Towards A Topological Framework for Integrating Semantic Information Sources

Cliff Joslyn
Data Sciences and Analytics
Pacific Northwest National Laboratory
Seattle, WA 98103
Email: cliff.joslyn@pnnl.gov

Emilie Hogan
Computational Mathematics
Pacific Northwest National Laboratory
Richland, WA 99352
Email: emilie.hogan@pnnl.gov

Michael Robinson
Mathematics and Statistics
American University
Washington, DC 20016
Email: michaelr@american.edu

***Abstract*—In this position paper we argue for the role that Topological Data Modeling (TDM) principles can play in providing a framework for sensor integration. While used successfully in standard (quantitative) sensors, we are developing this methodology in new directions to make it appropriate specifically for semantic information sources, including keyterms, ontology terms, and other general Boolean, categorical, ordinal, and partially-ordered data types. Given pairwise information source integration principles, TDM can measure overall consistency, and most importantly, reveal cyclic dependencies amongst data sources where conflicts might not be able to be identified. We illustrate the basics of the methodology in an extended use case/example, and discuss path forward.**

I. Introduction and Motivation

There is a need to develop systems to establish situational awareness of events based on multiple real-time information feeds. Information about a typical public event may be available from published news reports, cameras, audio streams, eyewitness blog posts, public twitter feeds, and police information. What are the characteristics of such an information integration problem? What is the significance for modeling choices of the fact that (some of) these input feeds are of a semantic nature? How can we then integrate such multiple feeds to form a holistic operational picture of the relevant situational characteristics, such as participants, identities, attitudes, and preferably content? How do we assess consistency of data values given overlapping measurements (different units, vocabularies, numerical types)? How can we identify locally or globally consistent or inconsistent data, or regions of the network where such conflicts may not be able to be identified?

At present there are no rigorous mathematical techniques deployed to integrate qualitative, semantic information (e.g. from textual analytics) with traditional quantitative signals and sensors data [7]. But there is a provably well-justified mathematical approach to approach this problem. We propose Topological Data Modeling (TDM) techniques from topology, combinatorics, and category theory to address problems in information integration, extending to semantic data sources. The mathematical tools of TDM include machinery like topological spaces, set systems, cell complexes, simplicial complexes, delta complexes, homology and co-homology, and sheafs and co-sheafs to represent both the properties of each analytic, and, most importantly, their pairwise and multiway interactions.

While initially proven to be tremendously valuable in traditional signal analysis (e.g. radar networks or collections of optical cameras [9]), TDM methods using persistent homology, finite topology, and sheaf theory are increasingly penetrating data analytics and knowledge discovery [3]. Extensive theoretical work in sheaf theory [6] leads to powerful detection and inference methodologies in the abstract. When we cast applications into sheaves, the theory does the work of providing a systematic, algorithmic way to globalize data. These methods promise the possibility of critical new capabilities, including:

- Modeling the structural connectivity of information networks, representing multi-way interactions and information overlaps among heterogeneous sensors.
- Modeling the data content flowing within such networks, so that given knowledge of each sensor individually, and knowledge of locally consistent interactions, it can be possible to automatically generate a candidate global view of the integrated sensor network.
- Measuring the overall topology of the resulting network, providing critical information about where cyclic dependencies may hide latent inconsistencies.
- And finally, measuring a network's sensitivity to variation, perturbation, or reliability of the constituent sensors and their connections.

We abstract the concept of a "sensor" from an instrument generating a quantified signal to a generic information process returning a stream of observations, either direct measurements, derived measurements, or the output of an analytic process. We then cast "semantic sensors" more specifically as computational analytics which return *symbolic* information such as keywords, topics, handles, hashtags, proper nouns (individuals, groups, places), and sentiment, including polarity and intensity.

Mathematically, it may be sufficient to distinguish semantic sensors as those whose data types are less than numeric or scalar (integers or real-values, or vectors of these types). This would include Boolean values (polarity), categoricals (keywords), small cardinality ordinals (intensity, sentiment, e.g. high, medium, low), partially ordered entities (ontology nodes), and semantic graph nodes (social network role). Semantic sensor data live in mathematical spaces which are relatively impoverished compared to the richer vector spaces or higher-order structures normally used in TDM. While this has made penetration of topological approaches into data analytics more difficult, more modern methods like sheaf theory, and its use of "categorification", do have the ability to build the needed machinery to support topological representation of these simpler structures. This provides a method to build

integrated information networks which combine semantic and quantitative data in a principled way.

More strongly, TDM promises a mathematical approach which is not only sound, but axiomatically necessary, in that theorems indicate that *any* methods for consistency-checking and global modeling of linked sensor networks will recapitulate these TDM methods [1]. TDM promises to support a range of new capabilities such as 1) automatically generating a global model of how sources can be integrated; 2) assessing consistency within the model; 3) measuring the degree of fit of the two models given only partial information about each; and 4) testing for sensitivity with respect to the presence, absence, or credibility of certain sources.

II. Expositional Use Case

Our version of TDM proceeds by specifying some situation in the world about which we have some questions; and where there are many information feeds, "sensors", or "analytics" of different modalities (text, numeric, symbols, ontologies, places) which inform those questions. We require that the user specify only the mathematical form of each input, a mathematical mapping between them pairwise, and which sensors inform which world variables. So while no free lunch, TDM has the ability to handle both quantitative and semantic sources. TDM methods then promise the ability to calculate global and local consistencies. Additionally, and quite importantly, a topological analysis can identify cyclic dependencies amongst information sources, around which it may not be possible to resolve such inconsistencies, requiring intervention or recognition from the modeler.

We now introduce the following true story to drive the example information network. Appropriate for a short position paper, this example was deliberately constructed to be realistic while also illustrating the most important features of our TDM approach. Technical details and example data analysis will await a larger paper in another venue.

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.[1] While a fine time was being had by all that evening, one of us (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 additionally had access to overhead video, police scanner audio, Seattle urban transit cams at major intersections, and the feed from the Seattle Times. Fig. 1 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.

$P =$ **Place:** A categorical variable like "1st and Pine" or "Broadway".

$I =$ **Intensity:** An ordinal variable: "high", "medium", "low".

$L =$ **vioLence:** A Boolean variable: "present" or "absent".

$R =$ **Role:** Another categorical variable, reflecting the kind of person present, for example "protester", "police", "bystander", or "press".

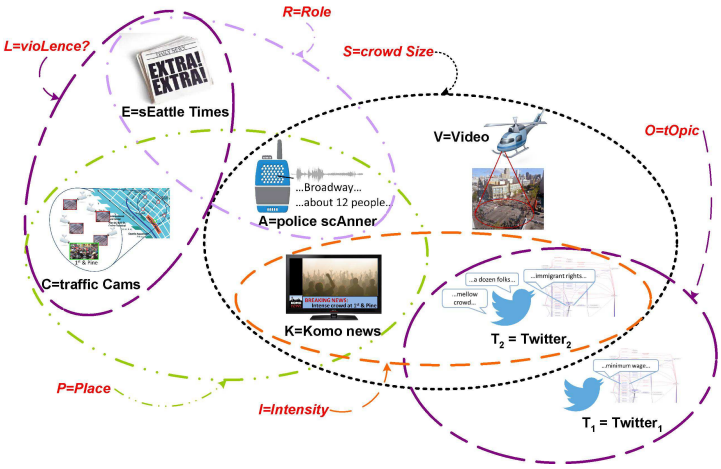

Fig. 1. An information integration scenario: multiple sensors partially informing multiple state variables.

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

A**= 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**= transit Cameras:** Cameras at specific intersections can show when the crowd has reached those locations, and whether violence is present.

E**= sEattle Times:** An anlaytic 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.

$K =$ **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.

$T_1 =$ **Twitter**$_1$**:** A text analytic extracts keywords to identify protest topics.

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

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

We model the sensors 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 **sensors**. Then table I shows the relationships between these sensors and the state variables they inform. We cast Table I as a binary relation $B \subseteq X \times Y$. Then Fig. 2 shows B as a set system (undirected hypergraph) $B(X) \subseteq 2^Y$ on the sensors Y. The variables $x \in X$ (i.e., the columns of B) are represented (in red) as subsets $B(x) \subseteq Y$ of the sensors (in black) which inform them.

[1] http://www.huffingtonpost.com/2014/05/02/seattle-may-day_n_5253707.html

	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	√					

TABLE I. SENSOR STRUCTURE.

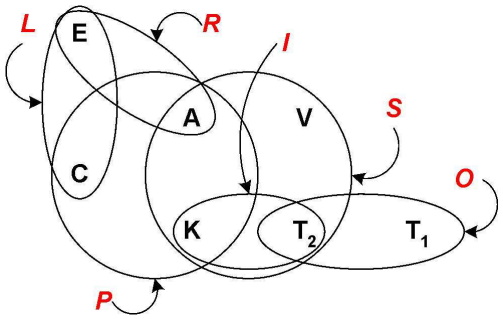

Fig. 2. Representation of sensor structure B as a set system.

Fig. 3 shows $B(X)$ as a combinatorial structure called an "abstract simplicial family" [4] 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 sensors 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 $\{A, C\}$).

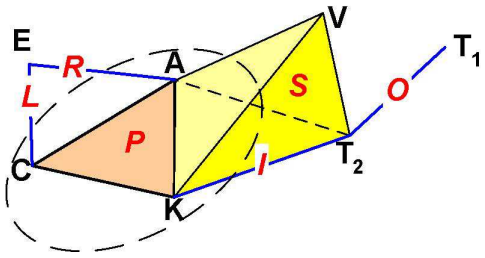

Fig. 3. Representation of B as a simplicial family. The dashed ellipse is exploded in Fig. 4.

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). The characteristic f-polynomial $x^3 + 5x^2 + 11x + 7$ of the ASC serves to enumerate the counts and dimensionalities of all 24 faces present, not just the "listed" ones read off the table: one (3D) tetrahedron, five (2D) triangles, eleven (1D) edges, and seven (0D) vertices. Abstraction to an ASC allows easy tracking of all k-way interactions dually amongst sensors and variales. 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 sensors.

We can represent the 24 faces (interactions) distinctly, but for brevity, we only show the 7 associated with the variable P in a "sheaf" diagram in Fig. 4. Here each node shows some combination of the sensors A, C, and K above (black), and the corresponding variables they inform below (red).

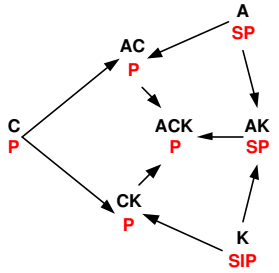

Fig. 4. Sheaf representation of the sensors A, C, K informing the variable P. The arrows are functions transforming data on faces into a common form.

The sheaf in Fig. 4 shows not just all the combinations of sensors, but how they can be mapped into each other to measure consistency. Continuing our drilldown, Fig. 5 shows this in detail for just the $A \mapsto AK$ edge of the sheaf, showing the sensor A (the police scanner) in interaction with the sensor K. Since sensor A reads off in crowd size, role, and location, this is in the form of a three-way data tensor as shown on the right. A and K share only S and P in common, so the matrix projects over R and aggregates S into the two-way tensor shown on the left, reported up from K. Here we can see that there is a match at City Hall (20 total people); a match for Main Street (5 police and 5 dozen bystanders yields 65 total people, which is in the interval $[50, 100]$); but finally a failure at Broadway ($26 \notin [5, 10]$).

Note how the semantic information is encoded in the various linear objects. Through the process of "categorification", the semantic variables of role and place (both categoricals) have unique positional assignments, as reflected in the block structure of the central matrix, called the "restriction map". This kind of categorification supports the integration of quantitative data with the mathematically weaker data types typically used for semantic information.

An assignment of data to the sensors which yields consistency over some of the faces is called a "local section" over

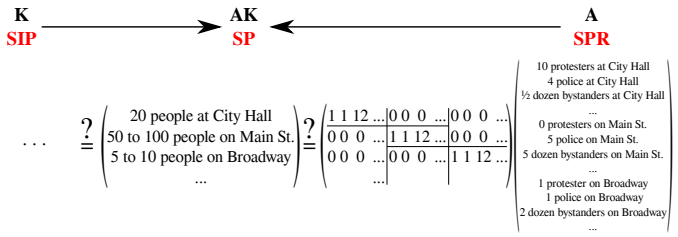

Fig. 5. One edge of the sheaf in detail, showing the mapping of the sensor A against K, checking for consistency.

those faces. Fig. 6 shows a local section over the AC edge and the isolated vertex K, but no data linking AC to K, which is just reporting the weather. Fig. 7 shows a global section over the whole P triangle, indicating agreement of all Place sensors. Both the degree of *consistency* and the degree of *completeness* can thus be measured over this whole portion of the sheaf.

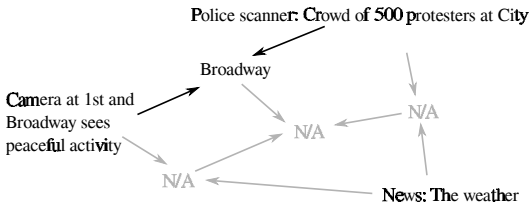

Fig. 6. A local section only over AC and K.

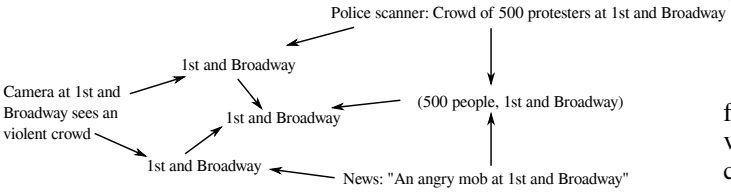

Fig. 7. A global section over all sensors informing P.

Can this approach be expanded across the entire ASC to measure consistency globally? The A, C, E triangle has no three-way interaction, only the three pairwise interactions. Thus fixing data at one vertex (say A), can constrain another (say C), which in turn can constrain E, but there is no way to guaurantee that there will then be final consistency needed between E and A. Knowing this loop is present is critical to the modeler, and identifying its presence (and those of more complexity) involves calculating the homology of the ASC, or the co-homology of the sheaf. Once identified, the modeler can be informed of the risk, and allowed to mitigate or address it.

III. Path Forward

This small position paper is intended to evoke the spirit and flavor of our TDM approach to semantic information integration. The path forward to a more complete expression of this idea obviously begins with encoding realistic datasets to demonstrate operaion of actual algorithms in our example.

We are showing Boolean methods for local sections: quantities and qualities either match exactly, or satisfy some crisp condition like $65 \in [50, 100], 26 \notin [5, 10]$. We are also exploring a mathematical theory of "approximate sections", which could provide more robust inferences in the presence of uncertainty. We will establish distances between numerical and non-numerical quantities, which can be aggregated to provide a quantitative degree of match. We will then additionally require the user specify distances measures between data types in addition to their types. In the case of fully semantic data, like the keyterms of ontologies, we could use order metrics on class hierarchies [8], which we have previously established in the context of ontology management [5].

Where sheaves provide a bottom-up view of integrating existing sensors covering certain variables, their dual "co-sheaves" (where the arrows of Fig. 4 are reversed) support "world models" which can specify the structure of sensors *needed* to cover variables of interest (see Fig. 8). Linear duality between sheaves and cosheaves corresponds to the duality between sensor-centric and world-centric modeling disciplines. Recent results on "sheaf and co-sheaf duality" [2] allow construction of explicit joint world/source models, so given a partial world model and a partial source model we may measure degree of fit and seek sensitivity analysis to source variations using "topological persistence".

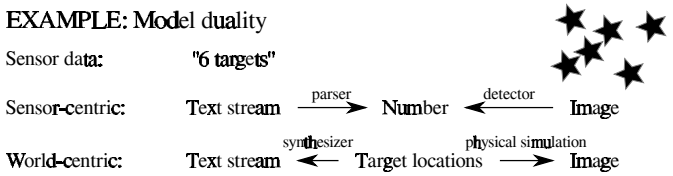

Fig. 8. A sensor-centric sheaf model of a text-image integration, together with its dual world-centric model as a co-sheaf.

IV. Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings contained in this article/presentation are those of the author(s)/presenter(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited). This is PNNL-SA-105411.

References

[1] G. Bredon. *Sheaf theory*. Springer, 1997.

[2] Justin M Curry. Sheaves, cosheaves, and applications, 2014. http://arxiv.org/abs/1303.3255

[3] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:1:61–75, 2007.

[4] Jeffrey Johnson. *Hypernetworks in the Science of Complex Systems*. Imperial College press, 2013.

[5] 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.

[6] M. Kashiwara and P. Schapira. *Categories and sheaves*. Springer, 2006.

[7] B. Khaleghi, A. Khamis, and F. Karray. *Multisensor data fusion: a review of the state-of-the-art*. Information Fusion, 2011.

[8] Chris Orum and Cliff A Joslyn. Valuations and metrics on partially ordered sets, 2009. http://arxiv.org/abs/0903.2679v1

[9] M Robinson. *Topological Signal Processing*. Springer-Verlag, 2014.