Relational Models of Complex Systems: Hierarchy and Topology of High Order Interactions

Cliff Joslyn
Math, Stats, and Data Science
Pacific Northwest National Lab

Systems Science and Industrial Engineering
Binghamton University

PNNL-SA-190675
So Many Contributors.

- Sinan Aksoy
- Dustin Arendt
- Dan Best
- Aivett Bilbao
- Alyson Bittner
- Mark Bonicillo
- Corey Broeckling (Colorado State U)
- George Chin
- Sean Colby
- Joseph Cottam
- Song Feng
- Robert Green (SUNY Albany)
- Tobias Hagge
- Emily Heath (Iowa State U)
- Greg Henselman-Petrusek
- Brett Jefferson
- Helen Jenne
- Bill Kay
- Henry Kvinge
- Andy Lin
- Patrick Mackey
- Jason McDermott
- Hugh Mitchell
- Audun Myers
- Carlos Ortiz-Marrero
- Brenda Praggastis
- Sumit Purohit
- Emilie Purvine
- Michael Rawson
- Prof. Michael Robinson (American U)
- Madelyn Shapiro
- Stephen Young
- Ji-Young Yun
Mathematics and Methodology:

- Hypergraphs for hypernetwork science: Hypergraph walks, centrality, connectivity, Laplacians, clustering
- Computational topology and multidimensional data analysis: Homological hypergraph analysis, topological data analysis, topological sheaves for data integration

Software

- HyperNetX (HNX, Python): Human scale
  - Proving ground for methods
  - User interfaces: Visualization
- Chapel Hypergraph Library (CHGL): HPC scale
  - Data parallel language

Applications

- Cyber: DNS, Netflow
- OSINT
- Computational virology
- Combinatorial chemistry
- Scientometrics, open source analysis
- Multi-criteria decision analysis
Hypernetworks and Computational Topology

Abstract Simplicial Complexes

(Grammar) Topological Spaces

(Persistent) Homology (TDA)

Topological Sheaves

Data layer Quantitative weights

Hypernetworks Science

Network Science


HyperNetX (HNX) 2.0 (May 2023)!

https://github.com/pnnl/HyperNetX

Python package for modeling complex data as hypergraphs

- Latest release 2.0 is now available!!!
- First release 2018, 24 releases
- Sponsor/Project driven
- Multiple contributors

- Combinatorics – Statistics
- S-metrics, S-linegraphs
- Topology – Simplicial Homology
- Generative models
- Laplacian Clustering
- Clustering and Modularity
- Contagion
- Cell and Object Property support
- Internal Vis and HNXWidget package
- Multiple tutorials, demos
- Built on Pandas DataFrames
- Highly interoperable with Networkx, Matplotlib, and other hypergraph libraries
- ReadTheDocs page available https://pnnl.github.io/HyperNetX/index.html

Important Things About Hypergraphs

While all graphs \( G \) are \( (2\text{-uniform}) \) hypergraphs \( H \), since they're very special cases, general hypergraphs have some important properties which really stand out in distinction, especially to those already conversant with graphs. The following issues are critical for hypergraphs, but “disappear” when considering the special case of \( 2 \)-uniform hypergraphs which are graphs.

All Hypergraphs Come in Dual Pairs

If our incidence matrix \( I \) is a general \( m \times n \) Boolean matrix, then its transpose \( I^T \) is an \( n \times m \) Boolean matrix. In fact, \( I^T \) is also the incidence matrix of a different hypergraph called the dual hypergraph \( H^* \) of \( H \). In the dual \( H^* \), it’s just that vertices and edges are swapped: we now have \( H^* = (E,V) \) where \( E \) is a set of vertices, and the new edges \( e \subseteq E \) are subsets of those vertices.

Fig. 3 The dual hypergraph \( H^* \)
Today’s Story

• How can we relate together mathematical models of complex systems involving:
  1. **Complex Networks**: (Multi-way) connections of items
  2. **Hierarchies**: Arrangements of items in levels
  3. **Topologies (finite)**: Gluing together structures of different dimensionalities

• Rooted in mathematical systems theory


Systems Foundations

• Some systems concepts

<table>
<thead>
<tr>
<th>Order</th>
<th>Organization</th>
<th>Control</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representation</td>
<td><strong>Structure</strong></td>
<td><strong>Hierarchy</strong></td>
<td>Growth</td>
</tr>
<tr>
<td>Information</td>
<td>Development</td>
<td>Adaptation</td>
<td>Evolution</td>
</tr>
<tr>
<td>Heterarchy</td>
<td><strong>System</strong></td>
<td><strong>Network</strong></td>
<td>Aggregate</td>
</tr>
<tr>
<td>Emergence</td>
<td>Constraint</td>
<td>Function</td>
<td>Goal</td>
</tr>
<tr>
<td>Purpose</td>
<td>Stability</td>
<td>Subsystem</td>
<td>Supersystem</td>
</tr>
<tr>
<td>Scale</td>
<td>Environment</td>
<td>Distinction</td>
<td>Relation</td>
</tr>
<tr>
<td>Input</td>
<td>Output</td>
<td>Throughput</td>
<td>State</td>
</tr>
</tbody>
</table>

• Grounded in rigorous modeling
• Mappings among mathematical formalisms (hint: category theory)
• Applied across disciplinary boundaries
AN ALGORITHM FOR FINDING ALL FUNCTIONS EMBEDDED IN A RELATION

JAMES L. SNELL

Department of Systems Science, Thomas J. Watson School of Engineering, Applied Science and Technology, State University of New York at Binghamton, Binghamton, New York, U.S.A.

(Received February 16, 1984; in final form June 19, 1984)

The problem is posed: find an algorithm which for any given n-dimensional relation $R \subseteq A_1 \times A_2 \times \ldots \times A_n$, defined on a set family $A = \{A_1, A_2, \ldots, A_n\}$, $n = 1, 2, \ldots$, determines all functional dependences between disjoint subsets of $A$ which are embedded in $R$. A solution algorithm is presented, a theorem is proved that allows a simplification in the algorithm, and an efficient computer implementation (available through the General Systems Depository) is demonstrated.

INDEX TERMS: Algorithm, computer algorithm, relation, function, embedded function.
0. Mathematical Systems Theory

- **System**: Multivariate relation
  \[ S \subseteq X_1 \times X_2 \times \ldots \times X_N \]

- **Dimension**: Each \( X_i \) can be “anything”
  - **Scalar quantity**: Integer, float, etc.
  - **Boolean**: 0/1
  - **Categorical variable**: A, B, C
  - **Ordinal variable**: \( \alpha \leq \beta \leq \gamma \)
    - Time! Dynamics!
  - **String**: “abz”
  - **Arbitrary structure**: List, vector
  - Etc.
Network Models of Complex Relational Data

- Many real-world data sets have complex relational structure
  - Cyber security: Domains x IP Addresses x MAC Addresses x Malware IDs x …
  - Social networks: People x Groups
  - Bibliometrics: Authors x Papers x Keywords
  - Biology: Proteins x Pathways, Complexes
  - CBP: Airline Passengers x Border Crossings x Cargo Shipments
  - Multi-Criteria Decision Analysis (MCDA): Products x Capabilities

- Modellable as e.g. pandas data frame:
  - Columns: Dimensions $X_i$
  - Rows: Points or vectors $\vec{x} \in S \subseteq X_1 \times X_2 \times \ldots \times X_N$

- Relational network structures:
  - Graph: Self-relation
  - Hypergraph: Binary relation
  - Tensor: Multi-way relation

---

![Diagram of network models](https://activednsproject.org/)

<table>
<thead>
<tr>
<th>House</th>
<th>Blood status</th>
<th>Species</th>
<th>Hair colour</th>
<th>Eye colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Gryffindor</td>
<td>Half-blood</td>
<td>Human</td>
<td>Black</td>
</tr>
<tr>
<td>1</td>
<td>Gryffindor</td>
<td>Pure-blood</td>
<td>Human</td>
<td>Red</td>
</tr>
<tr>
<td>2</td>
<td>Gryffindor</td>
<td>Muggle-born</td>
<td>Human</td>
<td>Brown</td>
</tr>
<tr>
<td>3</td>
<td>Gryffindor</td>
<td>Half-blood</td>
<td>Human</td>
<td>Silver, formerly auburn</td>
</tr>
<tr>
<td>4</td>
<td>Gryffindor</td>
<td>Part-Human</td>
<td>Half-Human/Half-Giant</td>
<td>Black</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>135</td>
<td>Unknown House</td>
<td>Unknown Blood status</td>
<td>Human</td>
<td>Grey</td>
</tr>
<tr>
<td>136</td>
<td>Unknown House</td>
<td>Unknown Blood status</td>
<td>Werewolf</td>
<td>Grey</td>
</tr>
<tr>
<td>137</td>
<td>Unknown House</td>
<td>Unknown Blood status</td>
<td>Human</td>
<td>Blond</td>
</tr>
<tr>
<td>138</td>
<td>Unknown House</td>
<td>Unknown Blood status</td>
<td>Unknown Species</td>
<td>Unknown Hair colour</td>
</tr>
<tr>
<td>139</td>
<td>Unknown House</td>
<td>Unknown Blood status</td>
<td>Unknown Species</td>
<td>Unknown Hair colour</td>
</tr>
<tr>
<td>103.86.122.130</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.148</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.149</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.152</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.154</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.160</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.169</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.173</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.181</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.192</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.195</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.222</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.223</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.225</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.238</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>103.86.122.242</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Dom = \{ IP \}

IP = \{ Dom \}

https://activednsproject.org/
Projections of Multivariate Data

- Mass spectrometry features in an \( n \)-dimensional space: MS-LC-IMS (ion mobility)
- Projections into lower dimensional spaces
- Nested spectra
- Discretized (peak-picked) data

<table>
<thead>
<tr>
<th>Retention Time (R)</th>
<th>Drift Time (D)</th>
<th>m/z (M)</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>700</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>900</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>700</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>800</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>850</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>850</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>900</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>900</td>
<td>1</td>
</tr>
</tbody>
</table>
A Discrete Relation

- Boolean tensor, incidence tensor
- 2D projections
- Duals: Matrix transposes
e.g. \( P \times A \)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Authors</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Andrews, Davis</td>
<td>Graphs</td>
</tr>
<tr>
<td>2</td>
<td>Andrews, Carter, Davis</td>
<td>Topology</td>
</tr>
<tr>
<td>3</td>
<td>Davis</td>
<td>Graphs, topology</td>
</tr>
<tr>
<td>4</td>
<td>Andrews, Bailey</td>
<td>Lattices</td>
</tr>
<tr>
<td>5</td>
<td>Bailey, Carter</td>
<td>Lattices, topology</td>
</tr>
</tbody>
</table>

\[
A \times P \times K
\]

<table>
<thead>
<tr>
<th>Authors</th>
<th>Lattices</th>
<th>Topology</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrews</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Bailey</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Carter</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

\[
A \times P
\]

\[
K \times P
\]

\[
A \times K
\]
Hypergraphs Instead of Graphs

$A \times P$

Coauthorship Matrix $A \times A$

<table>
<thead>
<tr>
<th>Paper #</th>
<th>Authors</th>
<th>Andrews</th>
<th>Bailey</th>
<th>Carter</th>
<th>Davis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Andrews, Davis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>Andrews, Carter, Davis</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>Davis</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>Andrews, Bailey</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>Bailey, Carter</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Hypergraph representation

Graph representation
Graphs, Hypergraphs, and Relations

- **A binary relation**: Incidence, not adjacency, information
- **Bipartite network**: Bijective
- **Graph on rows**: Pairwise relations
- **Graph on columns**: Pairwise relations

**Hypergraph on rows ("primal"):** Multiway relations

**Hypergraph on columns ("dual"):** Multiway relations

\[ D^* = \text{edge size distribution} \]
\[
I^T I - D^* = \begin{pmatrix}
0 & 2 & 1 & 1 & 0 \\
2 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0
\end{pmatrix}
\]

2-section of primal = Clique expansion = Underlying graph

Line graph of primal = 2-section of dual

\[
I \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ a & X & X & X \\ b & & X & X \\ c & & & X & X \\ d & & & & X & X \end{pmatrix}
E \begin{pmatrix} 1 & 2 & 1 & 1 \end{pmatrix}
\]

\[ H = I \subseteq V \times E \]

\[ H^* = I^T \subseteq E \times V \]
Network Representations of Relational View Projections

- **Data tensor**

- **Projection**: Two (combinations of) dimensions
  - **Vertices**: For each retention time
  - **Hyperedges**: What $m/z$, drift values are seen?
  - **View**: $(R; <M,D>)$ determines a hypergraph

- **Isomers separated by chromatography**: $(R; <M,D>)$
  Different RT; same $m/z$, drift

- **Isotopic Peaks**: $(M; <R,D>)$
  Different $m/z$, same drift, same RT

- **Adducts, In-source-fragments, Dimers/trimers**: $(<M,D>; R)$
  Different $m/z$, different drift, same RT

- **Isomers separated by mobility**: $(D; <R,M>)$
  Same $m/z$, different drift, same RT

<table>
<thead>
<tr>
<th>Retention Time (R)</th>
<th>Drift Time (D)</th>
<th>m/z (M)</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>700</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>900</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>700</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>800</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>850</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>850</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>900</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>900</td>
<td>1</td>
</tr>
</tbody>
</table>
Basic Hypergraphs (undirected, unordered)

\[ H = \langle V, \mathcal{E} \rangle, \text{ Family } \mathcal{E} = \{e\}, e \subseteq V \]
\[ H = \{\{a, d\}, \{a, c, d\}, \{d\}, \{a, b\}, \{b, c\}\} \]
\[ = \{ad, acd, d, ab, bc\} \]
\[ = \{1:ad, 2:acd, 3:d, 4:ab, 5:bc\} \]

(Multi)Set System

\[
\begin{array}{c|ccccc}
       & 1 & 2 & 3 & 4 & 5 \\
\hline
a & \times & \times & & X & X \\
b & & & X & X \\
c & & X & & X \\
d & X & X & X & & \\
\end{array}
\]

Bipartite Graph

Euler Diagram

• Axioms matter!
  - Singletons \( v \) vs. \( \{v\} \), isolated vertices, empty edges, multi-edges, multi-vertices, self-loops

Simplicial Diagram

Incidence Matrix
\[ S = V \times \mathcal{E} \]
Categorical Hypergraph Foundations

- **Sets:** $X, |X| = n; Y, |Y| = m$
- **Axioms:** $X \cap Y = \emptyset$

**Theorem 1.** The following are categorically equivalent:

- $R \subseteq X \times Y$
  Binary Relations

- $H = (V, E, I)$
  $I: V \times E \to \{0, 1\}$
  Hypergraphs: Incidence function

- $G = \langle X \cup Y, F \rangle$
  $F \subseteq \binom{X \cup Y}{2} - \binom{(X \cup Y)}{2}$
  Bipartite Graphs

- $S = \langle V, E \rangle, E \subseteq 2^V$
  Hypergraphs: Set system

$E$ must be a multiset, or an indexed family of subsets
Graphs vs. Hypergraphs: Precis

A graph is 2-uniform hypergraph

**Graphs:** $G \subseteq V^2$
- Connections have length
- Simple
- Lossy for multi-way interactions
- Small (quadratic)

**Hypergraphs:** $H \subseteq 2^V$
- Connections have length and width
- Complex
- Lossless
- Large (possibly exponential)
- Advanced mathematical properties (topology)
Burgeoning Movement in Network Science


FIG. 1. Illustration of a hypergraph. Infected nodes (red) infect a healthy node (gray) via hyperedges of sizes 2 and 3 with rates $\beta_2$ and $\beta_3$ respectively.
1. Hypergraph Walks Have Length and Width

- **Hypergraph Paths Have Width:** Minimum edge intersection
- **s-walk:** Sequence \( \langle e_i \rangle_{i=1}^{n} \) when \( s \leq \min_{e_i, e_{i+1}} |e_i \cap e_{i+1}|, i = 1 \ldots n - 1 \)

**A Graph Path:**
(Edgewise) length = 2
Width (necessarily) 1

**Two Hypergraph Paths:**
Same (edgewise) length = 2

**Weak interactions:**
Width = s = 1

**Strong interactions:**
Width = s = 3

- **Extend generally:**
  - **s-distance:** \( d_s(e, f) = \begin{cases} \min |s\text{-walk}(e, f)| & \text{if exists} \\ \infty & \text{otherwise} \end{cases} \)
  - **s-components, s-centrality, s-diameter s-motifs, s-clustering coefficient**

s-Closeness centrality

- **Question:** Which nodes or edges are “close” to everything?

**Graphs**

Closeness Centrality

\[ C(v) = \frac{|V| - 1}{\sum_{u \in V} d(v, u)} \]

Harmonic Closeness Centrality

\[ HC(v) = \frac{1}{|V| - 1} \sum_{u \in V} \frac{1}{d(v, u)} \]

**Hypergraphs**

\[ E_s = \{ e \in E : |e| \geq s \} \]

\[ C_s(e) = \frac{|E_s| - 1}{\sum_{f \in E_s} d_s(e, f)} \]

\[ HC_s(e) = \frac{1}{|E_s| - 1} \sum_{f \in E_s} \frac{1}{d_s(e, f)} \]

s-Betweenness centrality

- **Question:** Which nodes or edges are on many shortest paths?

### Betweenness Centrality

**Graphs**

\[ B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \]

**Hypergraphs**

\[ B_s(e) = \sum_{g \neq e \neq f \in E_s} \frac{\sigma^s_{gf}(e)}{\sigma^s_{gf}} \]

Example: Biological Data

- **Mouse and human cells infected with viral strains:**
  - Ebola, Influenza, MERS, SARS, West Nile Virus
  - Samples analyzed at various time points post-infection
  - *Transcriptomics* data: measuring expression of gene transcripts
    - $\log_2(\text{sample} / \text{control})$ for each [sample, gene] pair

- **Hypergraph:**
  - Nodes = conditions (virus, strain, cell type, time point, …)
  - Edges = genes
  - Node/edge containment = genes with $\log_2(\text{fold change})$ z-score $\geq 2$ and p-value $< 0.05$ for a given condition

Hypergraphs for identifying important genes

- **Goal**: Find genes which are central in host response to viral infection

- **Hypothesis**: Hypernetwork science measures will rank known central genes (e.g., immune response) higher than network science in context likelihood of relatedness (CLR) graph, and higher than simple measures

- **Enrichment score (GSEA)**: Determine whether members of a known gene set tend to occur toward the top (or bottom) of a ranked list

---

Gene Enrichment Scores

2. Hypergraphs Have Hierarchy

Hypergraph

- $|A| = 3, |B| = 4$
- $A \cap B \quad |A \cap B|$
- $= \{v_1, \ldots, v_n\} \geq 1$
- $= \emptyset \quad 0$
- $A \quad |A|$

Graph

- $|A| = |B| = 2$
- $A \cap B \quad |A \cap B|$
- $= \{v\} \equiv 1$
- $= \emptyset \quad 0$

- Incident Edges
- Disjoint Edges
- Included Edges

Diagram: A and B are non-empty sets with varying cardinalities and intersection sizes.
Hierarchy Theory

- Systems admitting to descriptions in terms of levels: Height, depth
  - Necessary for viable organization of large complex systems
  - Natural scale dependencies and interactions
- The Systems community has attended less to mathematical formalism
  - Way more than trees
  - Avoiding ethical implications of authoritarian social hierarchies
- Partial order on set $P$: \( \leq \subseteq P^2 \)
  - Reflexive, symmetric, anti-transitive
- Poset:
  \[ P = \langle P, \leq \rangle \]
- Lattice: Unique pairwise common parent/child
Some Aspects of Hierarchies = Partial Orders

- Tree (Unique Parents)
- Nodes with multiple parents/children
- Dual Structure
- Totally Bounded
- Chain (Totally ordered) (Unique parents and children)
- Antichain (Totally unordered)
- Not a lattice (Pairs with multiple parents/children)
- Ungraded (Unequal chain lengths) (Ambiguous levels)

**Hypergraph Inclusivity**

- **Hypergraph:**
  - 2 included edges: 1, 3
  - 3 “toplexes”: Maximal hyperedges: 2, 4, 5
  - Inclusivity = 2/5

- **Simple hypergraph:**
  - Remove all inclusions
  - All toplexes
  - “Reduction”
  - Inclusivity = 0

- **Abstract simplicial complex (ASC):** Add all inclusions
  - Toplexes and all below
  - “Closure”
  - Inclusivity = 7/10

- **All share the same topological structure:**
  - Determined by toplexes

- $\mathcal{H}$ and $\mathcal{H}$ are one-to-one
Hypergraphs Are Inherently Ordered

- Hyperedges have an inclusion order
- But more completely an intersection structure: intersection complex

**Theorem:** Intersection complex is bijective to the concept lattice

  - “Galois notation” shows joint relationships of unions, intersections of vertices, edges

**Questions:** How are hypergraph operations mirrored in the concept lattice?

**Theorem:** Closing by subset yields the ASC in the HG, and the “Dowker cosheaf” in the lattice structure

Rawson, Michael G; Myers, Audun; Green, Robert; Robinson, M; Joslyn, Cliff: (2023) “Formal Concept Lattice Representations and Algorithms for Hypergraphs”, https://doi.org/10.48550/arXiv.2307.11681

Example Concept Lattice of a Hypergraph

- A = 1245
- B = 2356
- C = 4567
- E = 38

AF = 12  AB = 25  AC = 45  BE = 3
ABF = 2  ABC = 5  BC = 56

A
B
C
F
E
Ukraine 2014 (UKR14) Knowledge Base

- **DARPA/I2O/AIDA Performers, 2018:**
  - Entity, relation, event extraction
  - Graph integration

- **Open source information about 2014 Russian invasion of Eastern Ukraine**
  - Multi-value attributes exist such as ‘name’ and ‘type’
  - Temporal information exist for a subset of nodes

- **Richly Attributed:** Graph Ontology:
  - **Nodes:** Entity, event, relation types
  - **Edges:** Relationships (roles) of entities within events/relations

- **Real-world Data**
  - Noisy / many inaccuracies
  - Most noise seems to come from incorrect relationships between nodes

- Original data represented as **RDF triples**
- Converted to **property graph** by PNNL: Neo4J

Node and edge types associated with a small graph sample.
Center node is a “relationship node” connecting two entities together.

- **Node Types:** 307
- **Node Instances:** 406K
- **Edge Types:** 367
- **Edge Instances:** 302K
- **Connected Components:** 314K
Event Hypergraph Model

- **Ukr14 is broadly bipartite:** Events/relations valued on entities
- **Generally supports hypergraph representation:**
  - Event/relation node: Hyperedge
  - Entity node: Hypernode
UKR14 Example Concept Lattice

Rawson, Michael G; Myers, Audun; Green, Robert; Robinson, M; Joslyn, Cliff: (2023) "Formal Concept Lattice Representations and Algorithms for Hypergraphs", https://doi.org/10.48550/arXiv.2307.11681
3. Hypergraphs are Topological Objects

- Hypergraphs have topological properties
  - $\beta_0 = 1$
  - $\beta_1 = 1$
  - $\beta_2 = 0$

- (Simplicial) homology identifies multidimensional open structures
  - As hypotheses for missing data
  - Need for bridging metadata
Homologies Show Multidimensional Open Structures

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_{\geq 3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNS1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DNS1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>DNS1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**DNS2**: One generator of a 2-hole, tetrahedral void


https://activednsproject.org/
Temporal Hypergraph Analysis

- Temporal hypergraph
- Trajectory of temporal sub-hypergraphs
- Measure change in structure, homology, distributions
Zigzag Persistence Example

- Temporal sequences
  - Are there topological features that persist over time in a dynamically evolving system?

```
<table>
<thead>
<tr>
<th>t</th>
<th>K_0</th>
<th>K_1</th>
<th>K_2</th>
<th>K_3</th>
<th>K_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>K_0</td>
<td>K_1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>K_1</td>
<td>K_2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td>K_2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>K_3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
<td></td>
<td>K_3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>K_4</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
<td>K_4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>K_4</td>
</tr>
<tr>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
Operationally Transparent Cyber (OpTC) data set

- Created by the Defense Advanced Research Projects Agency (DARPA) as part of a mission to test scaling of cyber attack detection
- Flow and host logs from both benign and malicious activity plus ground truth document describing the attack events
  - Downloading malicious PowerShell Empire, privilege escalation, credential theft, network scanning, and lateral movement
- Example subset of OpTC flow data:

<table>
<thead>
<tr>
<th>Time</th>
<th>Action-Object</th>
<th>PID</th>
<th>Source IP</th>
<th>Destination IP</th>
<th>Dest. Port</th>
<th>Executable</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/23/2023 9:06</td>
<td>MESSAGE-FLOW</td>
<td>864</td>
<td>10.20.2.47</td>
<td>224.0.0.252</td>
<td>5355</td>
<td>svchost.exe</td>
</tr>
<tr>
<td>9/23/2023 9:06</td>
<td>MESSAGE-FLOW</td>
<td>864</td>
<td>10.20.2.47</td>
<td>224.0.0.252</td>
<td>5355</td>
<td>svchost.exe</td>
</tr>
<tr>
<td>9/23/2023 9:06</td>
<td>MESSAGE-FLOW</td>
<td>864</td>
<td>10.20.2.93</td>
<td>224.0.0.252</td>
<td>5355</td>
<td>svchost.exe</td>
</tr>
<tr>
<td>9/23/2023 9:06</td>
<td>MESSAGE-FLOW</td>
<td>864</td>
<td>10.20.2.93</td>
<td>224.0.0.252</td>
<td>5355</td>
<td>svchost.exe</td>
</tr>
<tr>
<td>9/23/2023 9:06</td>
<td>MESSAGE-FLOW</td>
<td>2236</td>
<td>10.20.2.66</td>
<td>225.0.0.1</td>
<td>5000</td>
<td>python.exe</td>
</tr>
<tr>
<td>9/23/2023 9:06</td>
<td>MESSAGE-FLOW</td>
<td>3980</td>
<td>10.20.4.133</td>
<td>10.20.2.66</td>
<td>5959</td>
<td>python.exe</td>
</tr>
</tbody>
</table>

Myers, Audun; Bittner, Alyson S; Aksoy, Sinan G; Best, Dan, Roek, G; Jenne, Helen; Joslyn, Cliff; Kay, Bill; Seppala, Garret; Young, Stephen; Purvine, Emilie AH: (2023) “Malicious Cyber Activity Detection Using Zigzag Persistence”, IEEE Dependable and Secure Computing Wshop on AI/ML for Cybersecurity (AIML 23), arXiv:2309.08010
Zigzag ML Experiment on OpTC

- **Goal:** identify source IPs responsible for malicious activity, and the time interval that activity occurred

- **Method:** construct temporal hypergraph sequence for each host, run zigzag persistence, train autoencoder on barcode summary
  - **Nodes:** Executable files
  - **Edges:** Destination ports
  - 10 minute time windows per HG
  - Dimension 0, 1 zigzag on hour of HGs
  - Adcock-Carlsson barcode coordinates
  - Autoencoder trained on hosts not found in ground truth document

Myers, Audun; Bittner, Alyson S; Aksoy, Sinan G; Best, Dan; Roek, G; Jenne, Helen; Joslyn, Cliff; Kay, Bill; Seppala, Garret; Young, Stephen; Purvine, Emilie AH: (2023) “Malicious Cyber Activity Detection Using Zigzag Persistence”, *IEEE Dependable and Secure Computing Wshop on AI/ML for Cybersecurity (AIML 23)*, arXiv:2309.08010
**Goal**: identify source IPs responsible for malicious activity, and the time interval that activity occurred

**Method**: construct temporal hypergraph sequence for each host, run zigzag persistence, train autoencoder on barcode summary

- **Nodes**: Executable files
- **Edges**: Destination ports
- 10 minute time windows per HG
- Dimension 0, 1 zigzag on hour of HGs
- Adcock-Carlsson barcode coordinates
- Autoencoder trained on hosts not found in ground truth document

Myers, Audun; Bittner, Alyson S; Aksoy, Sinan G; Best, Dan; Roek, G; Jenne, Helen; Joslyn, Cliff; Kay, Bill; Seppala, Garret; Young, Stephen; Purvine, Emilie AH: (2023) “Malicious Cyber Activity Detection Using Zigzag Persistence”, IEEE Dependable and Secure Computing Wshop on AI/ML for Cybersecurity (AIMA23), arXiv:2309.08010
Closing Thoughts

• Delighted to be back in the SSIE department
  ▪ Current work with Kevin Stoltz, Grant Generaux, Prof. Sayama
  ▪ Next work with you?

• PNNL also works extensively with universities in multiple roles and modes

• cajoslyn@binghamton.edu
• cliff.joslyn@pnnl.gov

https://cliffjoslyn.github.io
Thank you