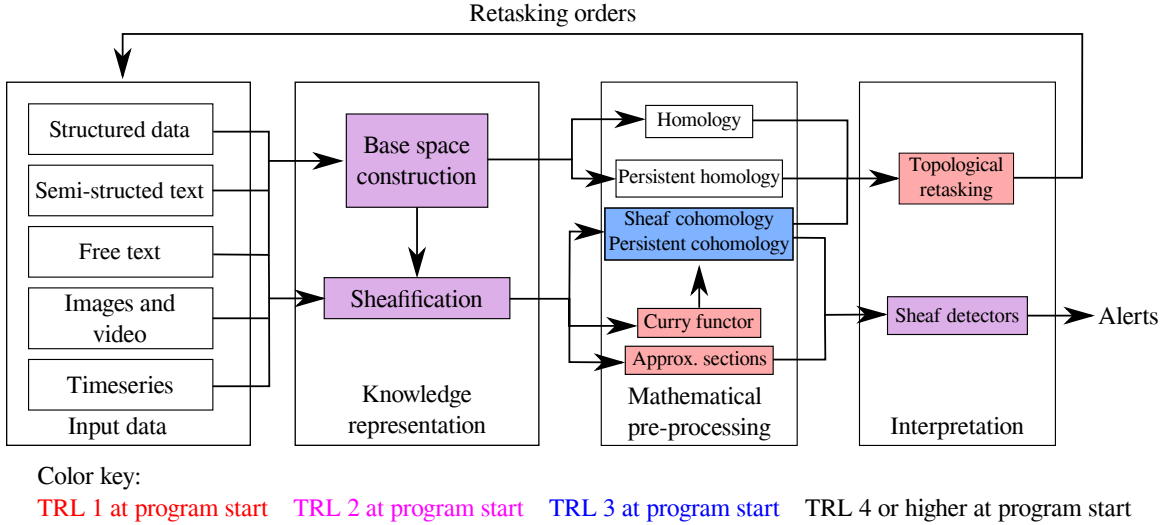


FIGURE 1. CHCLT data flow diagram, highlighting the maturity of various blocks at the start of the program

exclusively with either quantitative (more commonly) or qualitative (much more rarely) data and never both. Leveraging preliminary work by the team [20], we propose to bridge this gap by constructing sheaf models of data sources that incorporate both forms of data. These new sheaves necessitate new analysis methods, as the old methods are ill-suited to practical computation whenever qualitative data is used. We will therefore develop new mathematical methods to interrogate sheaves that store both qualitative and quantitative data.

3.1. Elements of a Sheaf-Theoretical Knowledge Representation. Until recent work by Prof. Robinson (culminating in [29]), sheaves were not used for explicit computations. We plan to remedy this deficiency by developing a sheaf-based knowledge representation and analysis methodology. Figure 1 shows the overall process we propose, starting from the data sources on the left and leading to query responses on the right. As Figure 1 shows, the proposed effort will involve the use of existing implementations of some algorithms (homology and persistent homology, see Section 3.6.4) and new implementations of some existing algorithms (sheaf cohomology, see Section 3.6.5). However, the primary focus will be the development and subsequent implementation of completely novel algorithms for analyzing sheaves (red and purple blocks).

We now list the essential aspects of a sheaf-theoretic knowledge representation. Section 3.2 introduces our first notional use case, which will be referred to in subsections 3.3–3.5 to explain these in detail. To emphasize that our approach is domain-agnostic, we include several other smaller notional use cases and one larger one in the following sections as well.

Knowledge is represented in a sheaf through the relationships between three structures:

Base Space: The base space represents the data sources and the expected relationships between them, including multi-way dependencies. The base space is formalized as multi-dimensional combinatorial object called an oriented simplicial or Δ -complex, and multi-dimensional simplices represent multi-way relationships between data sources. Vertices represent the individual data sources and faces represent relationships between data sources. An orientation can be assigned to each face to indicate dependency relations between sources.

Stalks: Each data source is assigned a stalk, which is a schema for the data provided by that source. Thus far, this is merely a representation of the knowledge to be stored. The analytic power of sheaves comes from their formalization of the relationships between data sources. To formalize the relationship between sources, each multi-way face in the base space is also assigned a stalk, which represents the observables that are common to all data sources implicated in that face.

Restriction Maps: Domain-specific functions – called *restriction maps* – are then used to translate the observables from every data source’s stalk to the stalks over each simplex including it.

An instance of the data within a sheaf is called an *assignment*. If an assignment is consistent with the models encoded in the restriction maps, it is called a *section* of the sheaf. The failure of an assignment to be a section indicates that either the data or the models (or both) are faulty.

We will be explicating two notional use cases. While the first is pedagogical, the second will be tested against real data inputs (some initial datafication). These will include numerical data, structured data (CSV and XML), semi-structured text, and text documents with significant linguistic constraint. Textual documents will also be permitted to include numerical data. The quantitative numerical data will also include several broad categories such as video, image, and timeseries data. For instance, the record of the hourly temperature at a given location would constitute “timeseries” data, while numerical tables would constitute “image” data.

Given an assignment or the sheaf itself, the CHCLT process will apply a collection of algorithms (labeled “mathematical preprocessing” in Figure 1) that enhance the quality of the inferences that can be made about the knowledge stored in the sheaf by summarization. Several of these algorithms (for instance *homology* [16]) are well-established in the mathematical literature and are becoming standard in engineering practice [5, 2, 7, 4, 34], though it will be necessary to develop new algorithms to manipulate sheaves containing both quantitative and qualitative data. After passing through the preprocessing stage, the resulting information must be interpreted in the context of experimental findings and subsequent iterative design as outlined in Section 3.6.

3.2. Notional use case 1: situational awareness of a public event. We will use the following true story to explain the construction of a sheaf datastructure and some of the analytics upon it. It should be noted that this example (like [12]) was deliberately constructed to be realistic while also illustrating the most important features of our approach, and is therefore understood to be notional.

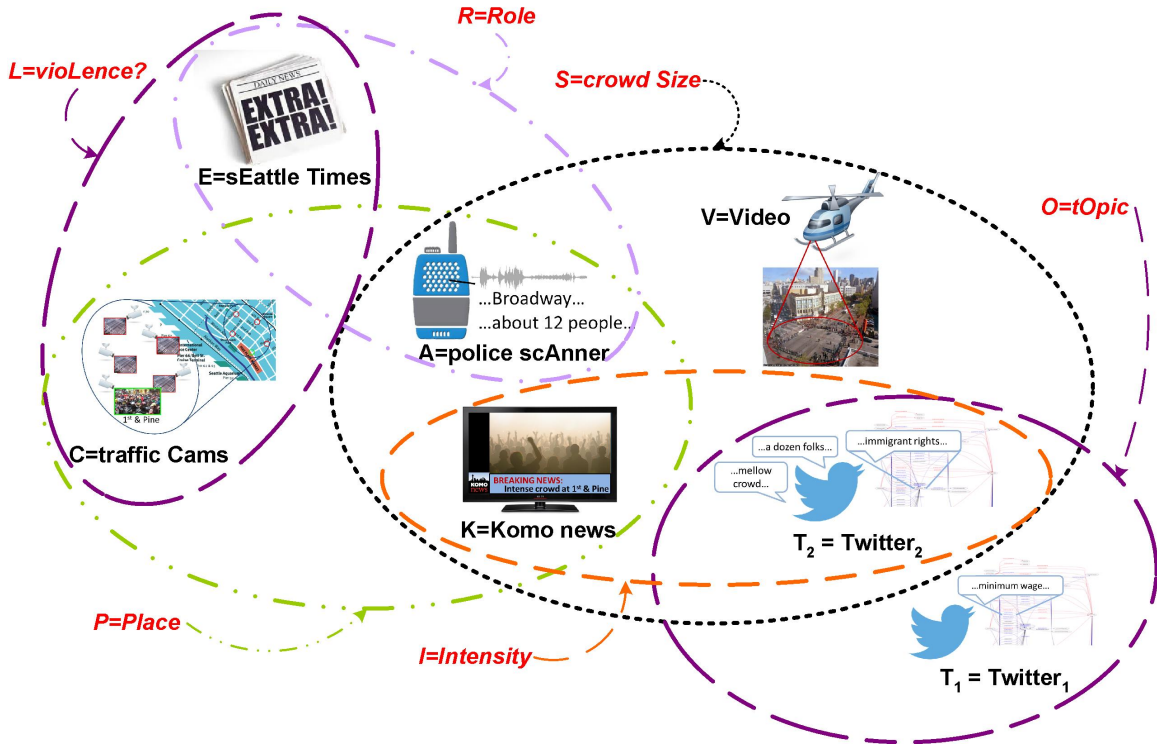


FIGURE 2. Notional set of data sources observing a public event

On Mayday, 2014, an exuberant group of protesters staged a peaceful demonstration in downtown Seattle in support of immigrant rights and an increased minimum wage. Shortly thereafter, a group of even more exuberant “anticapitalists” meandered through the city streets, from downtown to Capitol Hill, blocking intersections and lighting small fires. Police mostly watched or “escorted” the protesters, but towards the end a half dozen people were arrested, and some tear gas was deployed.² While a fine time was being had by all that evening, one of the PNNL team members (Joslyn) was spending a night in in Richland, Washington. There he followed the events of the day through the local KOMO TV news feed and a couple of twitter feeds.

Imagine that in addition to these sources, we had access to overhead video, police scanner audio, Seattle urban transit cams at major intersections, and the feed from the Seattle Times. The left panel of Figure 2 shows the overall situation, and how these means might inform our ability to track a collection of “state variables”:

S = **Size of the crowd**: An integer.

O = **tOpic being protested**: Terms like “immigrant rights”, “minimum wage”, or “anti big business” are normalized into an ontology, each being a node in a partially-ordered semantic class hierarchy.

²http://www.huffingtonpost.com/2014/05/02/seattle-may-day_n_5253707.html

$P = \mathbf{Place}$: A categorical variable like “1st and Pine” or “Broadway”.

$I = \mathbf{Intensity}$: An ordinal variable: “high”, “medium”, “low”.

$L = \mathbf{vioLence}$: A Boolean variable: “present” or “absent”.

$R = \mathbf{Role}$: Another categorical variable, reflecting the kind of person present, for example “protester”, “police”, “bystander”, or “press”.

We can cast each information source as a separate data source or analytic, with structure as follows:

$A = \mathbf{police\ scAnner}$: A speech recognizer has been trained to extract specific information about crowd size and location from speech like “I see about 12 people here at 1st and Pine, 4 police and 8 protesters”.

$C = \mathbf{transit\ Cameras}$: Cameras at specific intersections can show when the crowd has reached those locations, and whether violence is present.

$E = \mathbf{sEattle\ Times}$: An analytic deployed against the local newspaper web feed to parse out information about the presence of people in certain roles and the presence or absence of violence.

$K = \mathbf{KOMO\ News}$: The news broadcast shows a video feed of crowds with a chyron showing the specific intersections, and video analytics are trained to estimate crowd sizes and intensity.

$T_1 = \mathbf{Twitter}_1$: A text analytic extracts keywords to identify protest topics.

$T_2 = \mathbf{Twitter}_2$: A different text analytic extracts keywords to estimate topic, crowd size, and intensity.

$V = \mathbf{overhead\ Video}$: An algorithm is used to estimate the number of people shown in a live video stream.

We model the data sources and their overlapping coverage by letting $X = \{P, S, O, I, L, R\}$ be the set of **state variables** and $Y = \{A, C, E, K, T_1, T_2, V\}$ be the set of **sources**. Then Table 1 shows the relationships between these sources and the state variables they inform. Figure 2 shows this diagrammatically. The data sources are labeled in black, and include textual data (such as an analytic run on the Seattle Times or various analytics run on Twitter feeds) as well as quantitative data (such as video cameras). Each data source supplies information regarding the various state variables (labeled in red), such as crowd size (quantitative) and the topic of the event (qualitative). On the left of the diagram is the description of the data sources, and their domains of discourse.

3.3. Base space construction: relationships between analytics. The base space of a sheaf encodes the functional relationships between sources of data and their data fields. The construction of the base space proceeds by specifying the information sources and the state variables they provide. For instance, consider Use Case 1 outlined in Section 3.2. We cast Table 1 as a binary relation $B \subseteq X \times Y$. Then left Figure 3 shows B as a set system (undirected hypergraph) $B(X) \subseteq 2^Y$ on the source Y . The variables $x \in X$ (i.e., the columns of R) are represented (in red) as subsets $B(x) \subseteq Y$ of the source (in black) which inform them.

TABLE 1. Data source structure for notional use case 1

	S crowd Size Number Scalar	O tOpic Ontology term Partial Order	P Place Intersection Categorical	I Intensity Level Ordinal	L vioLence? T/F Boolean	R Role Police, Protester Categorical
A =Police scanner	✓		✓			✓
C =Transit cams			✓		✓	
E =sEattle times					✓	✓
K =KOMO News	✓		✓	✓		
T_1 =Twitter1		✓				
T_2 =Twitter2	✓	✓		✓		
V =Overhead video	✓					

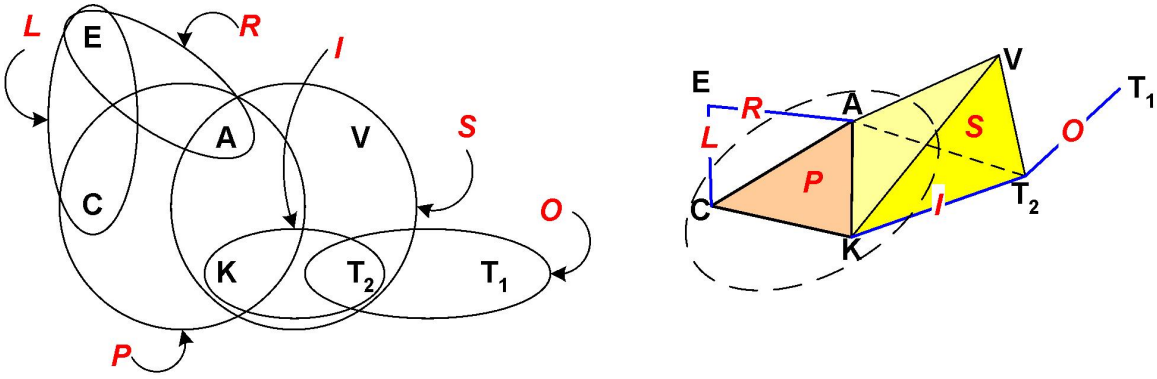


FIGURE 3. Overlapping set system (left) and simplicial complex (right) corresponding to the data sources in Use Case 1

The right frame of Figure 3 shows $B(X)$ as a combinatorial structure called an “abstract simplicial family” [19] with simplices $B(x), x \in X$ of dimension $|B(x)| - 1$. Note that $B(S)$ is the (solid) tetrahedron $\{A, K, V, T_2\}$, with the $\{A, T_2\}$ edge underneath, indicating the four-way interaction of the sources through the variable S . Similarly, $B(P)$ is the filled-in triangle $\{A, C, K\}$, while the triangle $\{A, C, E\}$ is *not* filled in, rather consisting of the three distinct edges $\{A, C\}$ (for P), $\{C, E\}$ (for L), and $\{E, A\}$ (for R). Also, the edges $\{K, T_2\}$ (column $B(I)$) and $\{T_1, T_2\}$ (the column $B(O)$) are called out from the table and shown in blue, as are the edges $\{C, E\}$ and $\{E, A\}$ (but not $\{A, C\}$).

While $B(I) \subseteq B(S)$, none of the other faces are pairwise inclusive, and so they comprise the maximal faces of an abstract simplicial complex (ASC), which further contains all the included sub-faces (all triangles, edges, and vertices). Topological features of the connectivity pattern can be identified, including loops, voids, etc., where potential informational feedbacks can result in faulty conclusions. In our case, the ACK triangle can establish consistency around place P ,

while the *ACE* loop may yield assignments which are impossible to resolve consistently amongst all three sources.

3.4. Stalks: encoding knowledge as algebraic types. Given a base simplicial complex, a stalk is then the assignment of data to each of its faces. In our example, this means the data held in common for each combination of interacting systems. For example, we can read off the sheaf diagram that the sources A and K both inform variables P and S . Thus the one-dimensional face AK holds data SP , as shown on the right frame of Figure 3. CHCLT will then represent knowledge by combining the values of the state variables stored in the stalks of a sheaf with relationships between them (stored in the restriction maps of a sheaf, Section 3.5).

The stalks of a sheaf will usually not be all of the same type, since different data sources will return state variables of different types. Additionally, when two different data sources inform a given state variable, they may choose to represent it differently, with different types. For instance, consider the problem of measuring the number of people in a crowd from Use Case 1. Several sources (the KOMO news K , the overhead video V , and one Twitter feed T_2) provide measurements of the crowd size state variable. Even though each of these are reporting on an ostensibly quantitative state variable, not all of them actually return quantitative values. For instance, the KOMO news could return a range of possible sizes, for instance “between 50 and 100 people.” An object detector run on the overhead video would instead return a single estimate, for instance “78 people.” Finally, the Twitter feed is unlikely to be that precise, returning instead a categorical – one of “small crowd”, “large crowd”, or “HUGE crowd”.

Integration methods for semantic information are substantially different from those of numerical data. Keyterm identification from free text suffers from ambiguities and errors around spelling varieties and ambiguities introduced by the idiosyncracies of local linguistic communities. Matching terms into more controlled vocabularies requires expensive management and integration of multiple such resources. For more fully semantically structured resources such as ontologies, a major research effort in alignment or matching efforts have produced results which are heuristically driven, modestly accurate, and computationally expensive [9]. But when a representation of the underlying semantic space is available which is mathematical, and equipped with at least a pseudometric space, then degrees of matching can be measured quantitatively. This is the case in our prior work in ontology alignment which uses the metric space over a single partial order used to represent ontological semantic hierarchies [8, 21]. When available, such metric approaches can provide substantial computational advantages, and are deployable in a sheaf-theoretical context when deployed in the context of “approximate sections” (see Section 3.6.5).

From a mathematical perspective, all of these types are instances of sets, and set theory has long been used as a formalism for representing all knowledge. However, this perspective does not scale well, and there are additionally issues with incomputability. CHCLT takes a different perspective, namely that being able to make *all* inferences about all knowledge is simply too much to ask for. Instead, CHCLT aims to compute weaker invariants about the information stored in a sheaf. These weaker invariants require more algebraic structure on the state variables than merely sets – as part of the encoding process, CHCLT will require all state

variables to be returned in vector spaces. This allows manipulation of all knowledge stored in the sheaf uniformly using linear algebra. This encoding of data in vector spaces is called *categorification*. Categorification has found recent use in a number of mathematical problems. For instance, Khovanov homology in knot theory is a categorification of the Jones polynomial, and Hopf algebras arise in the categorification of rooted trees. It has potential value for forming combined spaces of dissimilar measurements. Prof. Robinson used a particular categorification of boolean values to construct sheaf-based invariants for logic circuits that are sensitive to timing differences [30].

Continuing our example from Section 3.2, the values returned by the overhead video already form an algebraic ring, which can be extended to a vector space by permitting fractional values. However, the set of intervals (returned by the KOMO news) does not have additive inverses, so it is not a vector space. However, it can be formally extended to a vector space [25] via a categorification. This leaves qualitative, set-valued types, such as categoricals, partial orders, and booleans to be categorified. To this end, suppose that the set of values permitted by one of these types is the set A . The categorification of A is the vector space $\mathbb{R}(A)$ whose dimension is the cardinality of A , and that each element $a \in A$ corresponds to a basis element $1 \cdot a \in \mathbb{R}(A)$.

CHCLT will support at least the following types of state variables

- (1) Structured numerical types in vector spaces, such as images, timeseries, or lists of parameters for physical models,
- (2) Interval-valued numerical types,
- (3) Categorical variables, including boolean types, and
- (4) Hierarchical types such as partial orders and ontologies.

As mentioned above, sheaf-theoretical applications have been robustly developed against numerical data, while our ambition is to additionally incorporate semantic data. In our example, these semantic data present in data types which have less mathematical structure than numerical vector spaces, although they typically have metric spaces, requiring categorification.

Non-numerical data are provided by *semantic extraction* computational techniques drawing against a variety of sources with various levels and qualities of textual or linguistic data:

Structured Data: Frequently data sources contain textual information emnedded in a rich typing structure, including linguistic tags for columns, sections, and contents of data items in cells. Typical cases include tabular CSV files and hierarchically structured XML files. In these cases, robust extraction methods can be built to automatically populate a fact-base of recorded observations of those categories.

Semi-Structured Data: When not fully structured, documents, including social media inputs, may be sufficiently structured or restricted to be able to build high quality extraction methods for selected portions of the documents. This may require moderately complete models of the domain or knowledge-base intended, or knowledge of the kinds of document structure or restricted vocabularies are present. These techniques work particularly well when portions of those structures are repeated in a consistent manner [24, 15].

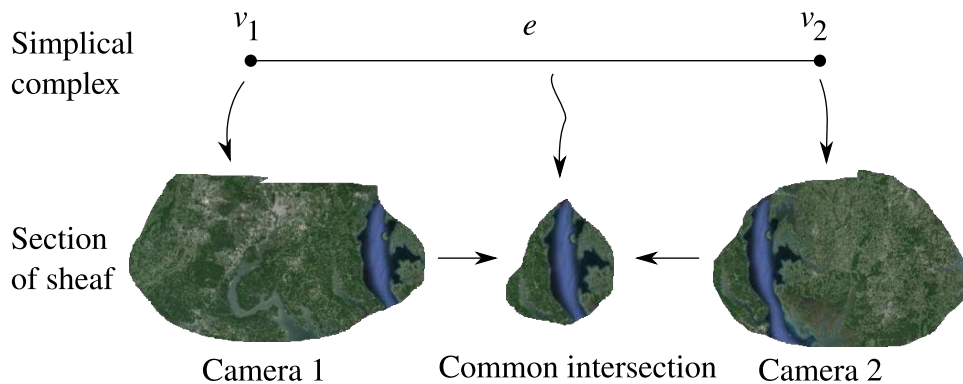


FIGURE 4. A simplicial model of two sensors (top) on which a sheaf containing image data is constructed. The images collected by the two sensors are consistent (bottom) when the portion of the images on the overlap agree.

Free Text: Finally, when documents are not explicitly structured at all, free text extraction methods can be deployed. While generally more computationally burdensome (see more discussion in Section 3.8 below) and less accurate, these deficiencies can be mitigated by exploiting restrictions in the actual vocabularies used in a particular technical discourse community, or the employment of more explicit knowledge models. These knowledge models in turn can be either built special-purpose, acquired from external sources, or a mixture. From the models, unsupervised and semi-supervised techniques for adding semantic structure and representations are readily available [32].

3.5. Restrictions: applying domain-specific models. The restriction maps translate information between two data sources into a common format, allowing it to be compared in that format. Necessarily, restrictions encode domain-specific information, and they make any underlying modeling assumptions explicit. Our approach to SIMPLEX requires that this domain specific information be supplied during the construction of the sheaf, and this explicitly includes the context of each information source. Fortunately, many information sources are easy to encode into restriction maps, and our team has identified several such types of restriction maps.

In our approach, sheaves store the information supplied by a number of sources. But this alone is not very useful – one is typically interested in the actual content of those information sources. An *assignment* is an instance of data from the stalk at each simplex in the base space of a sheaf. As such, an assignment only requires the specification of stalks. Using the restrictions, it is possible to test an assignment for self-consistency. Specifically, an assignment is a *section* whenever the assignment’s value on a simplex is determined by the values on all of its faces.

Thus restrictions play a crucial role, and can take different forms. We detail and illustrate with additional examples.

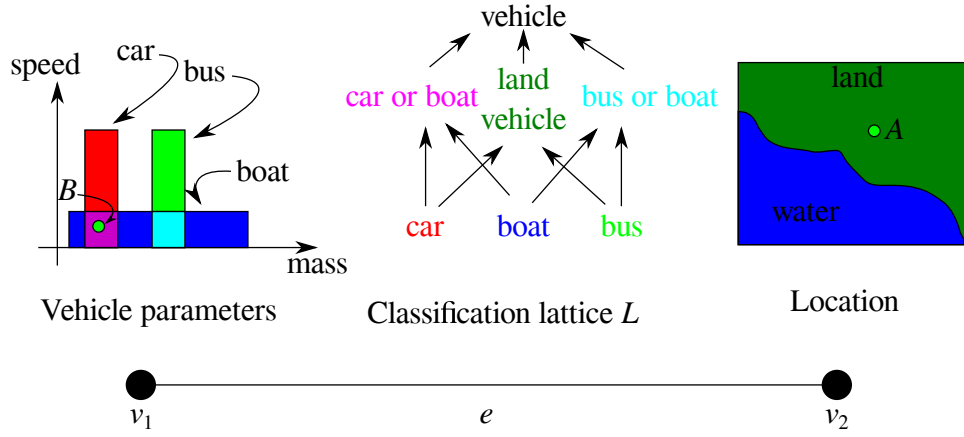


FIGURE 5. A collection of vehicle parameters (left), a lattice of vehicle types (middle), and a scene (right)

3.5.1. *Restrictions implicitly aggregate information along common domains.* Restrictions can represent the process of forming a mosaic from two images. For instance, two cameras whose coverage regions U_1 and U_2 overlap can be used to form a mosaic, as shown in Figure 4. The base space associated to this pair of sensors is a single line segment with two distinct endpoints. The sensor sheaf assigns the vector space of images collected by camera 1 to the vertex v_1 on the left, and the vector space of images collected by camera 2 to the vertex v_2 . The two collected images are compatible when they agree after cropping to the intersection region. Therefore, we assign the vector space of images over the intersection region to the edge e connecting the two vertices. The restriction maps (the horizontal arrows in Figure 4) merely crop the images from the two cameras.

Although we have not explicitly defined images over the union of the two coverage regions, the global sections of this sheaf are the mosaics representing a single image over the union. Therefore, *fused data is constructed implicitly by sheaves of sensors.*

3.5.2. *Restrictions translate between quantitative and qualitative information.* Consider the task of classifying vehicles into three target types: cars, buses, or boats (see Figure 5) based on measurement of vehicle parameters and their locations. This can be encoded by taking two quantitative information sources – a measurement of the vehicle’s location, and a paired measurement of its speed and weight – with a qualitative hierarchy of vehicle types. The information on the left identifies vehicles based on their mass and operating speed; boats do not travel fast, but can be very light or heavy. Buses and cars travel at similar speeds, but cars are much lighter. On the right is a map of locations where targets can appear, and in the middle is a lattice of vehicle types.

We encode this situation as a sheaf $\mathcal{S}$ over a base space with two vertices and one edge:

- (1) Vertex v_1 , whose stalk is $\mathbb{R}^2$ (quantitative) represents the speed and weight of the vehicle
- (2) Vertex v_2 , whose stalk is $\mathbb{R}^2$ (quantitative) represents its location on a map,

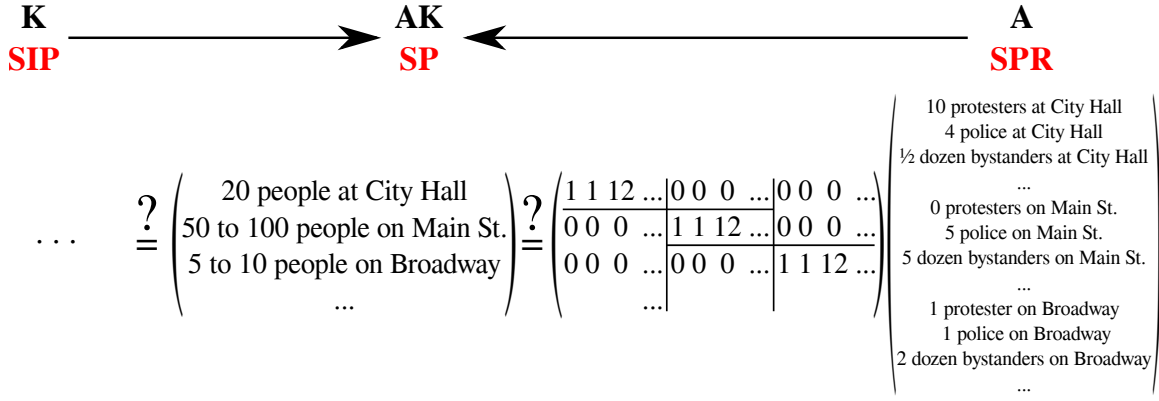


FIGURE 6. The translation of a richer datatype (right, A/SPR) to a poorer datatype (middle, AK/SP), to be used as the “common ground” between different datatypes

- (3) Edge e , whose stalk is L (qualitative) is a lattice that represents the possible classifications.

The restriction maps go from the vertices to the edge. In Figure 5, they are indicated by color. For instance in the restriction $\mathcal{S}(v_1 \rightsquigarrow e)$, a fast moving heavy vehicle is a bus, and is shown marked in light green. The other restriction map $\mathcal{S}(v_2 \rightsquigarrow e)$ takes all locations corresponding to water (dark blue) to boats, and all locations corresponding to land (dark green) to land vehicles.

When populated with information, inferences about the detections can be made using the sheaf $\mathcal{S}$. Suppose that there are two detections A and B , one from location information, and one according to vehicle parameters. If they are not associated to one another, both detections are ambiguous: A could be either kind of land vehicle, while B could be a car or a boat. However, if these detections are from the same target, they are no longer ambiguous: they must represent a car.

3.5.3. Restrictions mediate between hybrid quantitative and qualitative data. Returning to the Use Case 1 example defined in Section 3.2, consider the police scanner and KOMO news data sources. The police scanner provides the state variables Size, Place, and Role. Of these, only size is quantitative. Using our categorification approach (Section 3.4), these state variables can be encoded in a single vector whose entries specify the number of people observed in each possible place and role under discussion (see the right of Figure 6). The KOMO news variables can be similarly encoded in a vector describing the numbers of people in each Intensity state and Place (not shown). Notice that the Size and Place are common state variables between the two data sources, therefore the edge between these two data sources should have a stalk that represents both of these variables. A convenient categorification for the stalk over the edge merely lists the number of people at each location.

Since the data have been categorified, the restriction map from the vertex A (police scanner) to AK (common variables between the police scanner and KOMO) can be written as a matrix.

Since the restriction map removes all reference to Role, which is not represented by the KOMO data source, the matrix should be a projection matrix. Therefore, the matrix takes the block form shown in Figure 6. Additionally, the analytic run on the police scanner chooses to represent the number of “bystanders” in units of dozens, rather than individual people. In order to translate from dozens to individuals, the restriction matrix has a 12 in each column that represents bystanders to perform this conversion.

3.5.4. *Restrictions can express knowledge encoded in physical laws.* Sheaves can also be used to encode knowledge in physical laws and differential equations. For instance, consider the wave equation

$$\frac{\partial^2 u}{\partial t^2}(t, x) = c^2 \frac{\partial^2 u}{\partial x^2}(t, x) \quad (1)$$

with wave speed c , or the Helmholtz equation

$$\frac{\partial^2 U}{\partial x^2}(x) + k^2 U(x) = 0,$$

which arises if $u(t, x) = U(x)e^{ikct}$, where k is the *wavenumber*. This is a good approximation for wave propagation that happens along a complicated network of narrow channels, though the domain is better represented as a graph G instead of $\mathbb{R}$. (See [22, 23, 33, 26] for precise conditions under which this approximation holds.) This situation often arises in the study of radio propagation in urban canyons.

Since the lengths of edges in such a graph G play an important role in the propagation of waves, an important inverse problem is the reconstruction of these lengths from measurements of a small number of waves. From an urban propagation perspective, the reconstruction of edge lengths from signal measurements is a form of 1-dimensional topological “imaging” that produces a geometric map of an environment.

Existing approaches for performing this inversion usually rely on wideband spectral methods, which examine eigenfunctions of a differential operator on G that generalizes the Helmholtz operator $\left(\frac{\partial^2}{\partial x^2} + k^2\right)$. This has unfortunate consequences for most urban sensing applications, since the available bandwidth for sensing is limited.

In the context of wave propagation on a graph, the space of solutions forms a sheaf. This even holds if the propagation is lossy, or if we instead consider fundamental solutions, in which there are a number of sources [31]. If the topological structure of the graph is known, then Prof. Robinson found that this sheaf model of the differential equation organizes the information in a way that allows easy reconstruction of the edge lengths.

Suppose that X is the base space, represented as a 1-dimensional simplicial complex representing the propagation environment. Each vertex represents measurements taken at a junction of propagation channels, and each edge represents the traveling waves along that channel. Suppose that each edge e is assigned a positive real number $L(e)$, called its *length*. The *transmission line sheaf* $\mathcal{T}$ with wavenumber k is given by

$$(1) \quad \mathcal{T}(v) = \mathbb{C}^{\deg v} \text{ for all vertices } v,$$

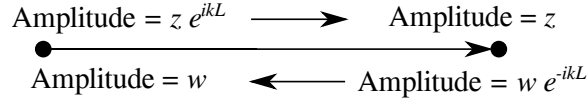


FIGURE 7. A section of a transmission line sheaf supported with value (w, z) on an edge with length L

- (2) $\mathcal{T}(e) = \mathbb{C}^2$ for all edges e , and
- (3) If e_m is the m -th edge attached to a degree n vertex v ,

$$\mathcal{T}(v \rightsquigarrow e_m)(u_1, \dots, u_n) = \begin{cases} \left(u_m, e^{-ikL(e_m)} \left(\frac{2}{n} \sum_{j=1}^n u_j - u_m \right) \right) & \text{if } e_m \text{ is inward at } v \\ \left(e^{ikL(e_m)} \left(\frac{2}{n} \sum_{j=1}^n u_j - u_m \right), u_m \right) & \text{if } e_m \text{ is outward at } v \end{cases}$$

The stalks of a transmission line sheaf represent the complex amplitudes of the travelling waves *exiting* each of its two endpoints, measured at the endpoints. Because each edge has an assumed orientation, the first component represents the wave amplitude at the outgoing vertex, while the second component represents the wave amplitude at the incoming vertex. It is important to note that the complex amplitudes measured at other points will be phase shifted with respect to these points. On an edge with length L , a travelling wave that exits with amplitude z will have complex amplitude ze^{ikL} at the entrance. Figure 7 summarizes this interpretation of the stalks over an edge in a transmission line sheaf. In contrast, the stalk over a vertex represents the complex wave amplitudes of waves *entering* the vertex.

The global sections of a transmission line sheaf correspond to solutions to the Helmholtz equation on each edge of X that satisfy the following *Kirchhoff conditions* at each vertex v

- (1) All solutions on the edges attached to v extend to a single continuous function at v
- (2) The sum of the derivatives (taken in the direction away from v) of each solution at v is zero.

3.6. Analytical methods. Having outlined and exemplified the key components of sheaf-theoretical modeling, mostly in the context of our first use case, we now go on to discuss further aspects of the overall technical approach. Specifically, referring back to Figure 1, we will address two main classes of queries against data in support of automated inference:

- (1) Detection: Given a description of some phenomenon or behavior of some entities encoded in the sheaf, can we certify that this is present in some measurement data? Once so identified, can we trace them back to a semantic report of the behaviors?
- (2) Scheduling the next best collections: Given a set of information sources, what additional sensing or semantic resources should be deployed to best constrain the observations to a particular target type or behavior?

In order answer these queries, we will exploit a number of existing topological and sheaf theoretic invariants, and also will develop several new classes of sheaf invariants. The framework we plan to develop is both modular and extensible, in that each sheaf-theoretic processing block will

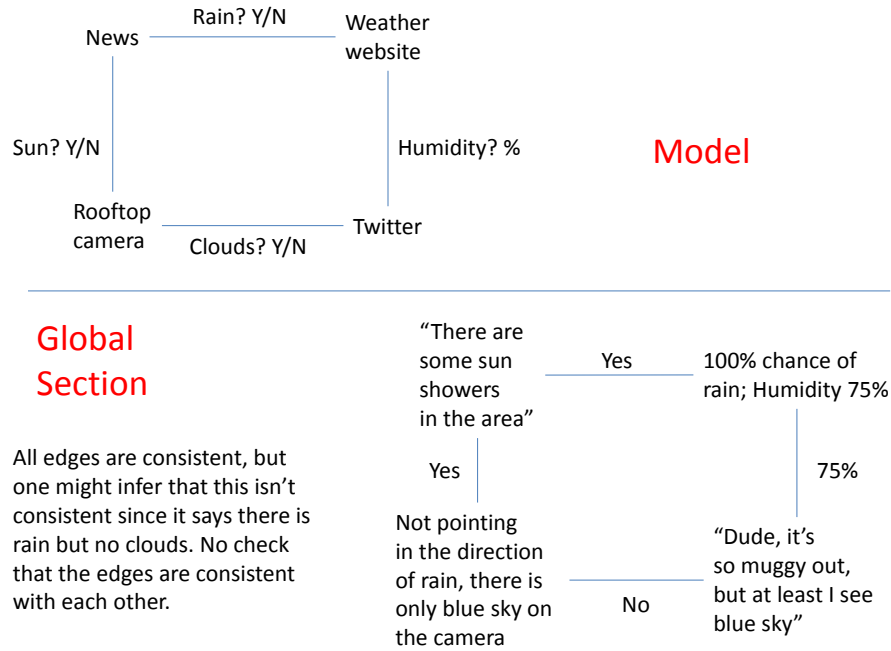


FIGURE 8. Example of global consistency that leads to an erroneous inference. The conclusion from these data is that it's raining without any clouds in the sky! This can be identified by the presence of nontrivial sheaf cohomology classes.

ingest some portion of the knowledge encoded as a sheaf, and will produce answers to the above queries. Additional blocks can be added in parallel to the ones we envision in this proposal.

3.6.1. *Query: topological detection of "blind spots".* Although sections of sheaves are self-consistent instances of information, they can still be misleading. This is a cause of failure in reasoning systems – information feedback loops can result in confounding and faulty conclusions. One major advantage of topological methods is that they can identify when there is a risk of drawing conclusions from a feedback loop. In order to ensure scalability of the process, this query will specify a certain domain of interest, or a certain list of information sources to be checked for feedback.

For instance, consider attempting to determine the cause of a nearby rainstorm. To this end, you take observations from several information sources:

- (1) A rooftop camera,
- (2) A news report,
- (3) A weather website, and
- (4) A friend's Twitter feed.

No one source of information happens to present you with a complete picture, but each has some common information with several of the others. This can be represented by the diagram in the upper frame of Figure 8. Given the actual instances of information from each of these sources, shown in the lower frame of Figure 8, you can conclude that rain is falling. From the available knowledge, you would be forced to draw the incorrect conclusion that rain is falling without any clouds! This is due to the fact that each source isn't providing complete information, *and the base space graph has a loop*. These loops have been detected using higher sheaf cohomology [29, 30] when the sheaves contain only quantitative information, but not for qualitative and quantitative information simultaneously. Therefore, our team will develop cohomological invariants to detect information feedback loops involving mixed quantitative and qualitative information.

3.6.2. Query: Topological retasking. The information feedback loop in Figure 8 can be easily resolved if one specifically requests information about the presence of clouds from all sources. This amounts to “filling in” the loop with additional simplices, representing additional consistency checks being applied to the information stored in the sheaf. Merely asking a more direct question about the clouds from one of the news or website sources amounts to enriching the information provided by those information source. A simplex would be added to the base space with restriction maps bringing information about clouds from all sources to be tested for consistency. In particular, this would detect a fault in the Twitter feed’s analytic that declares the lack of clouds.

Prof. Robinson has discovered that there is a Nyquist-like topological sampling criterion [28] for general quantitative information in sheaves that ensures exact reconstruction of phenomena. Clearly the weather example just given doesn’t meet such a sampling condition, even though Prof. Robinson’s result doesn’t technically apply. A goal of CHCLT would be to extend this theoretical Nyquist condition to include qualitative information as well.

Following the analogy of reconstruction from samples, one can attempt to increase the “sampling rate.” Specifically, the sheaf-theoretic Nyquist theorem discovered by Prof. Robinson can produce exemplars of ambiguous sets of measurements, and witnesses topological features in the base space corresponding to information feedback loops. CHCLT will produce these witnessed topological features in response to the reliability of answers to specific queries.

3.6.3. Interface with automated reasoning systems. The goals of SIMPLEX include development of a common application programming interface (API) through which all domain-specific users will access its framework. This will allow analytic overlays or compositions for automated inference, reasoning, hypothesis generation, and steering for context-aware data analysis and causal inference.

But the way that automated reasoning and inference is performed and imagined by different technical communities also speaks to the other primary SIMPLEX goal of integrating heterogeneous information, especially numerical and symbolic (qualitative, semantic). For example, propositional information derived from semantic data supports logical inferencing models exploiting rule bases and ontological structures. But quantitative modeling, on the other hand,

GLOBALIZATION:

Consistency between analytics allows fusion

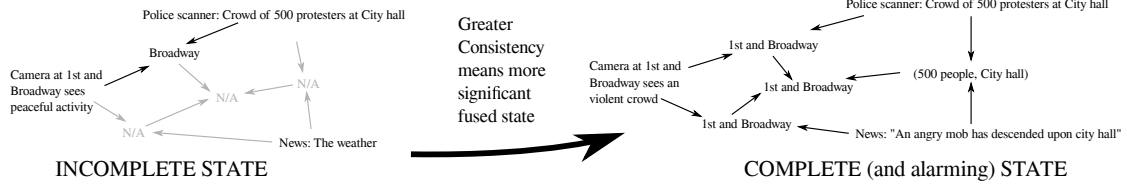


FIGURE 9. Consistency between disparate data sources allows for stronger inferences to be made

exploits mathematical models such as differential equations or other linear or non-linear mathematical systems over numerical vector spaces.

So prior to supporting any particular logical or reasoning model is the basic representation of the data and hypotheses to be integrated and reasoned over. CHCLT's team's perspective comes from mathematical modeling of sensors on the one hand, and the mathematical modeling of semantic data on the other. Thus our role within SIMPLEX is envisioned to be providing a common mathematical representational framework for all types of data integration. Thus CHCLT structures will be candidates to be included in the data model for any future SIMPLEX API. But rather than *performing* automated inference itself, our job is to provide the necessary mathematical *framework* on which such inference can be carried out, integrating quantitative and qualitative information.

3.6.4. Existing processing methods. There are numerous algebraic invariants that can be computed about a given sheaf - descriptions of these have formed a large set of mathematical literature focused around sheaf theory over the past 50 years. The easiest of these invariants to understand is the class of *sections*. A section is the assignment of data values to a set of faces (from their respective datatypes as assigned by the sheaf) in such a way that translation from one data value to another in the section agrees with the translation maps. The size of a given section is a good indication of its significance. Figure 9 shows two sections: on the right is a highly significant section, since all sources are reporting on the same phenomenon. On the left is a less significant section, in which sources are presenting inconsistent information.

Our initial algorithms will build on the well-established techniques of *homology* and *persistent homology* for studying the base space and on *sheaf cohomology* to aggregate information from all information sources stored in the sheaf, respectively. Table 2 explains some typical interpretations of these invariants.

Although homology and persistent homology are commonplace, they only treat the base space. In his recent book [29], Prof. Robinson has focused on the interpretation of sheaf cohomology, largely for quantitative data sources. Recent work by the team indicates that more is possible, and that sheaf cohomology can be interpreted in the context of qualitative/quantitative information sources as well. CHCLT will rely on both existing implementations of homology and persistent homology as well as new implementations of sheaf cohomology.

TABLE 2. Homological invariants and interpretation

Degree	Homology	Sheaf cohomology
0	Connected components	All possible consistent information states
1	Loops	All possible information feedback loops
2	Voids	(Unknown)
...	...	...

3.6.5. *New processing methods.*

(1) Persistent sheaf cohomology

Query: Used in responding to both query types

Inputs: (1) Knowledge database encoded in the sheaf; (2) Method of coarsening the quantitative data stored in the sheaf

Outputs: (1) Cohomology in all degrees at varying complexity scales; (2) Persistent diagrams

TRL: 2 at program start; 3 at Phase I end; 4-5 at Phase III end

Details: Persistent cohomology provides some measure of robustness against distortions in the stalks of a sheaf by segregating those features that “persist” for a large range of perturbations from those that are present only for a small collection of perturbations. Persistent *homology* is well-understood and efficient to calculate – existing implementations are TRL 4+. In contrast, CHCLT will feature the first ever computations of persistent sheaf cohomology, and the initial attempts to connect it to a knowledge database.

(2) Approximate sections

Query: Used in responding to both query types

Inputs: (1) Knowledge database encoded in the sheaf; (2) Pseudometric on each stalk to allow distances between data elements to be measured; (3) Partial information to be extended

Outputs: (1) Assessment of the level of consistency that partial information has within the sheaf; (2) List of approximate sections near the given partial information

TRL: 1 at program start; 2 at Phase I end; 4 at Phase III end

Details: Approximate sections were recently discovered by the team, and appear to be a different way to endow error tolerance to a sheaf that encodes both quantitative and qualitative information. Specifically, rather than declaring that a set of observations are consistent when different sources agree exactly, approximate sections merely require approximate agreement. The primary technical hurdle is the efficient calculation of classes of approximate sections, since they have not been tested on realistic datasets.

(3) Curry functor and statistical aggregation methods for heterogeneous knowledge

Query: Used in responding to both query types

Inputs: Knowledge database encoded in the sheaf

Outputs: Local and global statistics aggregated over portions over the base space

TRL: 1 at program start; 2 at Phase II end; 3 at Phase III end

Details: There are essentially two ways to address a sensor integration problem: start with either a model of the phenomenon being measured (a *world model*), or with a *sensor model* and instead infer the phenomenon. That quantitative sensor models are represented by *sheaves* has been known for some time, and is the unifying theme of Prof. Robinson's recent book [29]. Recent work by Curry [3] may be useful for formalizing disciplined ways to aggregate statistical measurements across sheaves.

As indicated in Figure 1, homological methods are intermediate steps to produce responses to the two main queries that CHCLT will address. The output of homological methods are precise indications and warnings of the existence of certain kinds of information faults. These are the raw material that will be used by the new algorithms we will develop to answer detection and retasking queries.

(4) Cohomological sheaf detector

Query: Used in responding to Detection queries. Usually will be focused on a specific set of knowledge sources to limit computational burden.

Inputs: Sheaf cohomology in all available degrees, though primarily degree 1 and higher

Outputs: Lists of sources that participate in feedback loops, as well as exemplars of consistent, but suspicious knowledge states

TRL: 2 at program start; 3 at Phase I end; 4-5 at Phase III end

Details: This algorithm extends Prof. Robinson's work on the sheaf-theoretic Nyquist theorem. It will reinterpret nontrivial sheaf cohomology classes as information feedback loops, and will identify potential sources of erroneous inferences. Given an assignment of the sheaf – information stored in the sheaf – this algorithm will assess the risk for feedback states being present.

(5) Topological retasker

Query: Used in responding to Retasking queries

Inputs: Sheaf cohomology in all available degrees, degree 1 and higher

Outputs: (1) List of variables to be obtained from existing information sources, (2) List of variables to be obtained from new information sources, (3) List of suspicious data elements that should be re-examined

TRL: 1 at program start; 2 at Phase I end; 4 at Phase III end

Details: This algorithm will identify cohomological loops of degree 1 and higher by finding minimal simplices to add to the base space and their associated stalks and restriction maps. These new simplices correspond to desired comparisons that will allow information feedback loops to be eliminated.

3.7. Knowledge updates. Since sheaves are constructed locally, the addition of new information sources will merely add additional simplices to the base space. This also enables computations with sheaves to be handled incrementally – the two queries supported by CHCLT are

local. Technical methods for manipulating sheaf cohomology calculations incrementally already exist and are well-established. For instance, the Mayer-Vietoris exact sequence specifies how the sheaf cohomology changes when a new information source is added.

However, information sources are not the only way that knowledge is encoded in a sheaf. Modelers can also update the restriction maps of a sheaf. This makes a local change to the sheaf, and the resulting cohomology can likely be updated incrementally; Prof. Robinson expects to discover an incremental algorithm to update cohomology when one of the restriction maps of a sheaf is changed.

3.8. Scalability, modularity, and extensibility. CHCLT will integrate heterogeneous information provided from multiple sources. For numerical and structured data, the methods to provide these are straightforward, simply ingesting data files or providing simple input filters. But for semi-structured and structured text (see more discussion in Section 3.9 below about some particular cases) some form of semantic extraction is required (see discussion above in Section 3.4). This process will begin by identifying certain fields in the structured data (e.g. CSV or XML files) which have semantic overlaps with some of the fields in the semi-structured documents, and some which have facts in the free text. Specific extractor modules will then be developed to the needs of the notional use cases.

The goal of CHCLT is to *consume* extracted facts, not to build or research extraction methods themselves. Thus this step presents both a cost and a potential computational challenge as well. First, semantic resources (such as vocabularies or ontologies) helpful in supporting extraction need to be acquired or modestly modified. One potential example relevant for Notional Use Case 2 is the Transportation Ontology³, representing transportation-related information in the CIA World Fact Book. Matching of terms from the structured data to terms in the semi-structured data and free text introduces some ambiguity. For example, “vehicle” is referenced in the structured data, while more specific types of vehicles are referenced in the semi-structured data and free text. This can be addressed through semantic resources such as the lexical database WordNet⁴, which includes hypernyms, hyponyms, and synonyms in a tree structure.

Given needed knowledge resources, algorithms and code needed to build specific extractors will be exclusively off-the-shelf. Extracting events from the free text can be accomplished with existing methods developed by PNNL. WordNet synonyms are used to extract events of a similar type, and statistical classification based on Naive Bayes or k -Nearest Neighbors with syntactic features is used to refine the event mapping. In-domain annotated training data would be very useful for this task.

Even within free text the level of difficulty posed can vary greatly. Formal reports usually have a number of features that aid in understanding. In such reports authors usually use a discernable structure and organization as well as well-formed grammar and spelling. They also exploit a specific and constrained vocabulary that is appropriate to the domain of discourse. As a result, formal reports are often amenable to lower effort and higher accuracy approaches. Informal text

³<http://www.daml.org/ontologies/409>

⁴<http://wordnet.princeton.edu>

is often more problematic because it lacks some or all of these supporting features. This text may contain extensive use of personalized or idiosyncratic shorthand including acronyms, jargon, sentence fragments, and numerous misspellings. As a result, extraction of specific targeted information from such texts can be much more resource intensive and still yield lower quality results.

Beyond these general observations, a number of specific techniques can also be used to increase accuracy and computational efficiency as needed. Often, knowledge can be extracted or bootstrapped in an accurate manner through a multi-level approach ([27]). The challenge to do extraction well is to have a fairly broad spread coverage without sacrificing the accuracy of extraction. In the case of the multi-sensor experiment described in Section 3.9 below, the computational techniques operate on a limited portion of the knowledge in the sources, and some degree of missing data can be tolerated.

Much of the structured and semi-structured data contains measurements for temperature, length, etc. These measurements can be mapped to measurements in the free text through direct matches of units and keywords as related through a dependency parse. To combat ambiguity, dictionaries of measurements, conversions, and meanings (e.g. centimeters are a measurement of length) can be used.

These techniques result in a set of computational models and resources with improved scalability. The extraction on a larger number of documents can often be performed as a parallelizable process of independent document extractions. Furthermore, the extraction techniques are typically set up in a series of modular components structured so as to extract structured and unstructured text, map the resultant extraction, and then finally disambiguate in the resultant knowledge representation. PNNL has considerable experience implementing such systems in a scalable, domain independent manner ([13]).

3.9. Notional use case 2: multi-sensor experiment analysis. Having introduced the basics of sheaf-theoretical modeling in the context of our first pedagogical use case, we now describe a second use case which is equipped with rich heterogeneous data. Our ability to encode this use case in a sheaf-theoretic manner and demonstrate initial datafication will provide critical verification of the soundness of our approach.

The use case concerns exercises for radar tracking and classification of vehicles in controlled tests performed by SRC Inc. and the Air Force Research Lab (AFRL).

In order to determine sensor and processing capabilities, controlled test and demonstration events are planned and executed. SRC has extensive experience with all stages of this process, and has data from AFRL available for immediate use from the Gotcha 2008 experiment as well as others. Although the experiments in question were tailored to answer specific questions about the performance of specific systems, the expense of conducting an experiment usually means that several experiments are conducted simultaneously with the same sensor platform.

Not all of these experiments go as planned, and it is not always clear exactly what happened on the ground, since the sources of ground truth are not always reliable. Additionally, unexpected or opportunistic phenomena play an important role in the capabilities discovery process, for instance

- Operator errors in tasking the sensors,
- Uncontrolled targets,
- Weather phenomena, and
- Mistakes in executing the experimental plan by the experimental team, to name just a few.

Although these sources of deviation from the experimental plan sometimes prove detrimental to the scientific value of the experiments, they are also serendipitous. Sadly, many experimental budgets do not include adequate resources for thorough post-experimental analysis. CHCLT could remedy this deficiency by identifying opportunistic aspects of these experiments – those that were unexpected deviations from the plan or models, yet have sufficient fidelity to be useful.

In order to do this, it is necessary to identify the correct sheaf-theoretic framework for encoding locations, behaviors, and intents of moving targets, develop appropriate models of sensor interaction, sheafify the datasets and their interaction into a sheaf-based datastructure, analyze the inferences that can be made using sheaf operations to demonstrate their decisive value on sensor data, and recommend which opportunistic aspects of the data should be studied in more detail.

The main two queries of CHCLT map directly to this setting, in the following way:

- (1) Detection: Given a description of a target or its behavior and some measurement data can we certify that a target is or is not present? Once a set of targets have been identified in the measurement data, can we trace them back to a semantic report of the behaviors they exhibit?
- (2) Scheduling the next best collections: Given a set of information sources, what additional sensing or semantic resources should be deployed in the next experiment to best constrain the observations to a particular target type or behavior?

The analysis of high-quality, well-understood experimental datasets that are available to our team will play an important role in constructing the pipeline described earlier in this proposal. These datasets were collected using a variety of different modalities, and therefore detect different properties about the targets in scene. They have a large spatial diversity, with a variety of background land covers and land uses, and the behaviors demonstrated by the targets are similarly diverse. Since the data contain high resolution synthetic aperture radar (SAR) and optical images with repeat passes of the same scene including changes, there is a large volume of data (gigabytes).

We plan to leverage existing datasets available from AFRL through a subcontract with SRC. These datasets will provide ground moving target indication (GMTI) tracks, SAR images, optical images, and analysts' metadata tags collected throughout the experimental collection process as shown at left in Figure 10. The right panel of Figure 10 shows an excerpt of AFRL's Gotcha 2008 collection over Wright Patterson Air Force Base that is already available to SRC, and is processed to a 3×3 km full-backprojection SAR image at 0.5 m resolution. In addition, SRC will supply the domain knowledge necessary to interpret such datasets. This processing will leverage Air WASP – an AFRL funded program – which is an extensible, distributed, scalable, real-time,

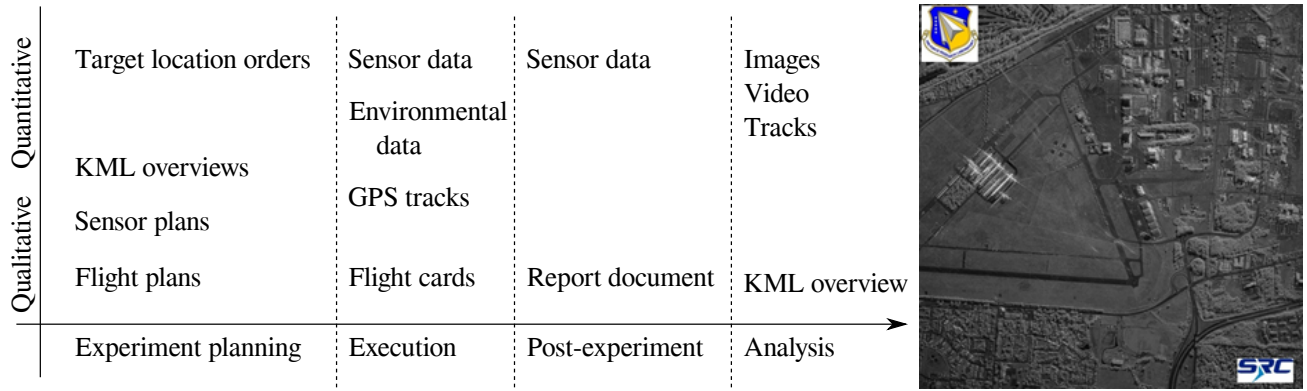


FIGURE 10. Timeline of experimental data sources for Use Case 2 (left); An X-band SAR image from the AFRL Gotcha 2008 collection over Dayton, OH (right)

processing framework for HPC systems. Air WASP is a very fast, configurable, software framework that users can utilize to rapidly create and multi-threaded, multi-node, multi-platform, distributed applications.

Although we plan to address Technical Area 1, we will be extremely well-positioned to datafy the datasets we have available, and therefore obtain a sheaf that encodes all relevant data. As part of this implementation, we will write the appropriate software necessary to parse the raw data and convert it into the native format used by our computational sheaf library developed in support of Technical Area 1.

The data sources that are either immediately available or presumed to be easily obtained for the AFRL Gotcha 2008 collection include the sensor data (quantitative), experiment planning materials (qualitative, including semi-structured and free text), and various reports (qualitative, mostly free text). The qualitative data varies considerably in fidelity. The target truth data is usually very detailed, for example:

- Target A1: Blue Toyota Camry
- Target A2: Desert Camouflaged ZIL-135 Military Transport

Target detections and classifications are derived from the sensor data, and are much less detailed

- Target 1: Small Car
- Target 2: Large Vehicle

In addition to target type, target information typically includes position, velocity, size (for instance “small”, “medium”, or “large”), and category (“car”, “truck”, “transport”).

Due to expense limitations, the number of targets that are carefully scripted and tracked by GPS is quite small – typically less than 30. However, the Gotcha 2008 collection is centered on Wright Patterson AFB and contains normal base traffic. This results in roughly 3500 moving targets of various types over the 4 hours of collection data. These largely unscripted targets

exhibit considerable variability in their intentions, behaviors, and overall activities. CHCLT would present an extremely flexible framework for studying models of target behaviors.

Classifier errors are a substantial problem when there are multiple targets in the scene [6]. Since the Gotcha 2008 dataset contains a number of different target types, it is possible to use that data to detect and possibly correct classifier errors. The two CHCLT queries can be used to identify situations where classifier errors may be further confounded due to topological features. For instance, consider an experiment plan that specifies that a large vehicle will traverse a particular path during the experiment. The experiment plan does not specify the particular kind of vehicle, but for the purpose of estimating a link budget specifies a nominal radar cross section (RCS). During the experiment, a particular vehicle is selected, surveyed with a GPS during its maneuver, imaged by a SAR, and tracked by a GMTI radar. This situation results in the base case shown in Figure 11. This base space consists of four triangles that form the boundary of a hollow three-dimensional simplex. These triangles represent the following combinations of sources:

Front left face: (Quantitative: asynchronous target path) Although the experiment plan usually does not specify the exact timing of the maneuver (down to the seconds), it does specify the locations fairly accurately. This allows the experiment plan to be compared with the GPS track and the GMTI track, but *not* the SAR image, which only has a specific location at the time of imaging.

Front right face: (Quantitative: RCS) The experiment plan, the SAR image, and the GMTI image all include nominal RCS measurements for the target. However, the GPS data will make no such assessment.

Back left face: (Qualitative: target type) The experiment plan and GPS track will list the target type (normalized into a common ontology, though the experiment plan will in general be less precise about the target), and a SAR target classifier will also resolve the target type into a broad classification as listed above.

Back right face: (Quantitative: instantaneous target location) The SAR, GMTI, and GPS are all synchronized to a common time reference, so an assessment of the target's location at the time of the SAR imaging can be made using these three sources.

Notice that there is no common measurement across all information sources in this example, which results in a hollow void in the base space. The presence of this topological feature means that certain kinds of classifier errors can go undetected. For instance, if the experiment plan stipulates that a convoy of similar vehicles move along a common path, it will be difficult to determine which vehicle in the scene corresponds to which vehicle in the plan.⁵

As this simple example shows, sheaves can be used to analyze dissimilar data types (radar, tracking, and plans) to support inferences about experiment quality and opportunistic aspects of the experiment. Although it is easy to locate the void in this example by merely by inspection

⁵Prof. Robinson has recently worked with a dataset exhibiting a maritime scenario with comoving targets – if the topological assessment performed by CHCLT would have been used before the experiment was executed, the data would have been considerably more valuable.

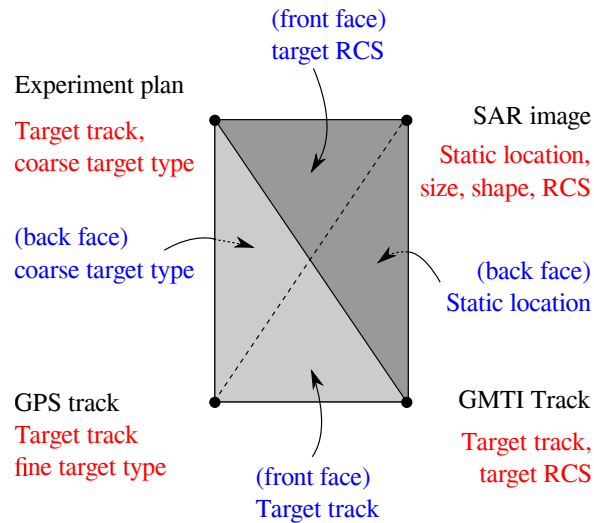


FIGURE 11. Notional use case 2: example base space for identifying classifier errors. Notice that the base space is a *hollow* tetrahedron, due to the lack of common timing between the experiment plan and the sensor data

– because it represents only one target – a similar void may be difficult to locate in more complicated experiment plans. Since CHCLT will perform topological computations algorithmically, it will provide the capability to identify these and more subtle issues with experimental data.

3.10. Interaction with teams focused on related areas. Our team is eager to collaborate with other teams working on both technical areas, and would be happy to host or visit other teams as appropriate to ensure overall program success. Additionally, since sheaf theory is not common within engineering practice, Prof. Robinson offers to provide tutorial sessions on sheaf theory at the SIMPLEX workshops or meetings. These tutorial sessions would be specifically tailored to explain the process of encoding knowledge in sheaves, and then interpreting the results. Prof. Robinson is currently preparing tutorial material to teach his own students applied topological methods, and recently graduated a student who wrote a master’s project in applied sheaf theory. Additionally, the team has a number of datasets that will be used to develop CHCLT and would be willing to share these to the extent that their particular distribution limits allow.

3.11. Technical risks and challenges. The primary technical hurdle in the proposed effort is the specification of the translations between datatypes, since these require intimate domain knowledge. To mitigate this risk, we plan to employ data sources that the team understands well in order to focus our development. SRC participated in all aspects of the AFRL Gotcha 2008 experimental effort that produced the track and imagery results we plan to use, and understands the data better than any other organization. This data is immediately available for use on this

program. However, there is a growing body of literature on “hard/soft fusions” [14] that will support our development.

We recognize that measuring consistency between disparate data sources may never result in perfect matches – there are both representational errors and noise present in real data. Prof. Robinson and his PNNL collaborators have developed the concepts of persistent sheaf cohomology and of approximate sections, which will allow the management of error budgets within a sheaf. In order for this to succeed, additional domain knowledge is needed to develop performance metrics for determining the quality of a match between data sources. Additionally, both approximate sections and persistent sheaf cohomology will need to be computed efficiently, though it is yet unclear how to perform these computations.

Most of the existing topological literature on sheaves is from the quantitative perspective, so applying these tools to qualitative data sources presents a new challenge. However, Prof. Robinson has found the technique of categorification described in Section 3.4 to be quite effective for boolean data [30]. Additionally, there is a formidable literature on the use of sheaves in exclusively qualitative settings – that of *topos theory* – which mathematically proves that sheaf theory is sufficiently expressive to encapsulate both traditional and nontraditional propositional logics [11]. Although this full expressive power is certainly unnecessary to ensure program success, it indicates that sheaf theory is the correct approach.

CHCLT focuses on integrating semantic data, which realistically requires semantic extraction. This can be very hard in general from free text, and CHCLT runs the risk of computational burdens. PNNL has extensive expertise in semantic processing, and will lead the development of the aspects of CHCLT that require textual ingestion. They are well positioned to mitigate risks by focusing on off-the-shelf technologies and exploiting semantic constraints in vocabularies and semi-structured files wherever possible. To complement extensive structured data (CSV and XML), we expect that far short of full NLP parsing, topic and keyword extraction will be sufficient to provide enough fidelity for strong inferences to be possible.

Prof. Robinson has the only known implementation of a persistence-capable sheaf library⁶. We will scale this to manipulate the larger datasets we have available. All three team members (AU, PNNL, SRC) have access to high performance computing resources, which should mitigate initial concerns of scalability as well. For image processing, we will leverage the baseline version of SRC’s Air WASP system. This provides real-time SAR processing using GPGPUs and multi-core CPUs and has been demonstrated on a variety of HPC and High Performance Embedded Computing (HPEC) systems without any modifications to code. Algorithms written in a variety of languages including Python, C/C++, or Matlab can be wrapped in the Air WASP distributed process component class.

4. CAPABILITIES

Our team has a broad range of expertise in several scientific and engineering domains, including the mathematics of sheaves. Prof. Robinson has written the only existing computational

⁶<https://github.com/kb1dds/pysheaf>

sheaf library, which will be leveraged to start the development of software on this project. Additionally, both subcontractors have extensive experience manipulating domain-specific data that will be used to develop and test our approach in advance of collaboration with other teams. This will reduce the risk of unanticipated effects from hidden assumptions in our approach.

Prof. Robinson's work is characterized by an aggressive application of sophisticated mathematical techniques to detailed, practical models of systems. He is uniquely qualified to undertake the work in this proposal, which requires combining practical models with sophisticated mathematical techniques. No other researcher combines extensive practical experience with the abstract theory of sheaves to develop both novel theory and effective prototype systems. He has found many new examples of practical, theoretically-motivated algorithms, and has demonstrated them on laboratory systems of his own design. Recently, he has completed two federally-funded projects that successfully applied sheaf theory to engineering problems, and developed new algorithms as a result.

PNNL has world-leading expertise in information visualization and semantic technologies. Joslyn and Hogan represent leadership on multiple projects in information integration for Battelle, DOE, and DoD, focusing on both foundational methods and domain-specific applications, including cyber systems, bioinformatics, and intelligence analysis. Technical support includes institutional and supercomputing facilities.

SRC is nationally recognized for the development and application of innovative radar, EW, and communication systems technologies. For over 50 years, SRC has performed research and development, primarily for the DoD and national intelligence agencies. SRC possesses an end-to-end RF system development capability, including concept formulation, system analysis and simulation, design, fabrication, and field evaluation, manufacturing, and support. SRC engineers have played key roles in developing over 50 ground, airborne, and space-based RF systems over the past 30 years, supporting a wide range of C4ISR and EW applications. SRC is successful in part because we take a modular approach to hardware and software design for reuse using well-established COTs standards (e.g., VME, VPX) and non-proprietary interfaces.

REFERENCES

- [1] Glen Bredon. *Sheaf theory*. Springer, 1997.
- [2] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.
- [3] J. Curry. Sheaves, cosheaves and applications, [arxiv:1303.3255](#). 2013.
- [4] V. de Silva and G. Carlsson. Topological estimation using witness complexes. In M. Alexa and S. Rusinkiewicz, editors, *Eurographics Symposium on Point-based Graphics*, 2004.
- [5] Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(339-358):24, 2007.
- [6] Joint Interoperability AT&L/DDR&E Director. Advanced mathematics for dod battlefield challenges, 2010.
- [7] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.
- [8] Jérôme Euzenat, Alfio Ferrara, P Hollink, A Isaac, Cliff Joslyn, Véronique Malaisé, Christian Meilicke, Andriy Nikolov, Juan Pane, Marta Sabou, Francois Scharffe, Pavel Shvaiko, Vassilis Spiliopoulos, Heiner Stuckenschmidt, Ondrej Iváb-Zamazal, Vojtech Svátek, Cássia Trojahn dos Santos, George Vouros, and

- Shenghui Wang. Results of the ontology alignment evaluation initiative 2009. In *Proc. 4th Int. Wshop. on Ontology Matching (OM-2009)*, CEUR, volume 551, 2009.
- [9] Jérôme Euzenat and P Shvaiko. *Ontology Matching*. Springer-Verlag, Hiedelberg, 2007.
 - [10] R. Godement. *Topologie algebrique et théorie des faisceaux*. Herman, Paris, 1958.
 - [11] Robert Goldblatt. *Topoi, the Categorical Analysis of Logic*. 2006.
 - [12] Jacob L Graham, David L Hall, and Jeffrey Rimland. A coin-inspired synthetic dataset for qualitative evaluation of hard and soft fusion systems. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011.
 - [13] Michelle L Gregory, Liam McGrath, Eric Belanga Bell, Kelly O’Hara, and Kelly Domico. Domain independent knowledge base population from structured and unstructured data sources. In *FLAIRS Conference*, 2011.
 - [14] David L Hall, Michael McNeese, James Llinas, and Tracy Mullen. A framework for dynamic hard/soft fusion. In *Information Fusion, 2008 11th International Conference on*, pages 1–8. IEEE, 2008.
 - [15] Joachim Hammer, Hector Garcia-Molina, Junghoo Cho, Rohan Aranha, and Arturo Crespo. Extracting semistructured information from the web. 1997.
 - [16] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.
 - [17] John H. Hubbard. *Teichmüller Theory, volume 1*. Matrix Editions, 2006.
 - [18] B. Iverson. *Cohomology of Sheaves*. Aarhus universitet, Matematisk institut, 1984.
 - [19] Jeffrey Johnson. *Hypernetworks in the science of complex systems*, volume 3. Imperial College Press London, 2013.
 - [20] Cliff Joslyn, Emilie Hogan, and Michael Robinson. Towards a topological framework for integrating semantic information sources. In *STIDS 2014*, 2014. accepted.
 - [21] Cliff Joslyn, Patrick Paulson, and Amanda White. Measuring the structural preservation of semantic hierarchy alignments. In *Proc. 4th Int. Wshop. on Ontology Matching (OM-2009)*, CEUR, volume 551, 2009.
 - [22] V. Kostrykin and R. Schrader. Kirchhoff’s rule for quantum wires. *J. Phys. A: Math. Gen.*, 32:595–630, 1999.
 - [23] Peter Kuchment. Graph models for waves in thin structures. *Waves in random and complex media*, 12(4):R1–R24, 2002.
 - [24] Nicholas Kushmerick. Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118(1):15–68, 2000.
 - [25] Svetoslav Markov. Towards an axiomatization of interval arithmetic. *PAMM*, 6(1):685–686, 2006.
 - [26] Stanislav Molchanov and Boris Vainberg. Transition from a network of thin fibers to the quantum graph: an explicitly solvable model. In *Quantum graphs and their applications*, pages 227–240, 2006.
 - [27] Ellen Riloff, Rosie Jones, et al. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479, 1999.
 - [28] M. Robinson. The Nyquist theorem for cellular sheaves. In *Sampling Theory and Applications (SampTA)*, 2013.
 - [29] M. Robinson. *Topological Signal Processing*. Springer, 2014.
 - [30] Michael Robinson. Asynchronous logic circuits and sheaf obstructions. In *GETCO 2010*, January 2010.
 - [31] Michael Robinson. Inverse problems in geometric graphs using internal measurements, [arxiv:1008.2933](#), 2010.
 - [32] Sunita Sarawagi. Information extraction. *Foundations and trends in databases*, 1(3):261–377, 2008.
 - [33] Uzy Smilansky and Michael Solomyak. The quantum graph as a limit of a network of physical wires. In *Quantum graphs and their applications*, pages 283–292, 2006.
 - [34] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.

DEPARTMENT OF MATHEMATICS AND STATISTICS, AMERICAN UNIVERSITY, 4400 MASSACHUSETTS AVE
NW, WASHINGTON, DC 20016
E-mail address: michaelr@american.edu