HIGH PERFORMANCE DATA ANALYTICS

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Topological Data Modeling

CHALLENGE

We will address the challenge of analyzing big data with high complexity by demonstrating the value of Topological Data Modeling (TDM). We will address the scalability of TDM, both in terms of data volume and complexity using these new methods.

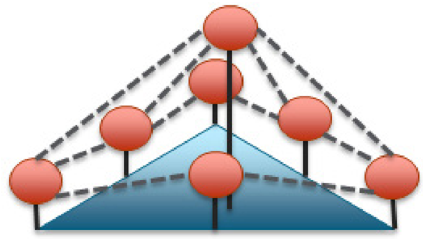
CURRENT PRACTICE

Graphical models (labeled nodes and arcs) are used to represent heterogeneous data with sparse connectivity. Graphs can scale massively, but they represent only simple interactions: an edge connects two nodes. Graph representations of complex relationships, where multiple entities are engaged in many relationships, require "reification" strategies to represent them as collections of binary edges. This introduces vast increases in graph size while additionally losing information.

TECHNICAL APPROACH

TDM is at the intersection of the mathematical fields of topology, combinatorics, and category theory and naturally represents information systems with complex, multi-way interactions. Because of this, our models are able to identify the topological properties of information systems, i.e., those of the highest-order and most robust structures present.

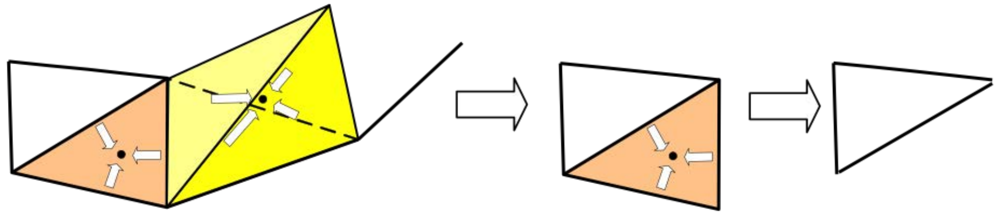
Analyze big data with high complexity using topological data modeling.



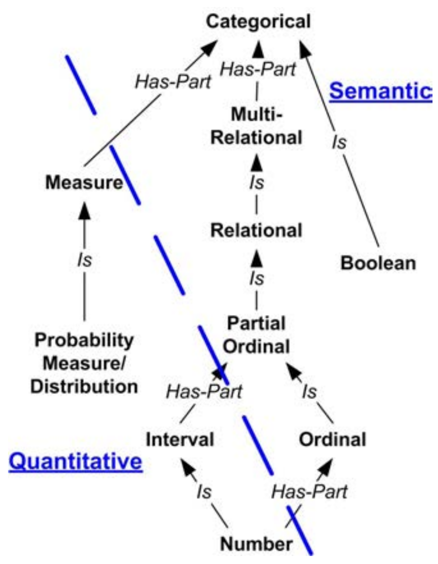
Simplicial sheaves model the data integrated on interacting collections of sensors.

First, TDM uses abstract simplicial complexes (ASCs), which are higher-order graphs where multi-way connections naturally represent both the k-way interacting groups in a data source and how those groups are connected structurally. Regular graphs stand as special cases of ASCs, where only binary interactions are included, and graph theoretical measures, such as degree and connectivity, generalize to ASCs.

Second, ASCs support measuring topological properties, such as homology, which reveal the highest order shapes of the data, e.g., their connectivity and loopiness. Most importantly, they reveal when portions of the database can be reduced or ignored, greatly increasing efficiency without destroying the most important information.



Morse reductions of an abstract simplicial complex representing complex interactions among heterogeneous input data.



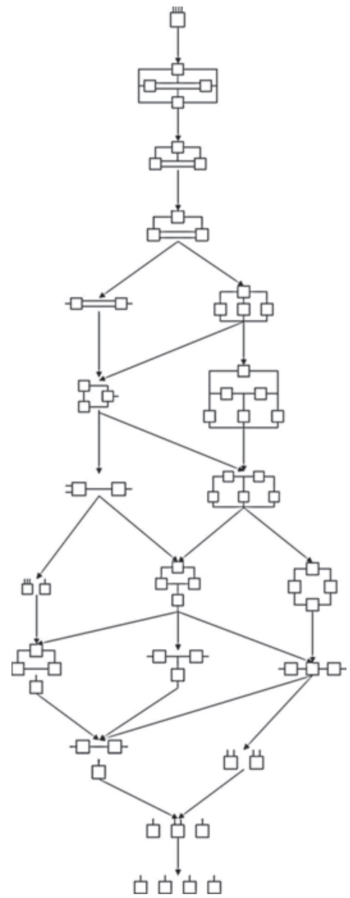
Hierarchy of heterogeneous data types.

TDM builds on topological methods to integrate heterogeneous information. We use category theory to lift disparate data types into a common mathematical structure, facilitating the interaction of quantitative, numerical data with qualitative, Semantic, and symbolic information.

Finally, using mathematical "sheaves," TDM represents both the structure of the interacting information sources and their content. We can identify not only when sources are connected in topological loops, but when there are "echo chamber" effects in the data themselves, indicating potentially difficult and confounding feedback loops among reporting sources.

Sheaves are canonical in that any principled specification of the interaction between multiple data sources will provably recapitulate some portion of sheaf theory. Given a sheaf model and knowledge of each source individually and their local interactions, it is possible to automatically generate a global view of the integrated system and measure sensitivity and reliability.

The initial test case is the inferred geolocation of a single group target from hybrid information. The experimental example is wildlife tracking, where animals equipped with a global positioning system and radio collars are tracked by field biologists, who additionally have Semantic data sources, e.g., text and map applications. The resulting complex structure manifests hybrid information integration among complex sources for overall situational awareness.



Hierarchy of possible interactions among four heterogeneous sensors.

Computational scaling is of special interest, as recent results from team members indicate efficient algorithms for calculating "sheaf cohomology." This critical aspect of TDM is well grounded with applications in sensor analysis. Efficient cohomology calculations will be a significant demonstration of TDM's ability to move into the big data area, and recent progress in methods such as Morse contractions point the way to efficient scaling.

IMPACT

TDM methods support the integration of heterogeneous information sources in networks, allowing the identification of complex interactions and confounding feedbacks. By extending recent work in computational scaling, we will enable using topological tools on larger, more diverse data sets.

Contacts

Cliff Joslyn
Principal Investigator
(206) 528-3042
cliff.joslyn@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Michael Robinson
(202) 885-3681
michaelr@american.edu

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965