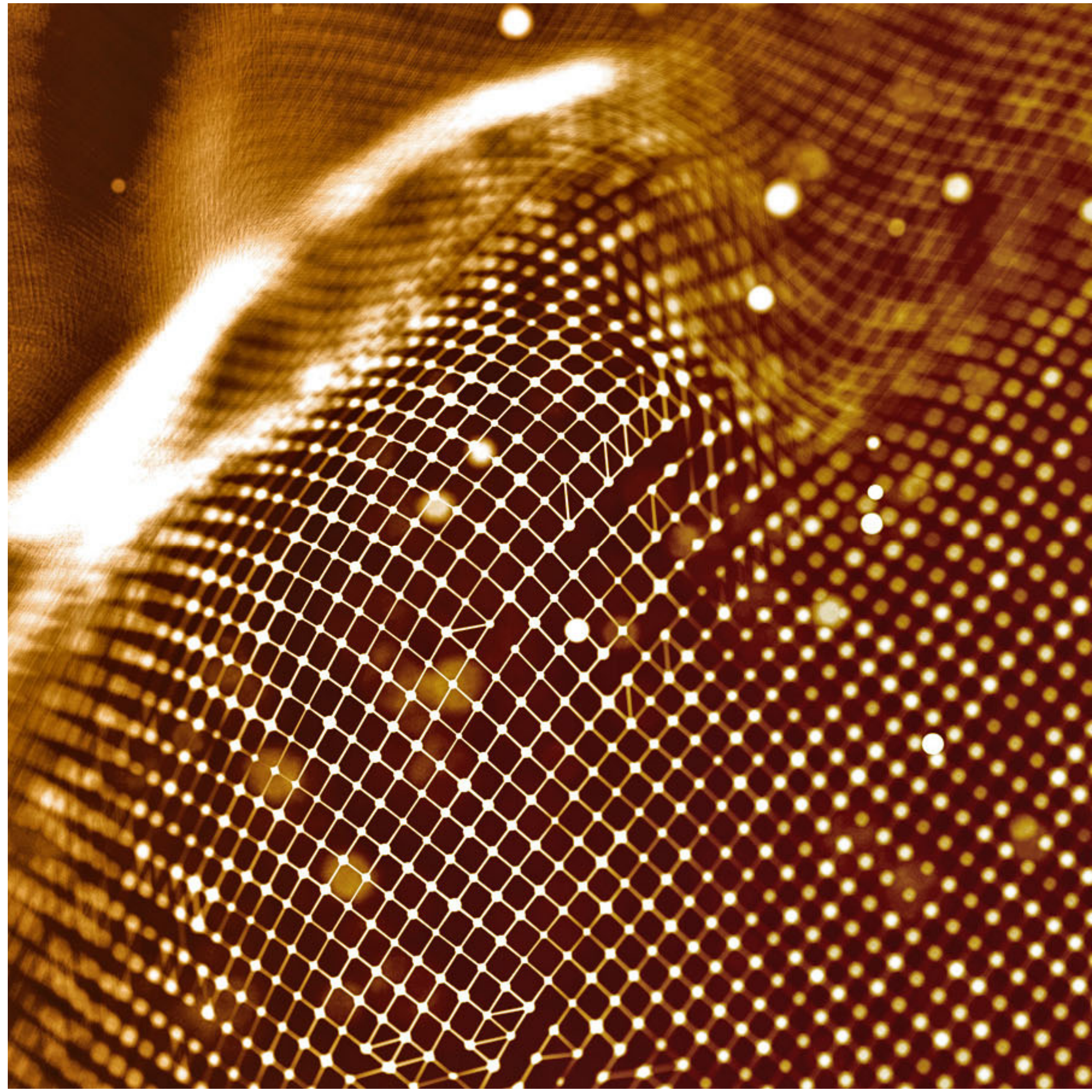
Relational Models of Complex Systems: Hierarchy and Topology of High Order Interactions

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

Systems Science and Industrial Engineering
Binghamton University

PNNL-SA-190675

PNNL is operated by Battelle for the U.S. Department of Energy



So Many Contributors. . .

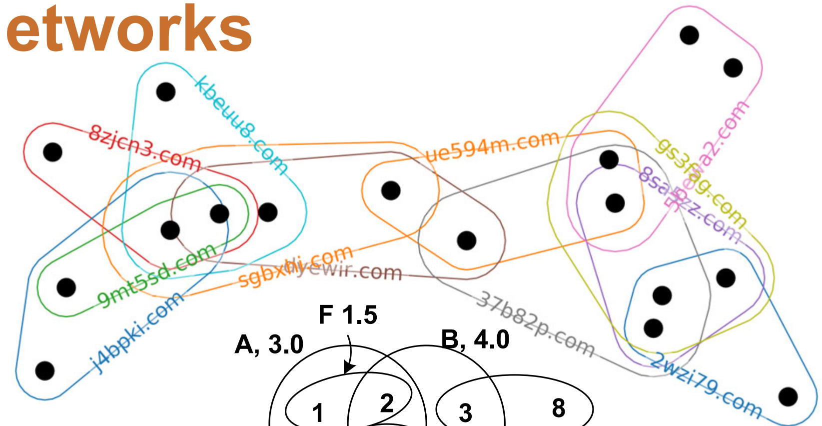
- Sinan Aksoy
- Dustin Arendt
- Dan Best
- Aivett Bilbao
- Alyson Bittner
- Mark Bonicillo
- Corey Broeckling (Colorado State U)
- George Chin
- Sean Colby
- Joseph Cottam
- Song Feng
- Robert Green (SUNY Albany)
- Tobias Hagge
- Emily Heath (Iowa State U)
- Greg Henselman-Petrusek
- Brett Jefferson
- 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

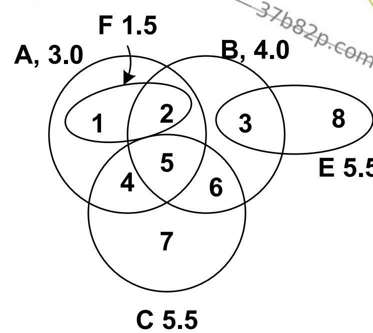
Pacific Northwest National Laboratory: Topology and High Order Networks

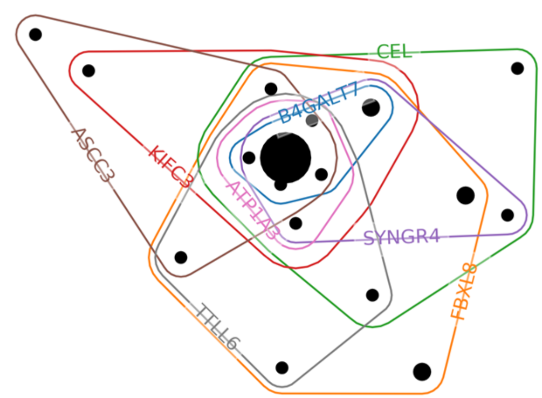
- **Mathematics and Methodology:**
  - **Hypergraphs for hypernetwork science:** Hypergraph walks, centrality, connectivity, Laplacians, clustering
  - **Computational topology and multidimensional data analysis:** Homological hypergraph analysis, topological data analysis, topological sheaves for data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

https://github.com/pnnl/HyperNetX

https://github.com/pnnl/chgl/

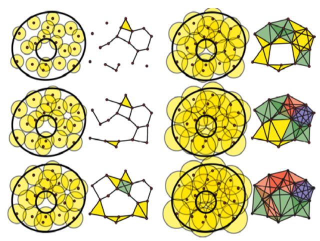




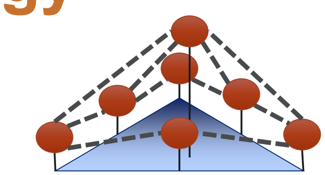


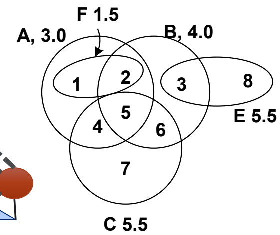
Hypernetworks and Computational Topology

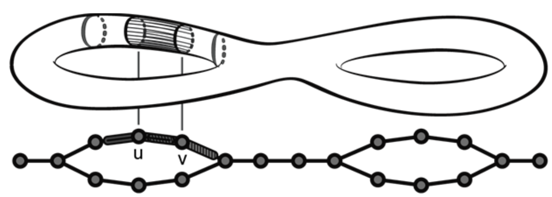
Joslyn, Cliff A et al.: (2021) "Hypernetwork Science: From Multidimensional Networks to Computational Topology", in: *Unifying Themes in Complex systems X: Proc. 10th Int. Conf. Complex Systems*, ed. D. Braha et al., pp. 377-392, Springer, https://doi.org/10 .1007/978-3-030-67318-5_25

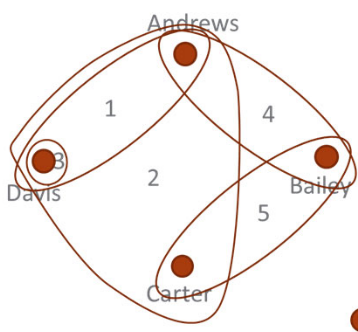


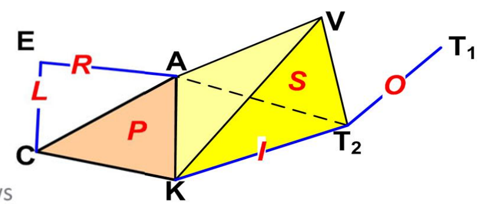
Ghrist, Robert: (2007) "Barcodes: The Persistent Topology of Data", *Bulletin of the American Mathematical Society*, v. 45:1, pp. 61-75

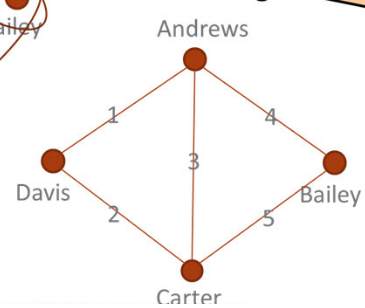












Topological Sheaves

Data layer Quantitative weights

Topological Spaces

(Persistent) Homology (TDA)

Abstract Simplicial Complexes

Included Edges

Hypergraphs

> 2 interacting elements

Graphs

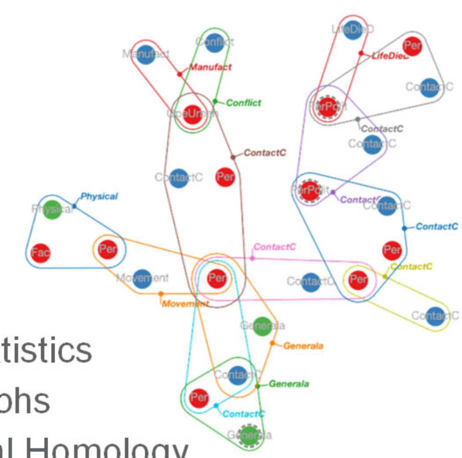
Computational Topology

Hypernetwork Science

Network Science

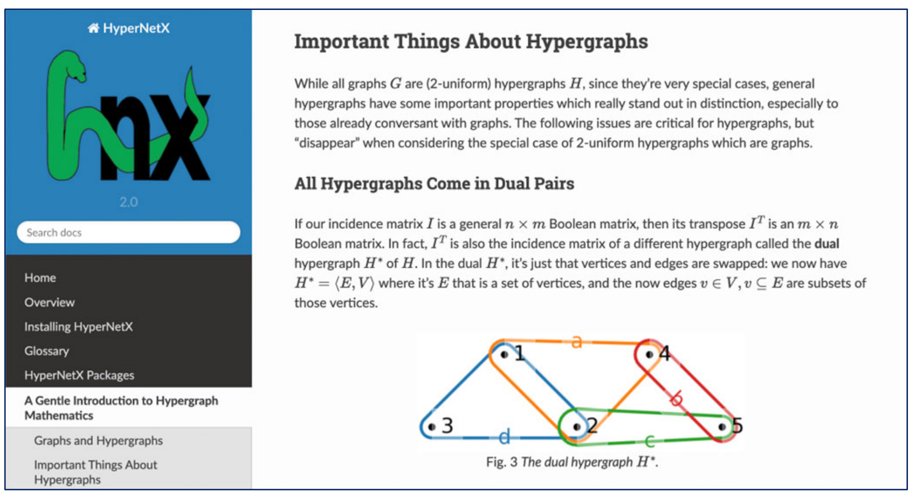
HyperNetX (HNX) 2.0 (May 2023)!

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



HyperNetX

2.0

Search docs

Home
Overview
Installing HyperNetX
Glossary
HyperNetX Packages
A Gentle Introduction to Hypergraph Mathematics
Graphs and Hypergraphs
Important Things About Hypergraphs

Important Things About Hypergraphs

While all graphs G are (2-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

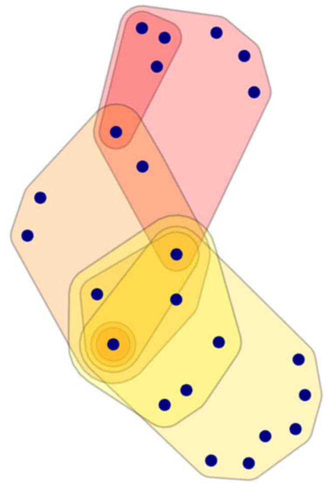
If our incidence matrix I is a general $n \times m$ Boolean matrix, then its transpose I^T is an $m \times n$ Boolean matrix. In fact, I^T is also the incidence matrix of a different hypergraph called the **dual** hypergraph H^* of H. In the dual H^*, it's just that vertices and edges are swapped: we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

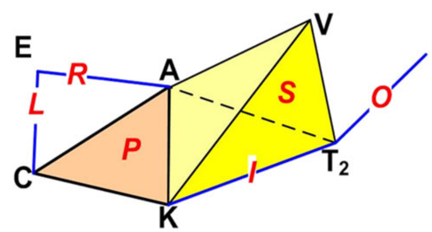
Fig. 3 *The dual hypergraph* H^*.

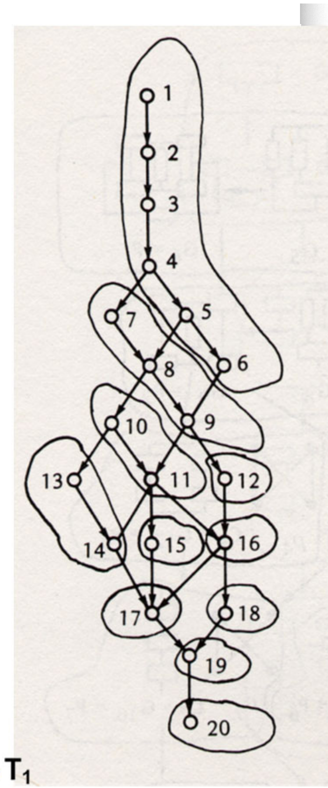
- Combinatorics – Statistics
- S-metrics, S-linegraphs
- Topology – Simplicial Homology
- Generative models
- Laplacian Clustering
- Clustering and Modularity
- Contagion
- Cell and Object Property support
- Internal Vis and HNXWidget package
- Multiple tutorials, demos
- Built on Pandas DataFrames
-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

Today’s Story

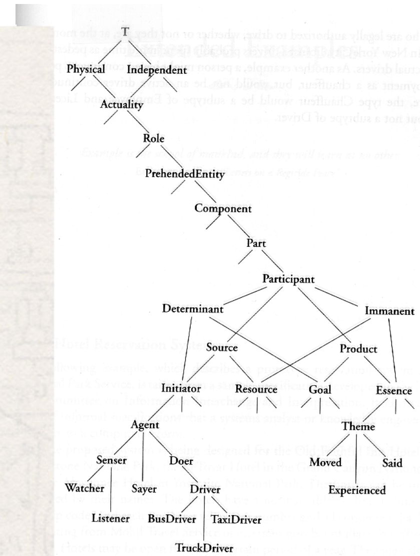
- **How can we relate together mathematical models of complex systems involving:**
  1. **<u>Complex</u> Networks:** (Multi-way) connections of items
  2. **Hierarchies:** Arrangements of items in levels
  3. **Topologies (<u>finite</u>):** Gluing together structures of different dimensionalities
- **0.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







Klir, George and Elias, Doug: (2003) *Architecture of Systems Problem Solving*, Plenum, New York, 2nd edition



Sowa, John F: (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole, Pacific Grove

Systems Foundations

- **Some systems concepts**

Order	Organization	Control	Complexity
Representation	Structure	Hierarchy	Growth
Information	Development	Adaptation	Evolution
Heterarchy	System	Network	Aggregate
Emergence	Constraint	Function	Goal
Purpose	Stability	Subsystem	Supersystem
Scale	Environment	Distinction	Relation
Input	Output	Throughput	State

- **Grounded in rigorous modeling**
- **Mappings among mathematical formalisms** (hint: category theory)
- **Applied across disciplinary boundaries**

Pacific Northwest
NATIONAL LABORATORY

A Binghamton Journey From 1985

Int. J. General Systems, 1985, Vol. 10, pp. 187–195
0308-1079/85/1003-0187 $18.50/0

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 $R \subset A_1 \times A_2 \times \ldots \times A_n$, 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.

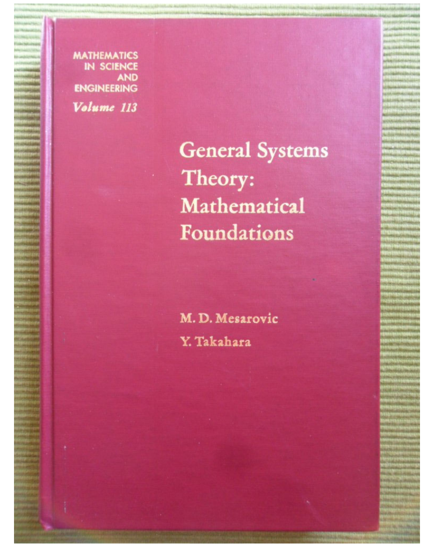
INDEX TERMS: Algorithm, computer algorithm, relation, function, embedded function.

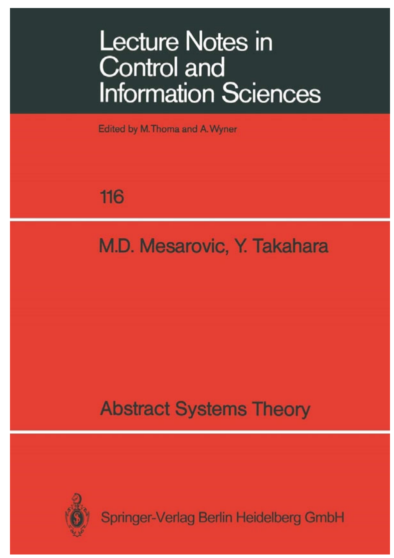
0. Mathematical Systems Theory

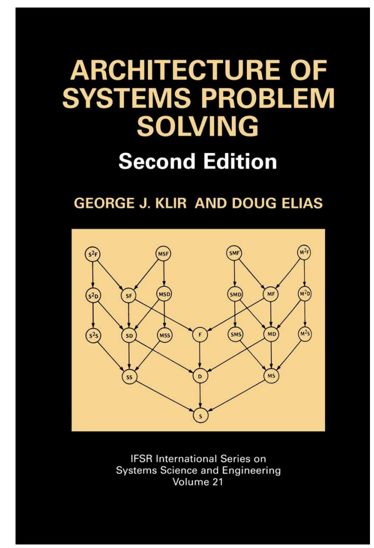
- **System:** Multivariate relation

  $$S \subseteq X_1 \times X_2 \times \ldots \times X_N$$

- **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.**





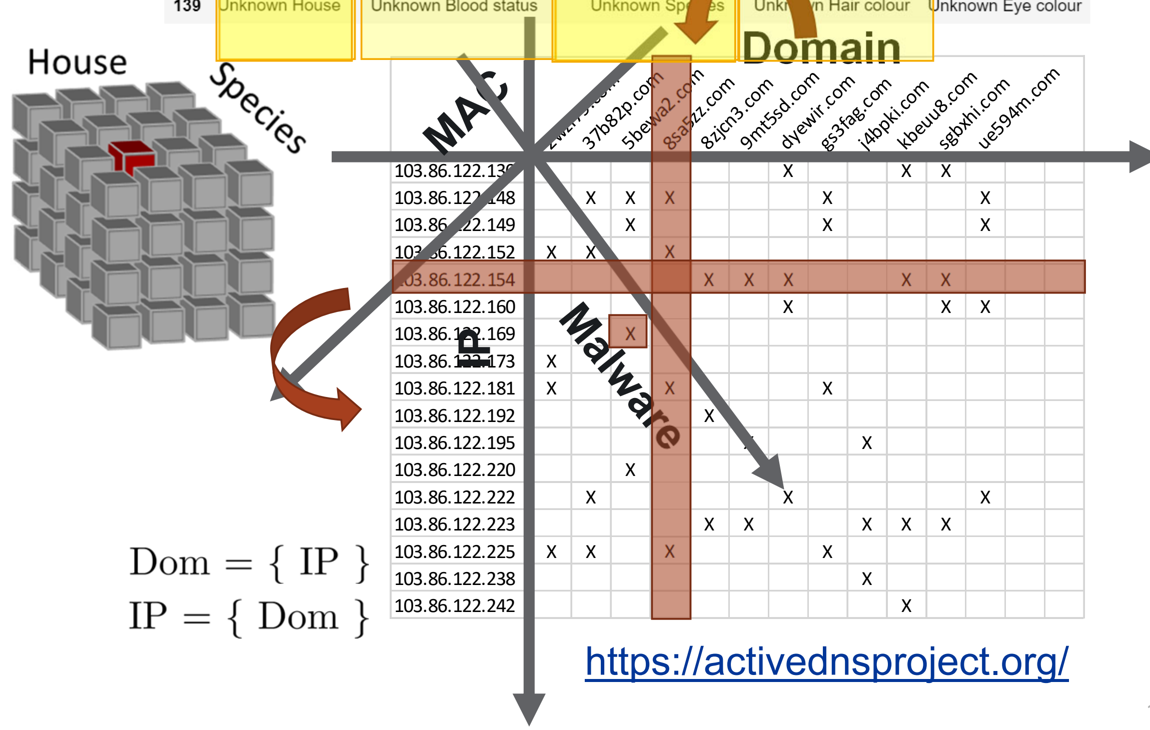


Network Models of Complex Relational Data

- **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

	House	Blood status	Species	Hair colour	Eye colour
0	Gryffindor	Half-blood	Human	Black	Bright green
1	Gryffindor	Pure-blood	Human	Red	Blue
2	Gryffindor	Muggle-born	Human	Brown	Brown
3	Gryffindor	Half-blood	Human	Silver| formerly auburn	Blue
4	Gryffindor	Part-Human	Half-Human/Half-Giant	Black	Black
...	...	...	...	...	...
135	Unknown House	Unknown Blood status	Human	Grey	Unknown Eye colour
136	Unknown House	Unknown Blood status	Werewolf	Grey	Unknown Eye colour
137	Unknown House	Pure-blood or half-blood	Human	Blond	Blue
138	Unknown House	Unknown Blood status	[illegible]	Unknown Hair colour	Unknown Eye colour
139	Unknown House	Unknown Blood status	Unknown Species	Unknown Hair colour	Unknown Eye colour



$\mathrm{Dom} = \{\ \mathrm{IP}\ \}$

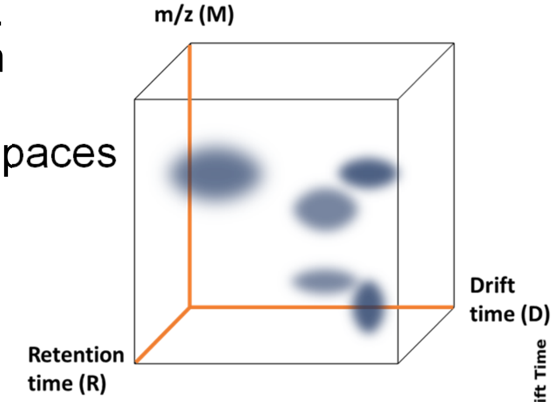
$\mathrm{IP} = \{\ \mathrm{Dom}\ \}$

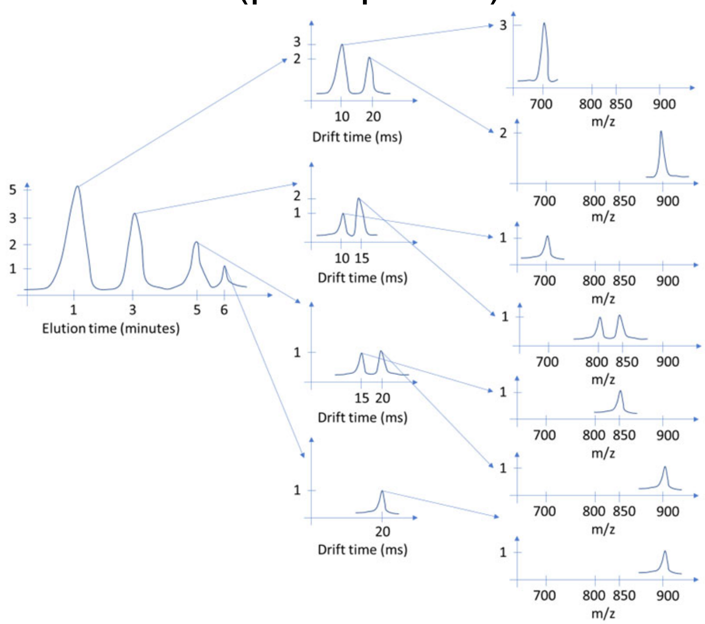
https://activednsproject.org/

Pacific Northwest
NATIONAL LABORATORY

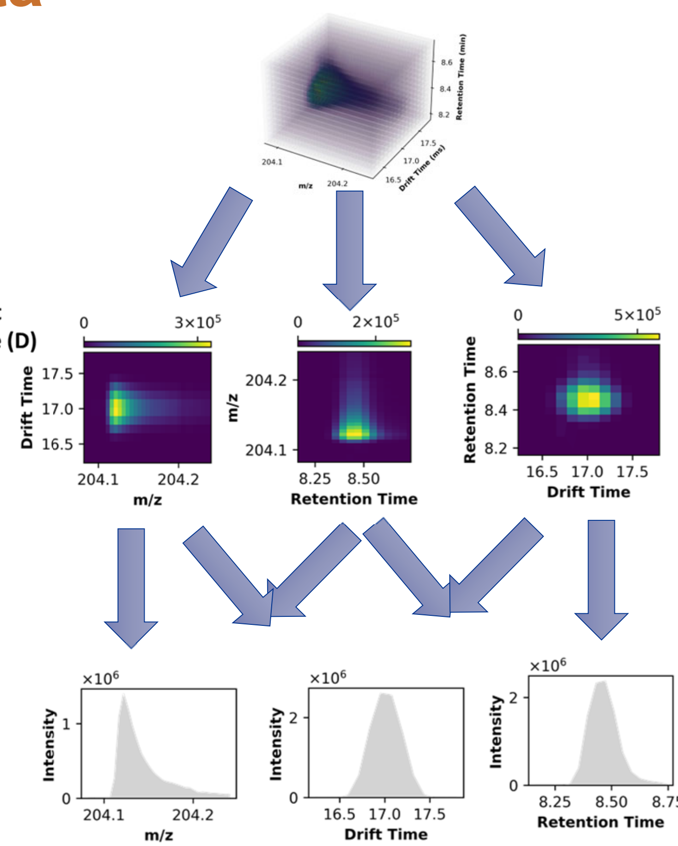
Projections of Multivariate Data

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





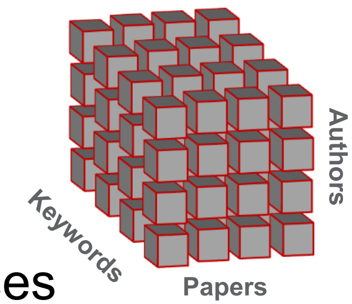
Retention Time (R)	Drift Time (D)	m/z (M)	Abundance
1	10	700	3
1	20	900	2
3	10	700	1
3	15	800	1
3	15	850	1
5	15	850	1
5	20	900	1
6	20	900	1



Colby, Sean; Shapiro, Madelyn; Bilbao, Aivett; Broeckling, C; Lin, Andy; Purvine, Emilie; Joslyn, Cliff A: (2023) ``Introducing Molecular Hypernetworks for Discovery in Multidimensional Metabolomics Data", submitted to *J Proteome Research*

A Discrete Relation

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



Paper	Authors	Keywords
1	Andrews, Davis	Graphs
2	Andrews, Carter, Davis	Topology
3	Davis	Graphs, topology
4	Andrews, Bailey	Lattices
5	Bailey, Carter	Lattices, topology

$A \times P \times K$

	1	2	3	4	5
Andrews		X		X	
Bailey	X			X	X
Carter		X			X
Davis	X	X	X		

$A \times P$

	1	2	3	4	5
Lattices				X	X
Graphs	X		X		
Topology		X	X		X

$K \times P$

	Lattices	Topology	Graphs
Andrews	X	X	X
Bailey	X	X	
Carter	X	X	
Davis		X	X

$A \times K$

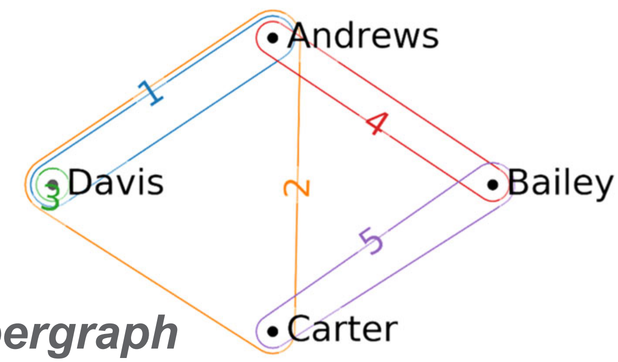
Hypergraphs Instead of Graphs

$A \times P$

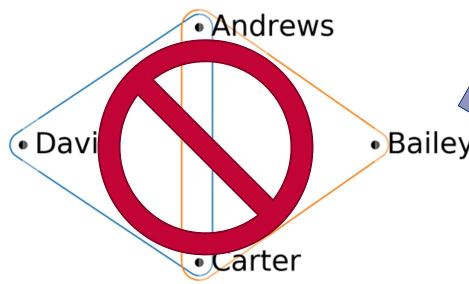
Paper #	Authors
1	Andrews, Davis
2	Andrews, Carter, Davis
3	Davis
4	Andrews, Bailey
5	Bailey, Carter

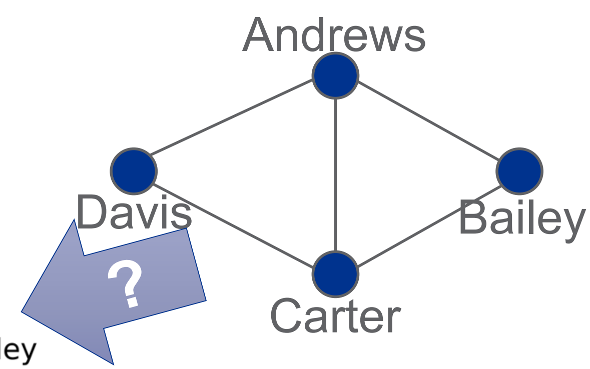
Coauthorship Matrix $A \times A$

	Andrews	Bailey	Carter	Davis
Andrews		X	X	X
Bailey	X		X	
Carter	X	X		X
Davis	X		X	



Hypergraph **representation**





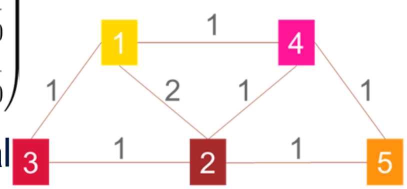
Graph **representation**

Graphs, Hypergraphs, and Relations

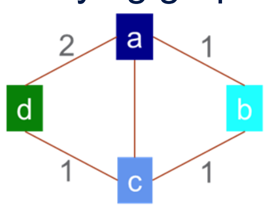
$D^* =$ edge size distribution

$$I^T I - D^* = \begin{pmatrix} 0 & 2 & 1 & 1 & 0 \\ 2 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Line graph of primal =
2-section of dual



2-section of primal
Clique expansion =
Underlying graph

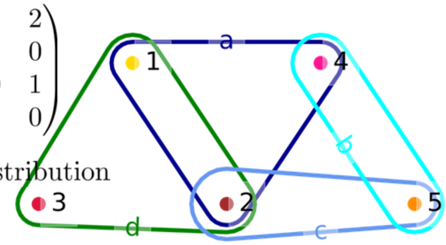


I E V

	1	2	3	4	5
a	X	X		X	
b				X	X
c		X			X
d	X	X	X		

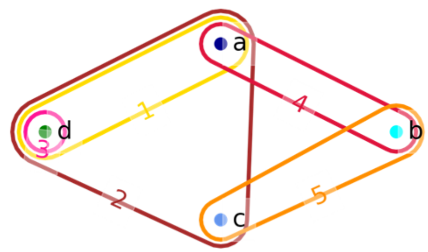
$$I I^T - D = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{pmatrix}$$

$D =$ vertex degree distribution

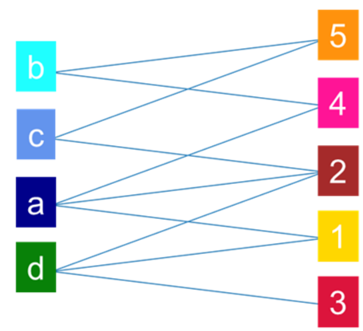


$H^* = I^T \subseteq E \times V$

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



$H = I \subseteq V \times E$

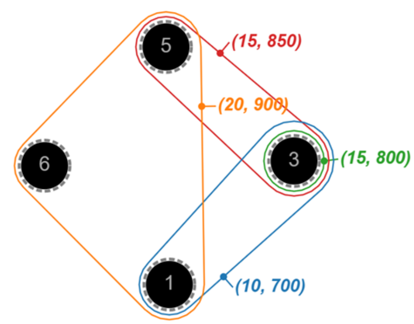


- **Hypergraph on rows ("primal"):** Multiway relations
- **Hypergraph on columns ("dual"):** Multiway relations

Network Representations of Relational View Projections

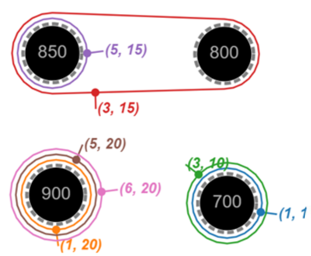
Colby, Sean; Shapiro, Madelyn; Bilbao, Aivett; Broeckling, C; Lin, Andy; Purvine, Emilie; Joslyn, Cliff A: (2023) ``Introducing Molecular Hypernetworks for Discovery in Multidimensional Metabolomics Data", submitted to *J Proteome Research*

- **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

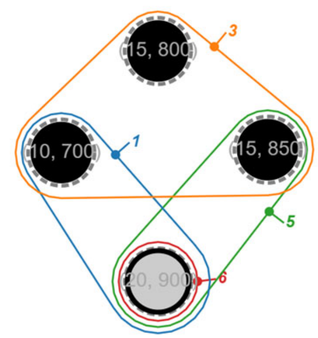


(R;<M,D>)

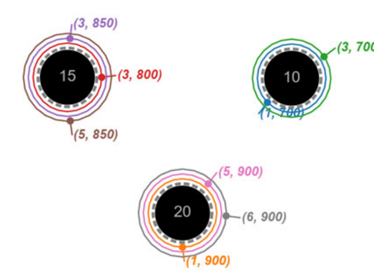
Retention Time (R)	Drift Time (D)	m/z (M)	Abundance
1	10	700	3
1	20	900	2
3	10	700	1
3	15	800	1
3	15	850	1
5	15	850	1
5	20	900	1
6	20	900	1



(M;<R,D>)



(<M,D>;R)



(D;<R,M>)

Basic Hypergraphs (undirected, unordered)

$H = \langle V, \mathcal{E} \rangle$, Family $\mathcal{E} = \{e\}, e \subseteq V$

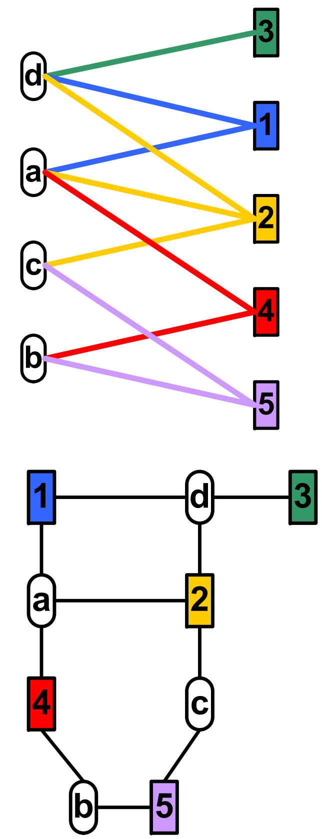
$$H = \{\{a,d\},\{a,c,d\},\{d\},\{a,b\},\{b,c\}\}$$
$$= \{ad, acd, d, ab, bc\}$$
$$= \{1{:}ad, 2{:}acd, 3{:}d, 4{:}ab, 5{:}bc\}$$

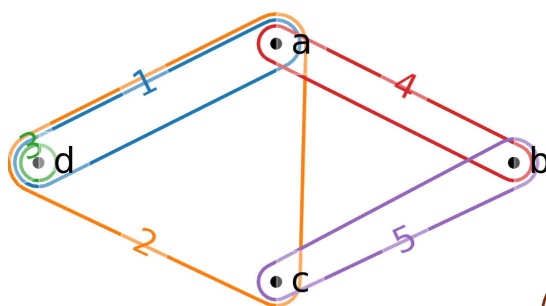
(Multi)Set System

	1	2	3	4	5
a	X	X		X	
b				X	X
c		X			X
d	X	X	X		

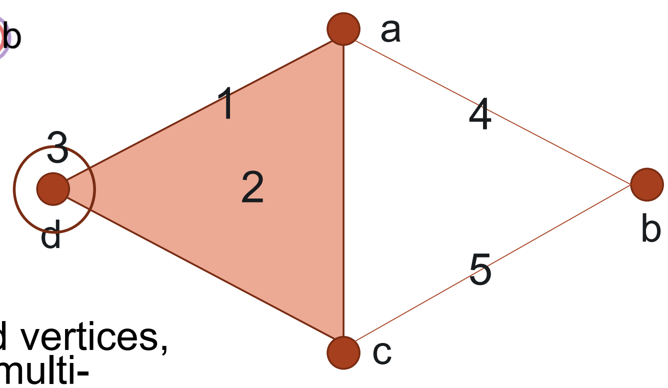
Incidence Matrix
$S = V \times \mathcal{E}$

Bipartite Graph





Euler Diagram



Simplicial Diagram

- **Axioms matter!**
  - Singletons *v* vs. *{v}*, isolated vertices, empty edges, multi-edges, multi-vertices, self-loops

Categorical Hypergraph Foundations

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

Theorem 1. *The following are categorically equivalent:*

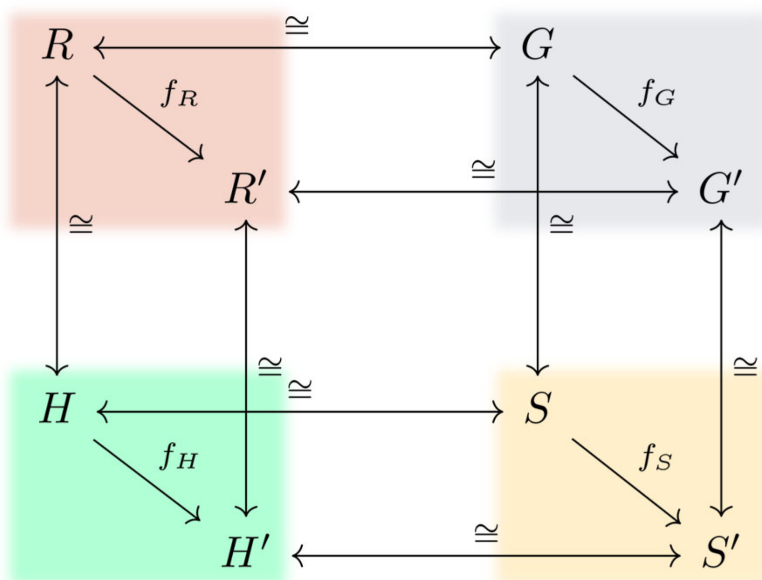
$R \subseteq X \times Y$

Binary Relations

$H = (V, E, I)$

$I : V \times E \to \{0, 1\}$

Hypergraphs: Incidence function



$G = \langle X \sqcup Y, F \rangle,$
$F \subseteq \binom{X \sqcup Y}{2} - \left(\binom{X}{2} \cup \binom{Y}{2}\right)$

Bipartite Graphs

$S = \langle V, E \rangle, E \subseteq 2^V$

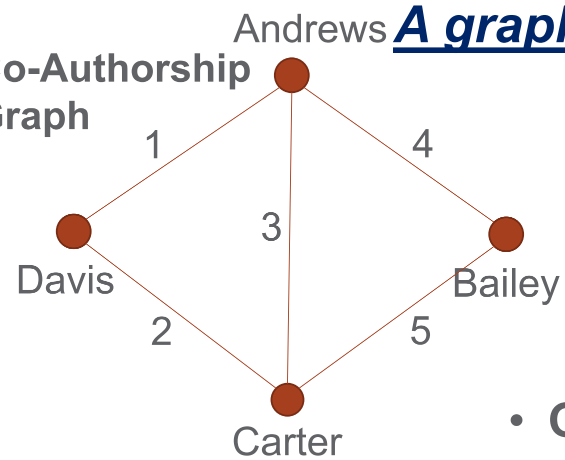
Hypergraphs: Set system

E must be a multiset, or an indexed family of subsets

Graphs vs. Hypergraphs: Precis

A graph is 2-uniform hypergraph

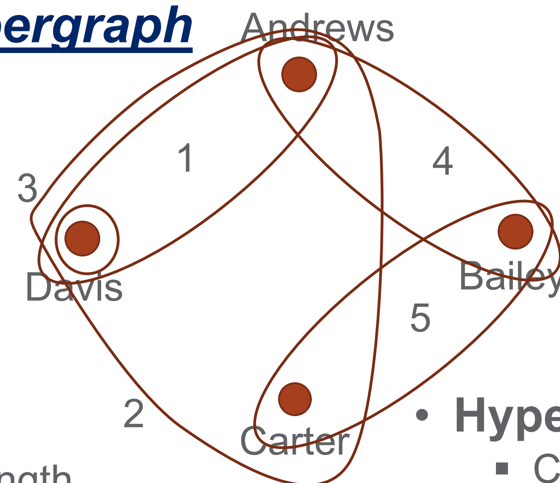
Co-Authorship Graph



	1	2	3	4	5
a	X		X	X	
b				X	X
c		X	X		X
d	X	X			

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

Collaboration Hypergraph



	1	2	3	4	5
a	X	X		X	
b				X	X
c		X			X
d	X	X	X		

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

Burgeoning Movement in Network Science

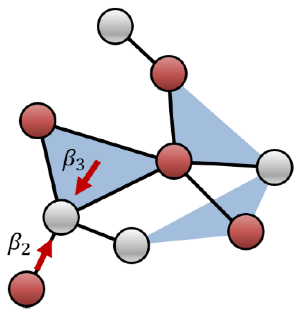
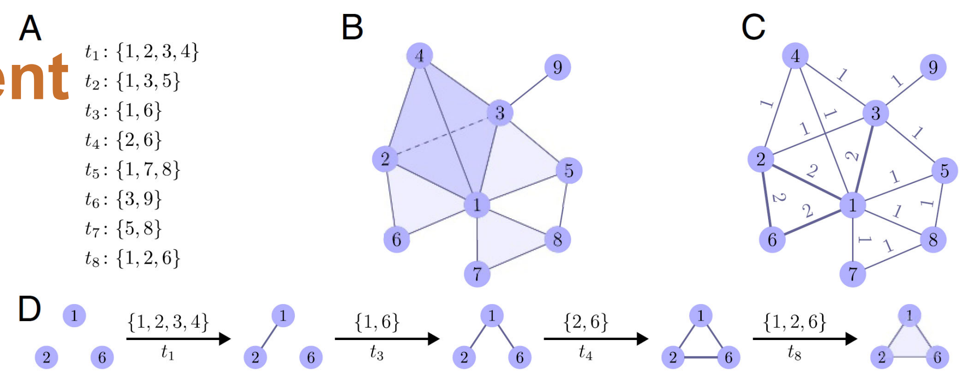


FIG. 1. Illustration of a hypergraph. Infected nodes (red) infect a healthy node (grey) via hyperedges of sizes 2 and 3 with rates β_2 and β_3 respectively.

Landry, Nicholas and Restrepo, Juan G: (2020) "The Effect of Heterogeneity on Hypergraph Contagion Models", *Chaos*, 30:10, pp. 3117, https://doi.org/10.1063/5.0020034

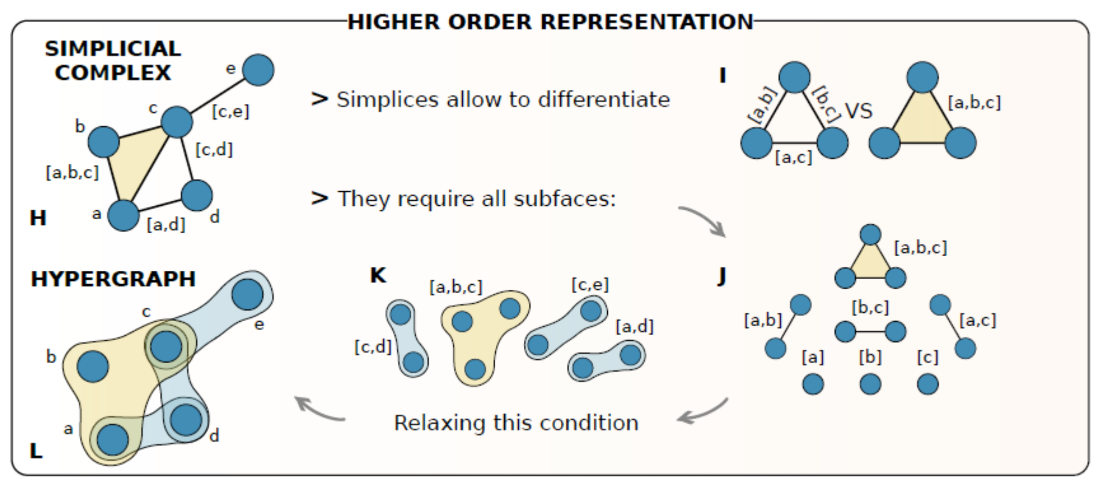


Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon Kleinberg: (2018) "Simplicial closure and higher-order link prediction", PNAS November 27, 2018 115 (48) E11221-E11230; https://doi.org/10.1073/pnas.1800683115

Bick, Christian; Gross, Elizabeth; Harrington, Heather A; and Schaub, Michael T: (2021) "What Are Higher Order Networks?", https://arxiv.org/abs/2104.11329

Leo Torres, Ann S. Blevins, Danielle S. Bassett, Tina Eliassi-Rad: (2021) "The why, how, and when of representations for complex systems", SIAM Review, 63:3, pp. 435–485

Federico Battistona, Giulia Cencettib, Iacopo Iacopini, Vito Latora, Maxime Lucash, Alice Pataniak, Jean-Gabriel Young, Giovanni Petri: (2020) "Networks beyond pairwise interactions: Structure and dynamics", Physics Reports, Volume 874, 25 Pages 1-92, https://doi.org/10.1016/j.physrep.2020.05.004

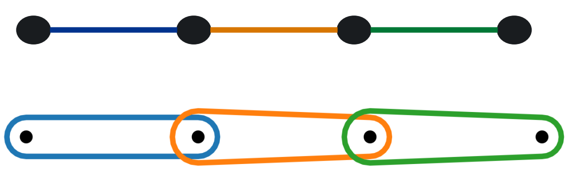


Iacopini, Iacopo; Petri, Giovanni; Barrat, Alain; and Latora, Vito: (2019) "Simplicial Models of Social Contagion", *Nature Communications*, v. 10, p. 2485

1. Hypergraph Walks Have Length *and Width*

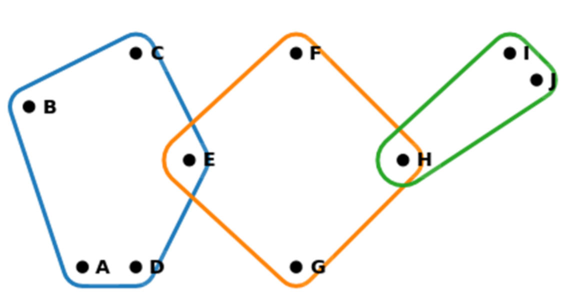
- **Hypergraph Paths Have *Width*:** Minimum edge intersection
- ***s*-walk:** Sequence $\langle e_i \rangle_{i=1}^{n}$ when $s \leq \min_{e_i, e_{i+1}} |e_i \cap e_{i+1}|, i = 1 \ldots n-1$

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

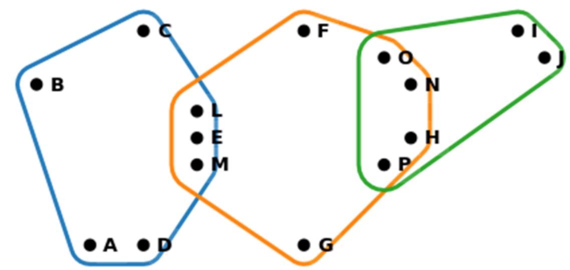


As a 2-uniform HG

Two Hypergraph Paths:
Same (edgewise) length = 2



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



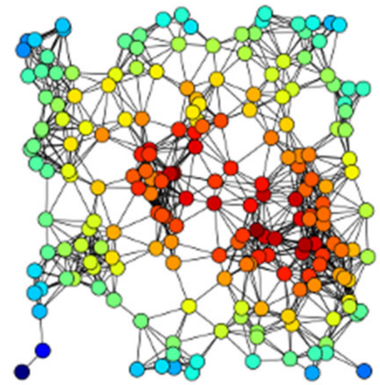
***Strong* interactions:**
Width=*s*=3

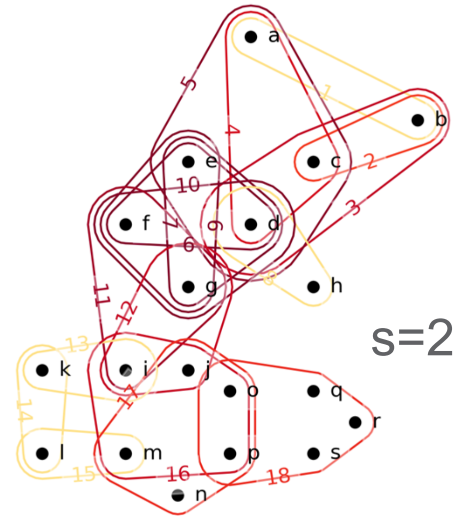
- **Extend generally:**
  - s-distance: $d_s(e,f) = \begin{cases} \min |s\text{-walk}(e,f)| & \text{if exists} \\ \infty & \text{otherwise} \end{cases}$
  - s-components, s-centrality, s-diameter s-motifs, s-clustering coefficient

SG Aksoy, CA Joslyn, CO Marrero, B Praggastis, EAH Purvine: (2020) "Hypernetwork Science via High-Order Hypergraph Walks", *EPJ Data Science*, v. 9:16, doi.org/10.1140/epjds/s13688-020-00231-0

s-Closeness centrality

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



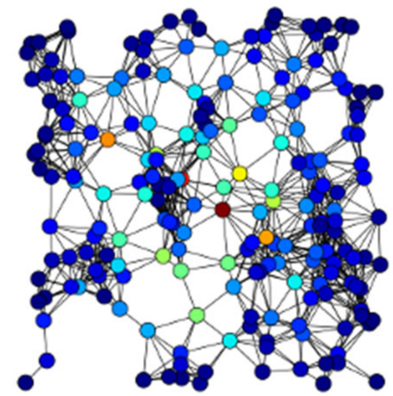


	Closeness Centrality	**Harmonic Closeness Centrality**
Graphs	$C(v) = \frac{\lvert V\rvert - 1}{\sum_{u \in V} d(v,u)}$	$HC(v) = \frac{1}{\lvert V\rvert - 1} \sum_{u \in V} \frac{1}{d(v,u)}$
Hypergraphs $E_s = \{e \in E : \lvert e\rvert \geq s\}$	$C_s(e) = \frac{\lvert E_s\rvert - 1}{\sum_{f \in E_s} d_s(e,f)}$	$HC_s(e) = \frac{1}{\lvert E_s\rvert - 1} \sum_{f \in E_s} \frac{1}{d_s(e,f)}$

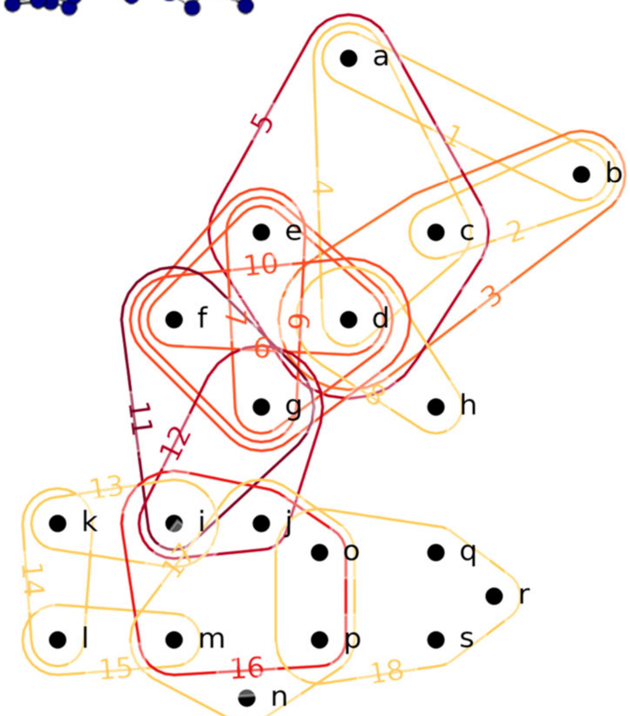
Image credit: Wikipedia user Tapiocozzo, https://en.wikipedia.org/wiki/Centrality

Pacific Northwest NATIONAL LABORATORY

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

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

Image credit: Wikipedia user Tapiocozzo, https://en.wikipedia.org/wiki/Centrality

Feng, S., Heath, E., Jefferson, B., Joslyn, C., Kvinge, H., Mitchell, H.D., Praggastis, B., Eisfeld, A.J., Sims, A.C., Thackray, L.B., Purvine, E., et al. 2021. Hypergraph models of biological networks to identify genes critical to pathogenic viral response. *BMC Bioinformatics*, 22(1), pp.1-21.

Pacific Northwest NATIONAL LABORATORY

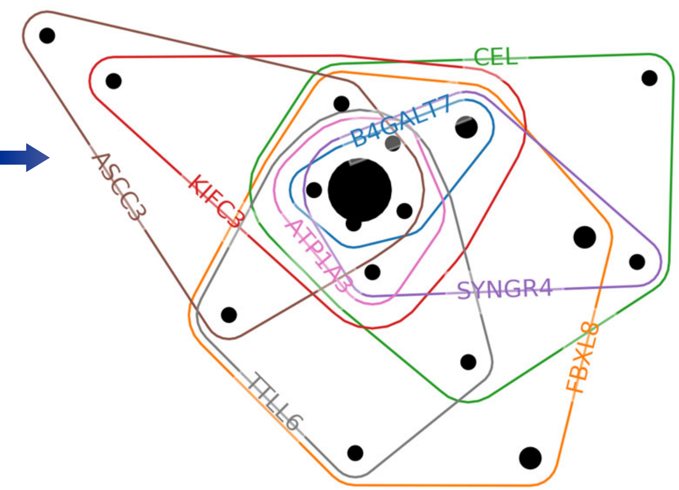
Example: Biological Data

- **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
    - ✓ Log2(sample / control) for each [sample, gene] pair
- **Hypergraph:**
  - Nodes = conditions (virus, strain, cell type, time point, …)
  - Edges = genes
  - Node/edge containment = genes with log2(fold change) z-score ≥ 2 and p-value < 0.05 for a given condition

Protein	EB1_WT_0h__b	EB1_WT_00h__b	EB1_WT_8h__b
AAAS	-0.053539	-0.021629	0.069450
AACS	0.031504	0.131252	0.309998
AADAC	-0.041660	0.031732	-0.106712
AAK1	0.139425	0.148185	0.251240
AAMP	0.139837	0.005684	0.077773
AARS	0.011162	0.106871	0.202427
AARS2	0.120218	0.040178	0.193941

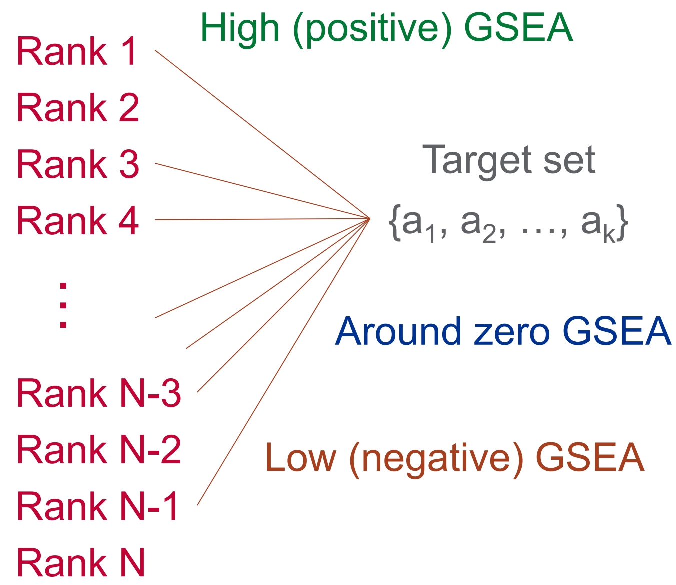
. . .

⋮



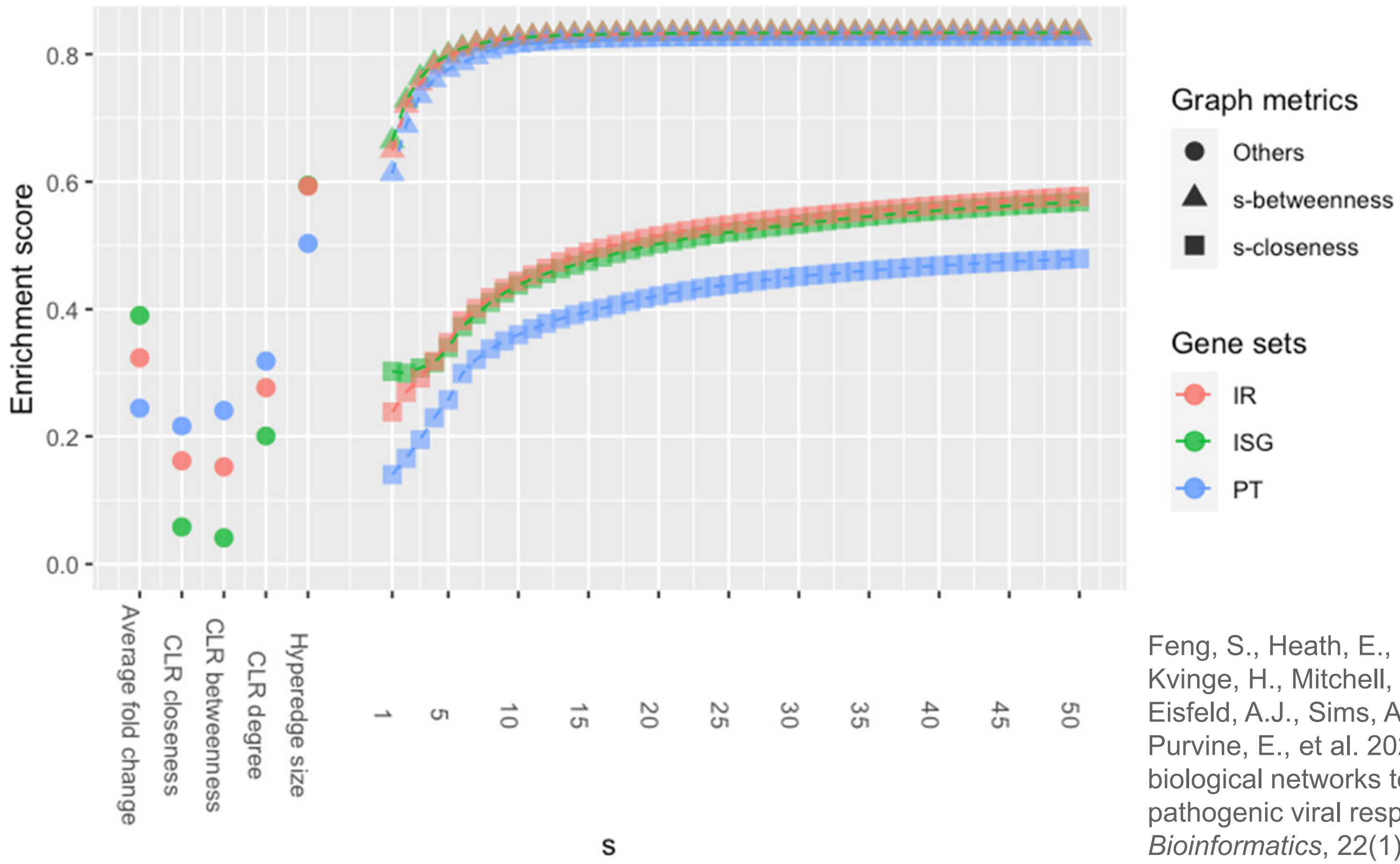
Hypergraphs for identifying important genes

- **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



Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences* 102.43 (2005): 15545-15550.

Gene Enrichment Scores



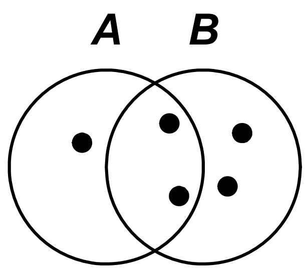
Feng, S., Heath, E., Jefferson, B., Joslyn, C., Kvinge, H., Mitchell, H.D., Praggastis, B., Eisfeld, A.J., Sims, A.C., Thackray, L.B., Purvine, E., et al. 2021. Hypergraph models of biological networks to identify genes critical to pathogenic viral response. *BMC Bioinformatics*, 22(1), pp.1-21.

2. Hypergraphs Have Hierarchy

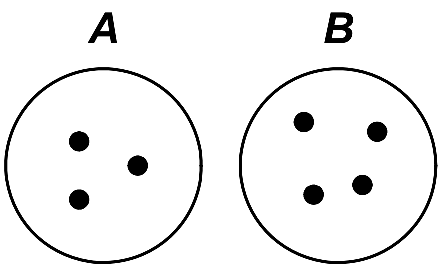
Hypergraph

$|A| = 3, |B| = 4$

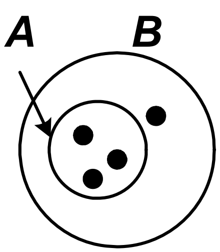
$A \cap B$ $\quad$ $|A \cap B|$



$= \{v_1, \ldots, v_n\}$ $\quad$ ≥ 1



$= \emptyset$ $\quad$ 0

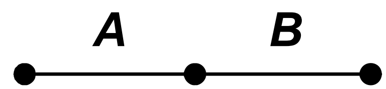


A $\quad$ $|A|$

Graph

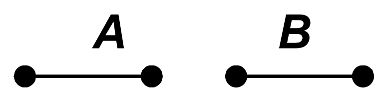
$|A| = |B| \equiv 2$

$A \cap B$ $\quad$ $|A \cap B|$



$= \{v\}$ $\quad$ $\equiv 1$

Incident Edges



$= \emptyset$ $\quad$ 0

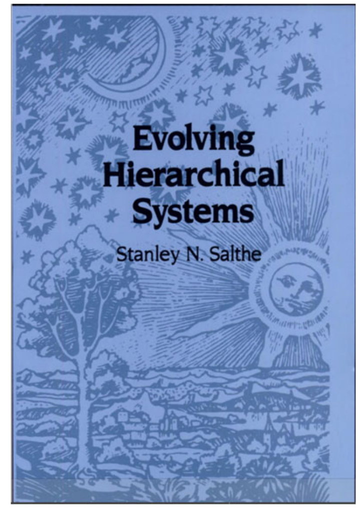
Disjoint Edges

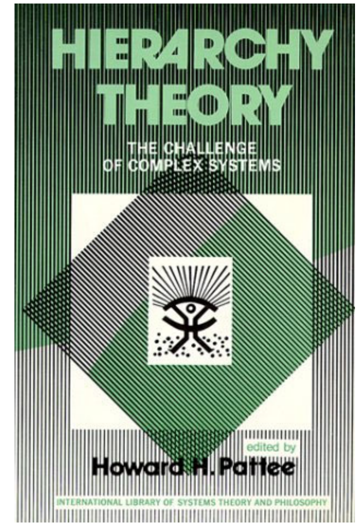
N/A

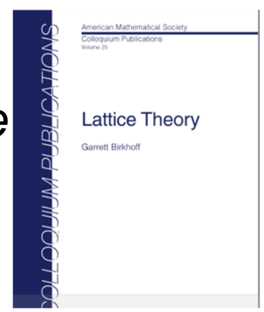
Included Edges

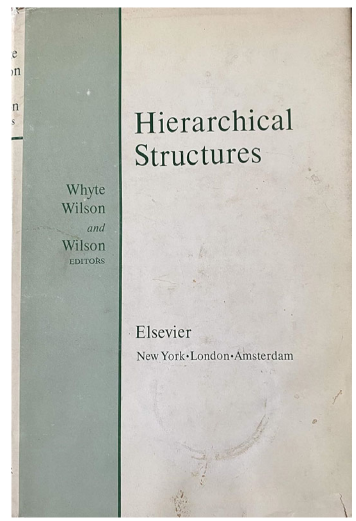
Hierarchy Theory

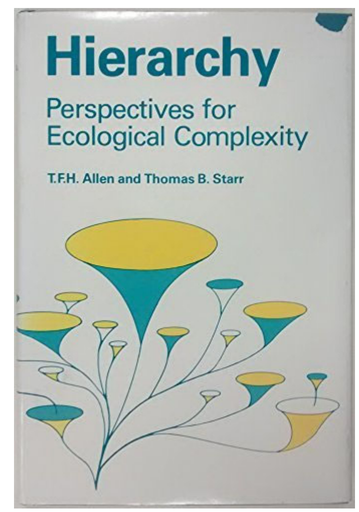
- **Systems admitting to descriptions in terms of *levels:*** Height, depth
  - *Necessary* for viable organization of large complex systems
  - Natural *scale dependencies and interactions*
- **The Systems community has attended less to *mathematical* formalism**
  - Way more than trees
  - Avoiding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



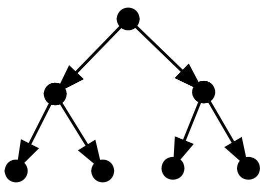




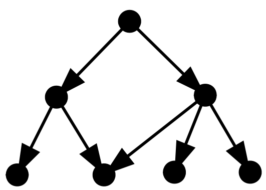




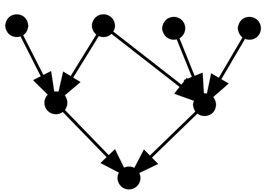
Some Aspects of Hierarchies = Partial Orders



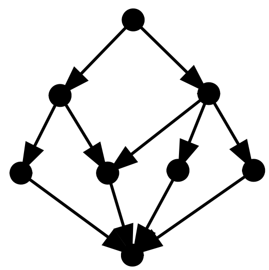
**Tree
(Unique Parents)**



**Nodes with
multiple
parents/children**

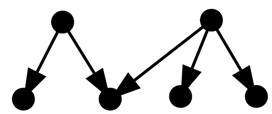


**Dual
Structure**



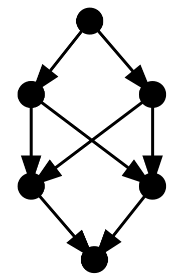
**Totally
Bounded**

**Chain
(Totally ordered)
(Unique parents
and children)**

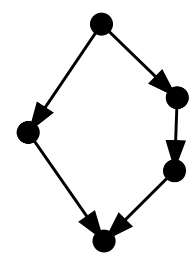


**Totally
Unbounded**

**Antichain
(Totally unordered)**



**Not a lattice
(Pairs with multiple
parents/children)**



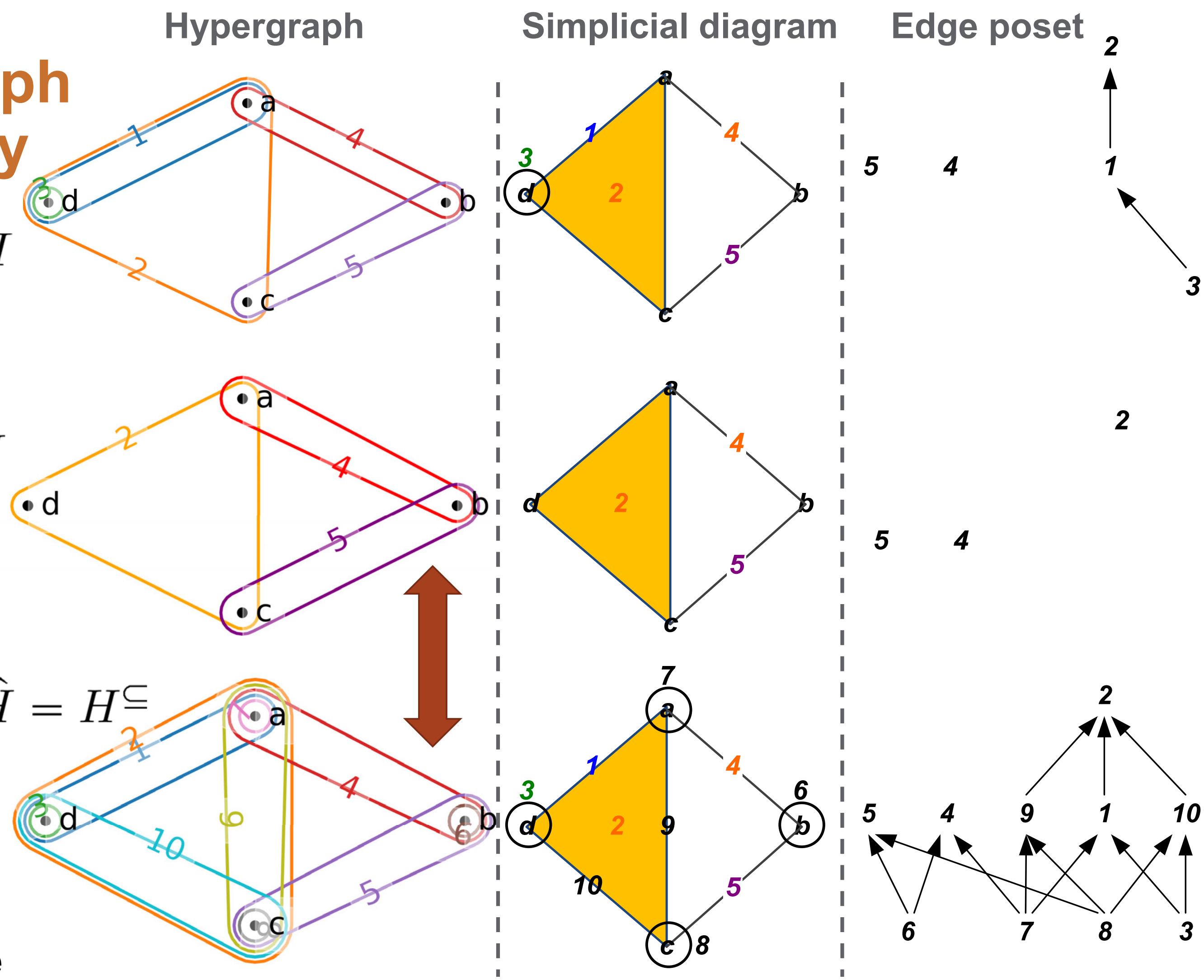
**Ungraded
(Unequal chain lengths)
(Ambiguous levels)**

Joslyn, Cliff A; Pogel, Alex and Purvine, Emilie A: (2017) "Interval-Valued Rank in Finite Ordered Sets", *Order*, v. **34**:3, pp. 491-512, https://doi.org/10.10 07/s11083-016-9411-2

Pacific Northwest
NATIONAL LABORATORY

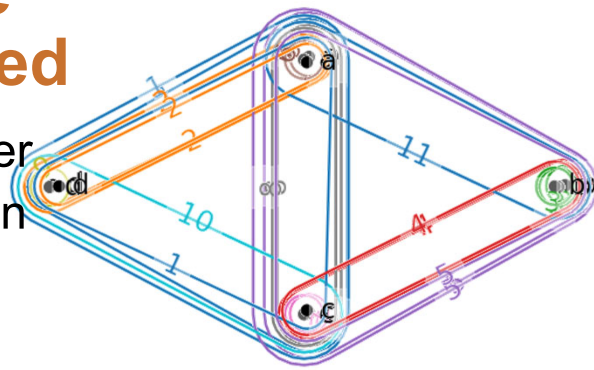
Hypergraph Inclusivity

- **Hypergraph:** H
  - **2 *included* edges:** 1, 3
  - **3 "toplexes":** Maximal hyperedges: 2, *4*, *5*
  - Inclusivity = 2/5
- **Simple hypergraph:** Remove all inclusions $\check{H}$
  - All toplexes
  - "Reduction"
  - Inclusivity = 0
- **Abstract simplicial complex (ASC):** Add all inclusions $\hat{H} = H^{\subseteq}$
  - Toplexes and all below
  - "Closure"
  - Inclusivity = 7/10
- ***All share the same topological structure:*** Determined by toplexes
- $\check{H}$ **and** $\hat{H}$ **are one-to-one**

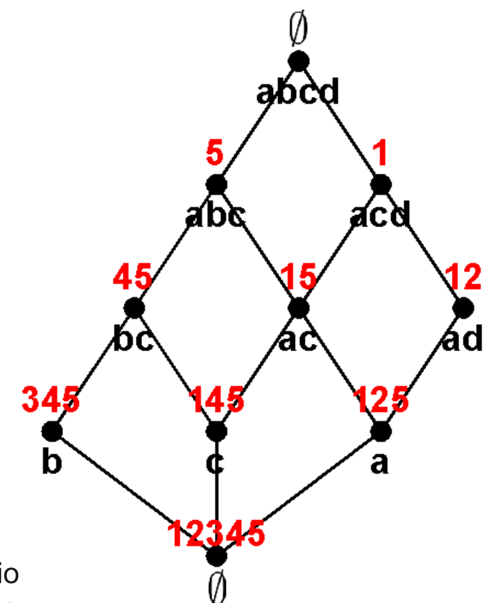


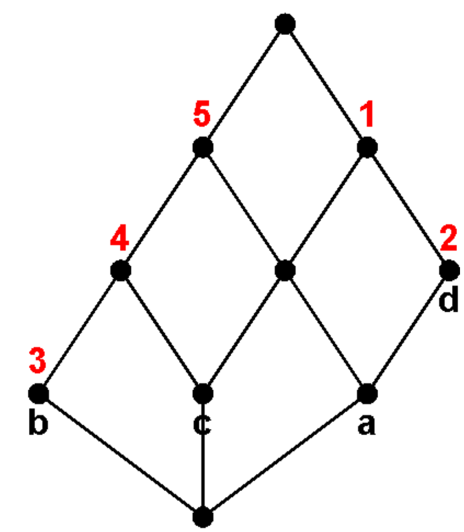
Hypergraphs Are Inherently Ordered

- 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



	1	2	3	4	5	6	7	8	9	10	11
a	x	x			x	x		x			x
b			x	x	x						x
c	x			x	x		x	x		x	
d	x	x							x	x	



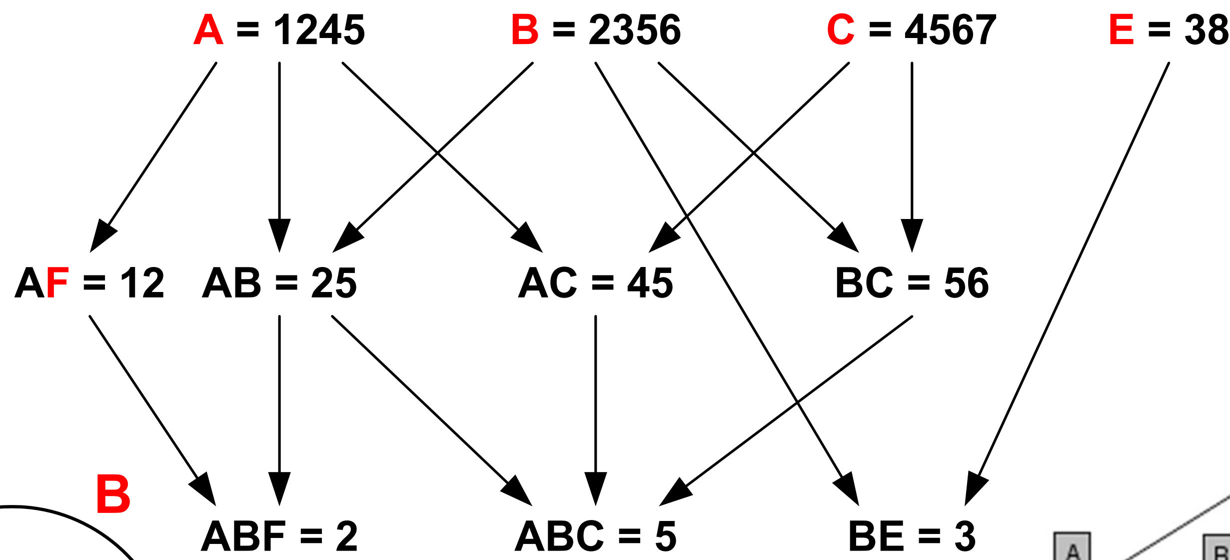


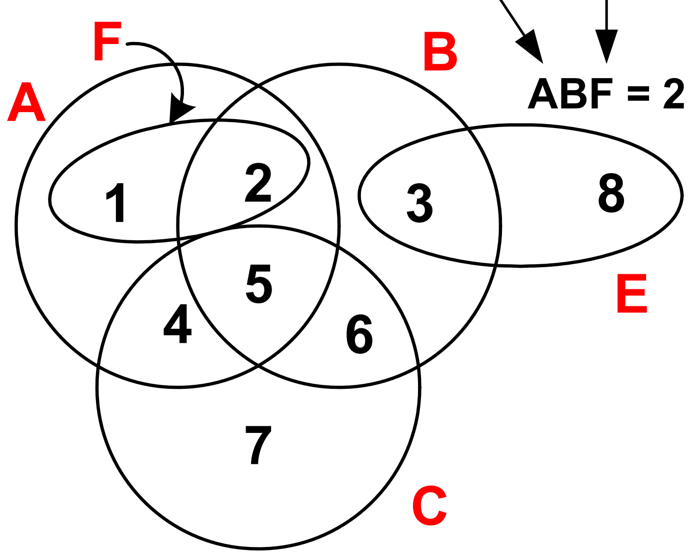
Rawson, Michael G; Myers, Audun; Green, Robert; Robinson, M; Joslyn, Cliff: (2023) ``Formal Concept Lattice Representations and Algorithms for Hypergraphs'', https://doi.org/10.48550/arXiv.2307.11681

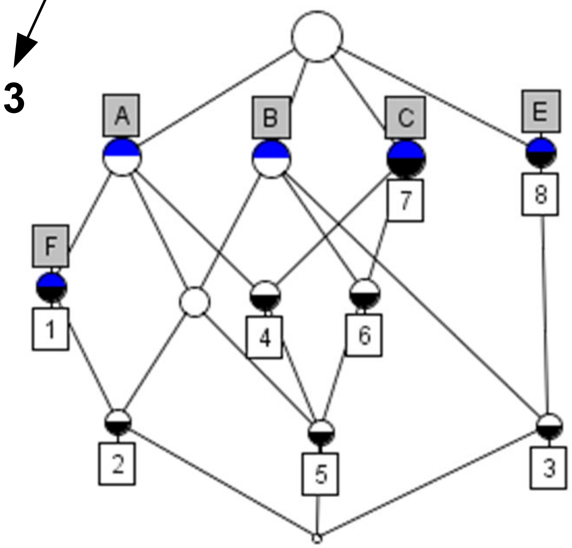
Robinson, Michael: (2022) ``Cosheaf Representations of Relatio and Dowker Complexes”, *J. Applied and Computational Topology*, v. 6, pp. 27-63

Pacific Northwest
NATIONAL LABORATORY

Example Concept Lattice of a Hypergraph







Ukraine 2014 (UKR14) Knowledge Base

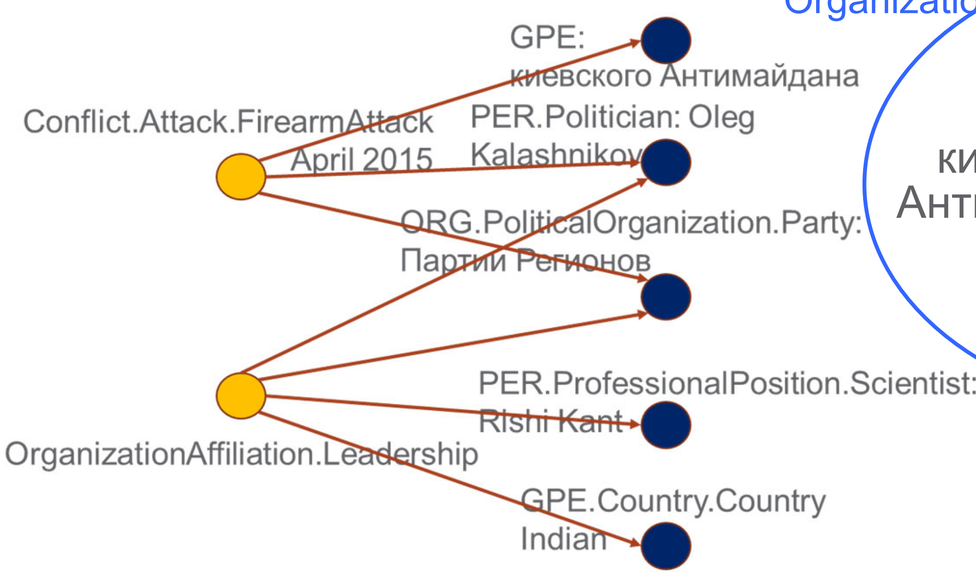
Node and edge types associated with a small graph sample.
Center node is a “relationship node” connecting two entities together.

- **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

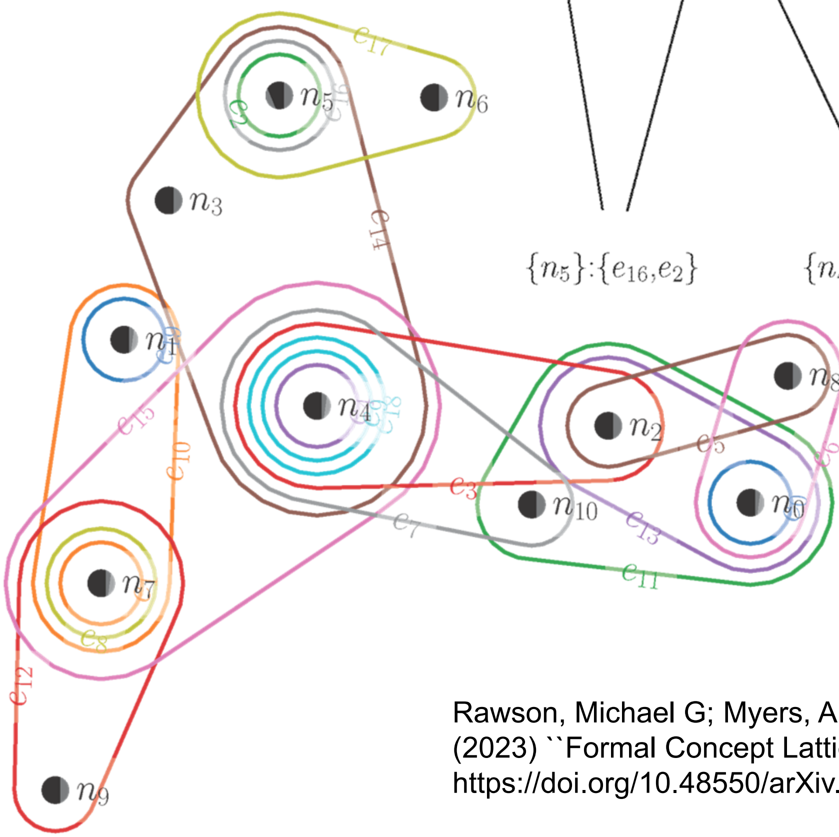
- **Node Types:** 307
- **Node Instances:** 406K
- **Edge Types:** 367
- **Edge Instances:** 302K
- **Connected Components:** 314K

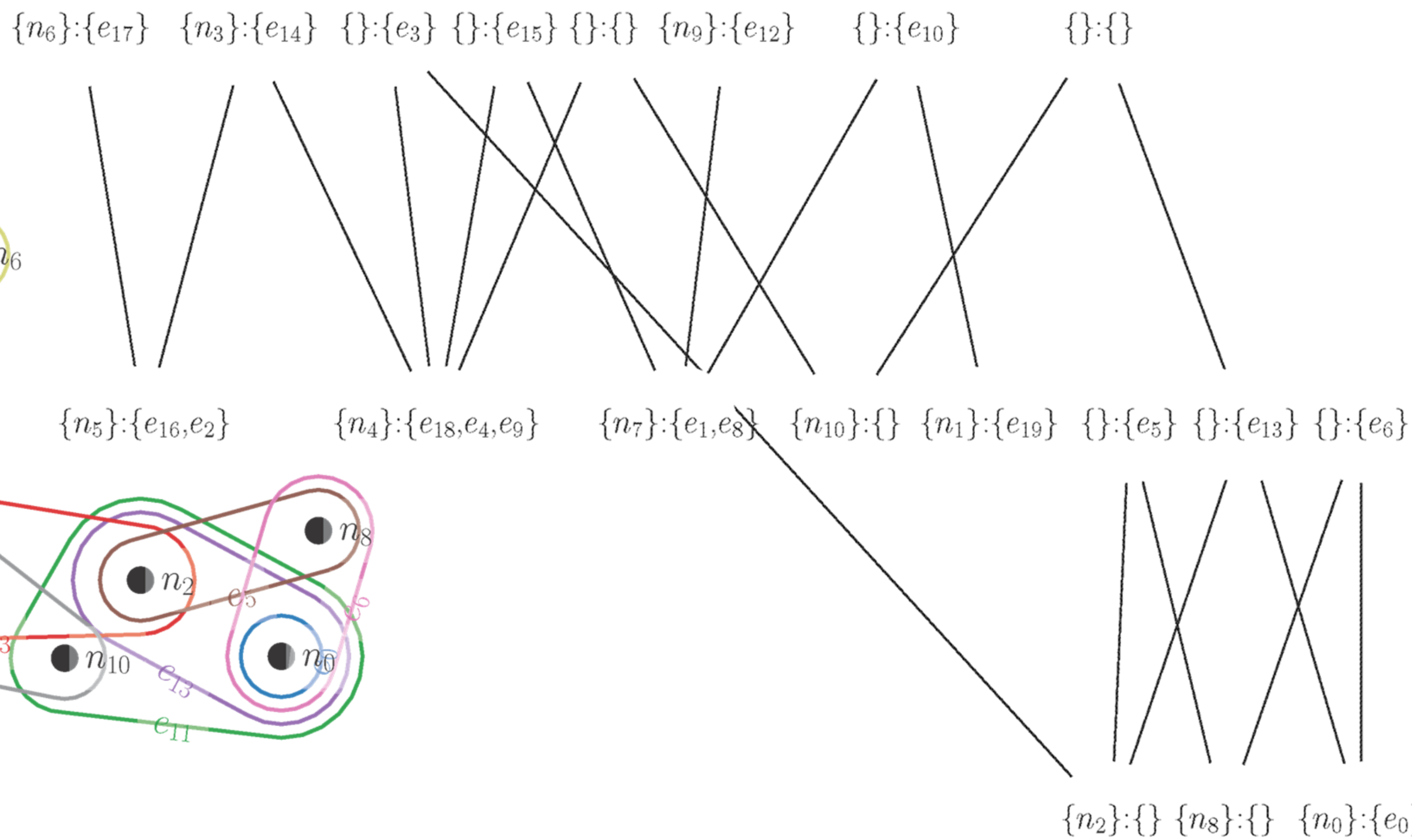
Event Hypergraph Model

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



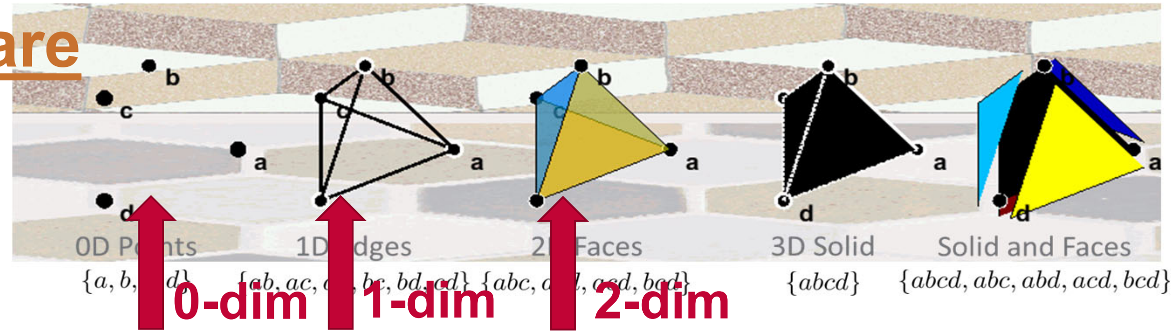
UKR14 Example Concept Lattice



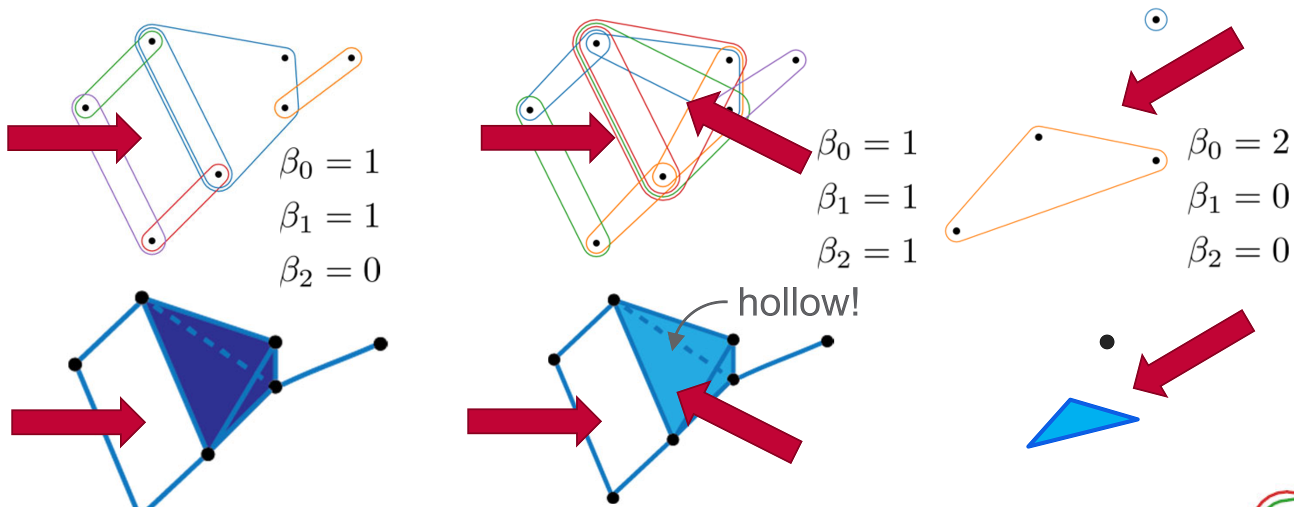


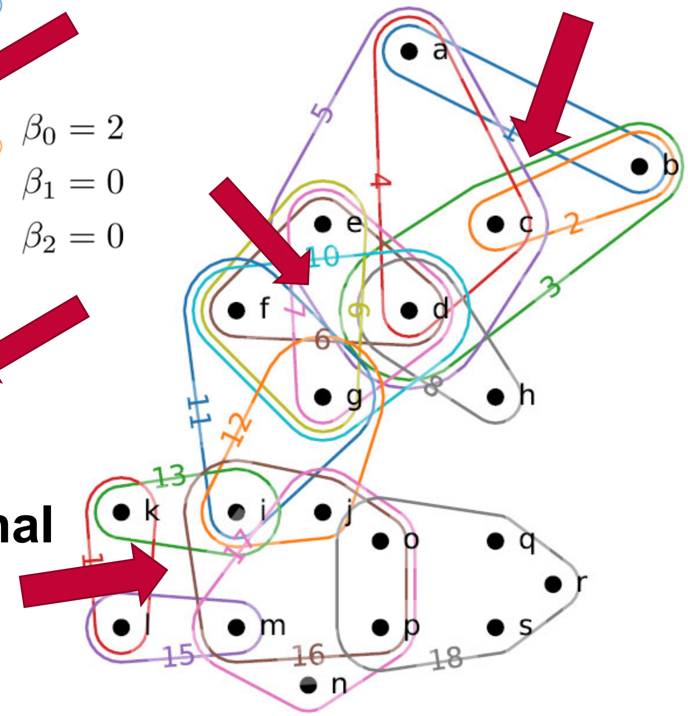
Rawson, Michael G; Myers, Audun; Green, Robert; Robinson, M; Joslyn, Cliff: (2023) ``Formal Concept Lattice Representations and Algorithms for Hypergraphs'', https://doi.org/10.48550/arXiv.2307.11681

3. Hypergraphs are Topological Objects



- **Hypergraphs have topological properties**



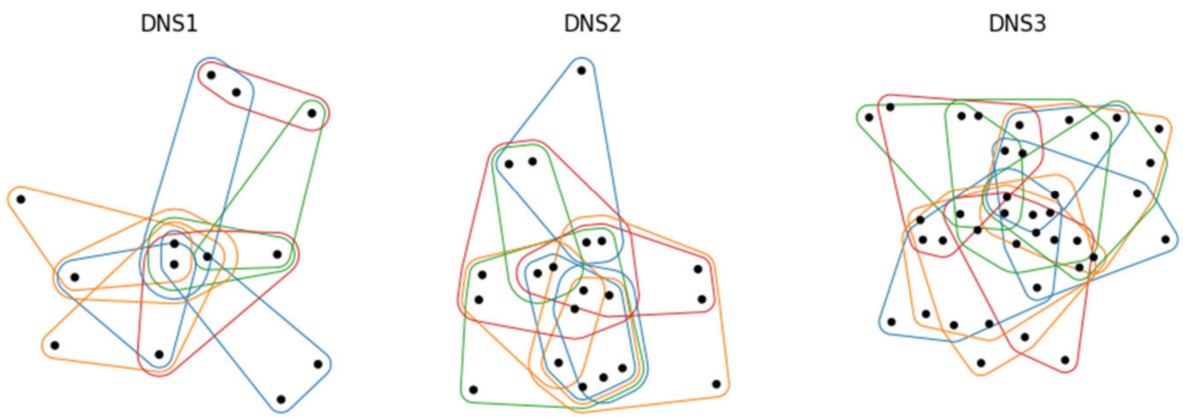


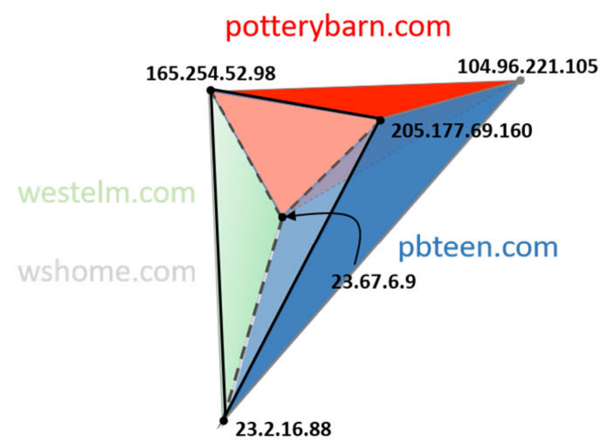
- ***(Simplicial) homology* identifies multidimensional open structures**
  - As hypotheses for missing data
  - Need for bridging metadata

Homologies Show Multidimensional Open Structures

	β_0	β_1	β_2	$\beta_{\geq 3}$
DNS1	1	1	0	0
DNS1	1	1	2	0
DNS1	1	3	1	0

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



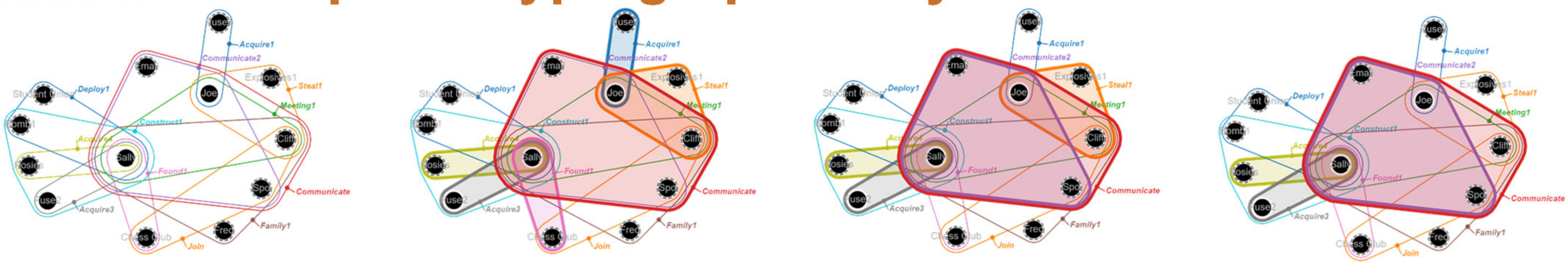


https://activednsproject.org/

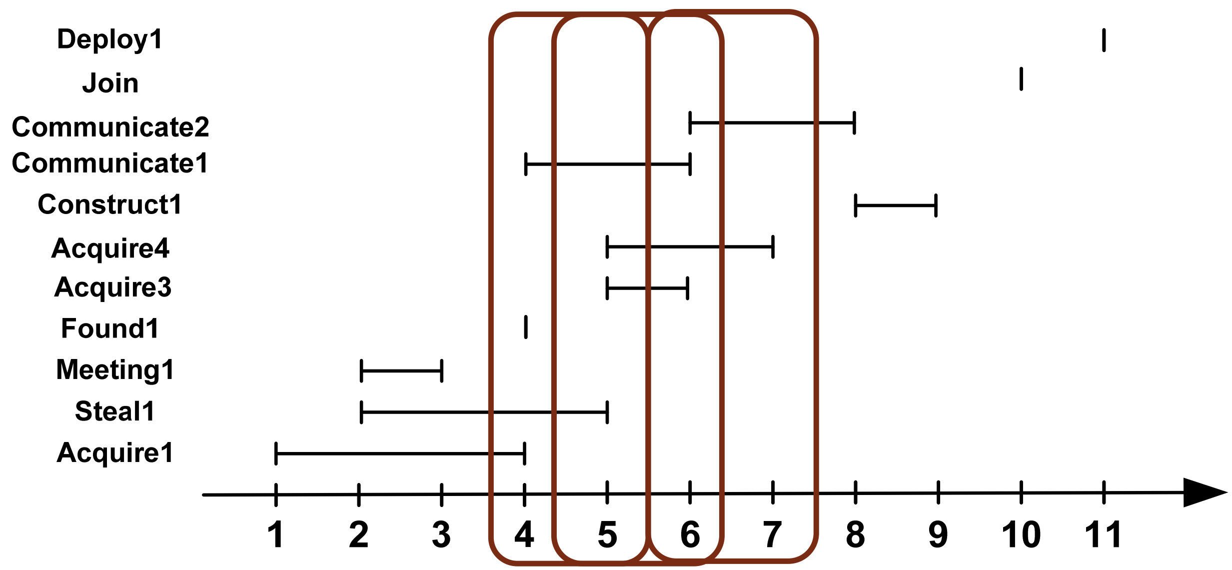
Joslyn, Cliff A; Aksoy, Sinan; Arendt, Dustin; Firoz, J; Jenkins, Louis; Praggastis, Brenda; Purvine, Emilie AH; Zalewski, Marcin: (2020) "Hypergraph Analytics of Domain Name System Relationships", 17th Wshop. on Algorithms and Models for the Web Graph (WAW 2020), *Lecture Notes in Computer Science*, v. 12901, ed. Kaminski, B *et al.*, pp. 1-15, Springer, https://doi.org/10.1007/978-3-030-48478-1_1

Pacific Northwest
NATIONAL LABORATORY

Temporal Hypergraph Analysis

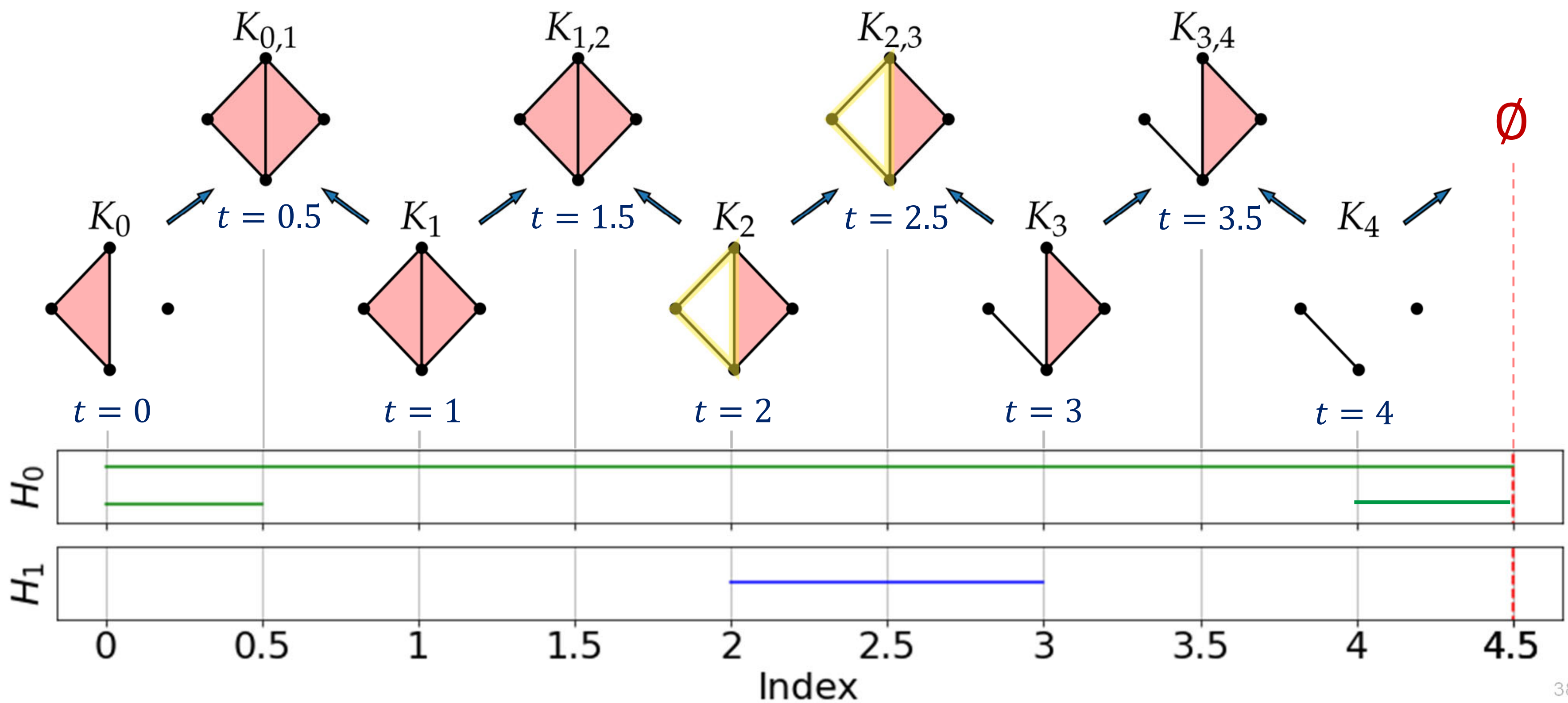


- Temporal hypergraph
- Trajectory of temporal sub-hypergraphs
- Measure change in structure, homology, distributions



Zigzag Persistence Example

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



Operationally Transparent Cyber (OpTC) data set

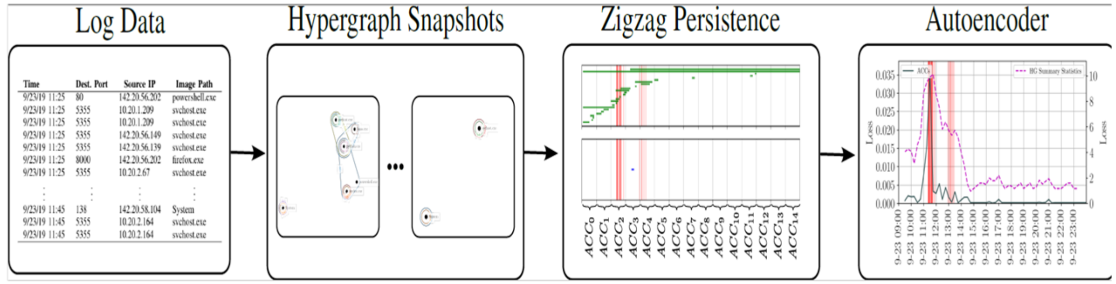
- Created by the Defense Advanced Research Projects Agency (DARPA) as part of a mission to test scaling of cyber attack detection
- Flow and host logs from both benign and malicious activity plus ground truth document describing the attack events
  - Downloading malicious PowerShell Empire, privilege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

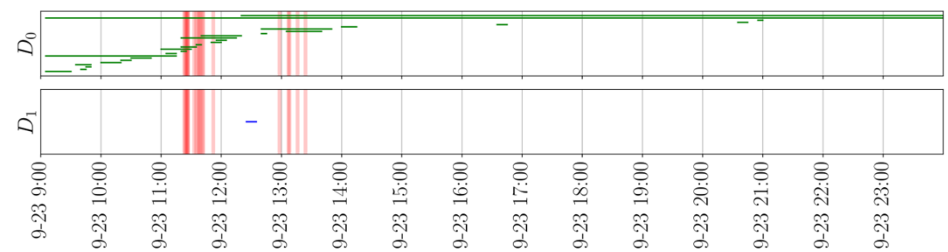
Time	Action-Object	PID	Source IP	Destination IP	Dest. Port	Executable
9/23/2023 9:06	MESSAGE-FLOW	864	10.20.2.47	224.0.0.252	5355	svchost.exe
9/23/2023 9:06	MESSAGE-FLOW	864	10.20.2.47	224.0.0.252	5355	svchost.exe
9/23/2023 9:06	MESSAGE-FLOW	864	10.20.2.93	224.0.0.252	5355	svchost.exe
9/23/2023 9:06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

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

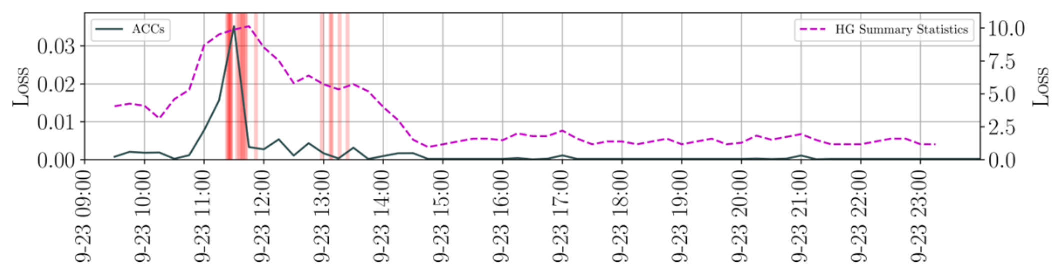
Zigzag ML Experiment on OpTC

- **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





Zigzag barcode for known ground truth IP, time windows of red team activity highlighted

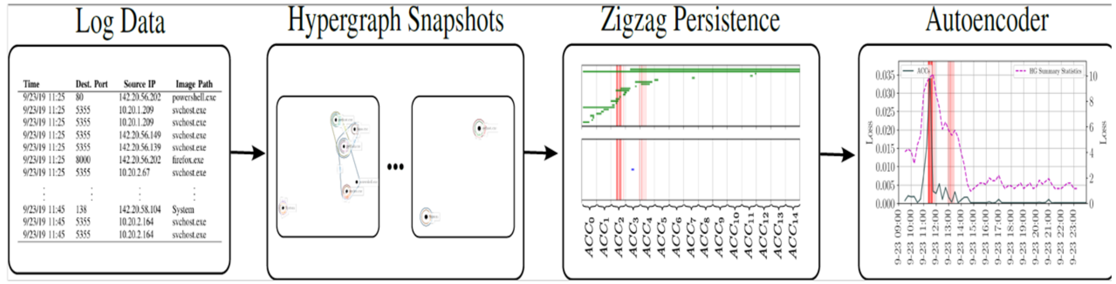


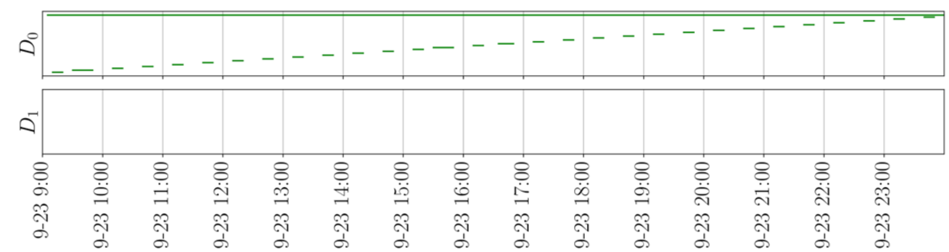
Autoencoder reconstruction loss

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

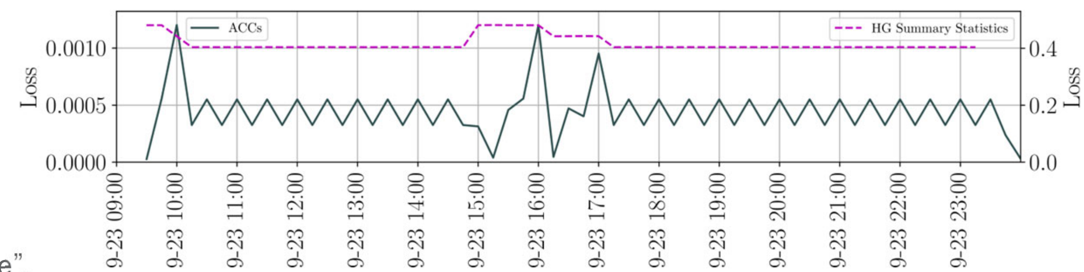
- **Goal:** identify source IPs responsible for malicious activity, and the time interval that activity occurred
- **Method:** construct temporal hypergraph sequence for each host, run zigzag persistence, train autoencoder on barcode summary
  - **Nodes:** Executable files
  - **Edges:** Destination 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





Zigzag barcode for known *benign* IP

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



Autoencoder reconstruction loss

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

Pacific Northwest
NATIONAL LABORATORY

Thank you