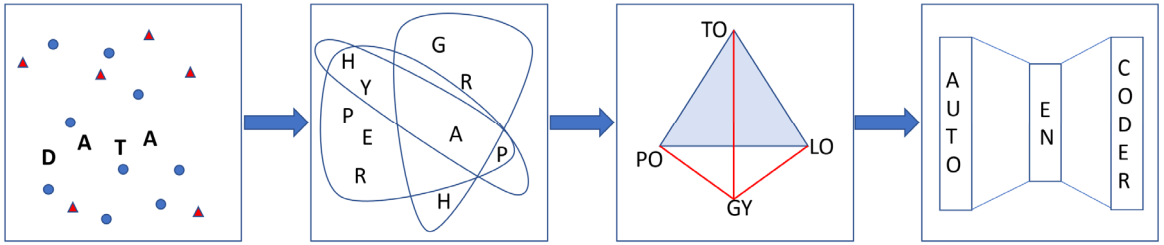
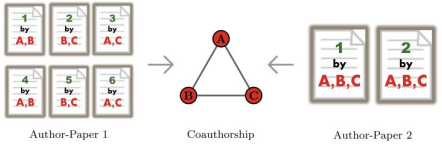
Hypergraph Topological Features for Autoencoder-Based Intrusion Detection for Cybersecurity Data

Bill Kay, Sinan Aksoy, Molly Baird,Daniel Best, Helen Jenne, Cliff Joslyn, Christopher Potvin, Gregory Henselman-Petrusek, Garret Seppala, Stephen Young, Emilie Purvine; Pacific Northwest National Laboratory



A workflow we propose for analyzing complex cybersecurity data:
•Construct hypergraphs which capture complex correlations.
•Use tools from topology to analyze relational structure.
•Use auto-encoders to detect anomalous behavior.

Hypergraphs capture multiway relational data.
•Strictly more information than graphs.
•Relationships give rise to complex structure.

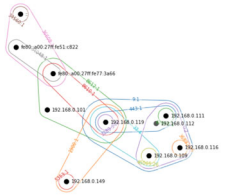


An example of information loss in a coauthor graph. Hypegraphs distinguish different multiway relations (pairwise, 3-way, etc.), while graphs do not.

Topology (homology): spatial properties known as "holes."
Connected components (dim-0), loops (dim-1), and voids (dim-k).
•Counts of these voids are known as the "Betti numbers," and provide insight into which regions of a hypergra

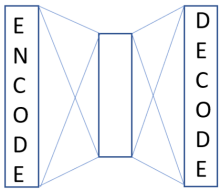
Cyber data to Hypergraphs: From log files, define useful sets of rules for constructing hypergraphs from cyber data.
•Useful for anomaly detection.
- Build hypergraphs on anomaly free data.
- Compare to hypergraphs from the wild.

sIP	dIP	sPort	dPort	pro	packets	flags	sTime
10.1.60.203	10.1.60.187	50398	80	6	5	FS PA	2009/04/20T11:35:19.439
10.1.60.187	10.1.60.203	80	50398	6	5	FS PA	2009/04/20T11:35:19.440
10.1.60.203	10.1.60.73	50189	5222	6	6	FS PA	2009/04/20T11:35:19.446



An example of one way to construct a hypergraph from cyber data
•Nodes are IP addresses.
•Edges are ports.
•A collection of IP addresses are in a given (port) edge if the port is accessed by the IP during a fixed time window.

Machine Learning (high level): A workflow for anomaly detection:
•Compute topological features of a time series of hypergraphs constructed from known non-anomalous data.
•Train ML on vectorized hypergraph topological data.
- ML has good performance on non-anomalous data.
- Experiences performance dropoff on anomalous data.



Autoencoders are a type of ANN which:
•Learn efficient encodings of inputs.
- By having a bottleneck layer.

•Given output you can guess at the input.
- On clean data, guesses are good.
- On attack data, guesses are bad.

For additional information, contact:
Sbill Kay | (509) 375-3624 | William.kay@pnnl.gov

U.S. DEPARTMENT OF ENERGY | BATTELLE
PNNL is operated by Battelle for the U.S. Department of Energy
10/9/2022 |
Pacific Northwest NATIONAL LABORATORY
www.pnnl.gov