Summer 7-29-2020

Hypergraph Analysis of Structure Models

Cliff A. Joslyn
Pacific Northwest National Laboratory, Cliff.Joslyn@pnnl.gov

Teresa D. Schmidt
Portland State University, tds@pdx.edu

Martin Zwick
Portland State University, zwick@pdx.edu

Pacific Northwest
NATIONAL LABORATORY

Hypergraph Analysis of Structure Models

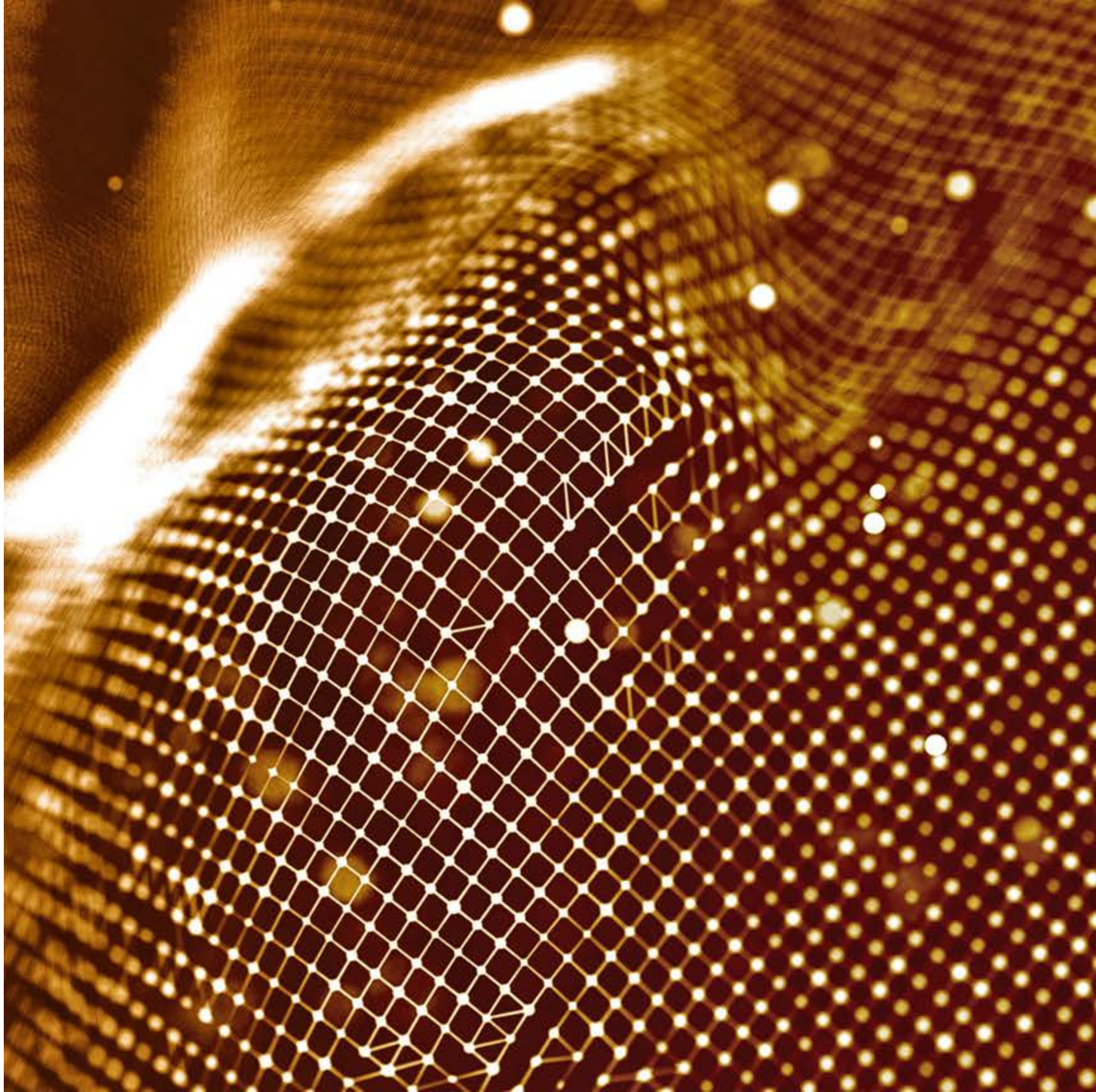
Cliff Joslyn

Teresa Schmidt, Martin Zwick

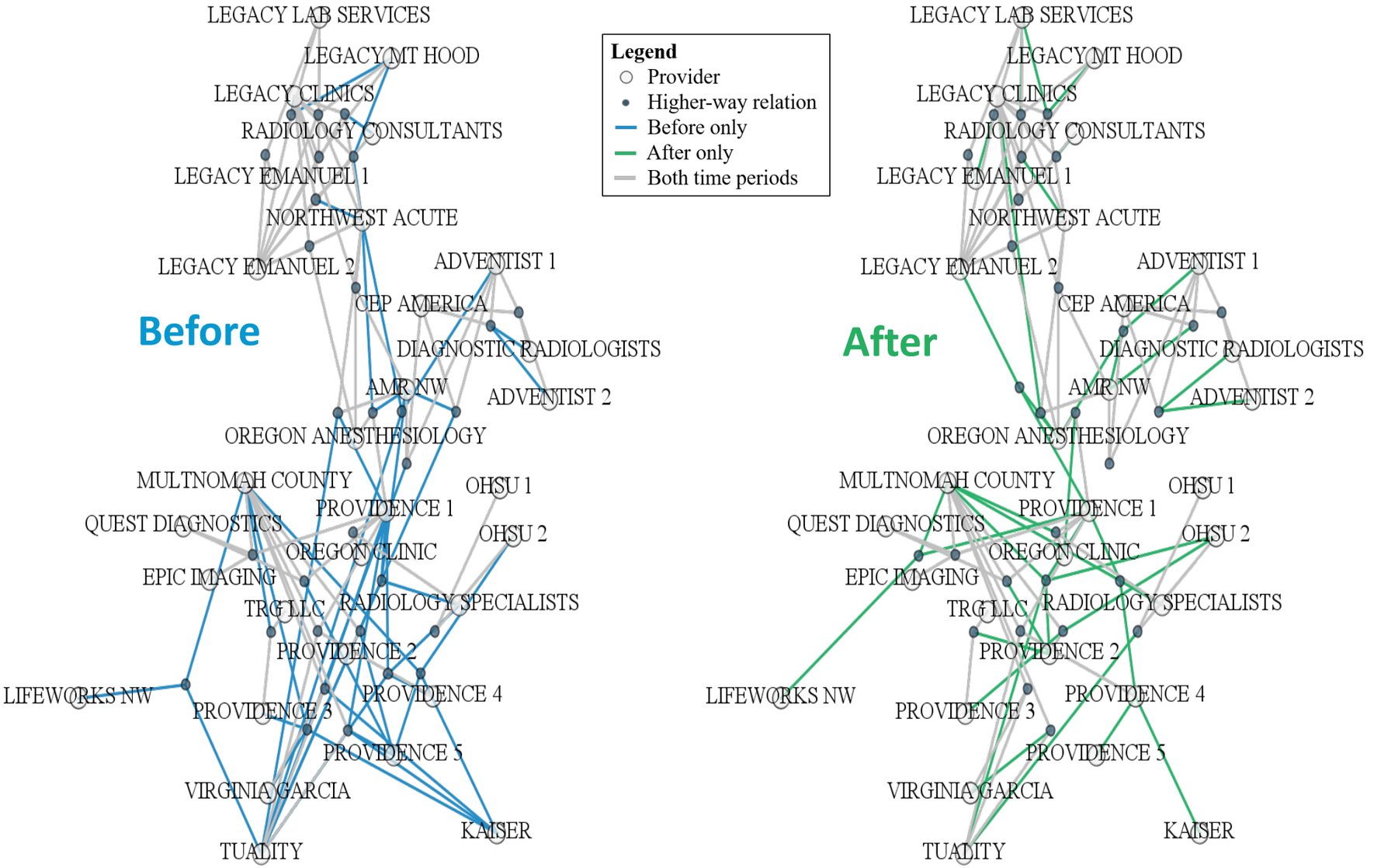
PNNL-30721

U.S. DEPARTMENT OF ENERGY BATTELLE

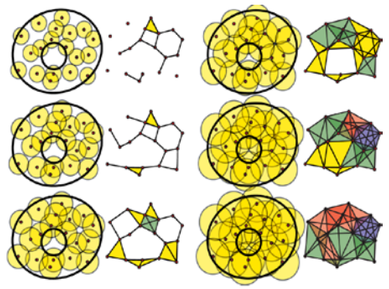
PNNL is operated by Battelle for the U.S. Department of Energy



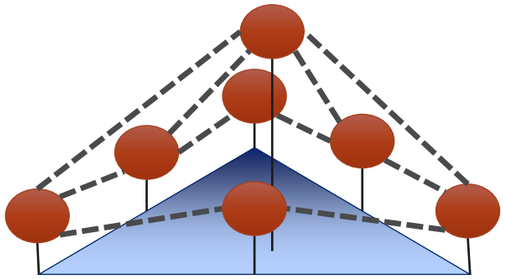
1. Introduction: what is RA
2. Input data to RA
3. Output model from RA

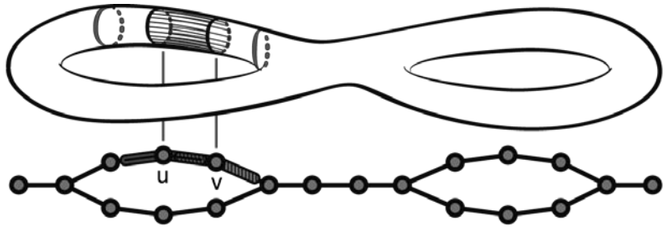

Legend
Provider
Higher-way relation
Before only
After only
Both time periods
Before
After
LEGACY LAB SERVICES
LEGACY MT HOOD
LEGACY CLINICS
RADIOLOGY CONSULTANTS
LEGACY EMANUEL 1
NORTHWEST ACUTE
LEGACY EMANUEL 2
ADVENTIST 1
CEP AMERICA
DIAGNOSTIC RADIOLOGISTS
AMR NW
ADVENTIST 2
OREGON ANESTHESIOLOGY
MULTNOMAH COUNTY
OHSU 1
PROVIDENCE 1
QUEST DIAGNOSTICS
OHSU 2
OREGON CLINIC
EPIC IMAGING
TRG LLC
RADIOLOGY SPECIALISTS
PROVIDENCE 2
LIFEWORKS NW
PROVIDENCE 4
PROVIDENCE 3
PROVIDENCE 5
VIRGINIA GARCIA
KAISER
TUALITY

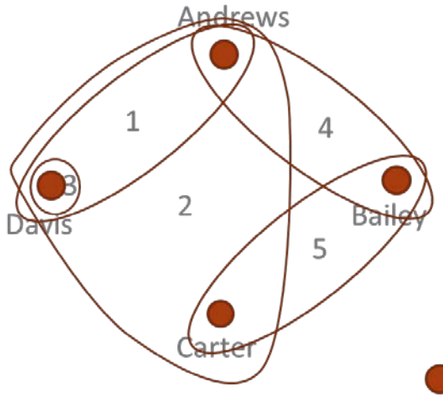
Computational Topology for Data Science

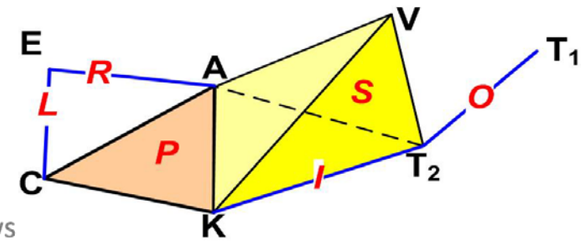


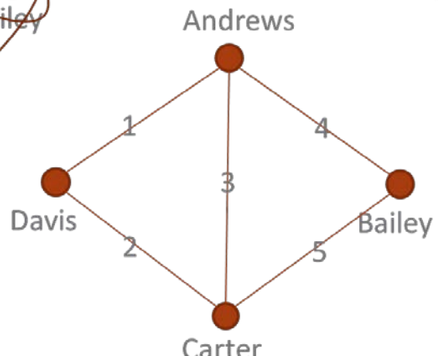
Ghrist, Robert: (2007) “Barcodes: The Persistent Topology of Data”, *Bulletin of the American Mathematical Society*, v. 45:1, pp. 61-75











Topological Sheaves

Attached Data

Topological Spaces

(Persistent) Homology

Abstract Simplicial Complexes

Included Edges

Hypergraphs

> 2 interacting elements

Graphs

Computational Topology

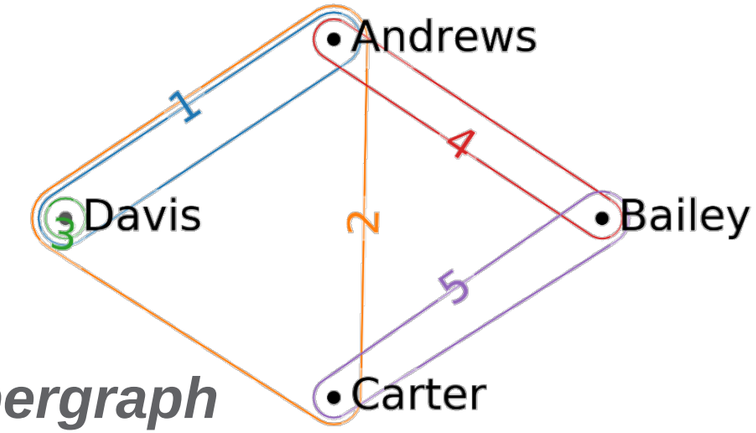
Hypernetwork Science

Network Science

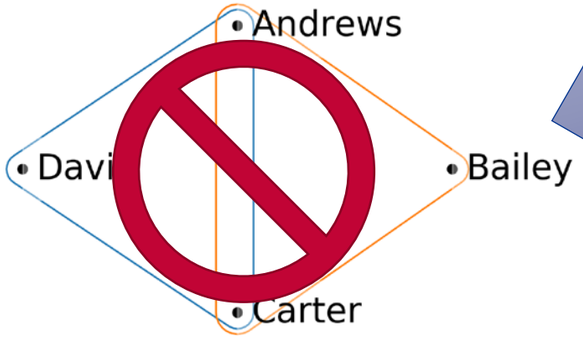
Hypergraphs vs. Graphs

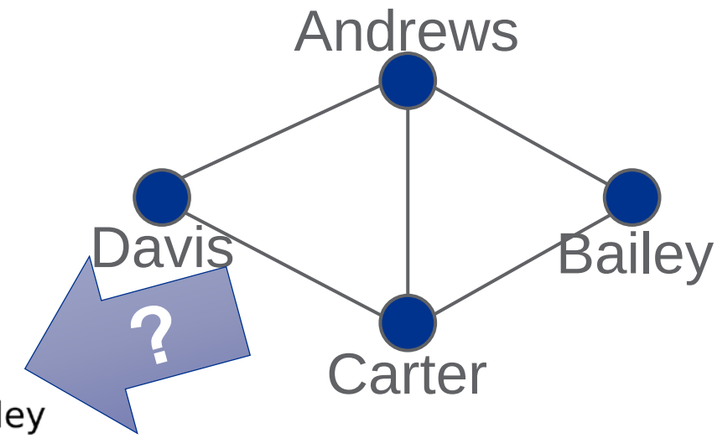
Paper #	Authors
1	Andrews, Davis
2	Andrews, Carter, Davis
3	Davis
4	Andrews, Bailey
5	Bailey, Carter

	Andrews	Bailey	Carter	Davis
Andrews		Y	Y	Y
Bailey	Y		Y	
Carter	Y	Y		Y
Davis	Y		Y	



Hypergraph **representation**





Graph **representation**

Hypergraphs: Complementary Representations

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

$H = \{\{a,d\},\{a,c,d\},\{d\},\{a,b\},\{b,c\}\}$

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

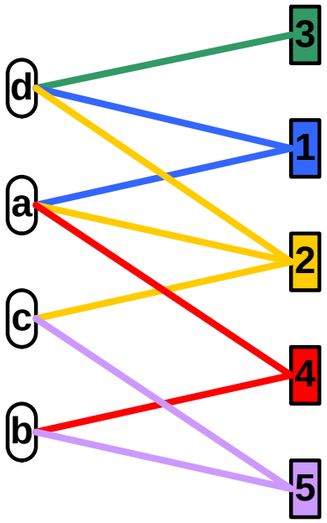
(Multi)Set System

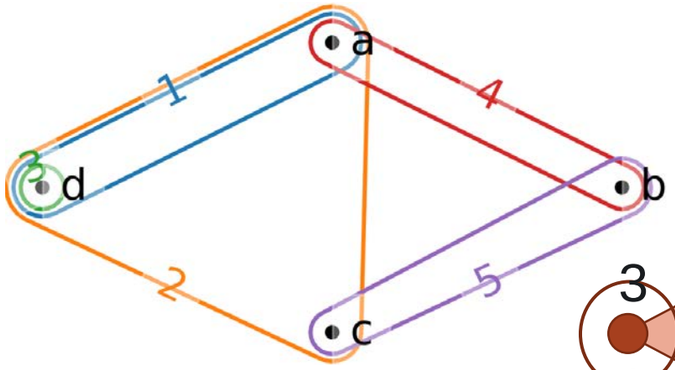
	1	2	3	4	5
a	X	X		X	
b				X	X
c		X			X
d	X	X	X		

Incidence Matrix

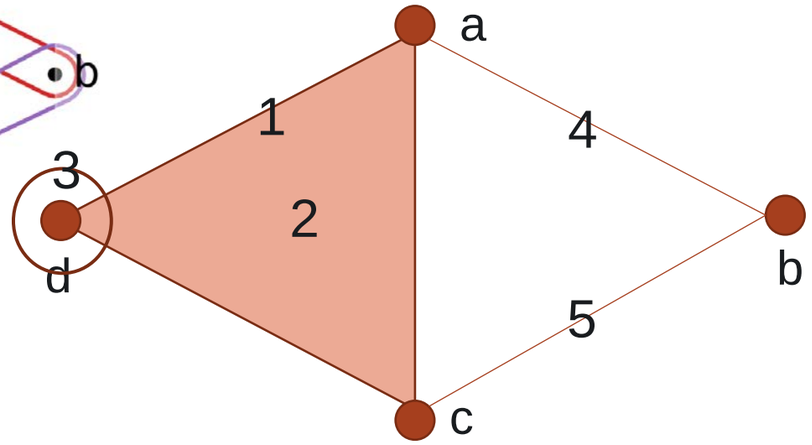
$M = V \times \mathcal{E}$

Bipartite Graph

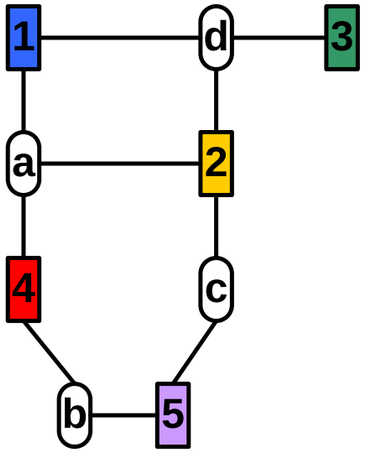




Euler Diagram

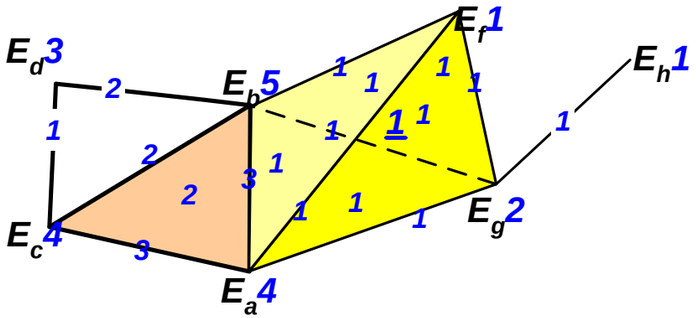


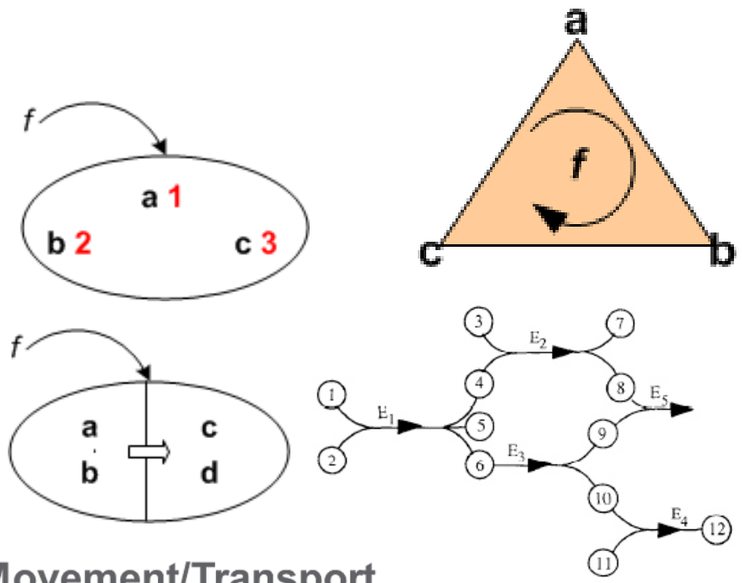
Simplicial Diagram

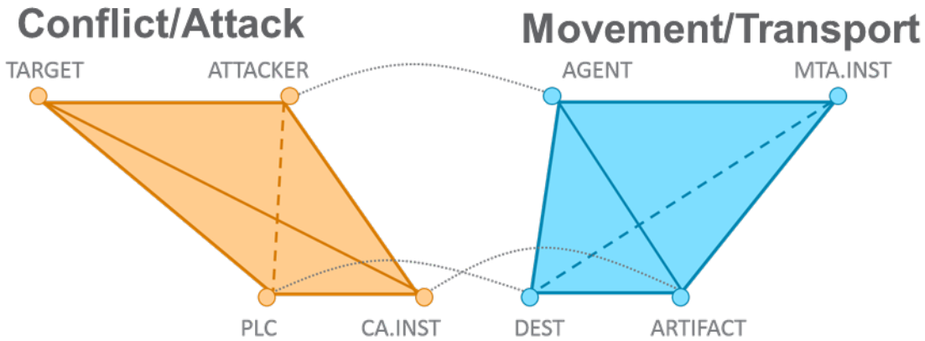


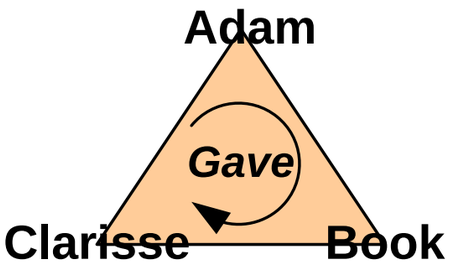
Other Extensions

- **Ordered HGs:** Permute vertices within edge
- **Directed HGs:** Bisect inputs and outputs
- **Property HGs:** Qualitative attributes
  - Semantic, categorical, ontologically typed
- **Weighted HGs:** Quantitative attributes
  - Numerical

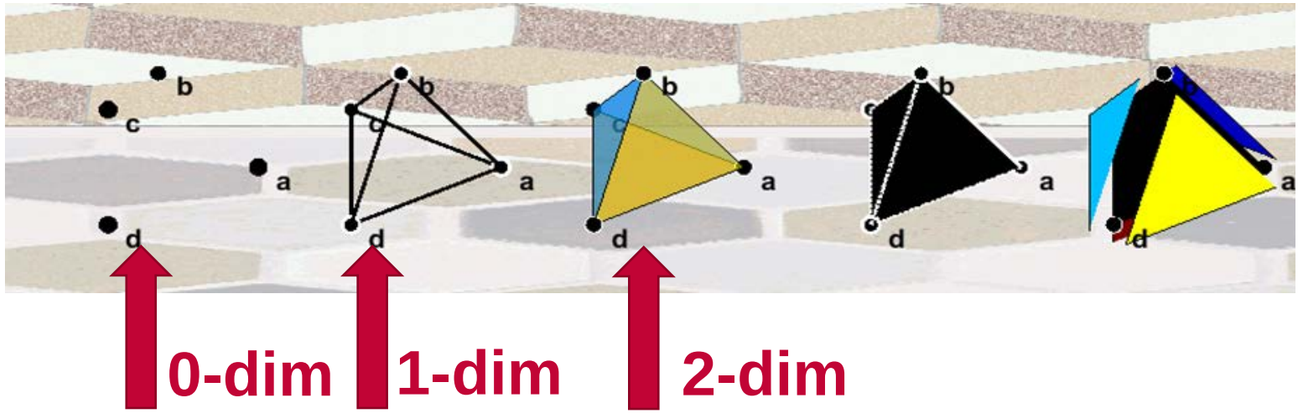




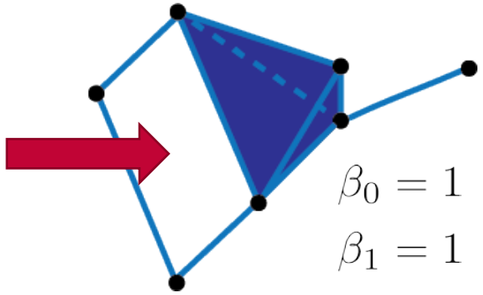


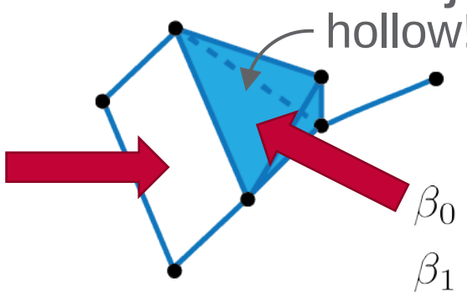


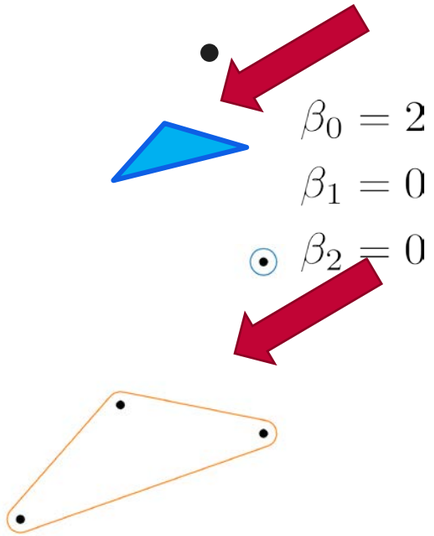
Hypergraphs as Topological Objects

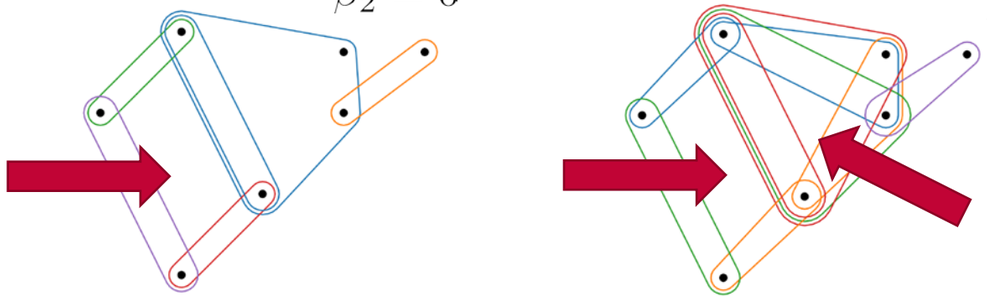


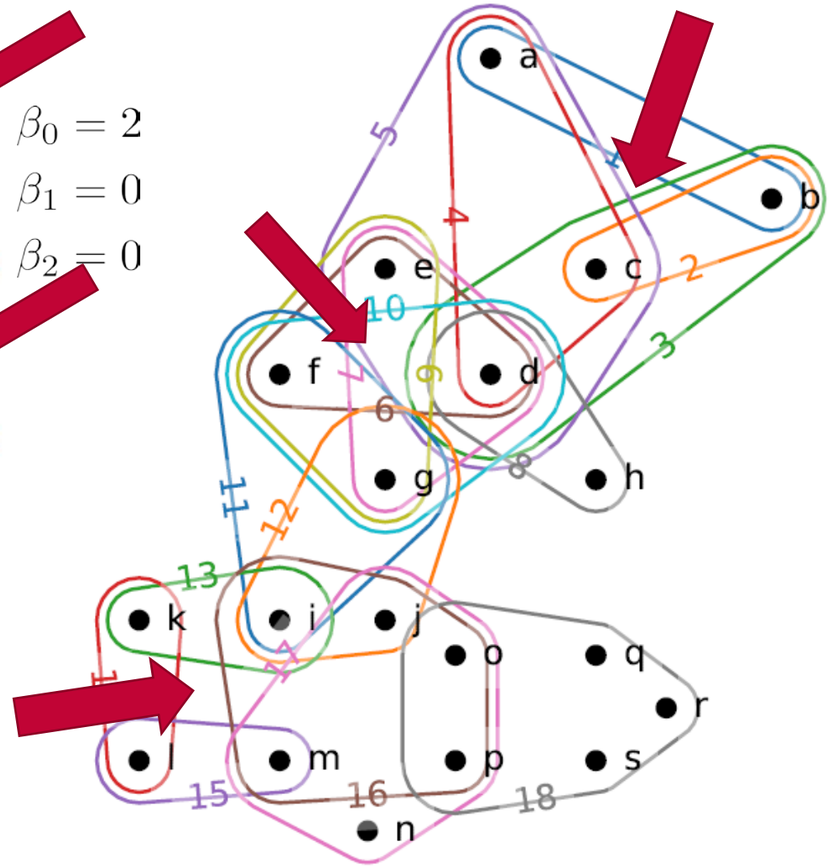
- Hypergraphs as multidimensional objects have topological properties







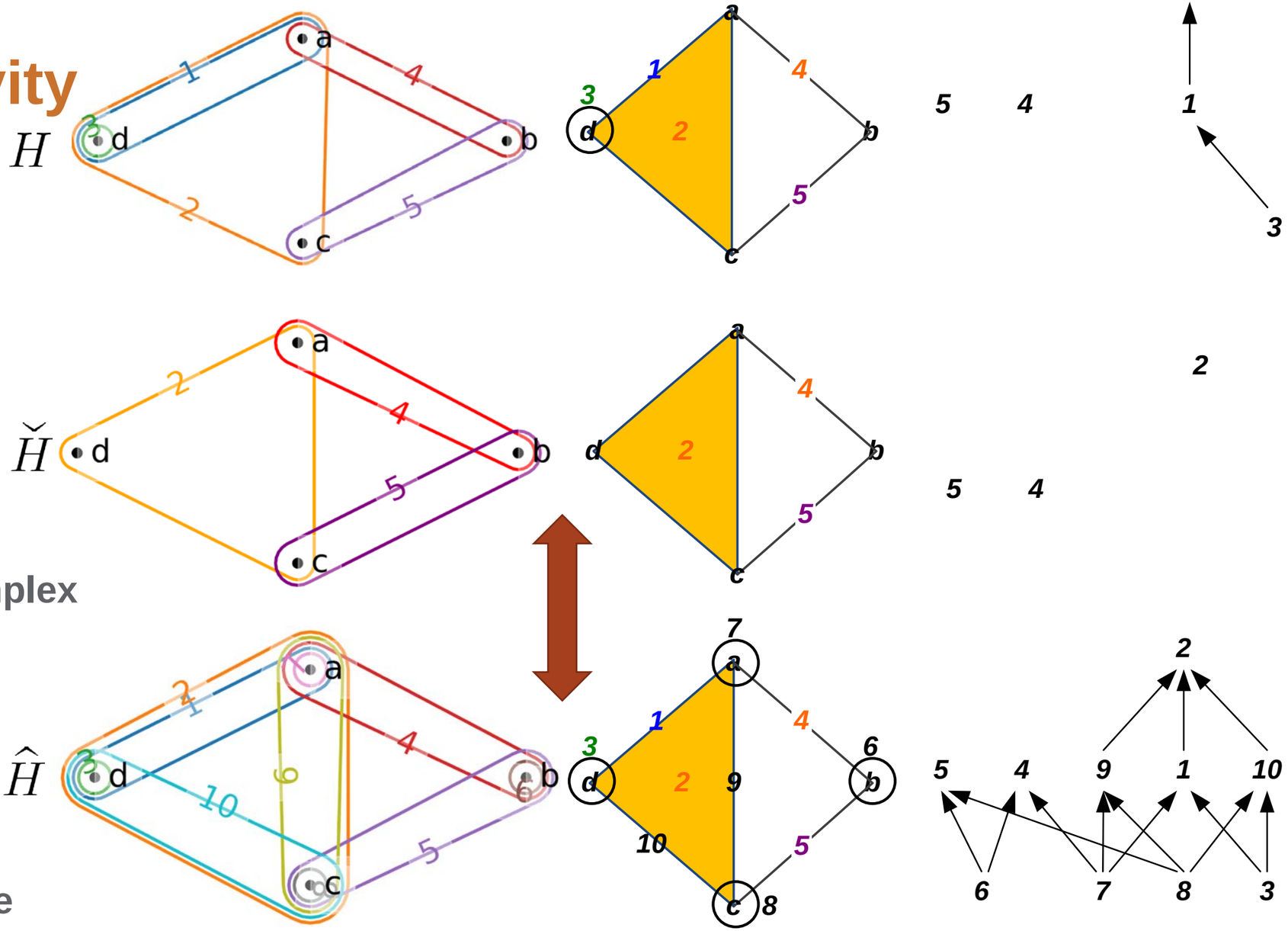




- **Homology to identify multidimensional holes**
  - As hypotheses for missing data
  - Need for bridging metadata

Inclusivity

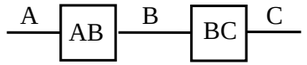
- **Hypergraph:**
- **Subedge order**
  - **2 *included* edges:** 1, 3
  - **3 “toplexes”:** Maximal hyperedges: 2, *4*, *5*
  - Inclusivity = 2/5
- **Simple hypergraph:** Remove all inclusions
  - All toplexes
  - “Reduction”
  - Inclusivity = 0
- **Abstract simplicial complex (ASC):** Add all inclusions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



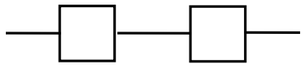
MODEL = STRUCTURE APPLIED TO DATA

A structure (graph or hypergraph) is a set of relationships (GT)

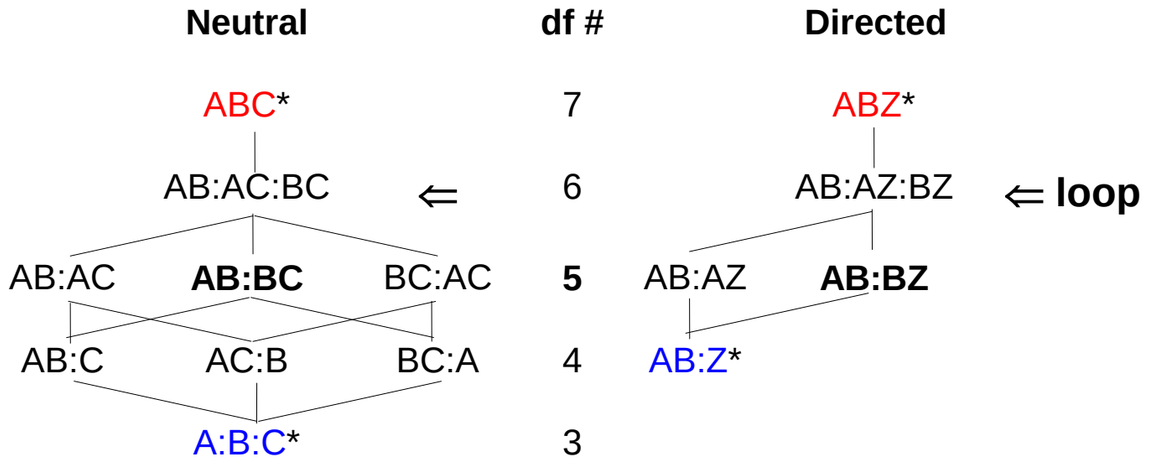
Specific structure **AB:BC**



General structure



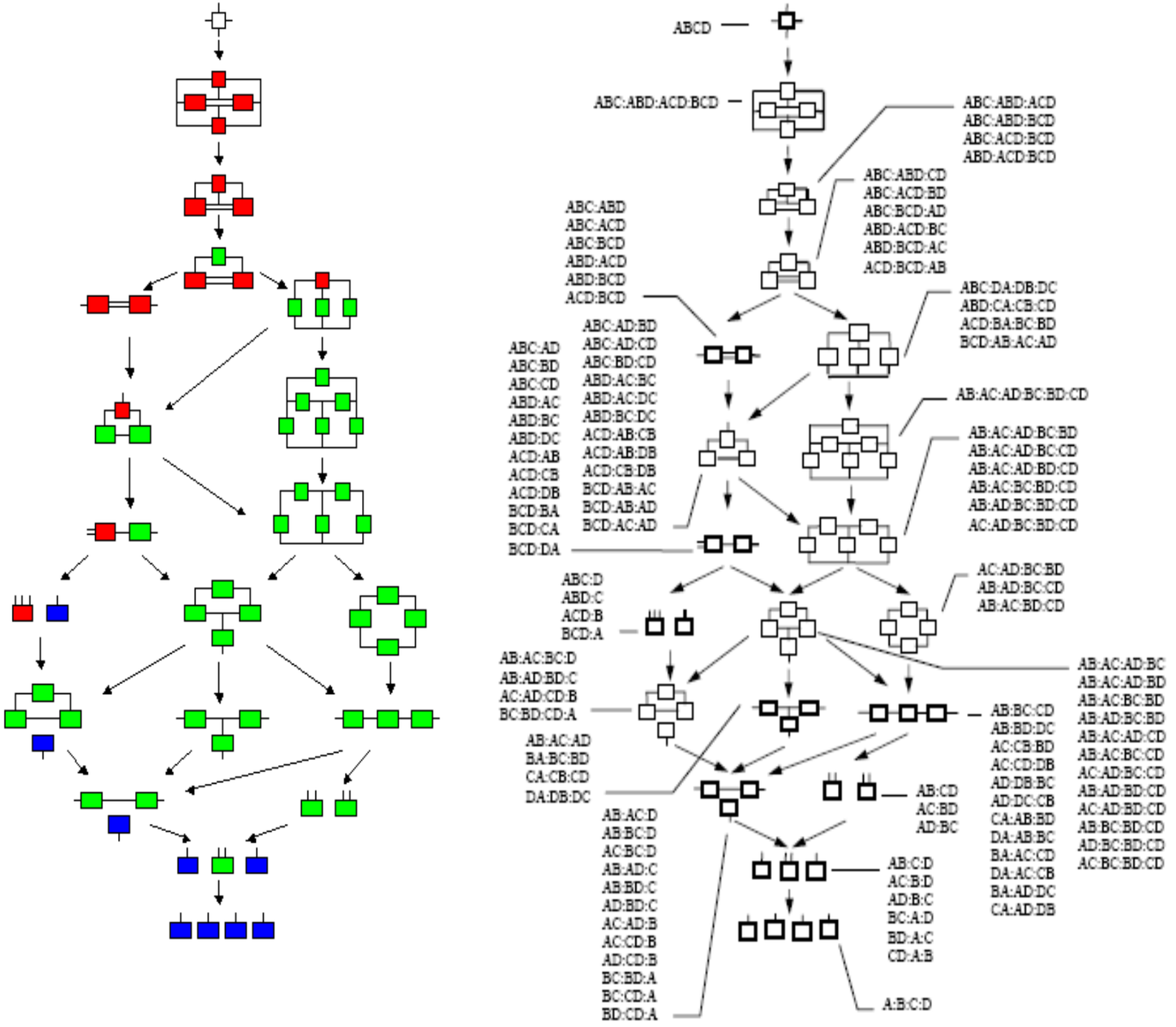
LATTICE OF *SPECIFIC* STRUCTURES (3 variables)



* **Reference model is** data or independence

\# df (degrees of freedom) values are for binary variables

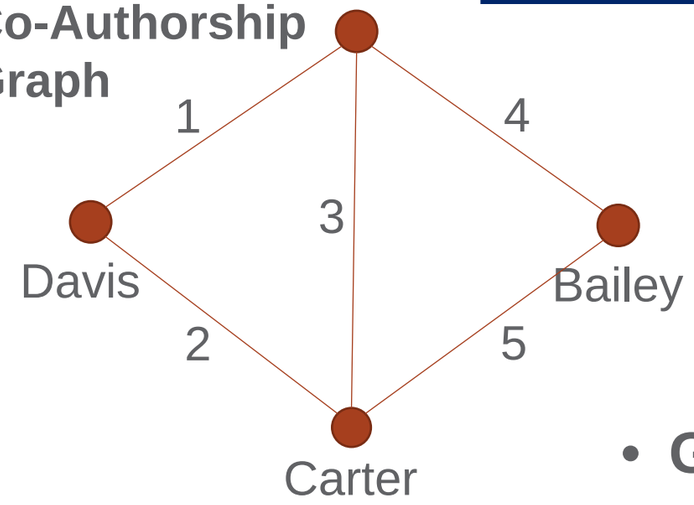
STRUCTURES 4 variables (GT)



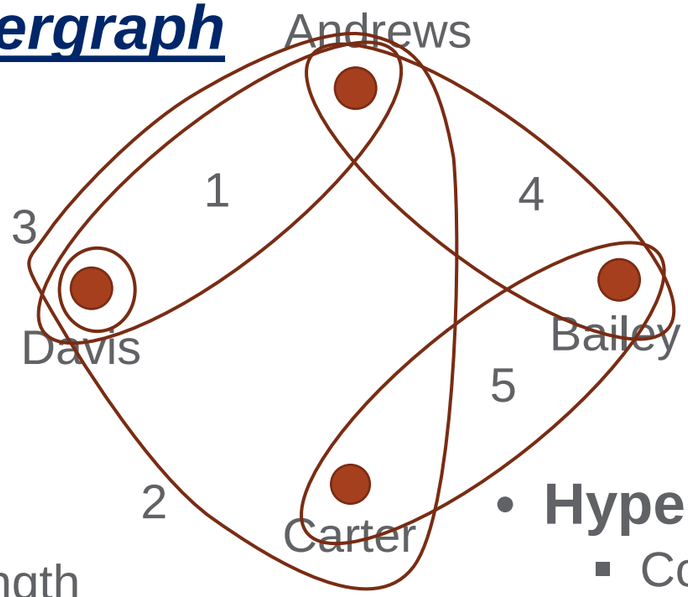
Graphs vs. Hypergraphs: Precis

A graph is 2-uniform hypergraph

Co-Authorship Graph



Collaboration Hypergraph



	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)

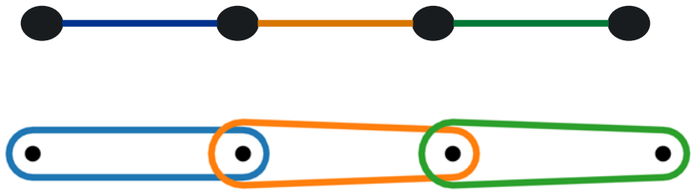
	1	2	3	4	5
a	X	X		X	
b				X	X
c		X			X
d	X	X	X		

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

Width in Hypergraph Structures

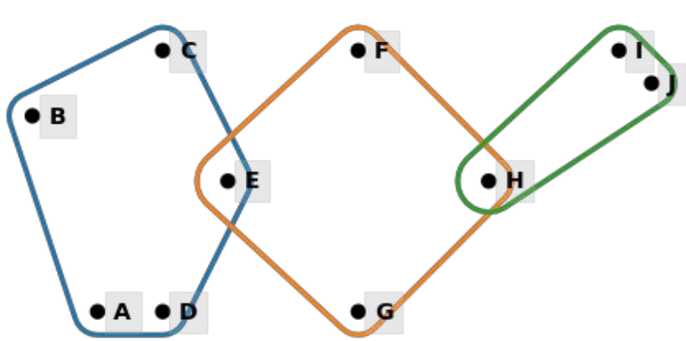
- **Hypergraph Paths Have *Width*:** Minimum edge intersection
- **s-walk:** Sequence $\langle e_i \rangle_{i=1}^{n}$ when $s \leq \min_{e_i, e_{i+1}} |e_i \cap e_{i+1}|, i = 1 \ldots n-1$

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

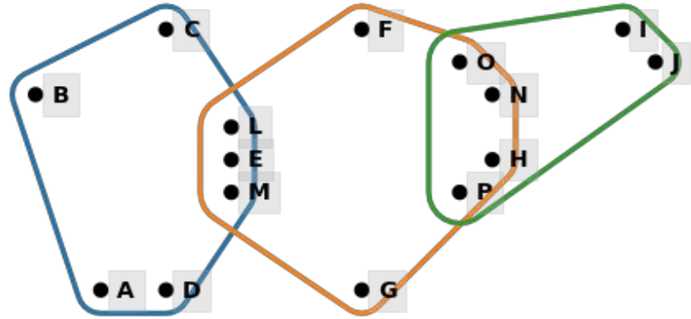


As a 2-uniform HG

Two Hypergraph Paths:
Same (edgewise) length = 2



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



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

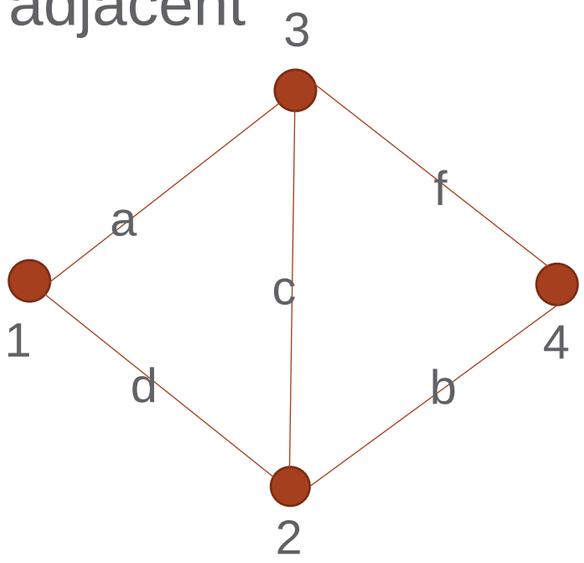
- **Extend generally:**
  - s-components, s-centrality, s-diameter s-motifs, s-clustering coefficient

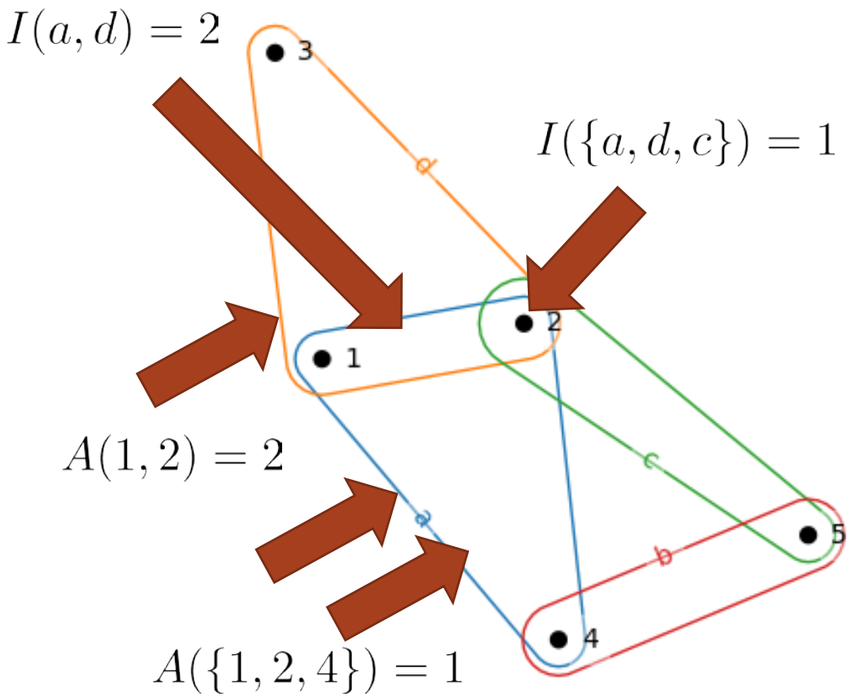
SG Aksoy, CA Joslyn, CO Marrero, B Praggastis, EAH Purvine: (2020) “Hypernetwork Science via High-Order Hypergraph Walks”, *EPJ Data Science*, v. 9:16, doi.org/10.1140/epjds/s13688-020-00231-0

Vertex Adjacency and Edge Incidence Are Generalized in Hypergraphs

- **Boolean in graphs:** *a* and *d* are incident; *1* and *2* are adjacent
- **Quantitative in hypergraphs:**

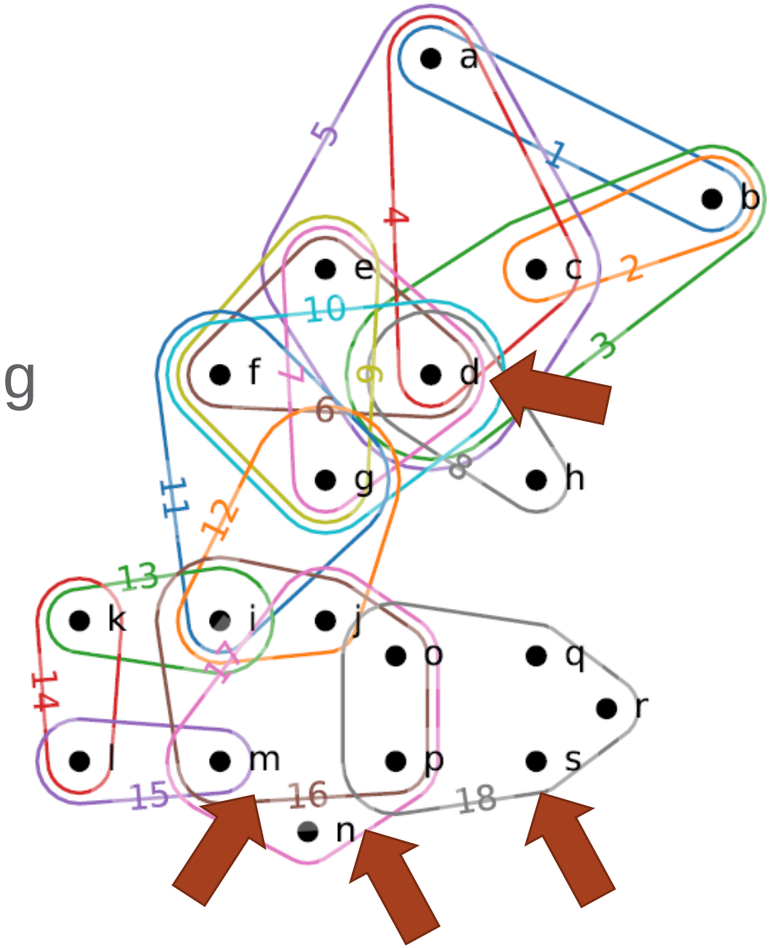
$$A(v, v') = |e \supseteq \{v, v'\}|, \quad A(U \subseteq V) = |e \supseteq U|$$
$$I(e, e') = s = |e \cap e'|, \quad I(F \subseteq \mathcal{E}) = |\bigcap_{e \in F} e|$$





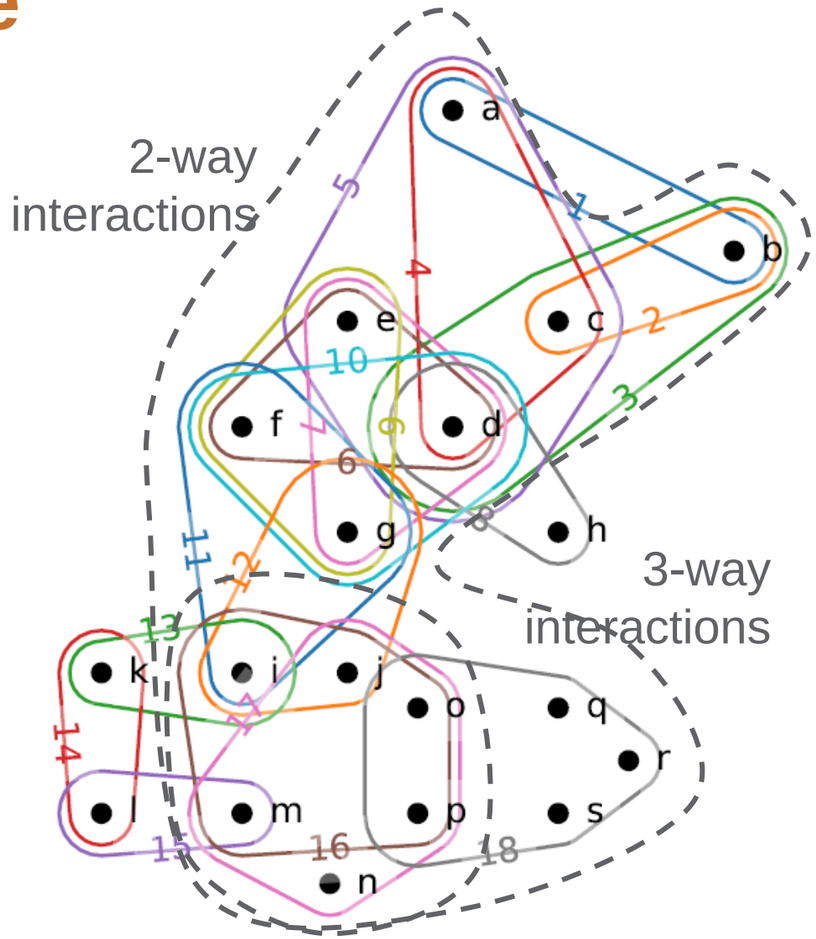
Towards Hypernetwork Science

- **Hypergraphs generalize network science to multi-way relationships**
  - For question of *community interaction*
  - *Multidimensional* connectivity, centrality, etc. among *groups* of entities
- **Who are most active authors?** *Max node degrees*
- **Which papers have most authors?** *Max edge sizes*



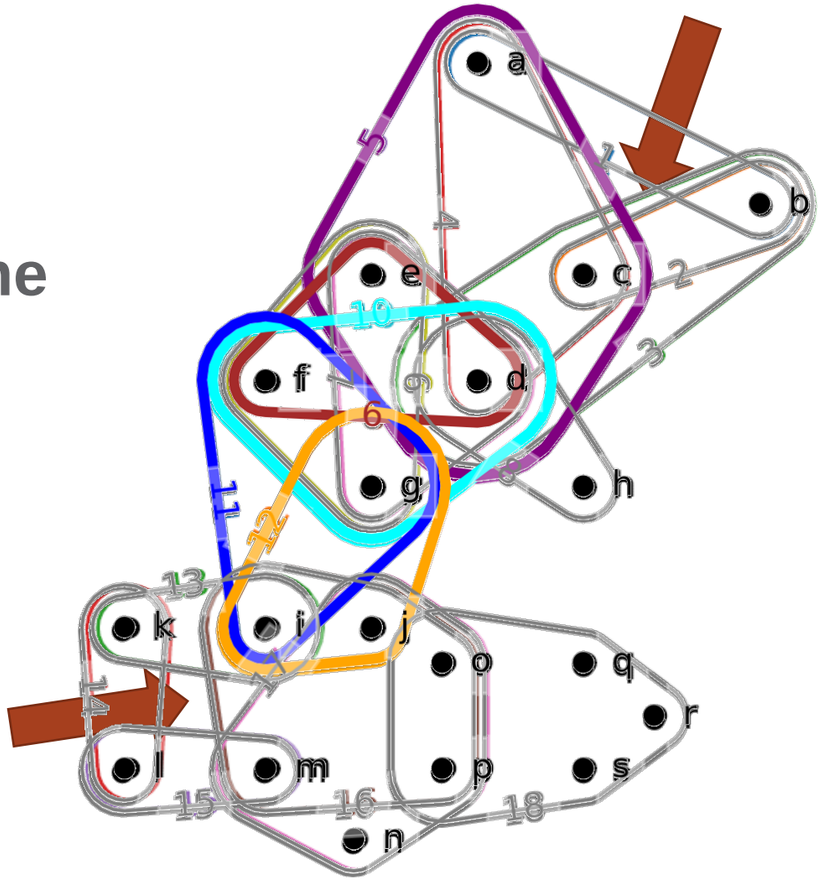
Towards Hypernetwork Science

- **What research communities are formed?**
  Connected components of different strengths



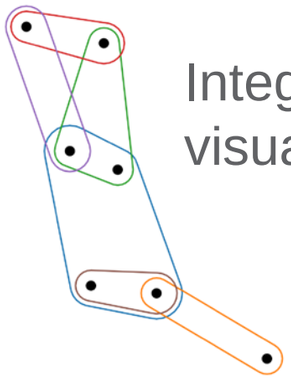
Towards Hypernetwork Science

- **What research communities are formed?**
  Connected components of different strengths
- **How many collaborations are there between some pair of papers?** “*s-Distance*”
  - ***5*** →***10*** → ***12*** *(intersections=1)*
  - ***5*** → ***6*** → ***10*** → ***11*** → ***12*** *(intersections=2)*
- **What is the most distant pair of papers?**
  “*s-Diameter*”
  - **2-diameter = 6:** 4 -> 5 -> 9 -> 11 -> 12 -> 16 -> 18
- **Are there groups of authors who aren’t working together but should?**
  Homology, Betti numbers
  - “Holes as hypotheses”



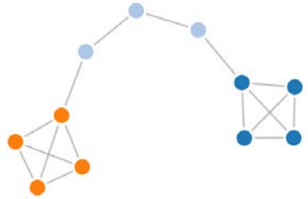
The HyperNetX (HNX) Library

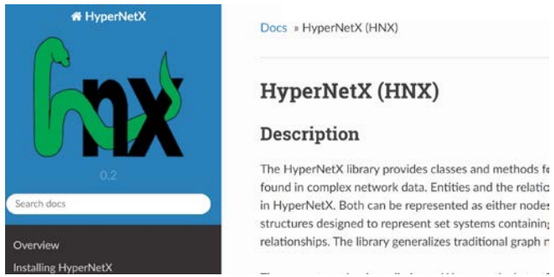
- HNX has various hypergraph constructors for: dictionaries, lists of lists, bipartite graphs…



Integrated visualization

Builds on NetworkX





Online documentation

- Current scale:
  - Hypergraph exploration for O(10K) vertices and hyperedges
  - Experimenting with CuPy for scaling up

Install with PyPI

Interactive tutorials

jupyter

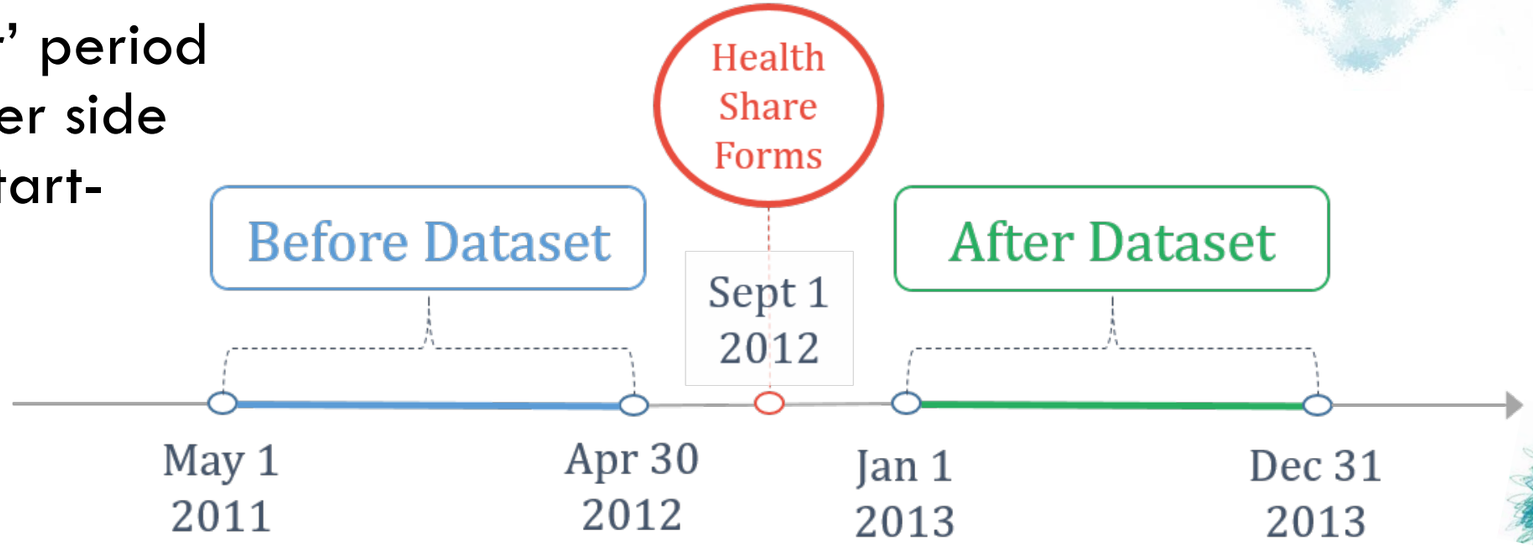
- Core Requirements:

Python >=3.6	SciPy
NetworkX	MatPlotLib
Numpy	Jupyter (for tutorials)

- Open Sourced on Github:
  https://github.com/pnnl/HyperNetX
- Developers: Brenda Praggastis (lead), Dustin Arendt (visualizations)

INFERRING A HEALTHCARE DELIVERY NETWORK

- Health Share of Oregon formed on September 1st, 2012, for the greater Portland area (Multnomah, Washington, and Clackamas counties)
- RA was applied to Medicaid insurance claims from 2011-2013, inferring the healthcare network before and after Health Share's Formation
- A four-month 'buffer' period was allowed on either side of the intervention start-date



USING RA FOR NETWORK INFERENCE

- RA finds a 'best' model of associations between members (e.g., billing providers)
- RA models contain calculated probabilities (q) for all variables at all states, e.g., $q(x_1^0 x_2^0 x_3^0 x_4^0 x_5^0), q(x_1^0 x_2^0 x_3^0 x_4^0 x_5^1), q(x_1^0 x_2^0 x_3^0 x_4^0 x_5^2)$
- These are compared to independence to identify the best RA model by BIC
- We define the distance between RA networks as the sum of absolute differences in the calculated probabilities (q) of the two RA models

	x_1	x_2	x_3	x_4	x_5
Patient 1	0	0	1	2	0
Patient 2	1	2	0	0	0
Patient 3	2	1	0	1	0
Patient 4	0	0	0	1	2
Patient 5	2	2	0	0	0
Patient 6	0	0	2	2	0
Patient 7	0	2	0	0	0
Patient 8	0	1	2	0	0
Patient 9	1	2	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮

- Network Distance:

$$\hat{\theta}_{RA} = \sum |q^2 - q^1|$$

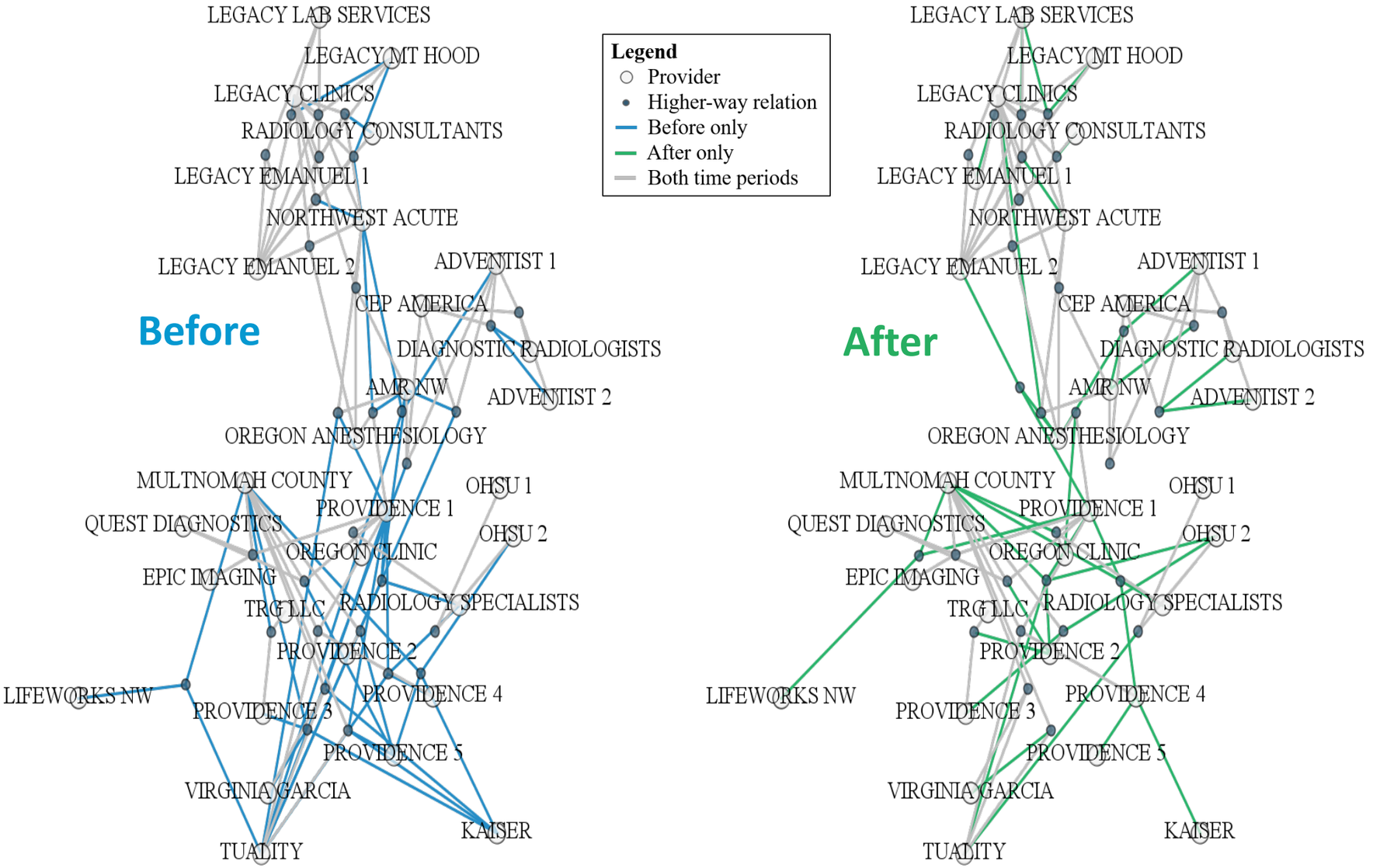
DATA PREPARATION

- Data was cross-tabulated, from claim-level records to a frequency count by patient and billing provider (BP).
- We had a total of 5,602,376 claims for 183,958 patients

Claim#	Patient	Provider	Date
⋮	⋮	⋮	⋮
1234	Patient A	BP3	2012
1235	Patient B	BP2	2012
1236	Patient C	BP1	2012
1237	Patient C	BP2	2013
1238	Patient D	BP1	2013
1239	Patient E	BP3	2013
⋮	⋮	⋮	⋮

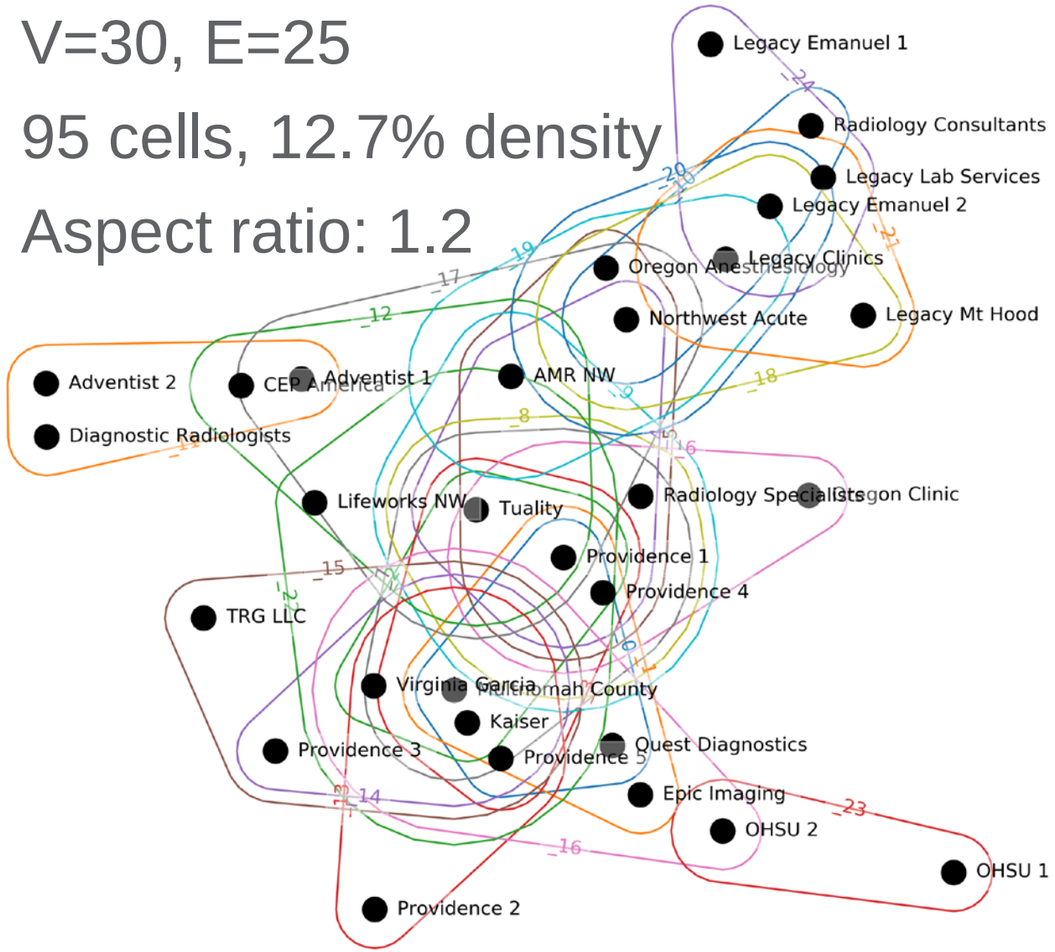
		BP1	BP2	BP3	BP4
Before	Patient A	0	0	1	0
	Patient B	1	**7**	0	2
	Patient C	2	1	0	1
	⋮	⋮	⋮	⋮	⋮
After	Patient C	0	1	0	0
	Patient D	**4**	0	0	0
	Patient E	0	0	1	1
	⋮	⋮	⋮	⋮	⋮

Values in **red** were recoded as 2 for I-DNA.

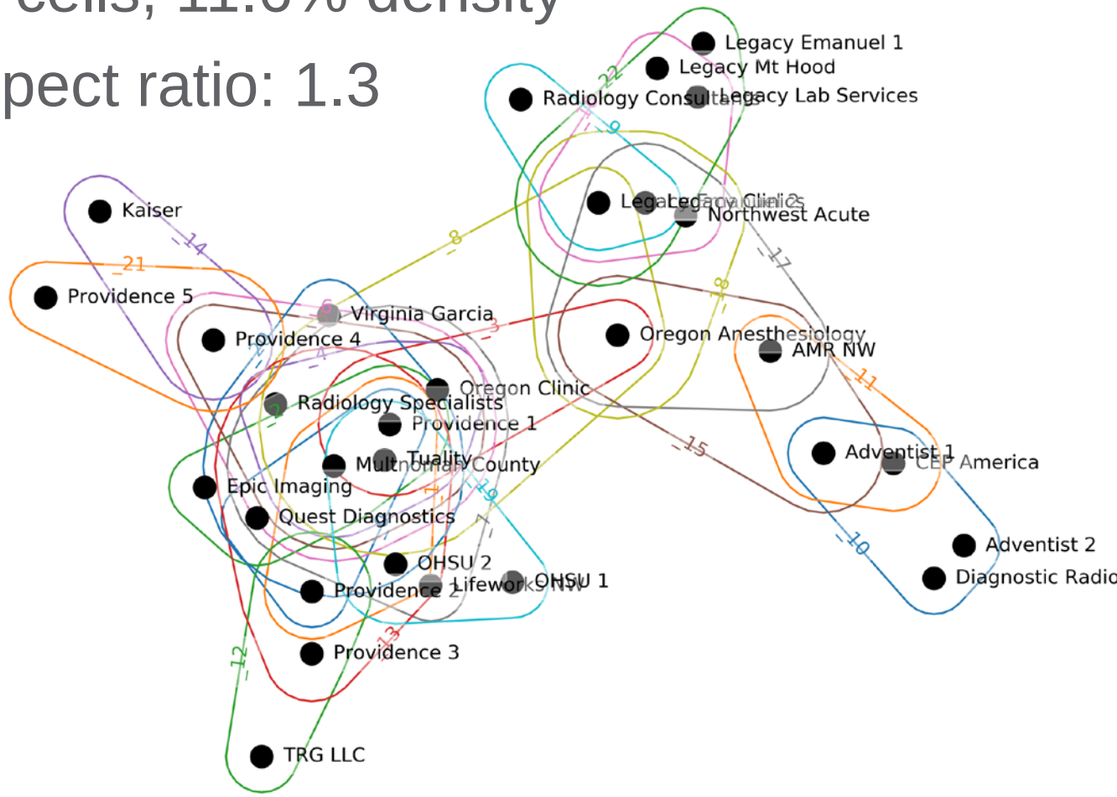

Legend
Provider
Higher-way relation
Before only
After only
Both time periods
Before
After
LEGACY LAB SERVICES
LEGACY MT HOOD
LEGACY CLINICS
RADIOLOGY CONSULTANTS
LEGACY EMANUEL 1
NORTHWEST ACUTE
LEGACY EMANUEL 2
ADVENTIST 1
CEP AMERICA
DIAGNOSTIC RADIOLOGISTS
AMR NW
ADVENTIST 2
OREGON ANESTHESIOLOGY
MULTNOMAH COUNTY
OHSU 1
PROVIDENCE 1
QUEST DIAGNOSTICS
OHSU 2
OREGON CLINIC
EPIC IMAGING
TRG LLC
RADIOLOGY SPECIALISTS
PROVIDENCE 2
LIFEWORKS NW
PROVIDENCE 4
PROVIDENCE 3
PROVIDENCE 5
VIRGINIA GARCIA
KAISER
TUALITY

Before and After

- V=30, E=25
- 95 cells, 12.7% density
- Aspect ratio: 1.2

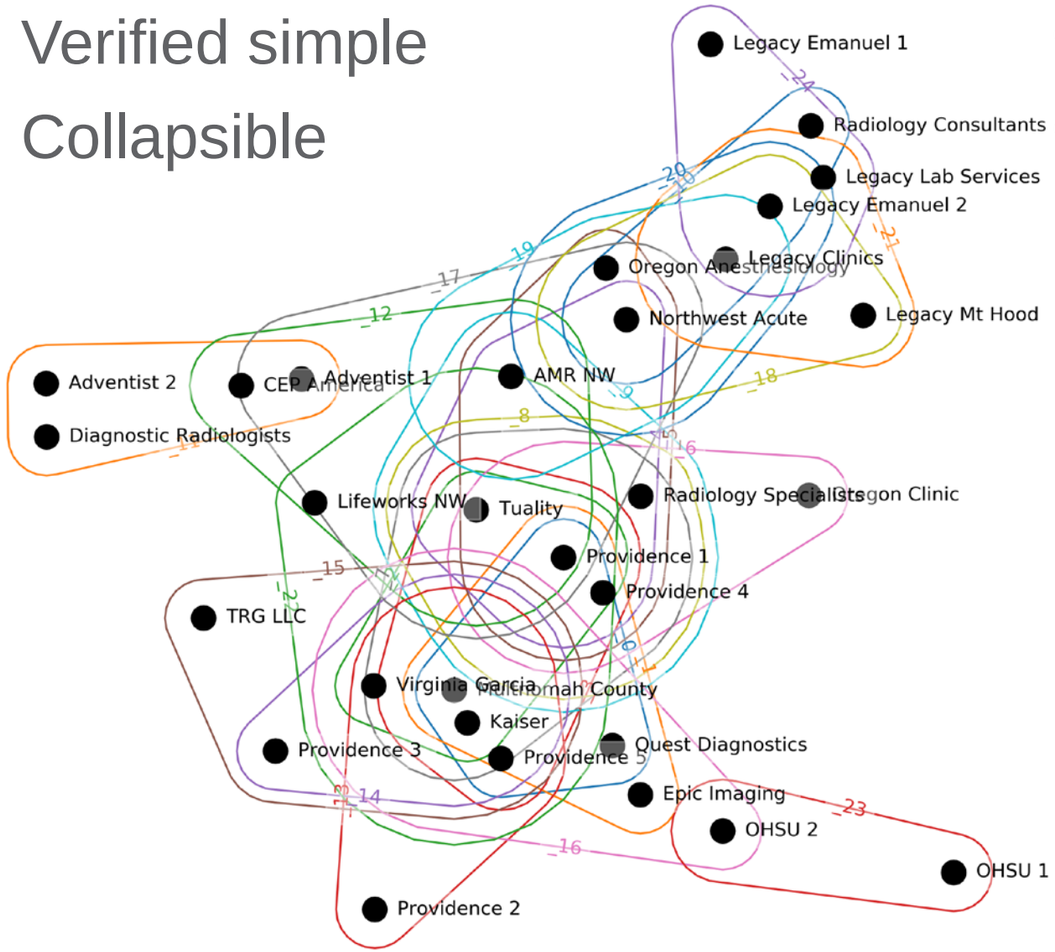


- V=30, E=23
- 80 cells, 11.6% density
- Aspect ratio: 1.3

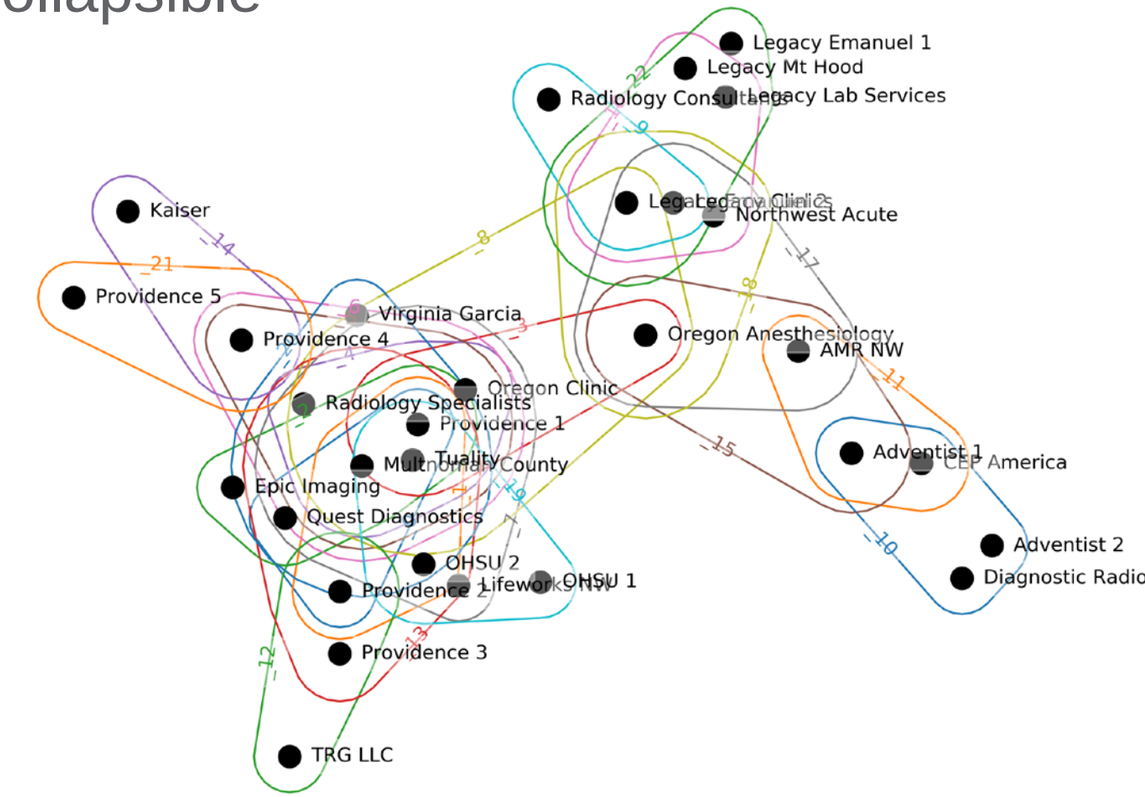


Topology

- Verified simple
- Collapsible

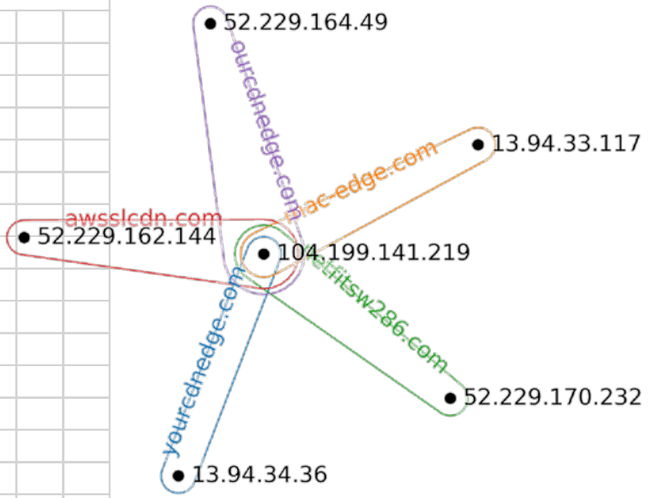


- Verified simple
- Collapsible



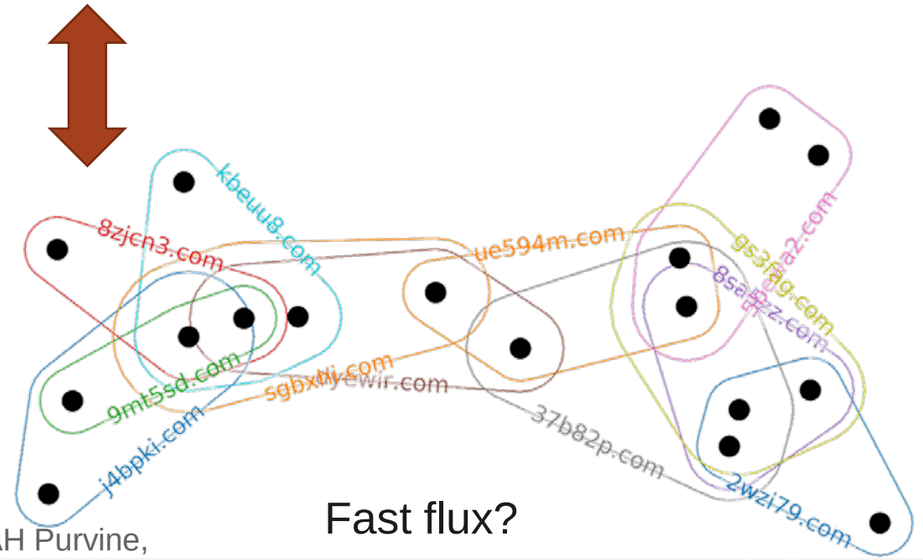
DNS Use Case

	2wzi79.com	37b82p.com	5bewa2.com	8sa5zz.com	8zjcn3.com	9mt5sd.com	dyewir.com	gs3fag.com	j4bpki.com	kbeuu8.com	sgbxhi.com	ue594m.com		
103.86.122.130							X			X	X			
103.86.122.148		X	X	X				X				X		
103.86.122.149			X					X				X		
103.86.122.152	X	X		X										
103.86.122.154					X	X	X			X	X			
103.86.122.160							X				X	X		
103.86.122.169			X											
103.86.122.173	X													
103.86.122.181	X			X				X						
103.86.122.192					X									
103.86.122.195						X			X					
103.86.122.220			X											
103.86.122.222		X					X					X		
103.86.122.223					X	X			X	X	X			
103.86.122.225	X	X		X				X						
103.86.122.238									X					
103.86.122.242										X				



Content delivery network?

- **Hypergraph:** IP X Domain
  - Nodes = IP addresses
  - Hyperedges = domains
- **When DNS is not one-to-one**:
  - Domain aliases
  - Hosting services to multiple web sites
  - Site management across IPs
  - Random IP assignment
- **ActiveDNS:** GA Tech https://activednsproject.org/
- **Analytical Questions:**
  - *General Exploration:* Abnormal IPs and domains
  - *Targeted Exploration:* Neighborhoods of known bad IPs or domains

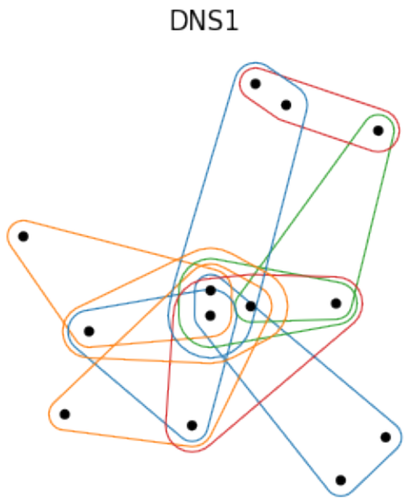


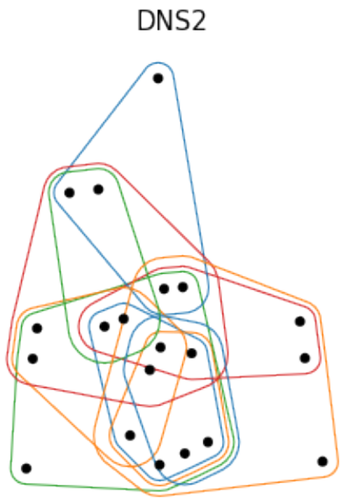
Fast flux?

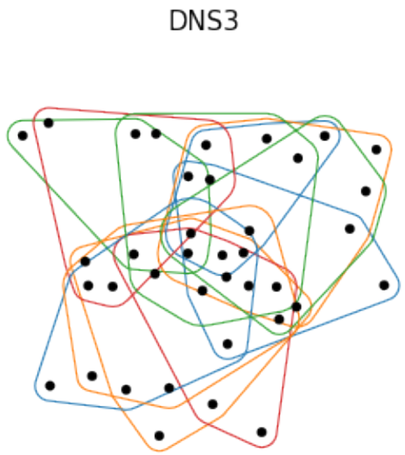
Joslyn, CA, S Aksoy, D Arendt, J Firoz, L Jenkins, B Praggastis, EAH Purvine, M Zalewski: (2020) “Hypergraph Analytics of Domain Name System Relationships”, in: 17th Wshop. on Algorithms and Models for the Web Graph (WAW 2020), LNCS 12901, pp. 1-15

Homologies Show Multidimensional Open Structures

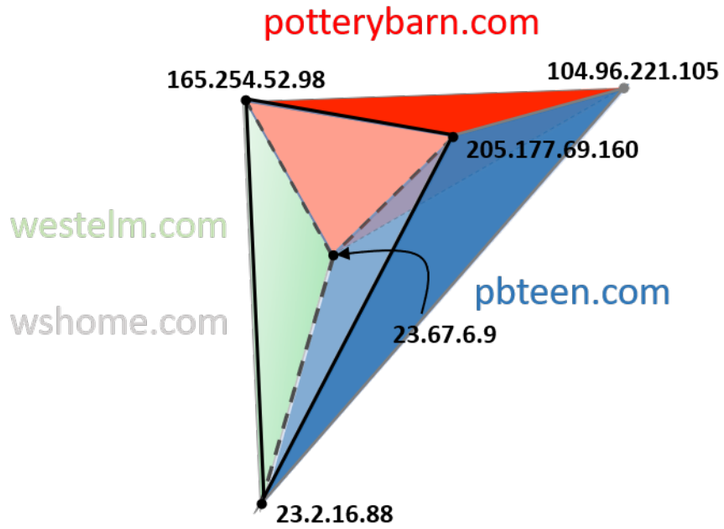
- **DNS1:** $\beta = \langle 1, 1, 0, 0, \ldots \rangle$
- **DNS2:** $\beta = \langle 1, 1, 2, 0, \ldots \rangle$
- **DNS3:** $\beta = \langle 1, 3, 1, 0, \ldots \rangle$







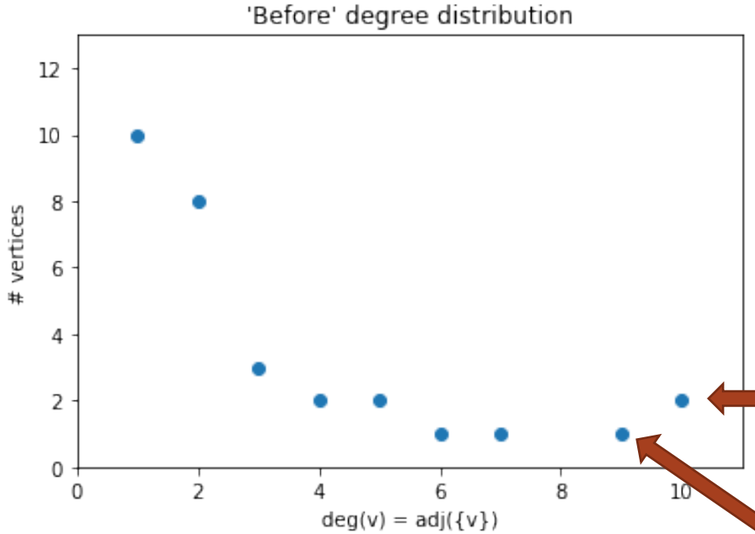
- **DNS2:** One generator of a 2-hole, tetrahedral void

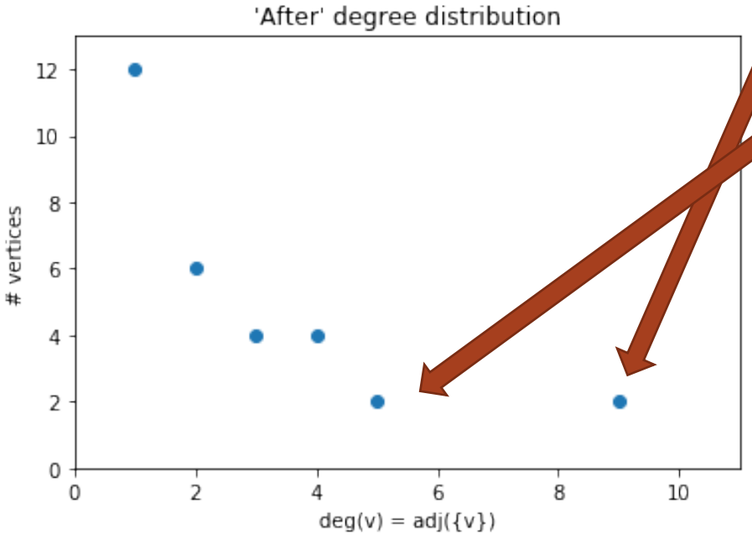


https://activednsproject.org/

Joslyn, CA, S Aksoy, D Arendt, J Firoz, L Jenkins, B Praggastis, EAH Purvine, M Zalewski: (2020) “Hypergraph Analytics of Domain Name System Relationships”, in: 17th Wshop. on Algorithms and Models for the Web Graph (WAW 2020), LNCS 12901, pp. 1-15

Distributions



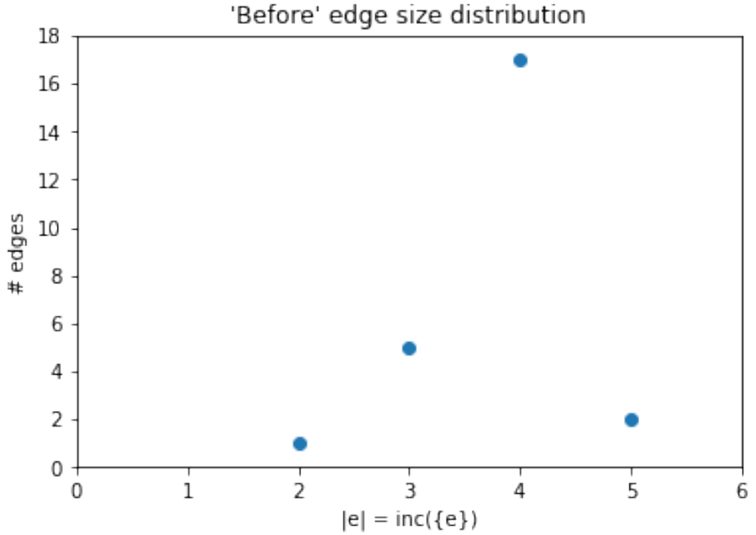


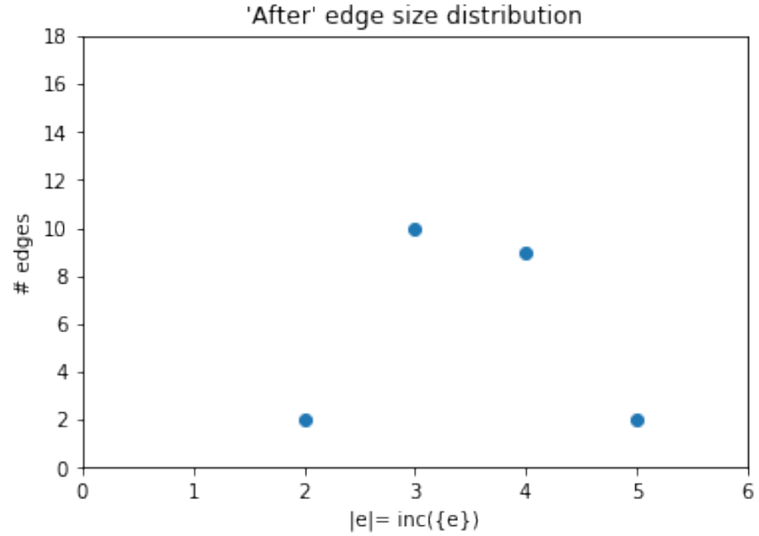
- **Before:**
  - Avg(deg(v)) = 3.17
  - Avg(|e|) = 3.91

Providence, Multnomah County

Tuality

- **After:**
  - Avg(deg(v)) = 2.67
  - Avg(|e|) = 3.48





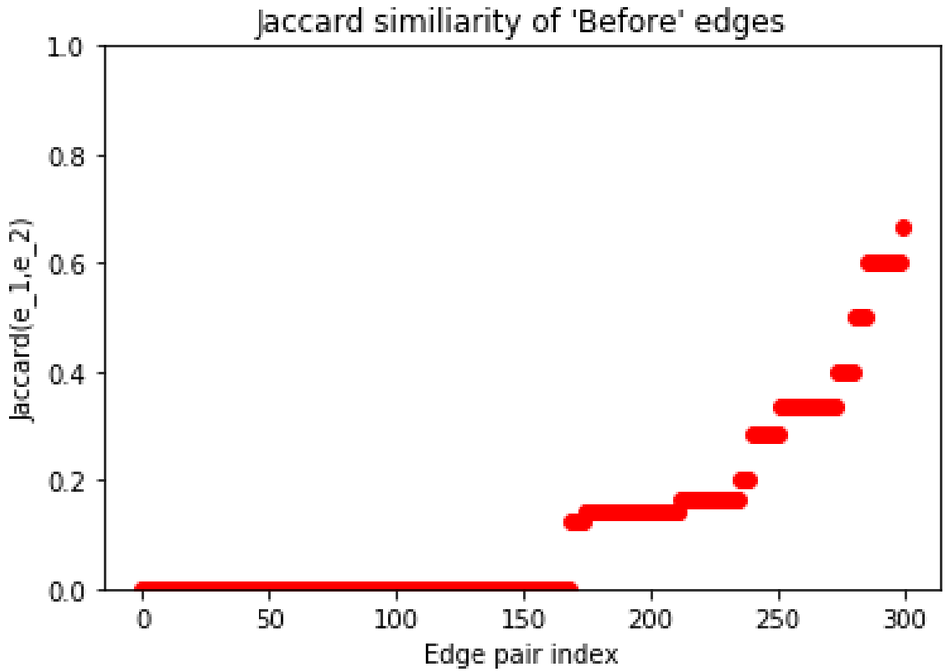
Edge Overlaps

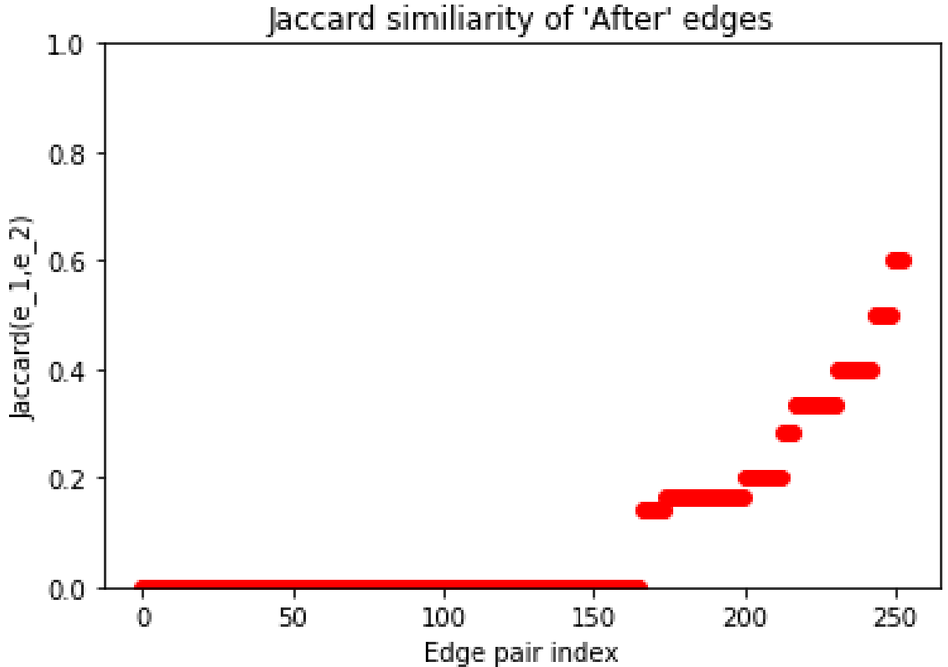
- Pairwise Jaccard edge similarity

$$J(e,f) = \frac{|e \cap f|}{|e \cup f|}$$

$$J(e,f) = 0 \leftrightarrow e \cap f = \emptyset$$
$$J(e,f) = 1 \leftrightarrow e = f$$

- **Before:** avg(J) = .118
- **After:** avg(J) = .095



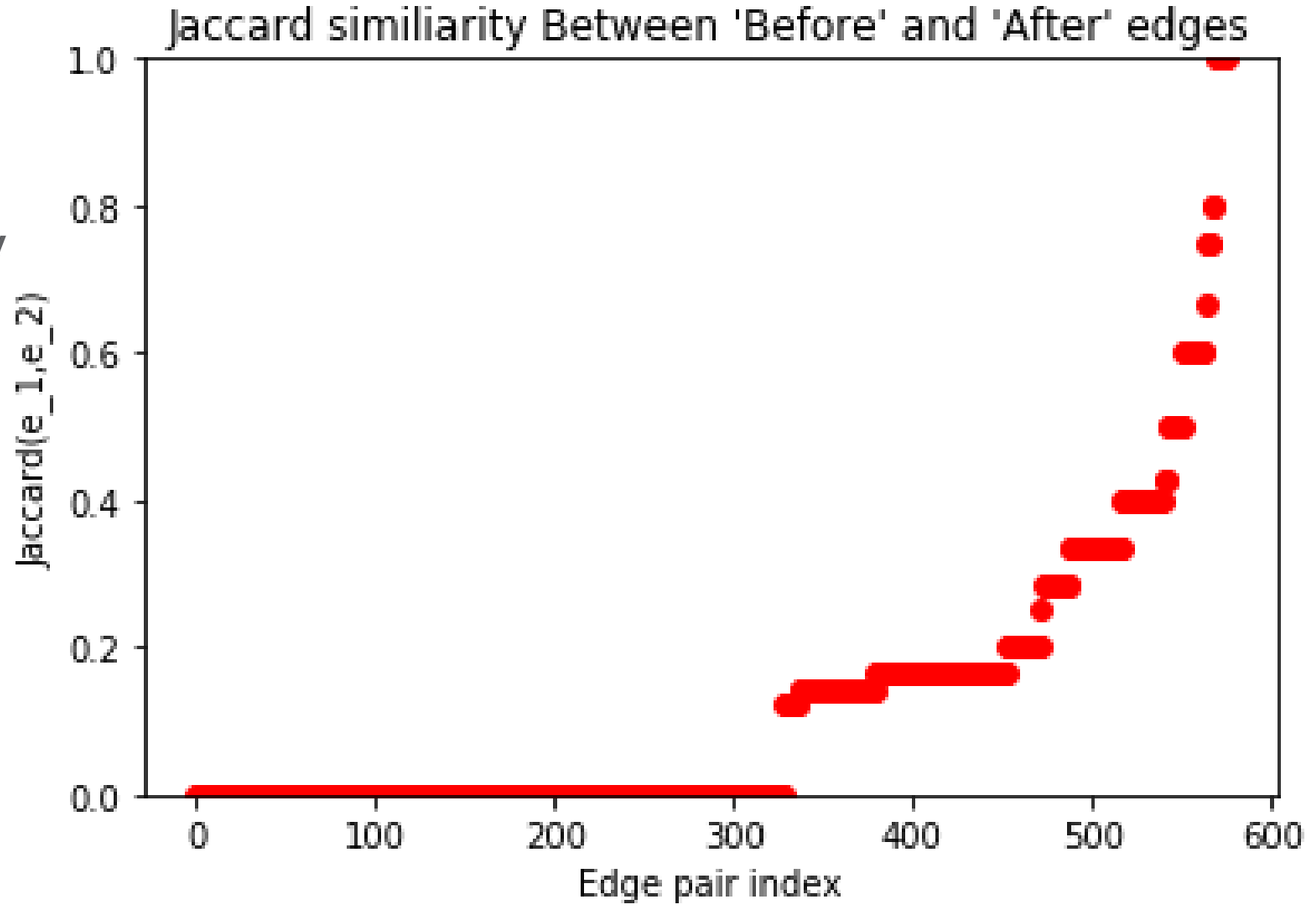


Edge Overlaps

- Pairwise Jaccard edge similarity

$$J(e,f) = \frac{|e \cap f|}{|e \cup f|}$$

$$J(e,f) = 0 \leftrightarrow e \cap f = \emptyset$$
$$J(e,f) = 1 \leftrightarrow e = f$$



- **Before:** avg(J) = .118
- **After:** avg(J) = .095
- **Before to After:** avg(J) = .121
- ***J(e,f)* = 1:** Epic Imaging, Multnomah County, Providence 1, Quest Diagnostics
  Multnomah County, Providence 1, Providence 4, Tuality
  Adventist 1, Adventist 2, CEP America, Diagnostic Radiologists
  AMR NW, Legacy Clinics, Northwest Acute, Oregon Anesthesiology
  Legacy Clinics, Legacy Emanuel 2, Northwest Acute, Oregon Anesthesiology,
  Legacy Emanuel 1, Legacy Emanuel 2, Legacy Lab Services

s-Component Structure

- **Before:**

s=1, 1: [25]
s=2, 2: [24, 1]
s=3, 8: [18, 1, 1, 1, 1, 1, 1, 1]
s=4, 24: [2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=5, 25: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

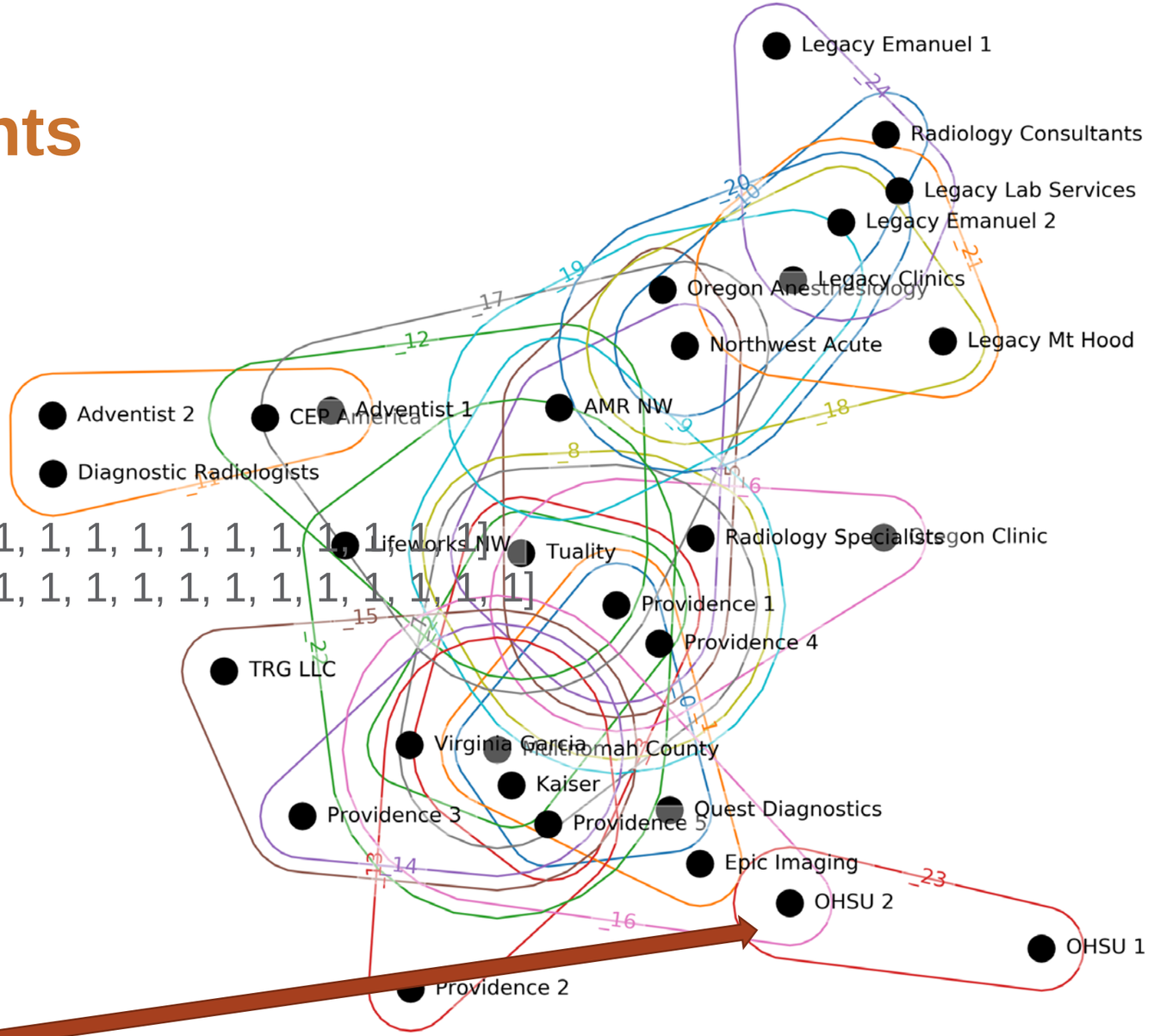
- **After:**

s=1, 1: [23]
s=2, 3: [21, 1, 1]
s=3, 15: [7, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=4, 23: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=5, 23: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Before Components

- **Before:** 1-diameter=6

s=1, 1: [25]
s=2, 2: [24, 1]
s=3, 8: [18, 1, 1, 1, 1, 1, 1, 1]
s=4, 24: [2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=5, 25: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

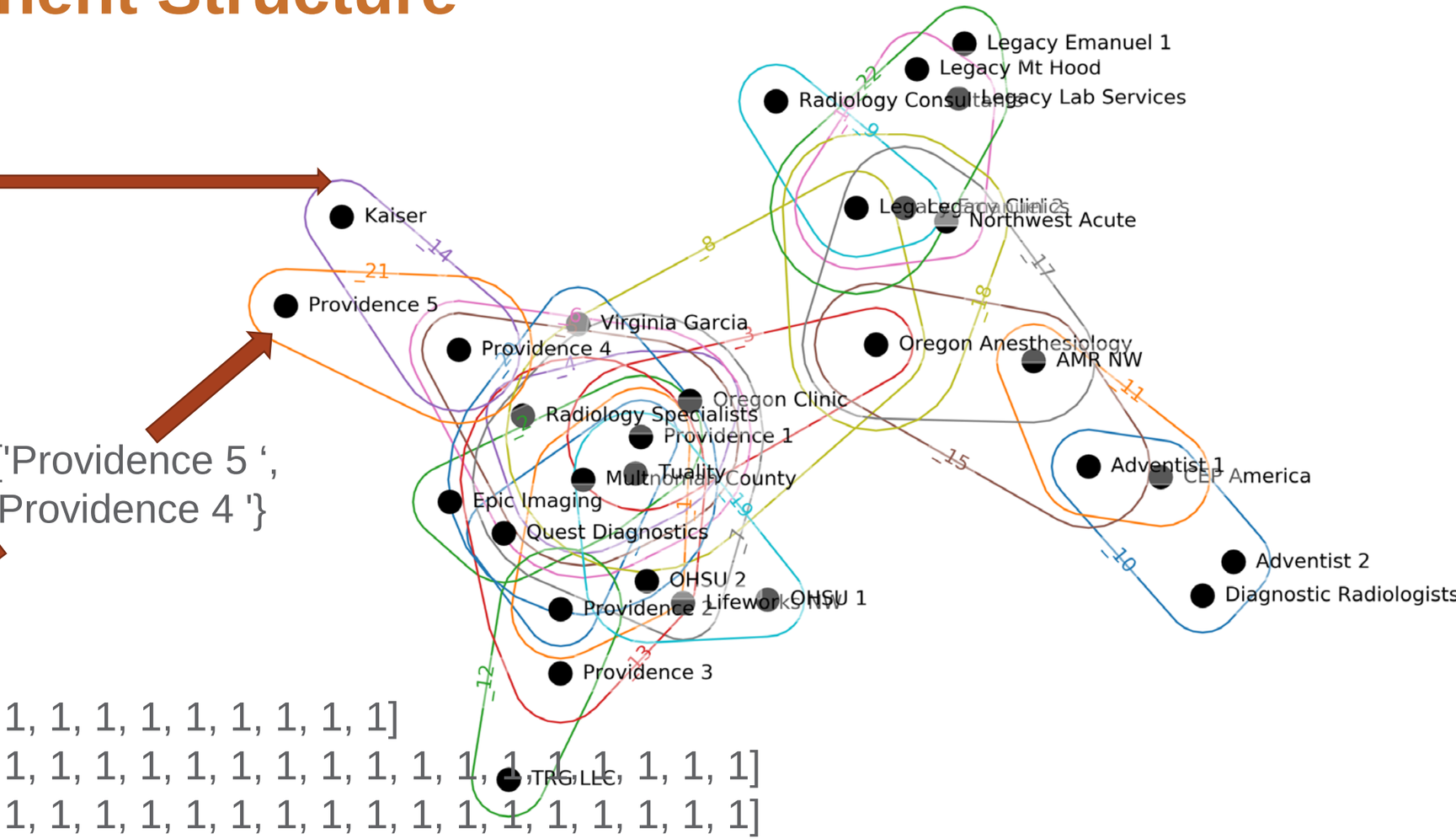


{'OHSU 1 ‘,
'OHSU 2 '}

Component Structure

{'Kaiser ‘,
'Providence 4 '}

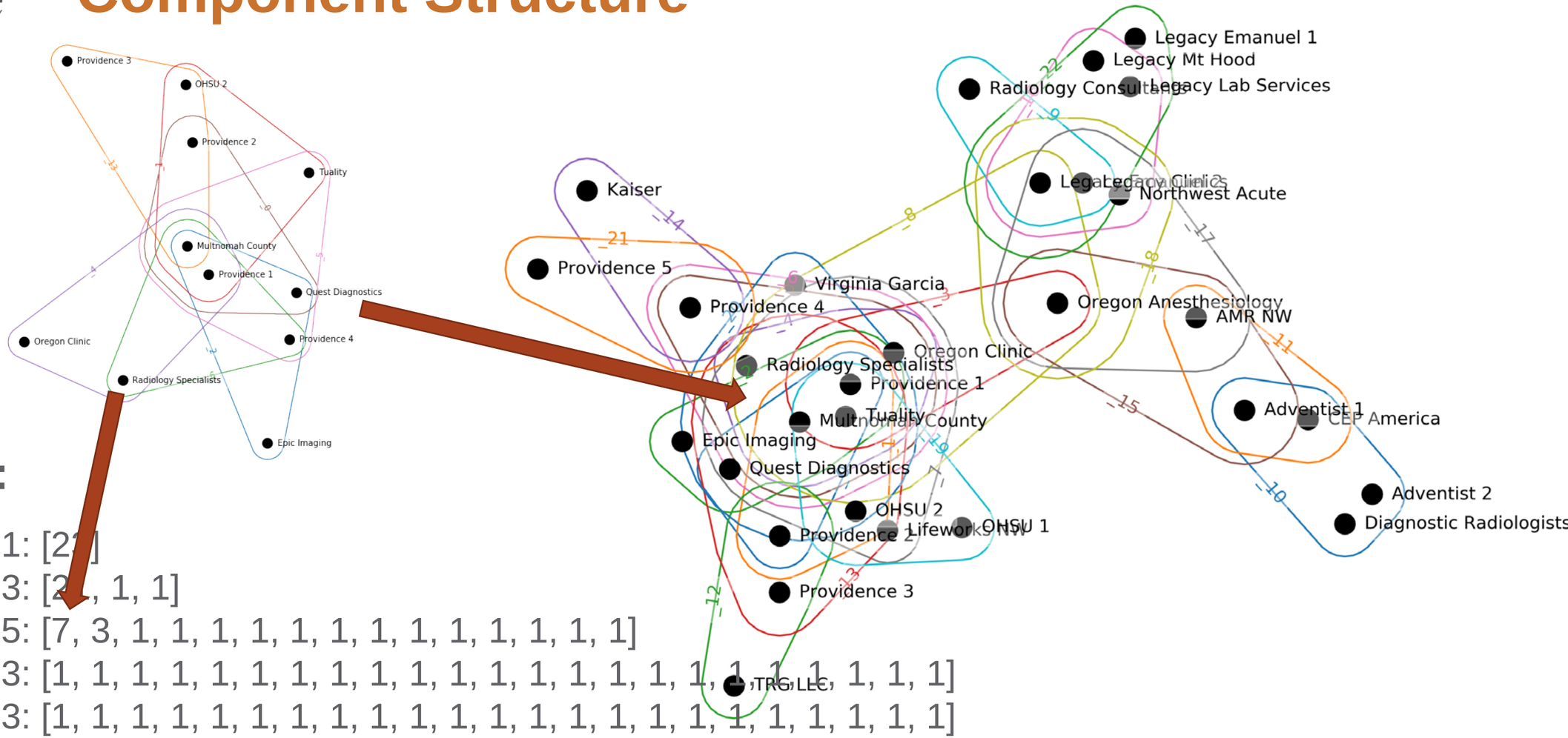
{'Providence 5 ‘,
'Providence 4 '}



- **After:**

s=1, 1: [23]
s=2, 3: [21, 1, 1]
s=3, 15: [7, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=4, 23: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=5, 23: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Component Structure



- **After:**

s=1, 1: [21]
s=2, 3: [2[illegible], 1, 1]
s=3, 15: [7, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=4, 23: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
s=5, 23: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Pacific Northwest
NATIONAL LABORATORY

Thank you