Pacific Northwest NATIONAL LABORATORY

Mathematics for Cyber Security

WSU CySER seminar
March 18, 2024

Emilie Purvine
Chief Data Scientist

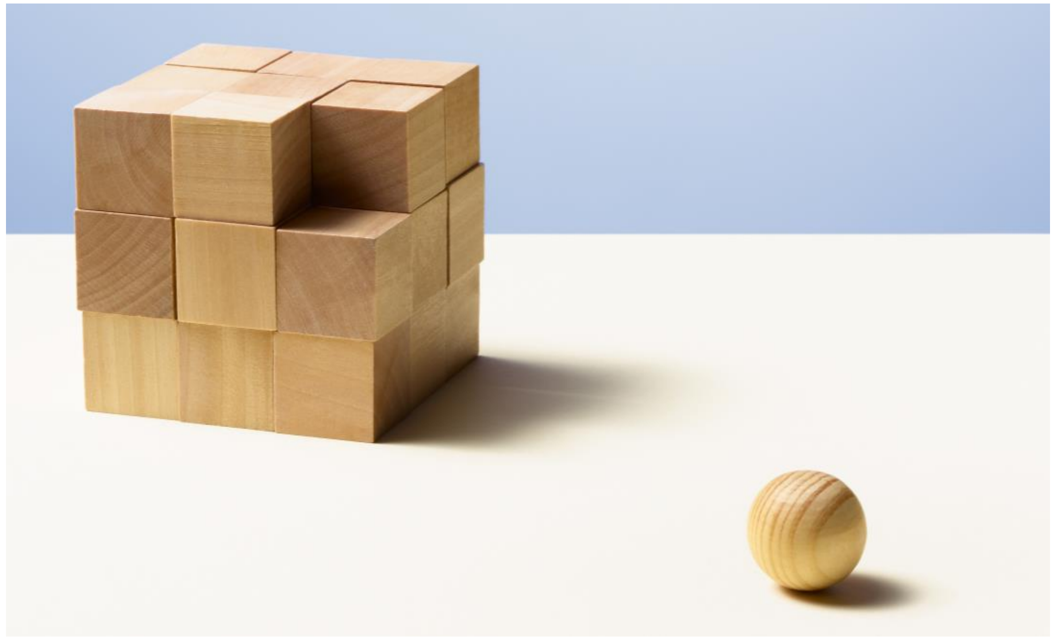
U.S. DEPARTMENT OF ENERGY BATTELLE

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

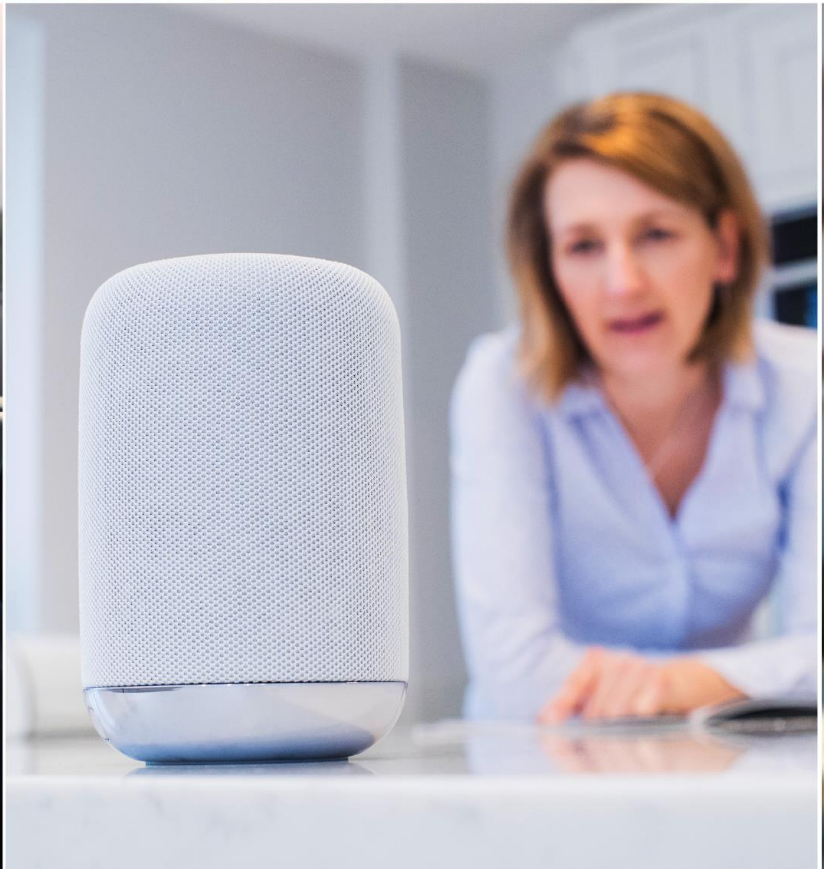
PNNL-SA-196388

Plan of the talk

- Computer network defense big picture
  - Cyber data and alignment to kill chain
  - Big challenges in cyber
- Opportunities for mathematicians!
  - Mathematical models of cyber data
  - Anomaly detection
  - Machine learning



Internet of things – how many do you have?

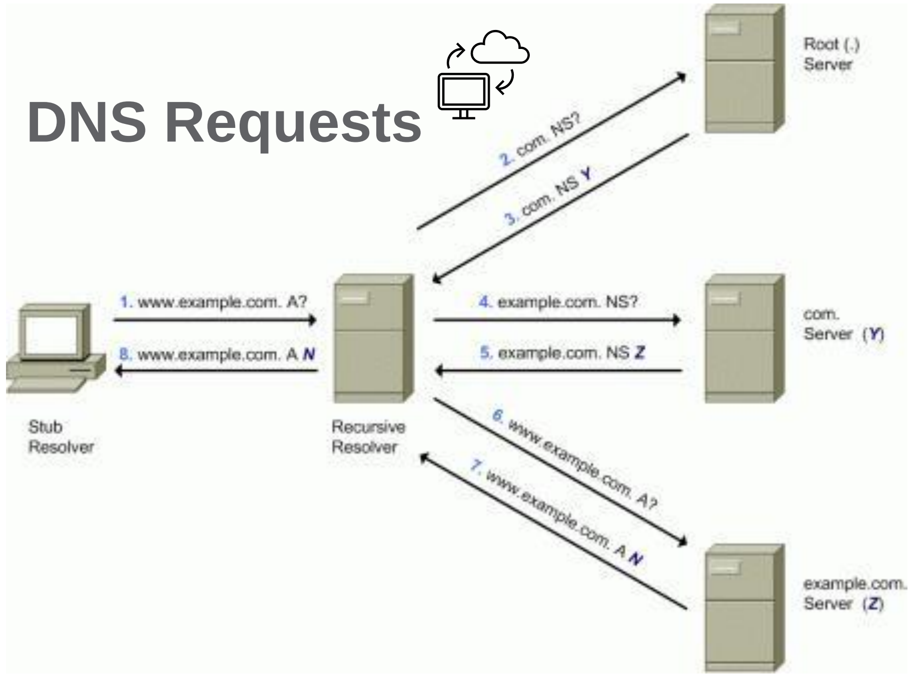


Pacific Northwest NATIONAL LABORATORY

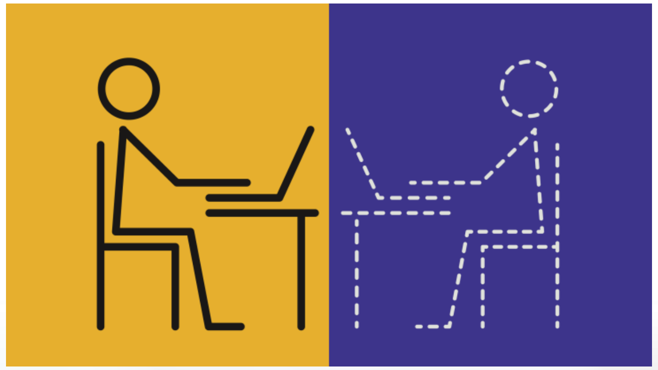
A Sea of Data

Flow record
- Source IP
- Source Port
- Destination IP
- Destination Port
- Packet count
- Byte count
- Start time
- End time

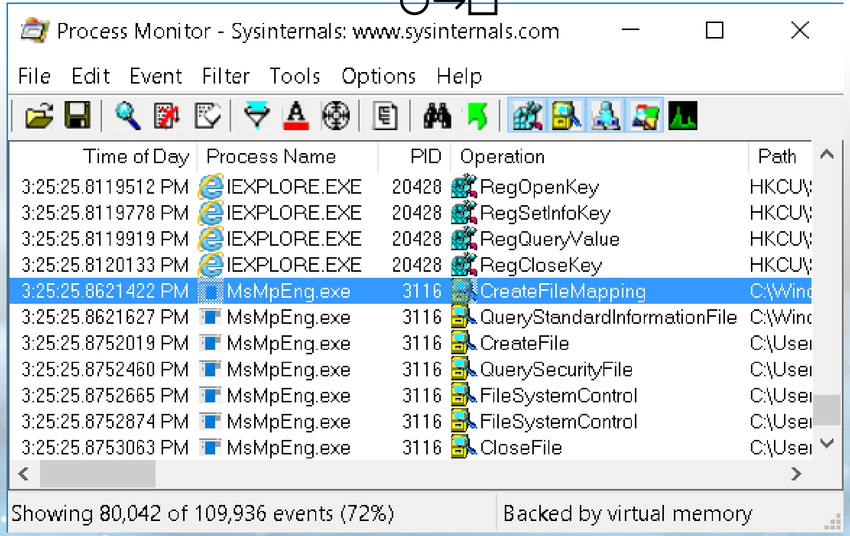
DNS Requests



Authentication: SSH, Kerberos, …



Process logs


Process Monitor - Sysinternals: www.sysinternals.com

File Edit Event Filter Tools Options Help

Time of Day	Process Name	PID	Operation	Path
3:25:25.8119512 PM	IEXPLORE.EXE	20428	RegOpenKey	HKCU\
3:25:25.8119778 PM	IEXPLORE.EXE	20428	RegSetInfoKey	HKCU\
3:25:25.8119919 PM	IEXPLORE.EXE	20428	RegQueryValue	HKCU\
3:25:25.8120133 PM	IEXPLORE.EXE	20428	RegCloseKey	HKCU\
3:25:25.8621422 PM	MsMpEng.exe	3116	CreateFileMapping	C:\Wind
3:25:25.8621627 PM	MsMpEng.exe	3116	QueryStandardInformationFile	C:\Wind
3:25:25.8752019 PM	MsMpEng.exe	3116	CreateFile	C:\User
3:25:25.8752460 PM	MsMpEng.exe	3116	QuerySecurityFile	C:\User
3:25:25.8752665 PM	MsMpEng.exe	3116	FileSystemControl	C:\User
3:25:25.8752874 PM	MsMpEng.exe	3116	FileSystemControl	C:\User
3:25:25.8753063 PM	MsMpEng.exe	3116	CloseFile	C:\User

Showing 80,042 of 109,936 events (72%) | Backed by virtual memory

Firewall logs

SNORT®

Signature-based alerts

```
alert tcp $EXTERNAL_NET $HTTP_PORTS -> $HOME_NET any
```

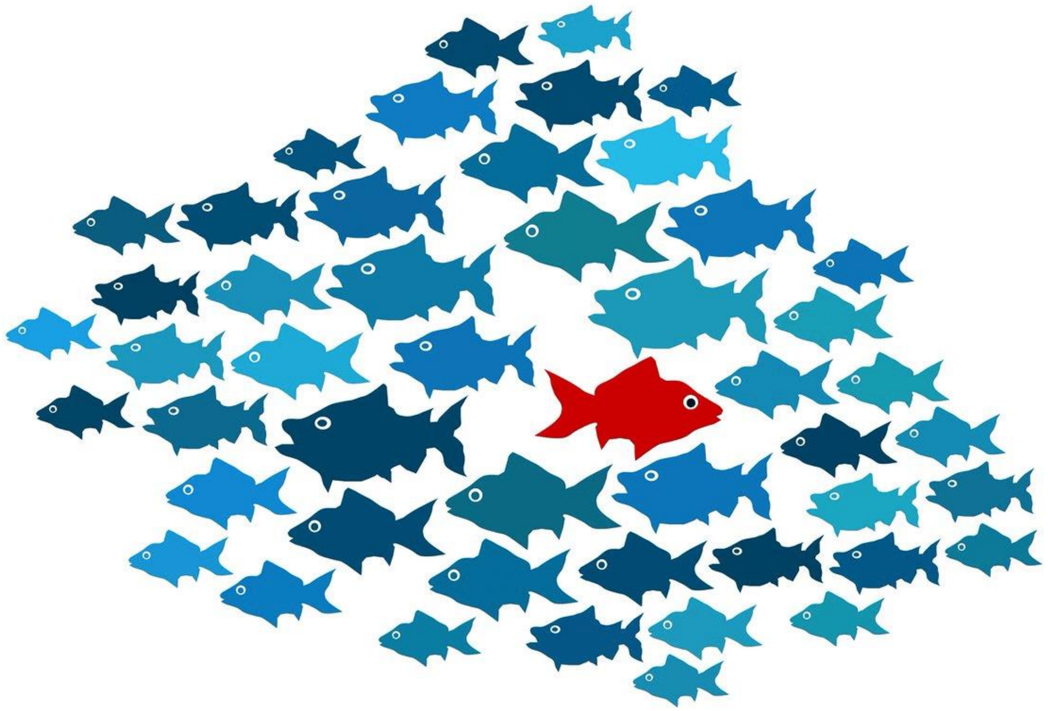
Aligning data with the cyber kill chain

- The *Cyber Kill Chain*® lays out the steps that an adversary goes through to compromise a system and get what they are looking for
  - This helps us organize how we think about detection – the earlier the better!
- How can we protect our networks?
  - Inspect the data we have in order to discover:
    - ✓ Known patterns of bad behavior
    - ✓ Unknown anomalies
  - Build in resilience

https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

Challenges in Cyber Defense

- Cyber systems do not have “laws of physics” type rules. Every rule or standard can be broken.
  - They can be broken by benign people that do not realize there is a rule, or by sophisticated adversaries.
- Adversaries are finding and exploiting vulnerabilities faster than defenders can identify them
- Signature-based alerts are still necessary, but threat hunting and anomaly detection are finding traction
  - Caution: An anomaly on one network is perfectly normal on another (e.g., off site backup vs. data exfiltration)



“OODA loop” – where do mathematicians fit?

Observe
- Collect data

Orient
- Develop situational awareness
- Run “analytics” on collected data

Decide
- What should we do based on observed data
- E.g., patch, segment, target additional data collection

Act
- Carry out the actions

Our Mathematical Toolbox

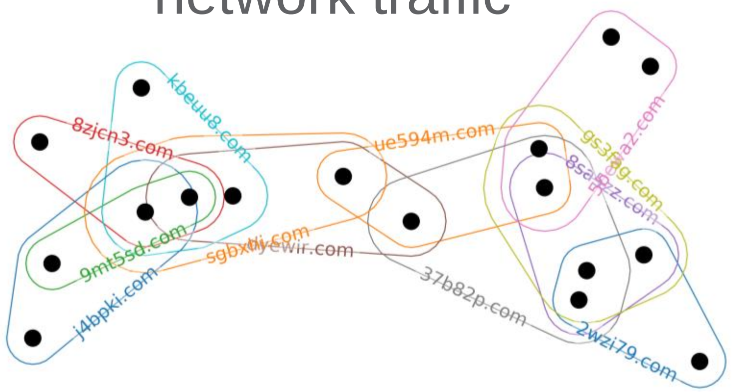
Models & Methods

Graphs

- Model *pairwise relationships* in data
    - **Edges** connect pairs of **vertices**
- Cyber examples:
    - Source IP to Destination IP in network traffic
    - Malware similarity

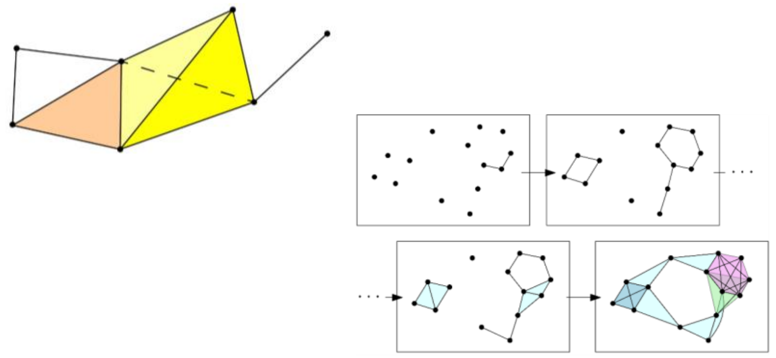
Hypergraphs

- Model *multi-way relationships* in data
    - **Hyperedges** associate groups of **vertices**
- Cyber examples:
    - IP vs. Domain in DNS
    - User vs. host in authentication logs
    - Source IP vs Destination port in network traffic



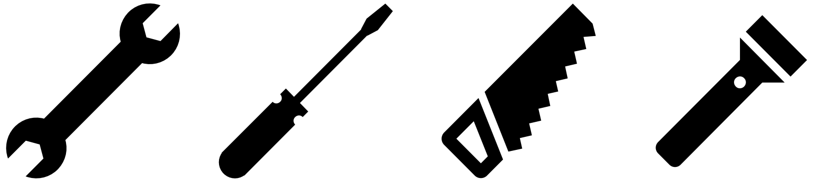
Topological Objects

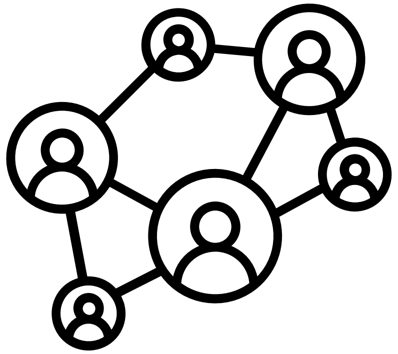
- Provide additional perspective on multi-way relationships
- Capture shape signature for high dimensional data
- Cyber examples:
    - Constructed from cyber hypergraphs
    - Derived using point clouds from features



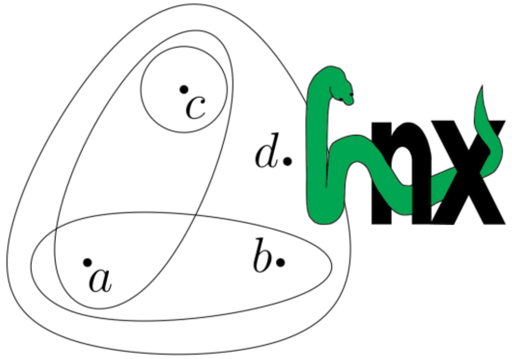
Our Mathematical Toolbox

Models & Methods

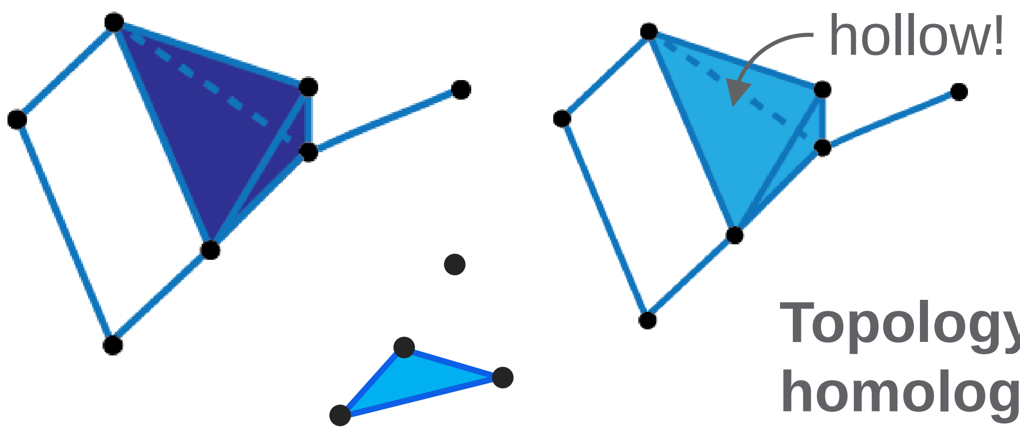




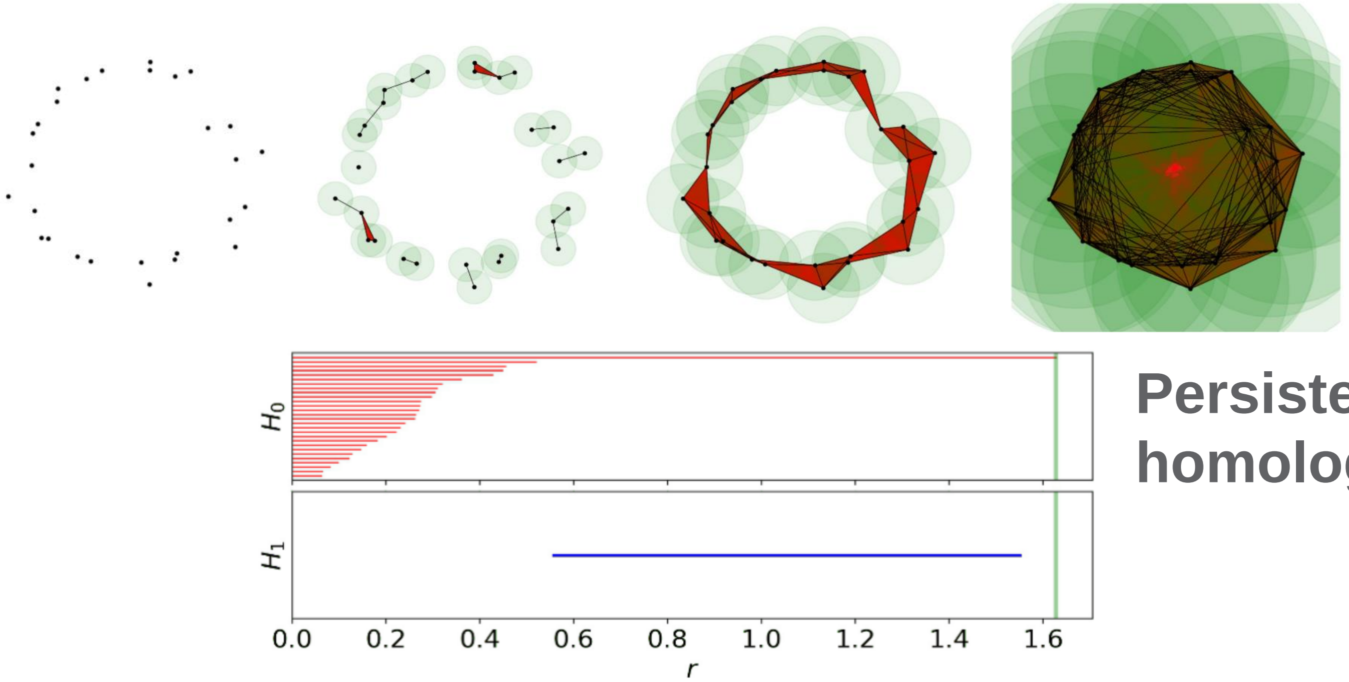
Network science



Hypernetwork science

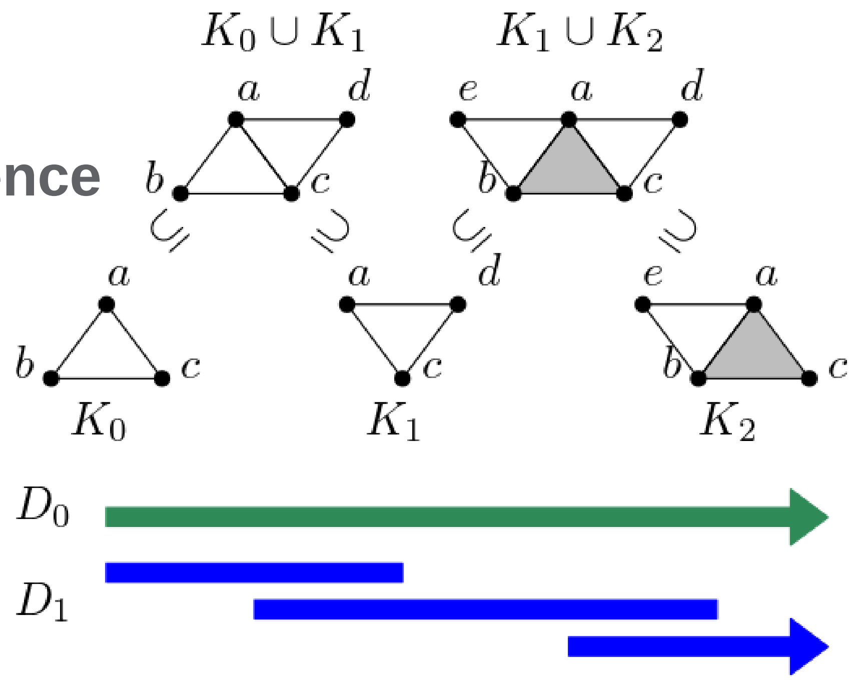


Topology / homology



Persistent homology

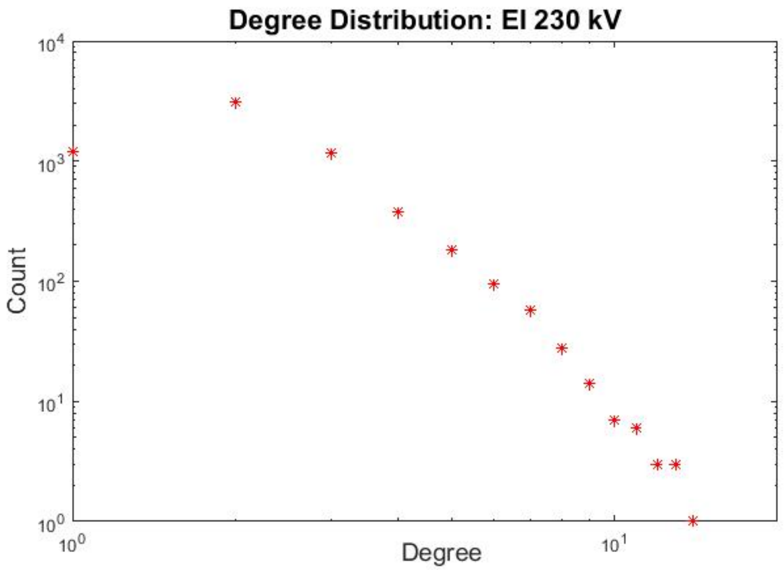
Zigzag persistence

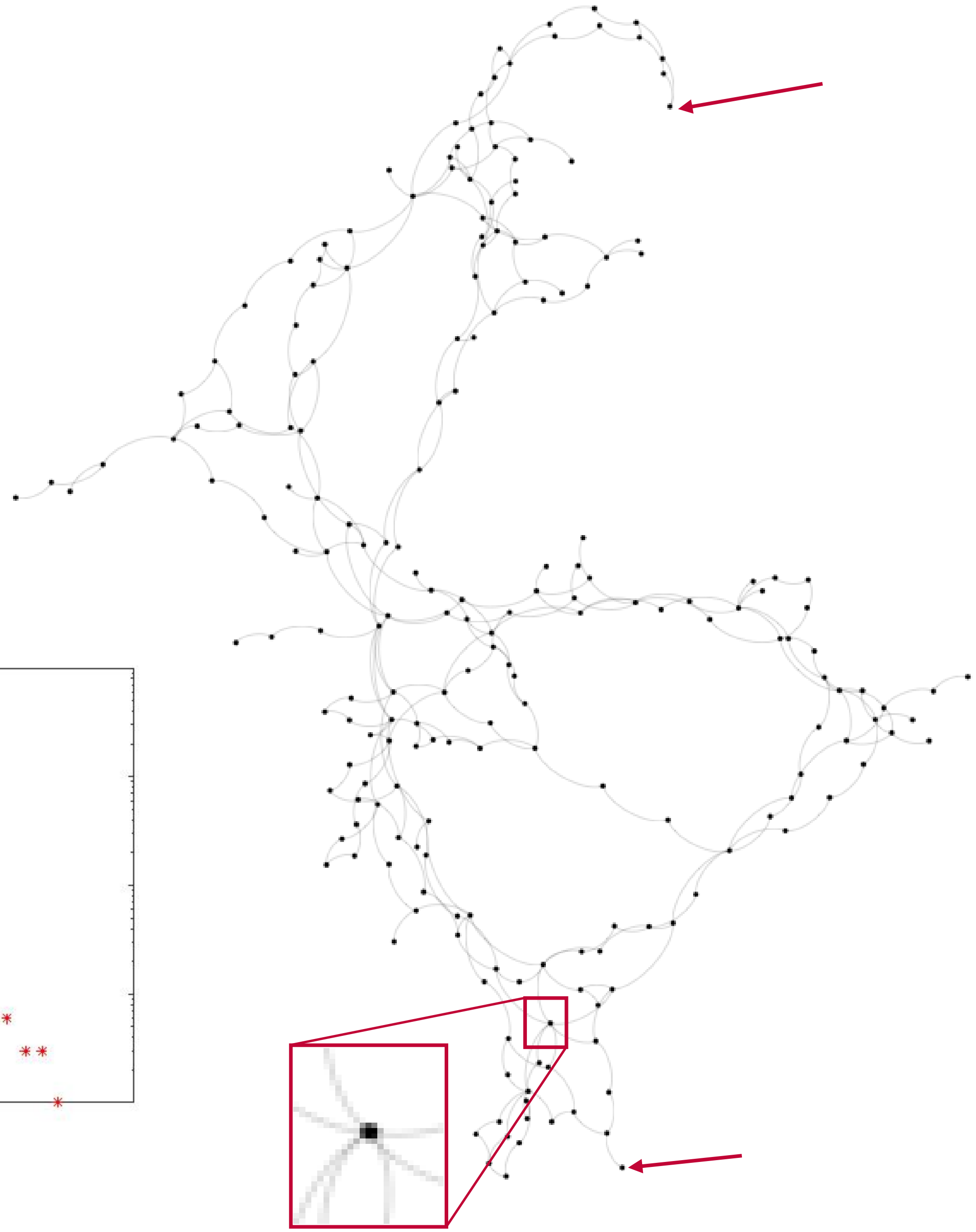


Network science: methods to study structure of graphs from real data

Graph properties

- Degree (distribution)
- Walk, Path, Diameter
- Connected components
- Centrality
- Clustering coefficient
- Triangle counting
- …

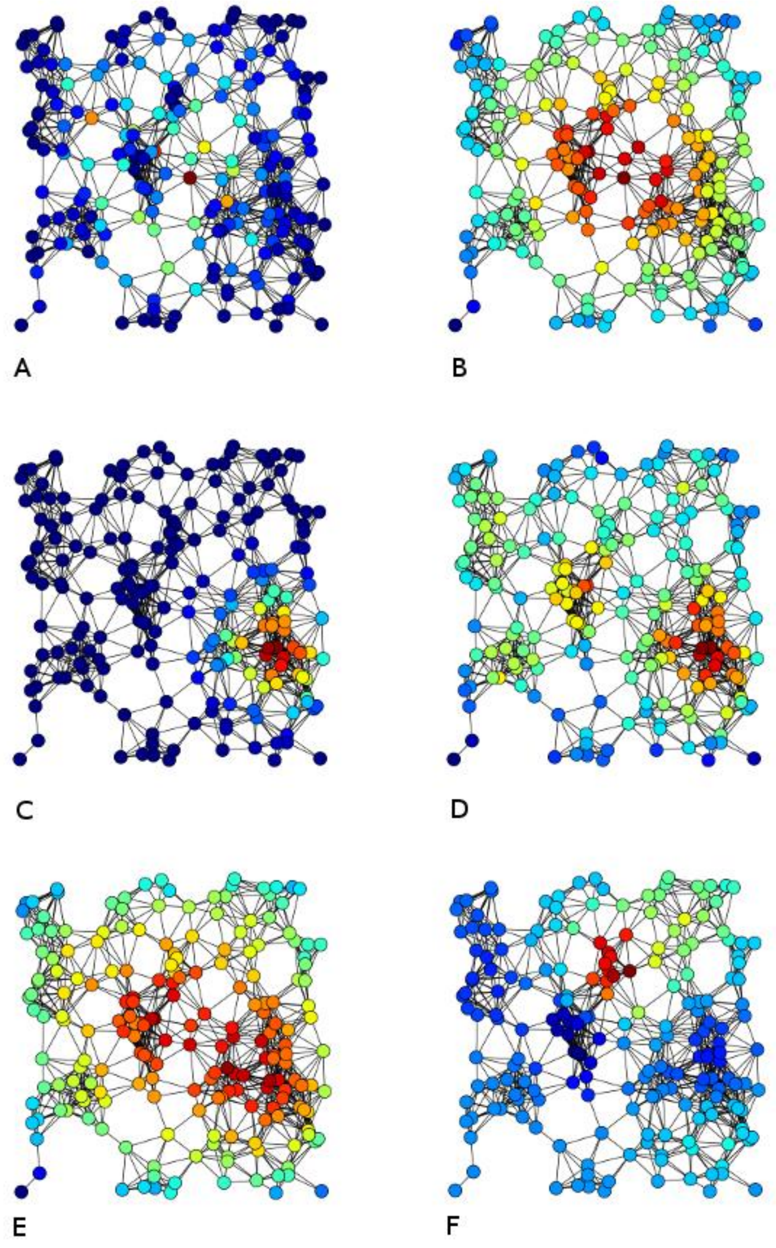




Network science: methods to study structure of graphs from real data

Graph properties

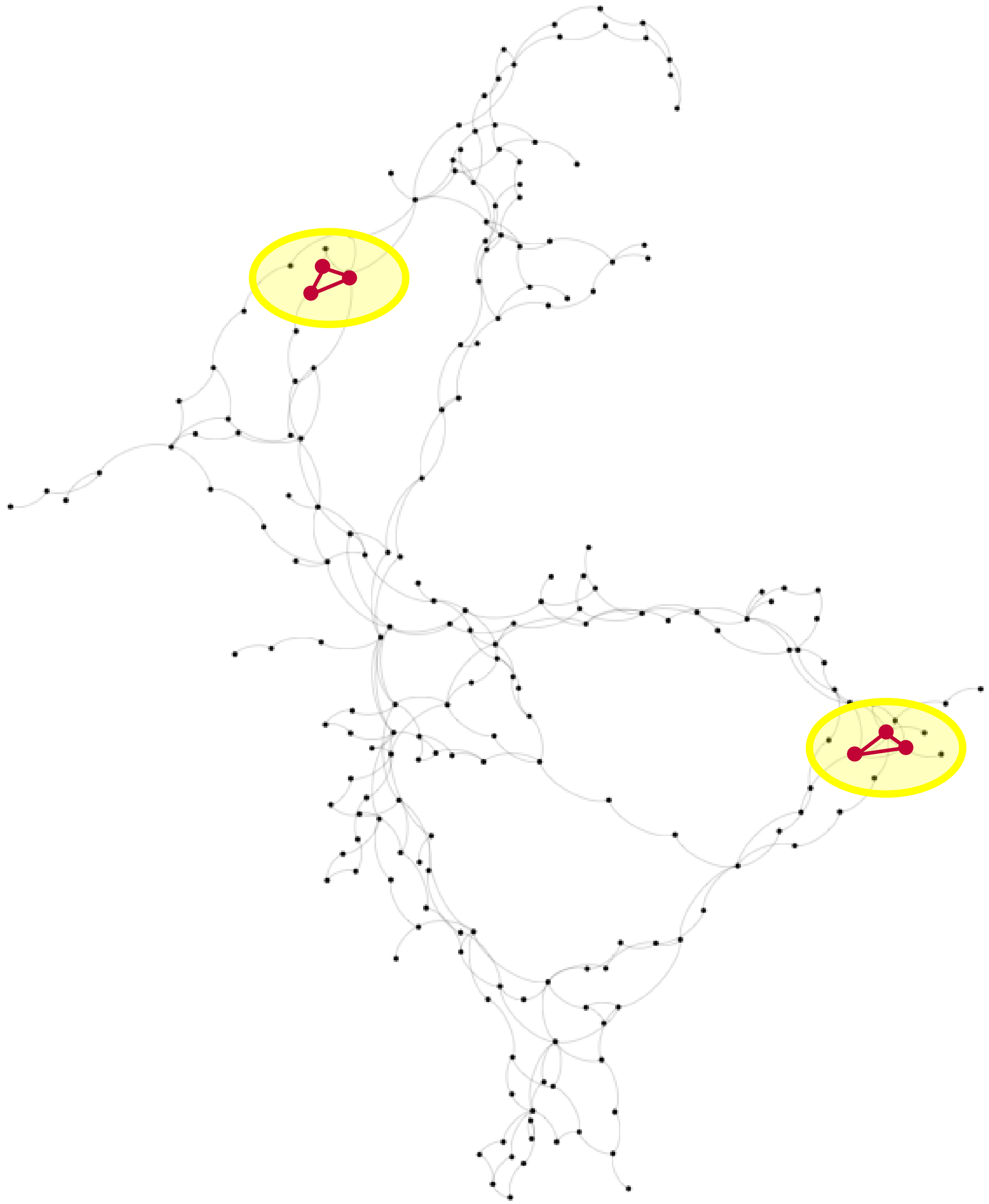
- Degree (distribution)
- Walk, Path, Diameter
- Connected components
- Centrality – measured for each vertex
  - Betweenness: measure of belonging to shortest paths
  - Closeness: measure of average distance to other vertices
  - Eigenvector: Solution to $Ax = \lambda x$
  - Degree: degree of vertex
  - Harmonic: measure of average distance, ok with disconnected graph
  - Katz: related to number of reachable vertices from, with farther vertices penalized



Network science: methods to study structure of graphs from real data

Graph properties

- Degree (distribution)
- Walk, Path, Diameter
- Connected components
- Centrality
- Clustering coefficient
- Triangle counting
- …*



* Number of edges, density, average distance, random graph models, link prediction

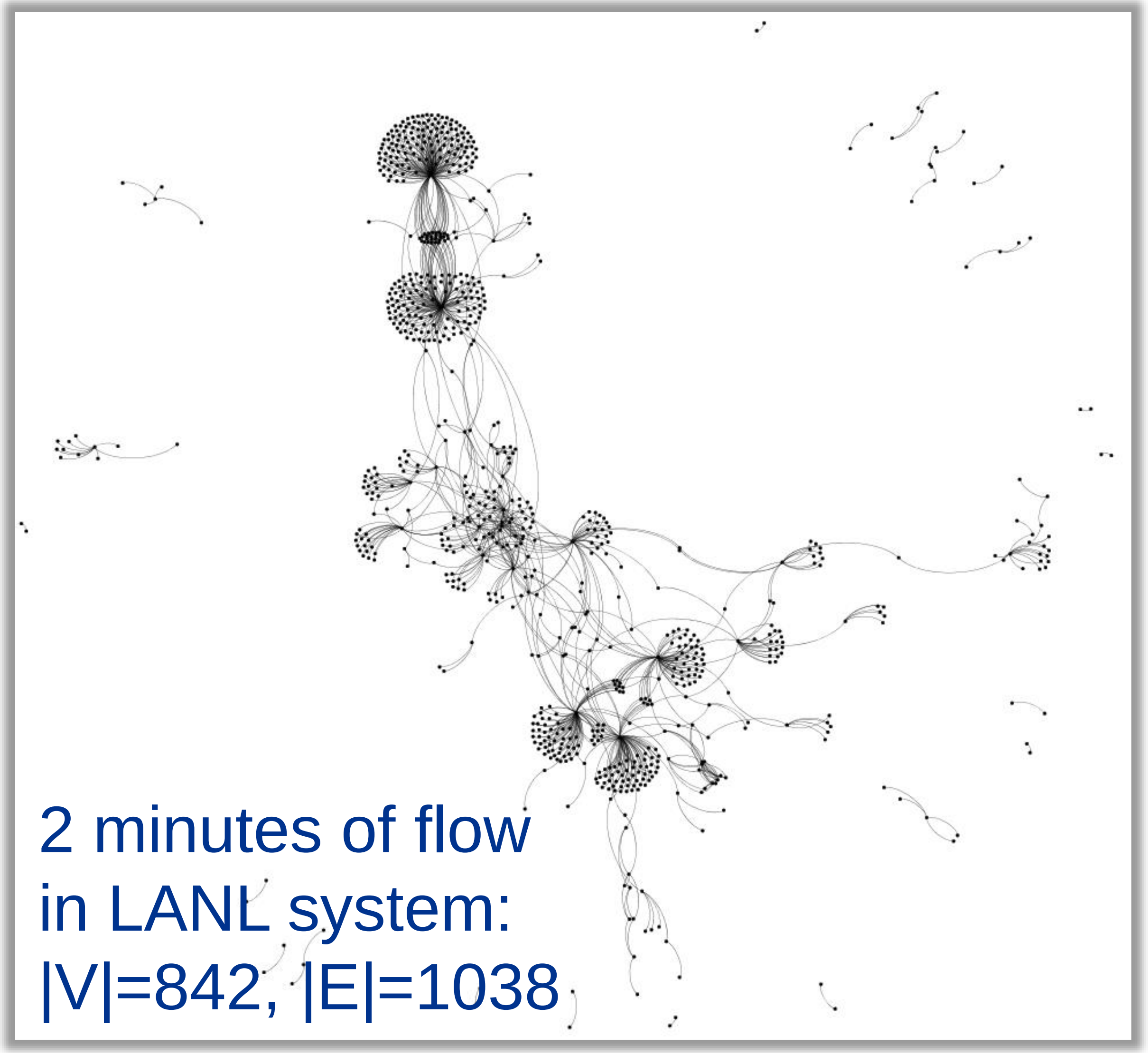
Network flow as a graph

One flow record:

Source, Destination IP	10.0.0.13, 10.0.0.1
Source, Destination Port	33165, 80
Start time	2016/04/15T16:44:41.948
End time	2016/04/15T16:44:41.950
# packets, # bytes	12, 714
Protocol	6 (TCP)

One edge in a graph:

10.0.0.13:33165 $\xrightarrow{\text{(6) 12, 714}}$ 10.0.0.1:80



Data from http://csr.lanl.gov/data/cyber1/

Pacific Northwest
NATIONAL LABORATORY

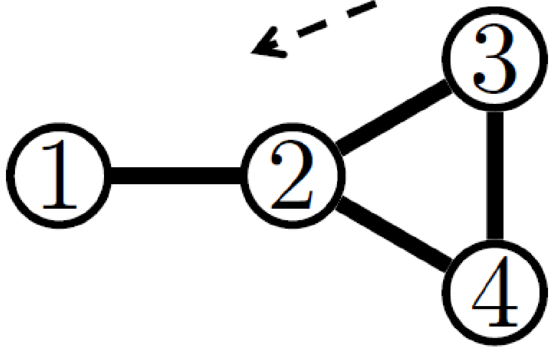
Graphs and Linear Algebra
aka, “Spectral Graph Theory”

Eigen-reminder

Matrix A, find
(λ, x) pairs such that

$$Ax = \lambda x$$

Eigen*vector* Eigen*value*

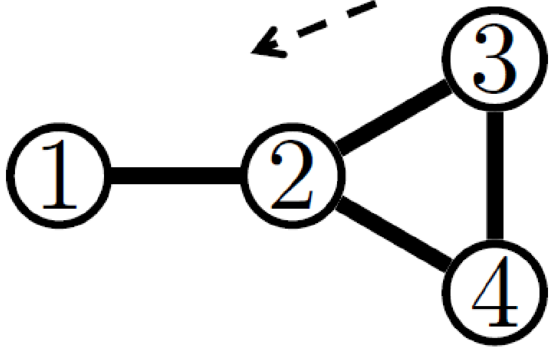


$\rightarrow$ $\begin{pmatrix} & \\ & \end{pmatrix}$ $\rightarrow$ $[\#, \#, \#, \#]$

Graph | Matrix representation | Eigenvalues = “Spectrum”

- Graph’s spectrum tells you something about graph’s structure
  - Connectivity
  - Expansion
  - Degrees
  - Average shortest path length
  - Diameter
  - Chromatic number
  - Random walk mixing time
  - Independence number
  - Number of spanning trees
  - Network flows and routing
  - …

Graphs and Linear Algebra
aka, “Spectral Graph Theory”



Graph

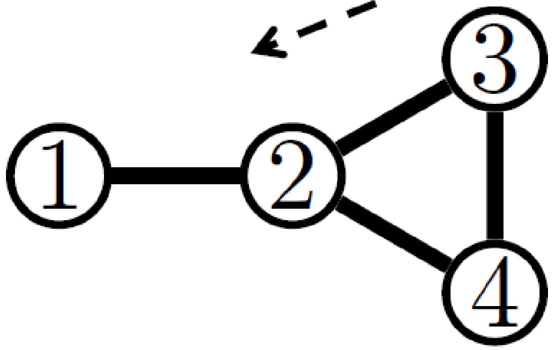
$$\to \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \to [-1.48, -1, 0.31, 2.17]$$

Adjacency matrix

Eigenvalues = “Spectrum”

- Graph’s spectrum tells you something about graph’s structure
  - Connectivity
  - Expansion
  - Degrees
  - Average shortest path length
  - Diameter
  - Chromatic number
  - Random walk mixing time
  - Independence number
  - Number of spanning trees
  - Network flows and routing
  - …

Graphs and Linear Algebra
aka, “Spectral Graph Theory”



Graph

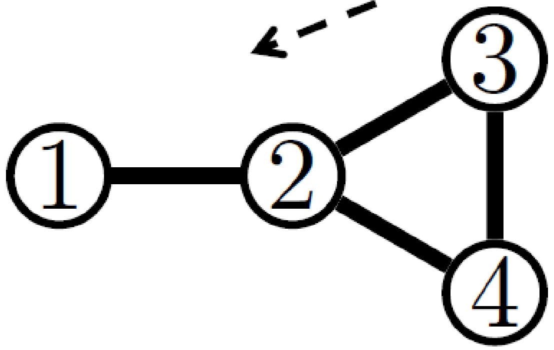
$$\rightarrow \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} \rightarrow [0, 1, 3, 4]$$

Combinatorial Laplacian matrix

Eigenvalues = “Spectrum”

- Graph’s spectrum tells you something about graph’s structure
  - Connectivity
  - Expansion
  - Degrees
  - Average shortest path length
  - Diameter
  - Chromatic number
  - Random walk mixing time
  - Independence number
  - Number of spanning trees
  - Network flows and routing
  - …

Graphs and Linear Algebra
aka, “Spectral Graph Theory”



$$\rightarrow \begin{bmatrix} 1 & \frac{-1}{\sqrt{3}} & 0 & 0 \\ \frac{-1}{\sqrt{3}} & 1 & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{-1}{\sqrt{6}} & 1 & \frac{-1}{2} \\ 0 & \frac{-1}{\sqrt{6}} & \frac{-1}{2} & 1 \end{bmatrix} \rightarrow 0, 0.77, 1.5, 1.73$$

Graph

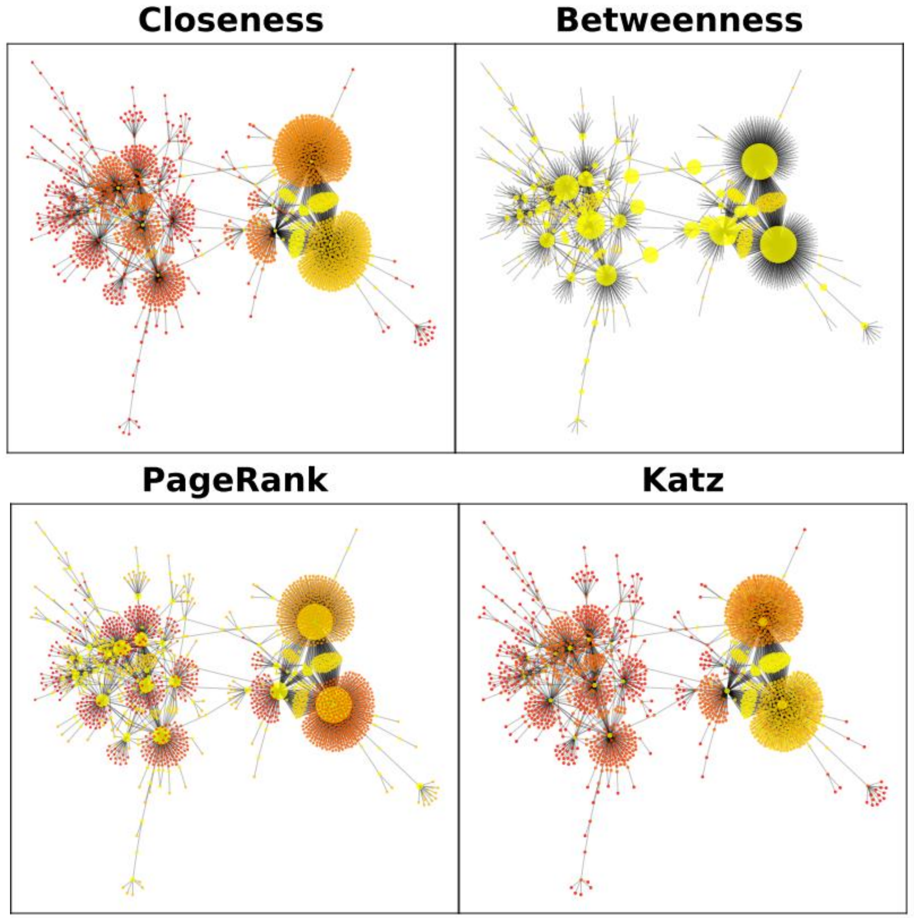
Normalized Laplacian matrix

Eigenvalues = “Spectrum”

- Graph’s spectrum tells you something about graph’s structure
  - Connectivity
  - Expansion
  - Degrees
  - Average shortest path length
  - Diameter
  - Chromatic number
  - Random walk mixing time
  - Independence number
  - Number of spanning trees
  - Network flows and routing
  - …

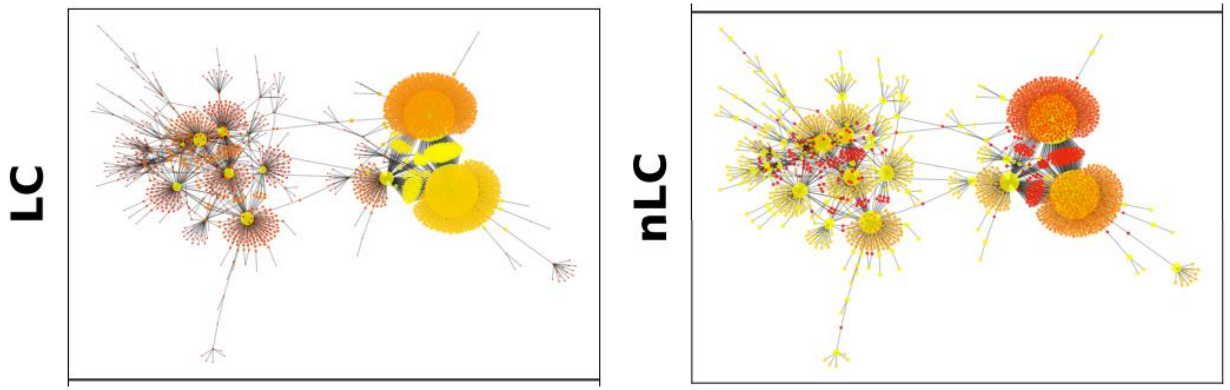
Measuring vertex importance

Vertex size is proportional to score, and color is proportional to score percentile, with yellow for higher percentiles and red for lower percentiles.



- Some common centralities
  - **Closeness**: measure of average distance to other vertices
  - **Betweenness**: measure of belonging to shortest paths
  - **PageRank**: related to stationary distribution of modified random walk
  - **Katz**: related to number of reachable vertices from, with farther vertices penalized
- Another one: **Laplacian centrality**[5] – change in spectrum from removing a vertex
  - Laplacian energy of a graph: $\mathcal{E}(G) = \sum_{i=1}^{n} \lambda_i^2$
  - Laplacian centrality of a vertex:

$$LC_G(v) = \frac{\mathcal{E}(G) - \mathcal{E}(G \setminus v)}{\mathcal{E}(G)}$$

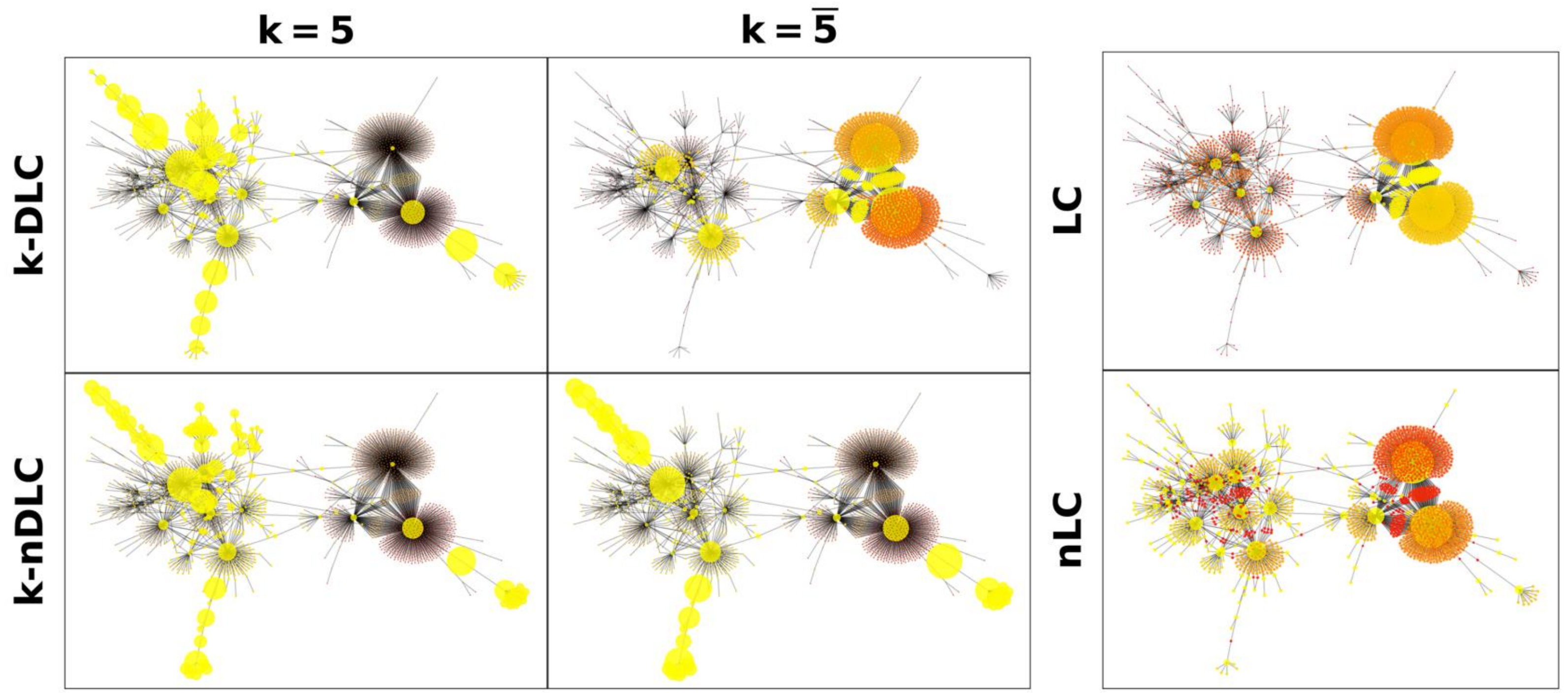


[5] Xingqin Qi, Eddie Fuller, Qin Wu, Yezhou Wu, and Cun-Quan Zhang. Laplacian centrality: A new centrality measure for weighted networks. Information Sciences, 194:240–253, 2012.

Aksoy, S.G., Purvine, E. and Young, S.J., 2021. Directional Laplacian Centrality for Cyber Situational Awareness. Digital Threats: Research and Practice (DTRAP), 2(4), pp.1-28. https://dl.acm.org/doi/10.1145/3450286.

Directional Eigenvalue Derivative – idea and adjacency matrix formulation

- Removing a vertex, as in Laplacian centrality, does damage to the graph
- Instead, make *infinitesimal* change to the graph structure – “derivative of an eigenvalue in the direction of a vertex”
  - k-DLC: uses smallest eigenvalues
  - $\bar{k}$-DLC: uses largest eigenvalues

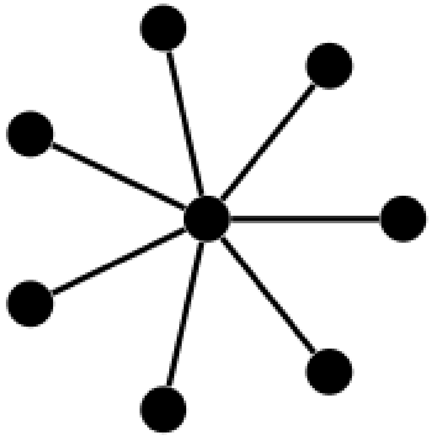


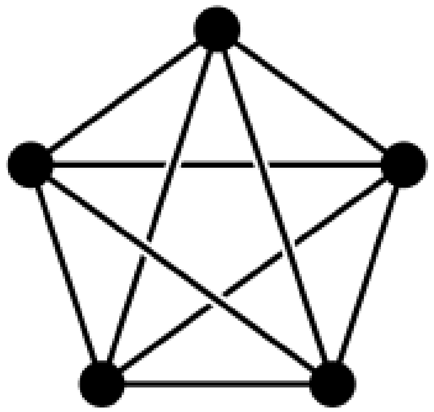
(a) Directional Laplacian Centrality

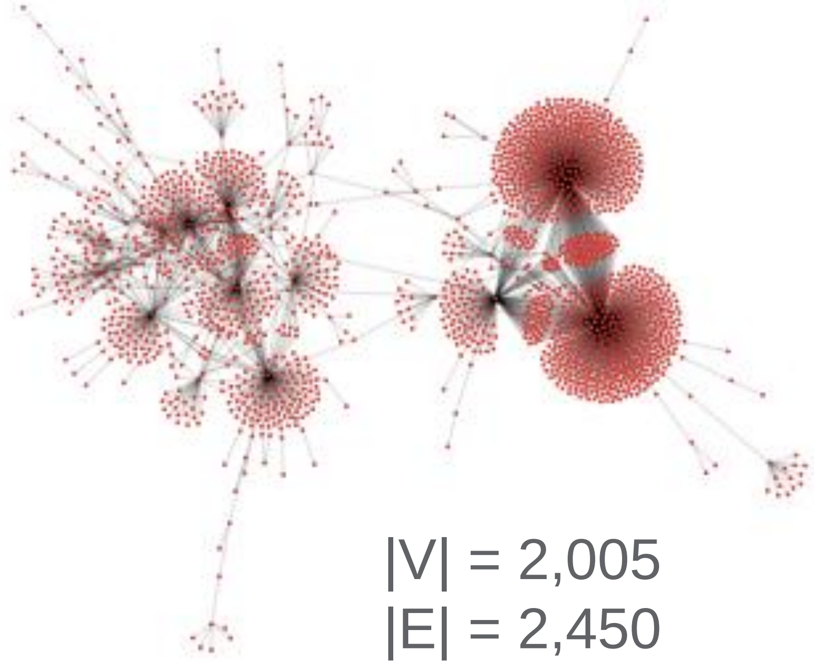
(b) Laplacian Centrality

DLC experiments for network data

- Base graph is snapshot from LANL
- Inject anomalies, measure perturbation of DLC measures
- Two injections
  1. Inject star and clique anomalies of increasing sizes **at low importance vertices**
  2. Inject star anomaly of increasing sizes **randomly**



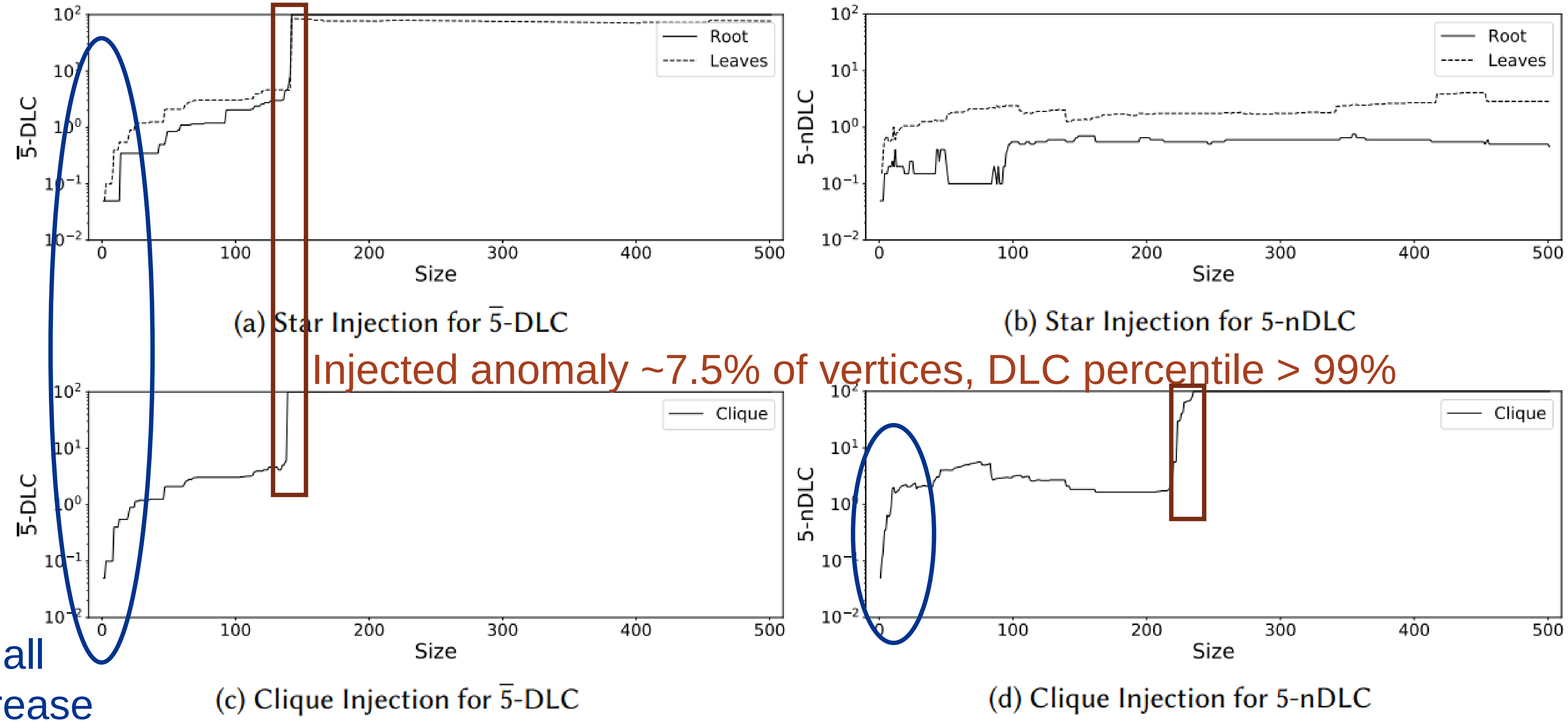




|V| = 2,005
|E| = 2,450

Star and clique at set of low importance vertices

Importance percentile for vertices in injected anomalies as a function of anomaly size



k-nDLC insensitive to size of injected star

Even very small injections increase percentile an order of magnitude

Star injected at random root

- Root connected to random k% of vertices
- Test detection (increase in importance score at root) and sensitivity (larger increases in score for larger values of *k*)

Results averaged over 500 trials		**0.1%** (2 edges)		**0.5%** (10 edges)		**1.0%** (20 edges)		**5.0%** (100 edges)		**10.0%** (201 edges)	
		Score	*PCTL*	*Score*	*PCTL*	*Score*	*PCTL*	*Score*	*PCTL*	*Score*	*PCTL*
$\bar{5}$-DLC	Before	2.30	44%	1.08	45%	1.11	45%	0.83	43%	2.04	46%
	After	2.31	55%	1.11	64%	1.19	67%	1.47	74%	205.68	99%
	Change	0.01	11%	0.03	20%	0.08	22%	0.64	31%	203.63	53%
5-nDLC	Before	-4e-5	50%	-7e-5	52%	-9e-7	50%	5e-5	49%	-9e-6	51%
	After	-8e-4	1%	-2e-3	1%	-4e-3	1%	-2e-2	1%	-3e-2	1%
	Change	-7e-4	-49%	-2e-3	-51%	-4e-3	-49%	-2e-2	-48%	-3e-2	-50%

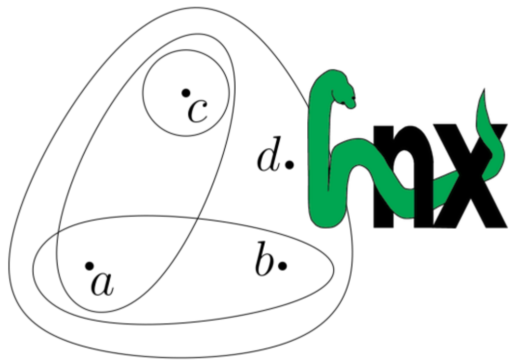
DLC shows detection and sensitivity

nDLC shows detection but not sensitivity

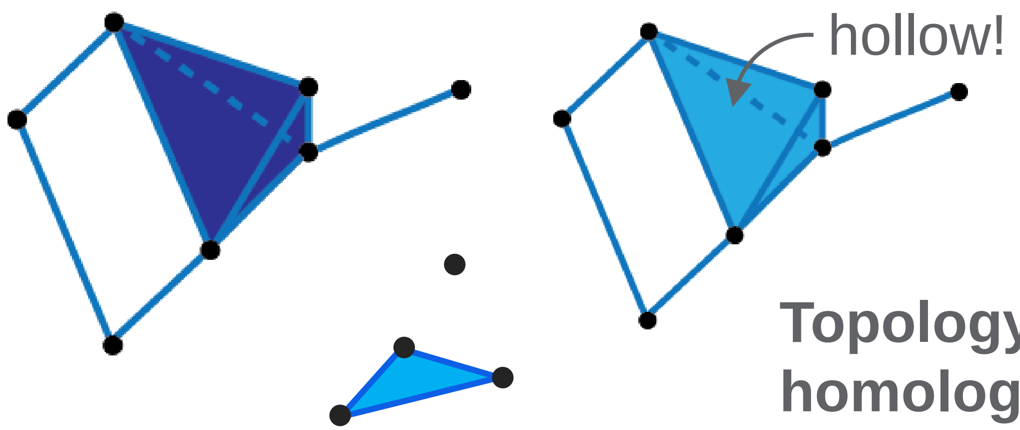
Our Mathematical Toolbox

Models & Methods

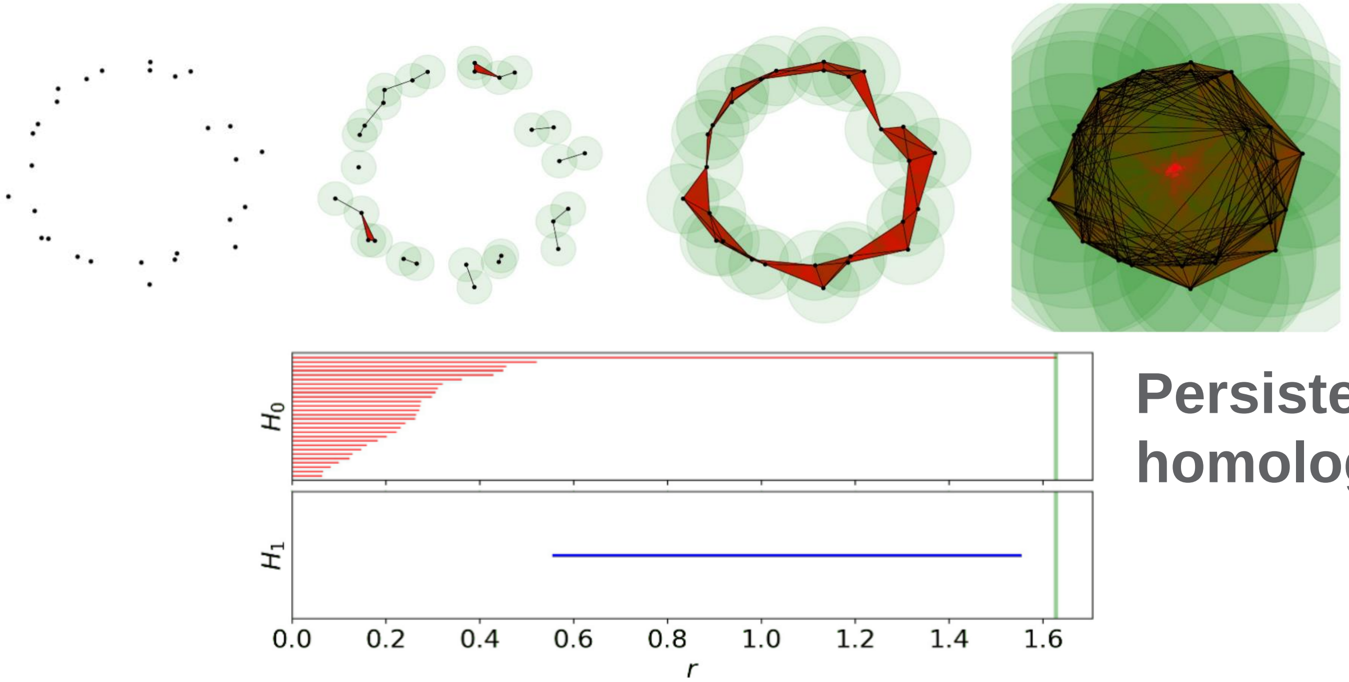
Network science



Hypernetwork science

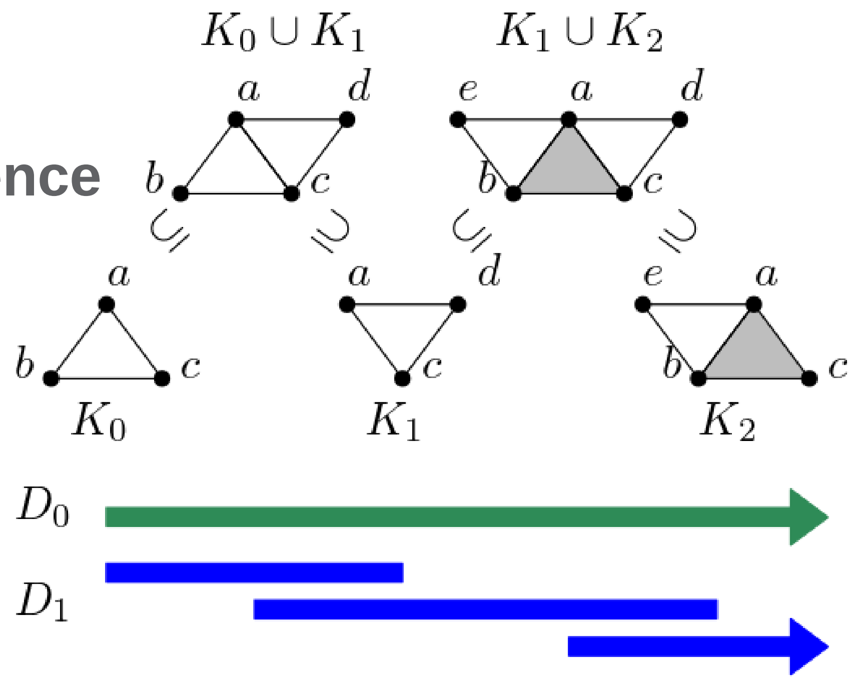


Topology / homology



Persistent homology

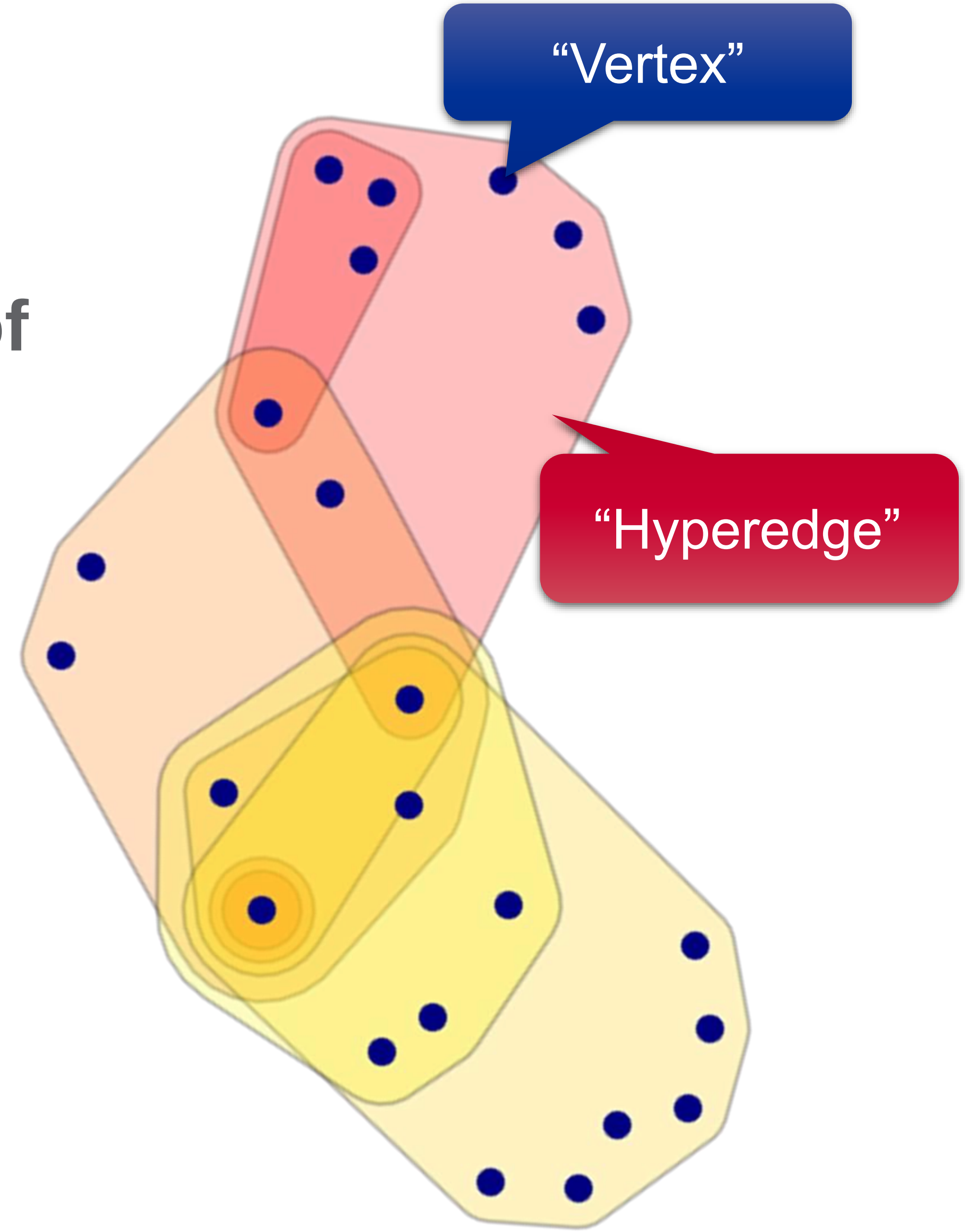
Zigzag persistence



Hypergraphs

- **Hypergraphs provide a mathematical model of data focused on multi-way relationships**
  - To ***ask*** certain kinds of questions
    - ✓Connectivity of entities
    - ✓Clustering structure
  - To ***model*** certain kinds of interactions
    - ✓Multi-way relationships

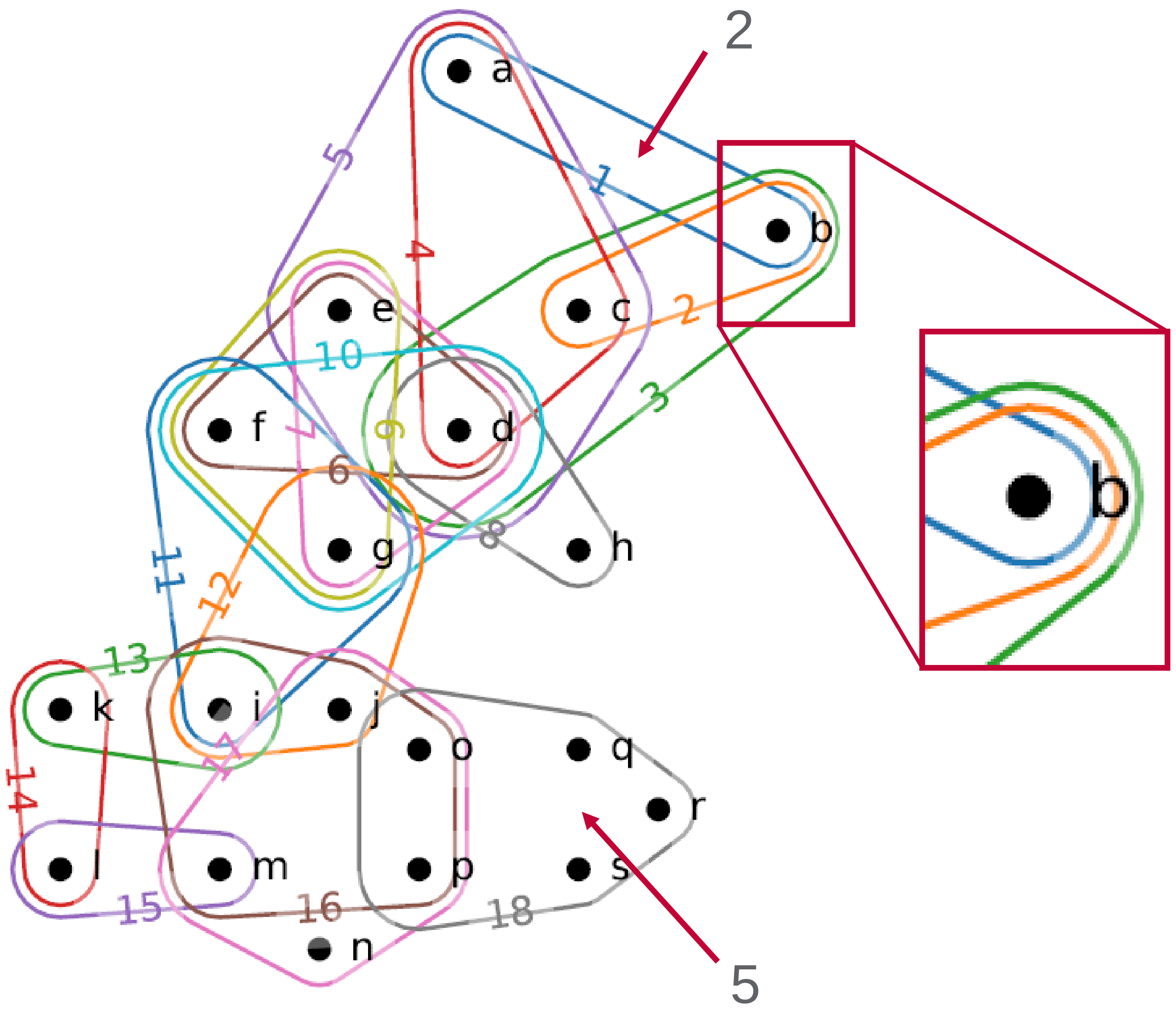
$$H = (V, E), E \subseteq 2^V$$



Co-occurrence of characters in Les Miserables, restricted to single character neighborhood. Image generated by HyperNetX.

Pacific Northwest NATIONAL LABORATORY

Hypernetwork science



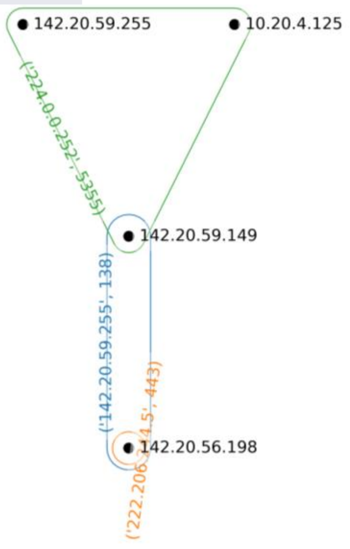
Hypergraph properties

- Degree (distribution)
- Edge size (distribution)
- s-Walk, s-Path, s-Diameter
- s-Connected components
- s-Centrality
- Clustering coefficient?
- Triangle counting?
- …

Operationally Transparent Cyber (OpTC) Data
And hypergraph construction method

- Released by DARPA to enable research that enhances understanding of and defense against APTs at scale
- 17 billion events generated from a simulated network consisting of ~500 hosts over 5 days of benign data plus 3 days that include red team behavior
  - Host and network logs: actions on file, process, flow, registry, module, thread objects

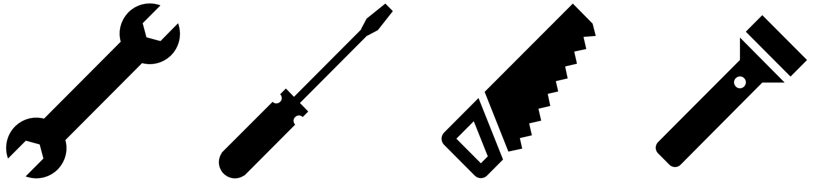
hostname	principal	pid	src_ip	dest_ip	dest_port	l4protocol	image_path
SysClient0201.systemia.com	NT AUTHORITY\SYSTEM	4	142.20.56.198	142.20.59.255	138	UDP	System
SysClient0201.systemia.com	NT AUTHORITY\NETWORK SERVICE	864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
SysClient0201.systemia.com	NT AUTHORITY\NETWORK SERVICE	864	142.20.59.255	224.0.0.252	5355	UDP	svchost.exe
SysClient0201.systemia.com	SYSTEMIACOM\zleazer	636	142.20.56.198	222.206.244.5	443	TCP	firefox.ex
SysClient0201.systemia.com	NT AUTHORITY\SYSTEM	4	142.20.59.149	142.20.59.255	138	UDP	System
SysClient0201.systemia.com	NT AUTHORITY\NETWORK SERVICE	864	142.20.59.149	224.0.0.252	5355	UDP	svchost.e



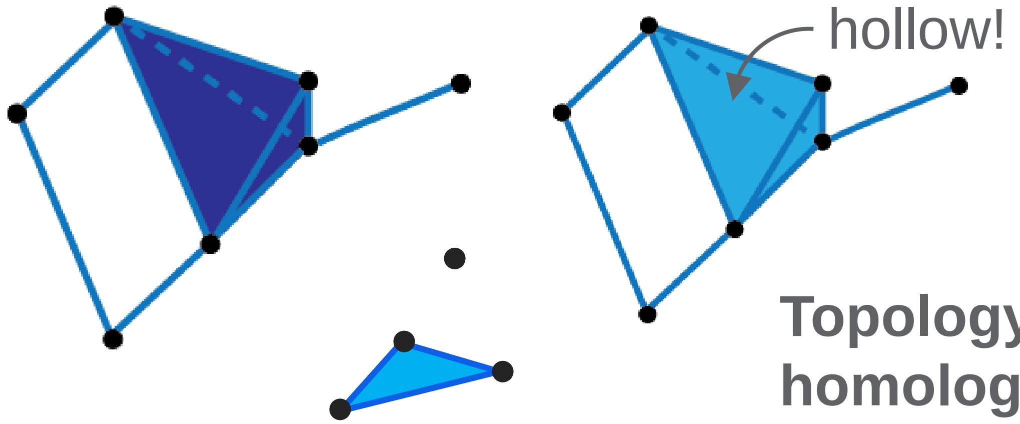
- Hypergraph construction from tabular data: choose hyperedge columns and node columns
  - A vertex is contained in a hyperedge if there is a record with that combination in the data. Think "hyperedges = common behaviors"

Our Mathematical Toolbox

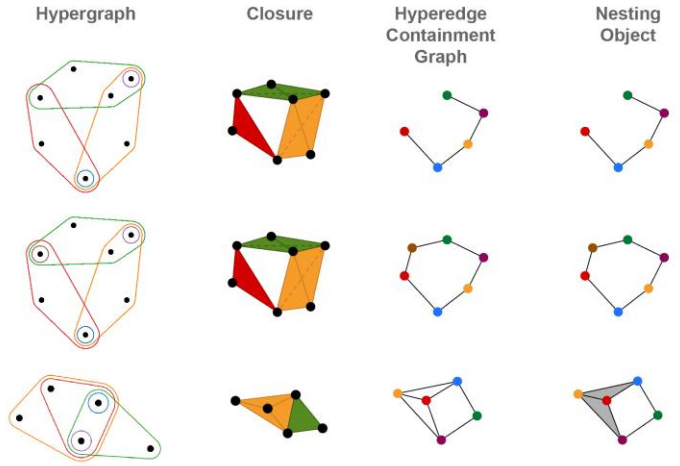
Models & Methods



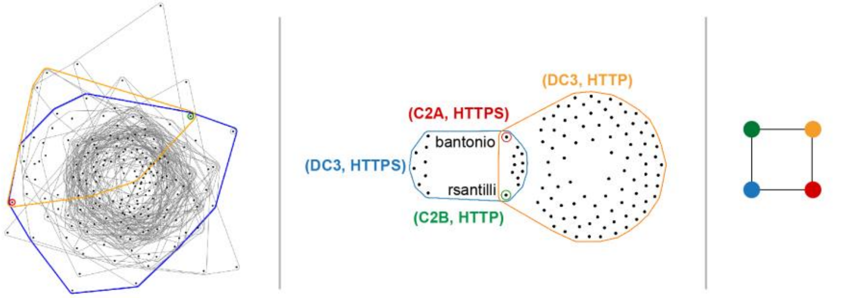
- Topology = stretchy geometry
  - Two shapes are considered “the same” if they can be continuously deformed into one another
- Homology = method to count holes in any dimension in a topological object
  - Connected components (0-dimensional holes), loops (1-dimensional holes), voids (2-dimensional holes), and higher dimensional analogs
  - Foundation of “Topological Data Analysis”
- We study homology of simplicial complexes derived from cyber hypergraphs
- **Cyber citation:** Helen Jenne, et al. "Stepping out of Flatland: Discovering Behavior Patterns as Topological Structures in Cyber Hypergraphs" https://arxiv.org/abs/2311.16154



Topology / homology



Two simplicial complex constructions from a hypergraph

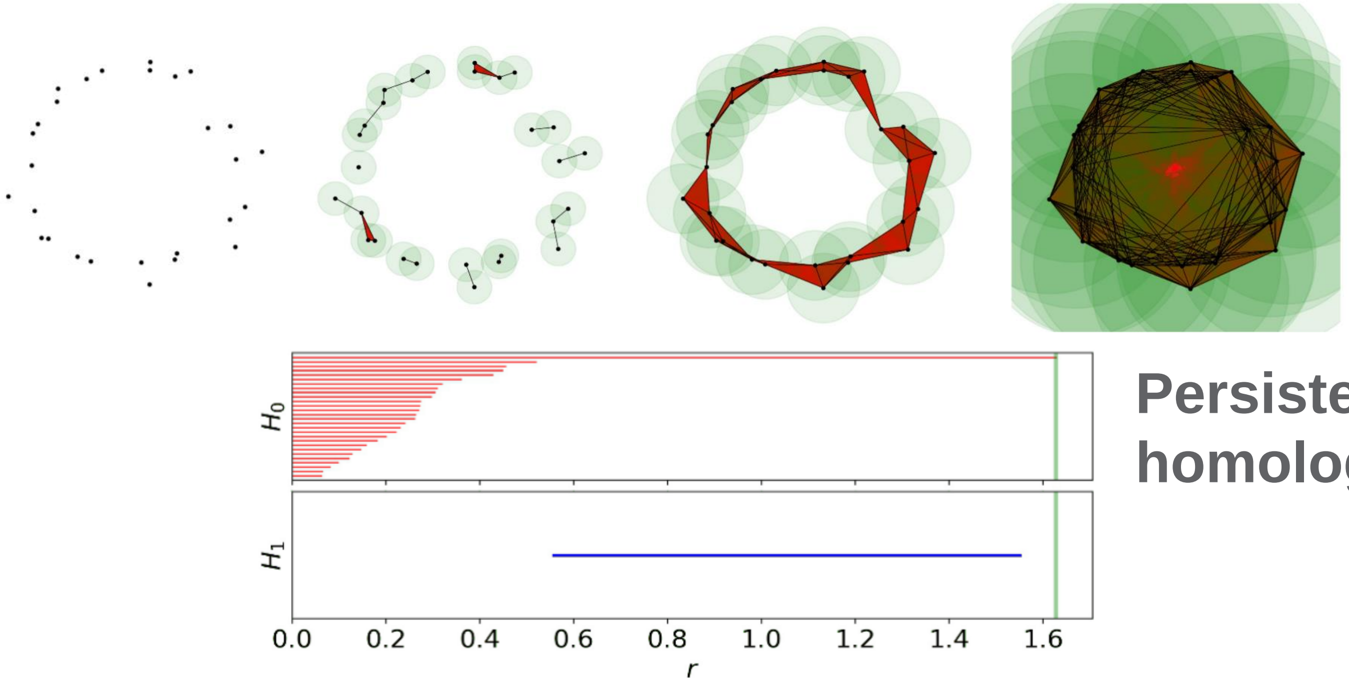


Homological feature of nesting object capturing adversary activity in OpTC data

Our Mathematical Toolbox

Models & Methods

- Persistent homology (PH) captures topological fingerprint of point cloud or metric space data (without PH point clouds have trivial homology)
  - Connect points within given distance threshold,
  - Compute homology of resulting simplicial complex,
  - Sweep across distance thresholds and record birth/death of homological features
- Can be applied to arbitrary nested sequence of simplicial complexes

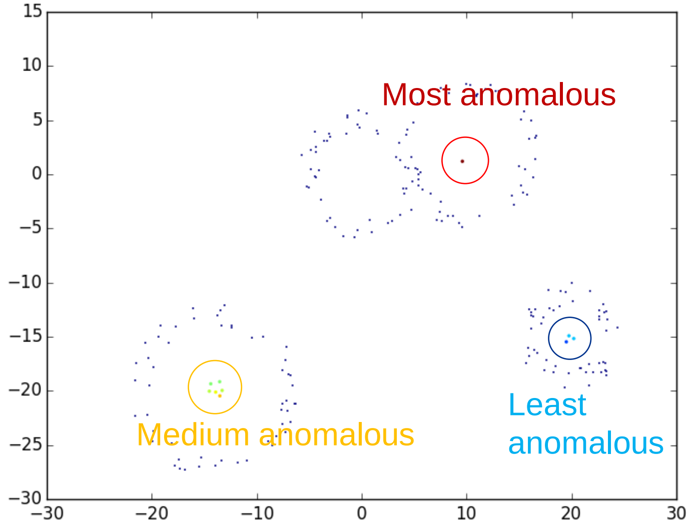


Persistent homology

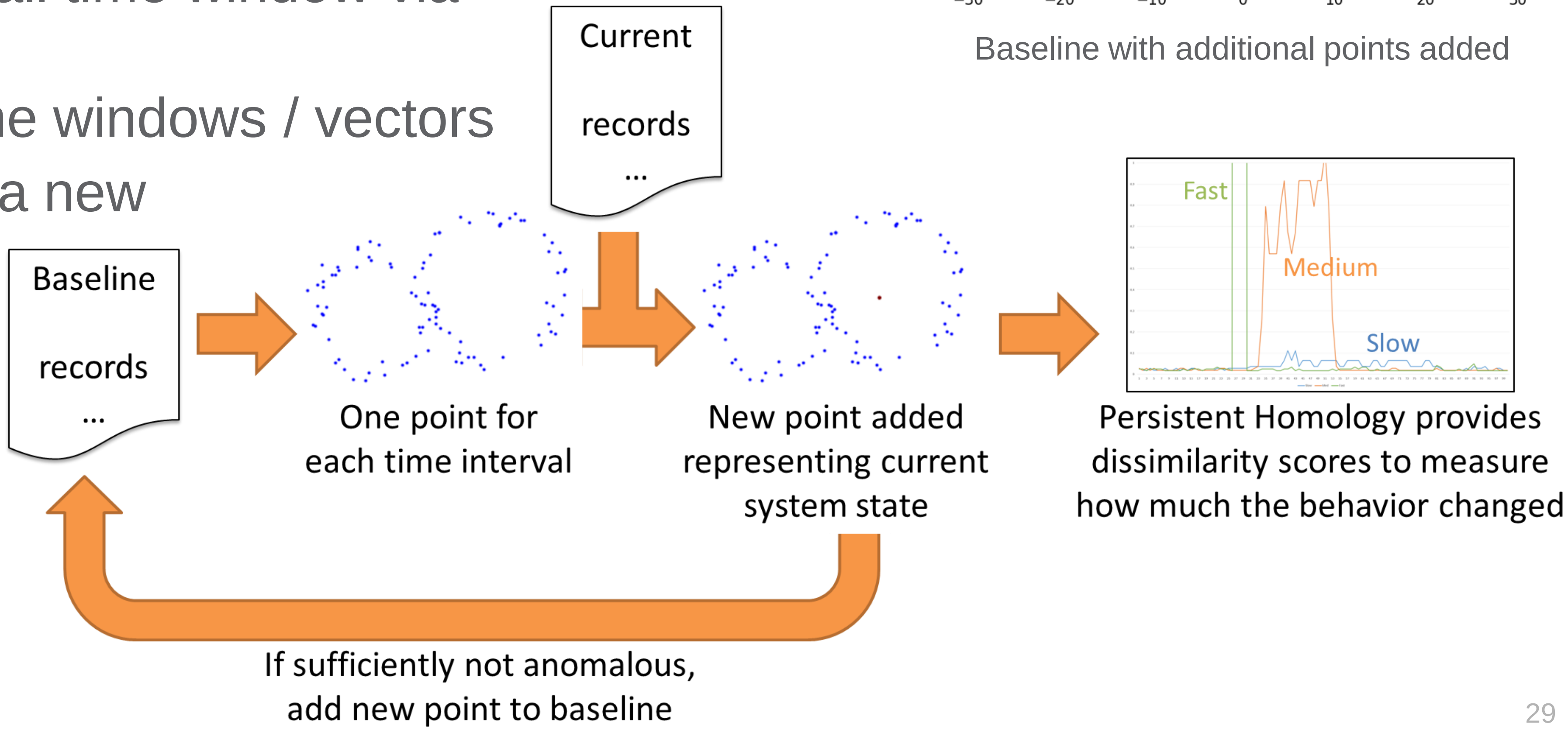
- Result is barcode or persistence diagram, can be used to compare metric spaces or as input to ML pipelines
- **Cyber citation:** Paul Bruillard et al. “Anomaly detection using persistent homology.” In 2016 Cybersecurity Symposium (CYBERSEC), pp. 7-12. IEEE, 2016.

PHANTOM = Persistent Homology Anomalous Traffic Observation Monitor

- **Main assumption:** Behavior varies smoothly from set of recent (or representative) small time windows to the next window
- PHANTOM algorithm built on this assumption
  - Behavior = summary of small time window via custom vectorization
  - Baseline contains many time windows / vectors
  - How, and how much, does a new time window / vector perturb the baseline?
- Perturbation measured using persistent homology

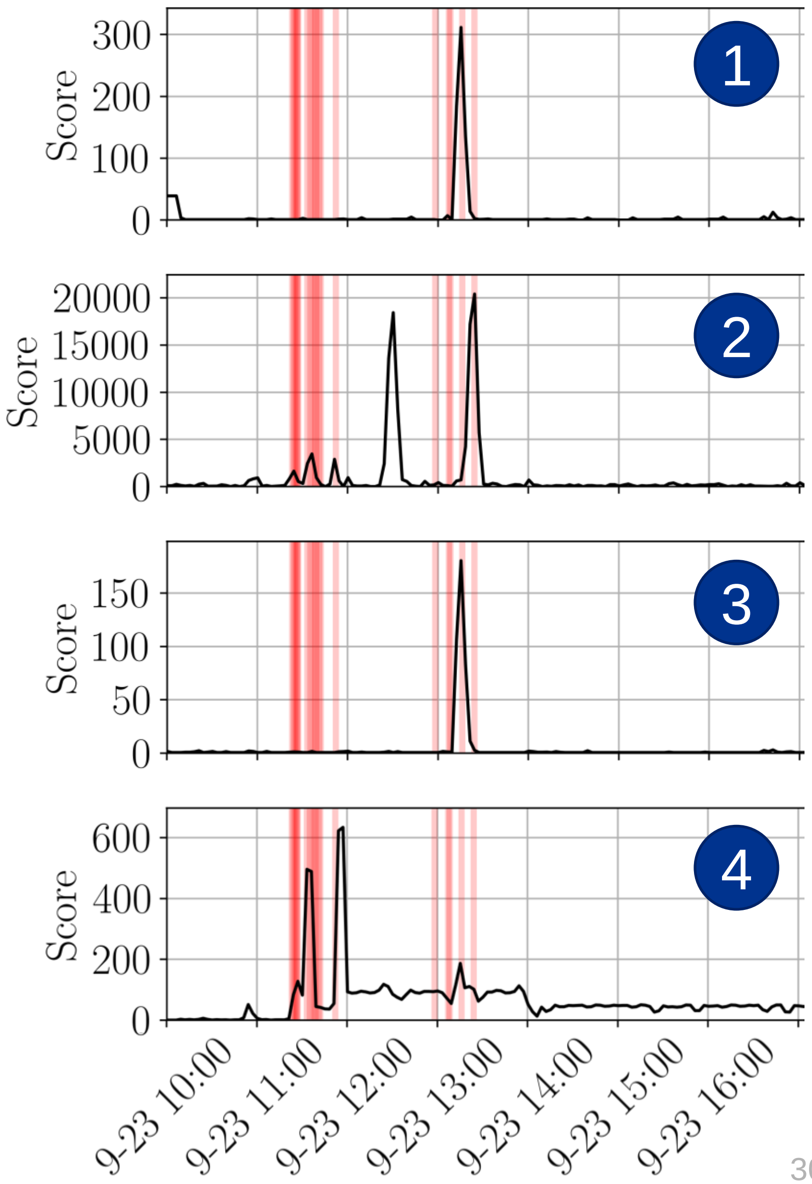


Baseline with additional points added



PHANTOM applied to OpTC

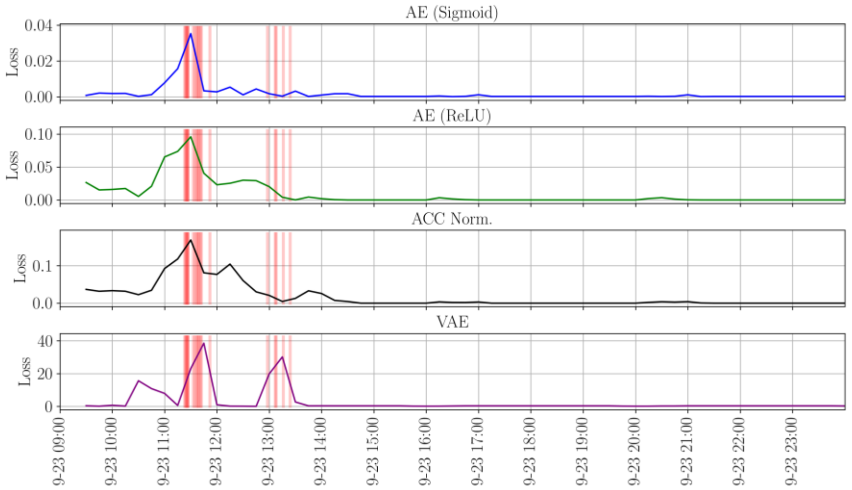
- Day 1 of OpTC included ping sweep on Host 201 around 13:15
  - ICMP traffic, Source port = 8, Destination port = 0
- Explored many vectorizations with varying levels of tailoring to this specific activity. Three examples:
  - 1. Ping sweep clearly seen
    - ✓ Count unique: process IDs
    - ✓ Count values: source port 0 and 8, destination port 0 and 8, protocol ICMP
  - 2. Looks like it aligns with ping sweep, but it doesn't!
    - ✓ Count unique: process IDs
    - ✓ Mean of size ← Anomaly score sensitive to size
    - ✓ Count values: source port 0 and 8 , destination port 0 and 8, protocol ICMP
  - 3. Ping sweep clearly seen
    - ✓ Count unique: process IDs, destination ports
    - ✓ Count values: protocol ICMP
  - 4. Ping sweep seen, overshadowed by earlier spikes globally
    - ✓ Count unique: process IDs, destination ports
    - ✓ Count values: protocol ICMP, TCP



Our Mathematical Toolbox

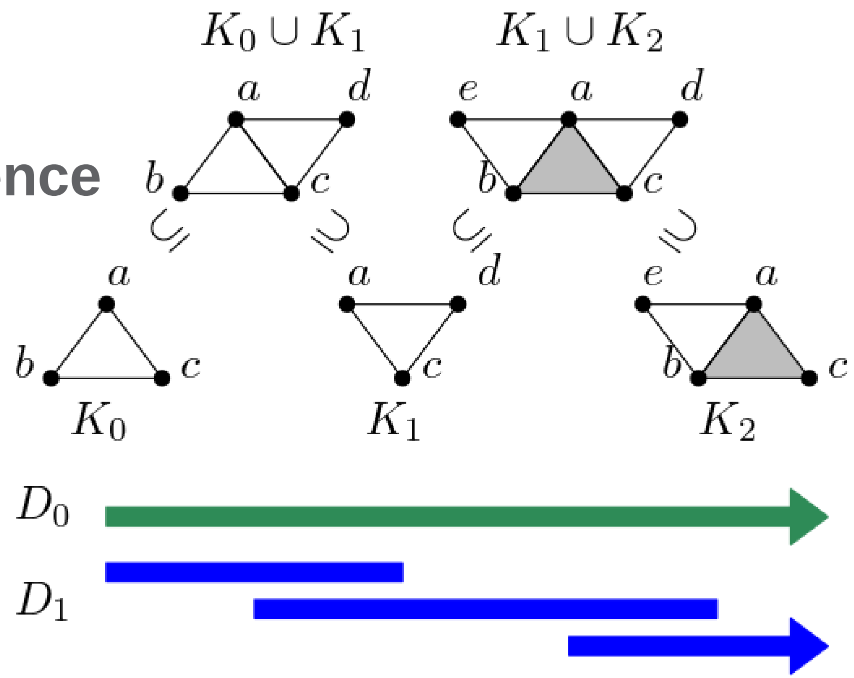
Models & Methods

- Homology of a single simplicial complex does not account for dynamics
- PH requires nested sequence of simplicial complexes
- Topological analysis of temporal sequence of simplicial complexes requires Zigzag Persistence approach
  - Fill out sequence to include sequential pairwise unions (or intersections)
  - Similar algorithm as PH
- **Cyber citation:** Audun Myers, et al. “Malicious Cyber Activity Detection Using Zigzag Persistence.” In IEEE Conference on Dependable and Secure Computing Workshop on AI/ML for Cybersecurity, 2023.



Autoencoders trained on stats derived from zigzag barcodes. High loss aligns with known adversary activity.

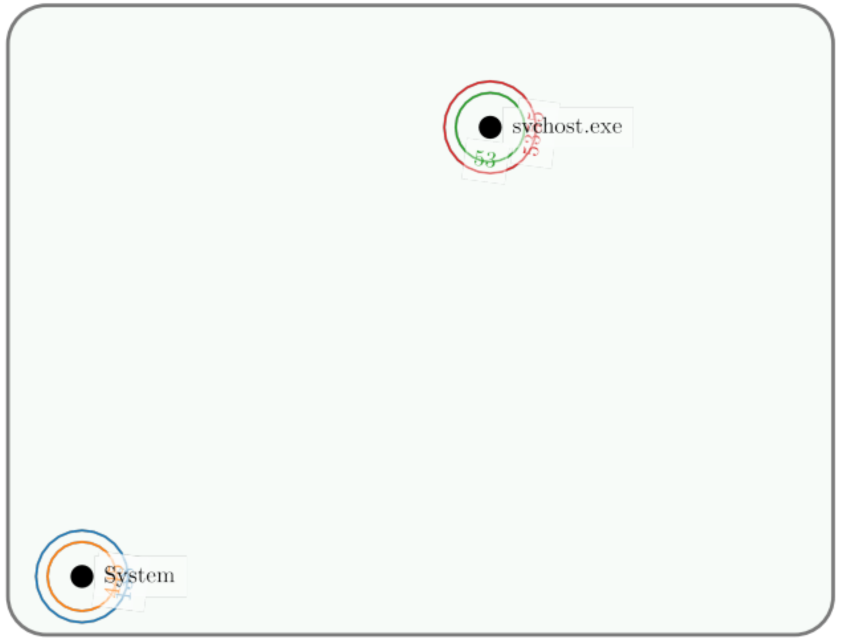
Zigzag persistence

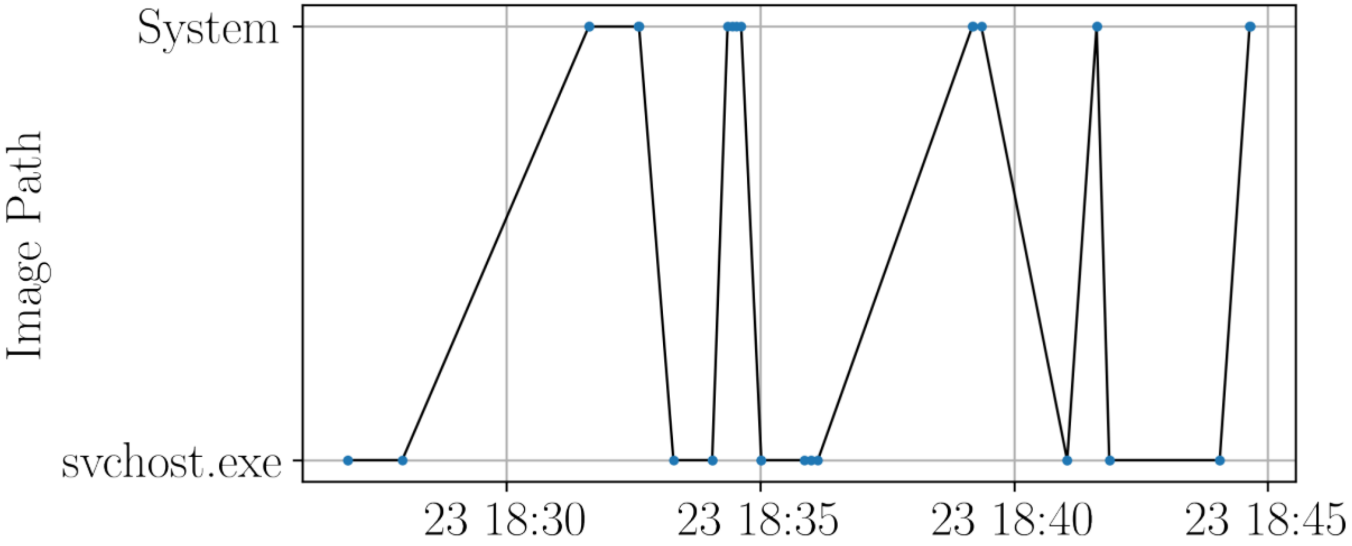


Pacific Northwest
NATIONAL LABORATORY

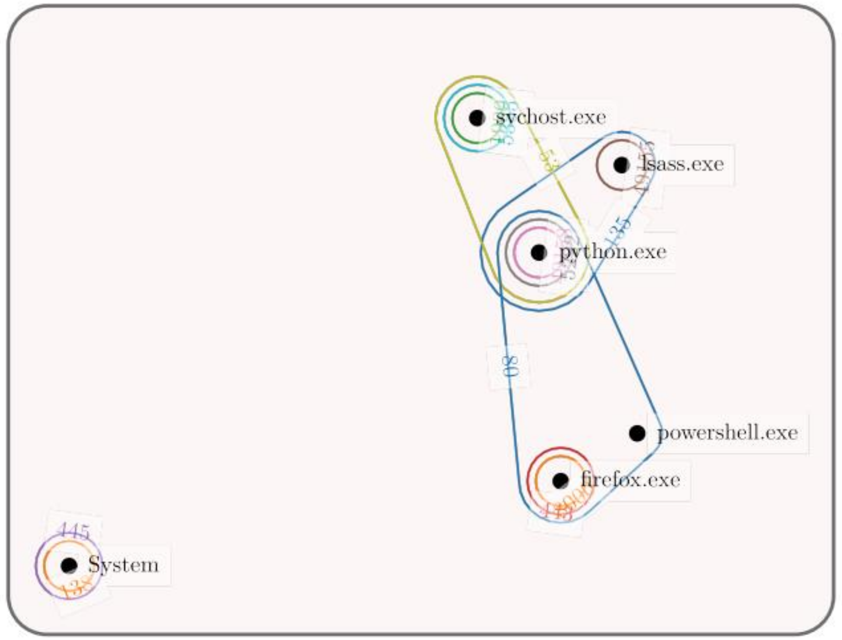
Dynamics of benign vs malicious activity

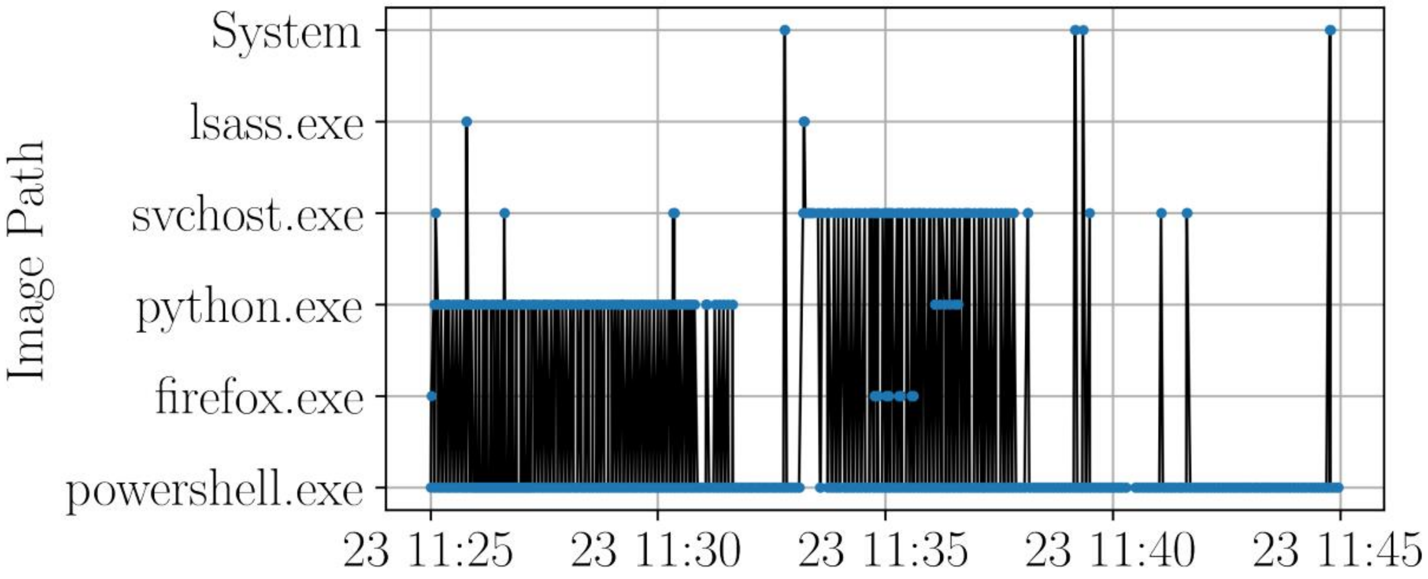
Benign



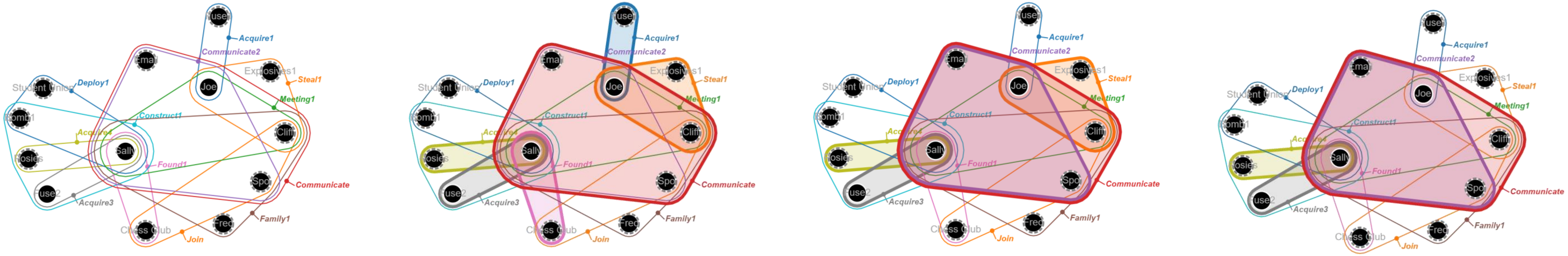


Malicious

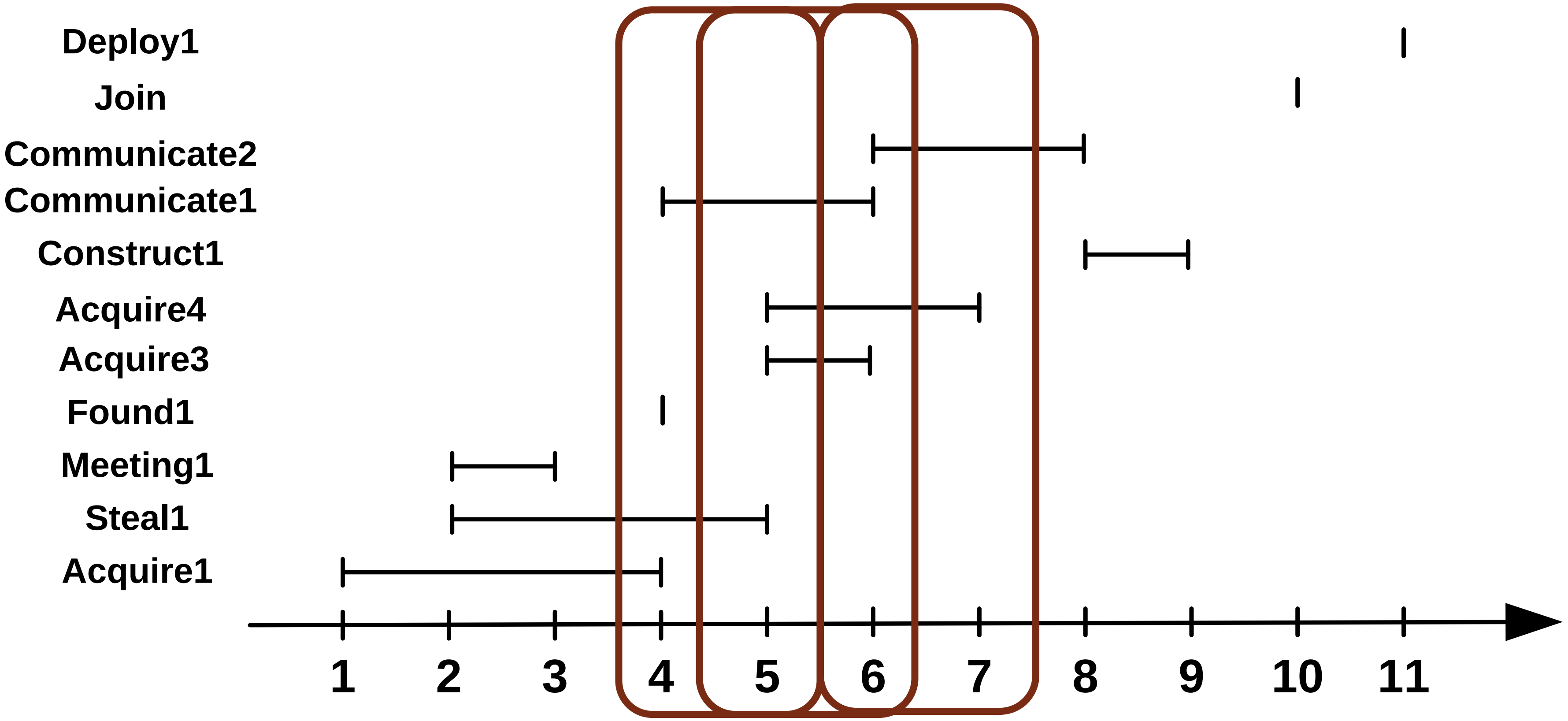




How can we track time evolving hypergraphs?

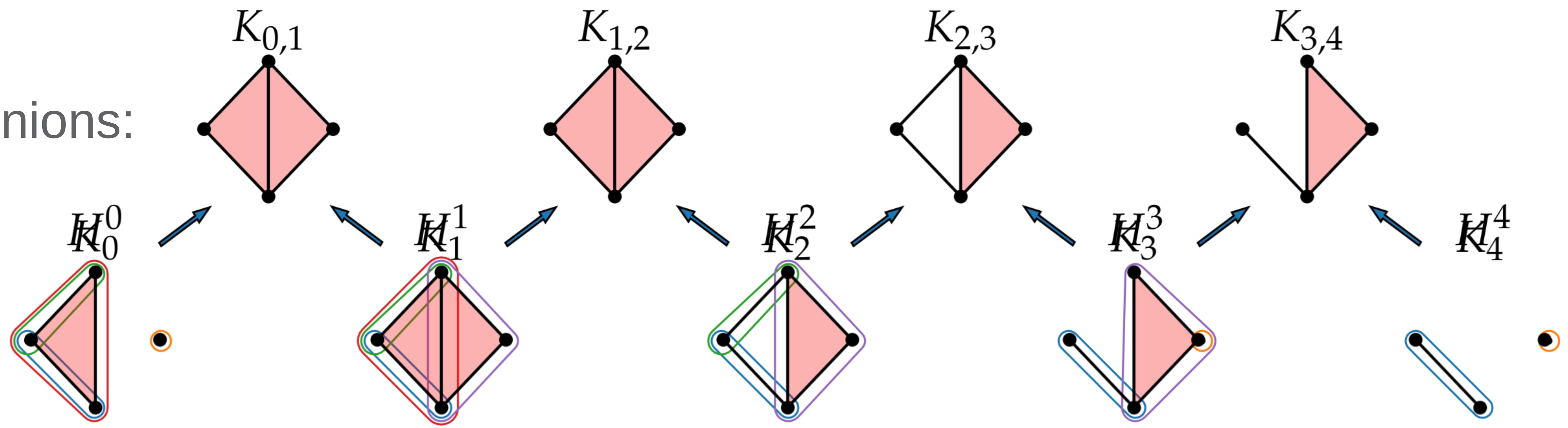


- Temporal hypergraph
- Trajectory of temporal sub-hypergraphs



Zigzag Persistence of Temporal Hypergraphs: *Associated Simplicial Complex*

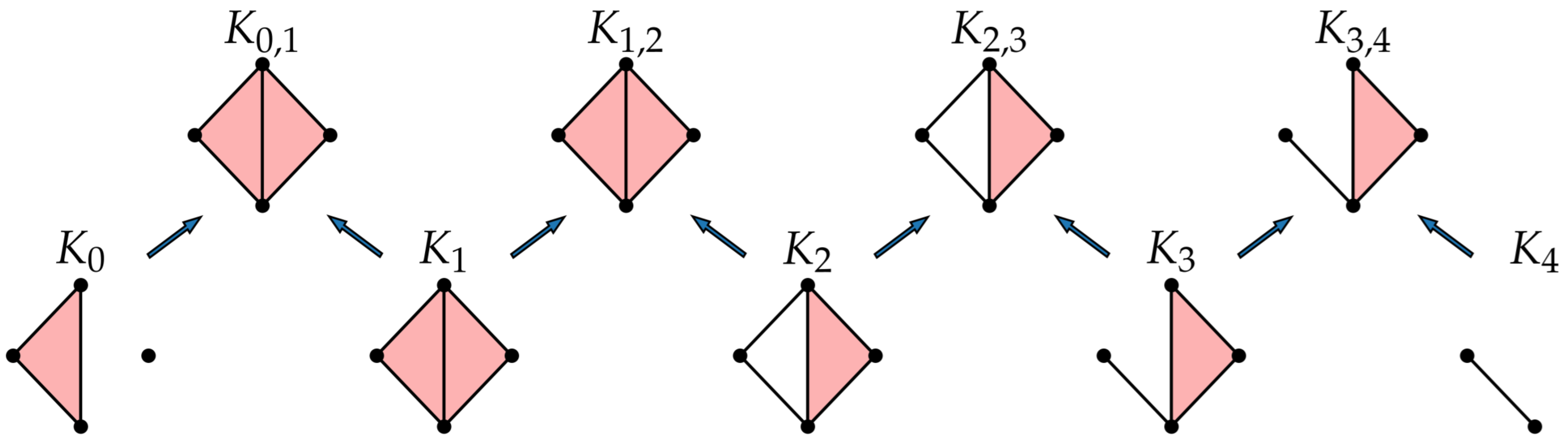
Unions:

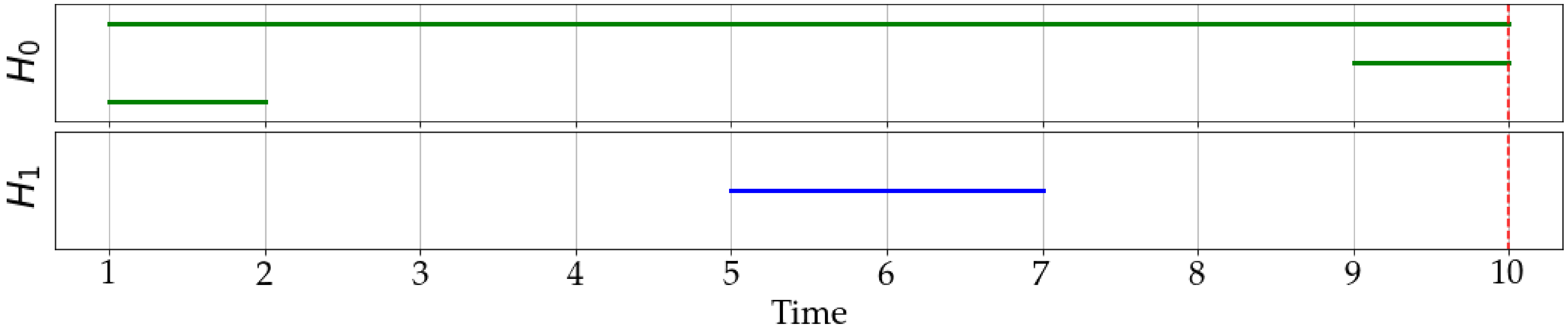


Associated Simplicial Complex:

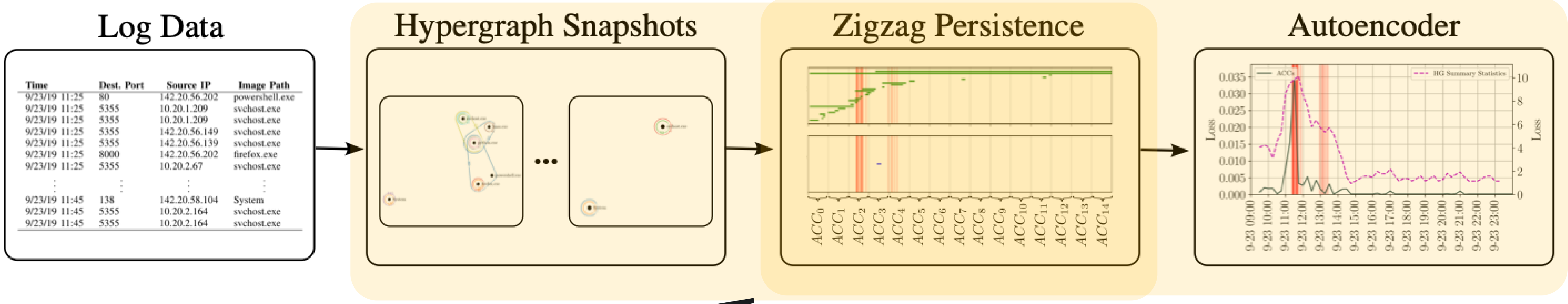
Ren, S. (2020). Persistent homology for hypergraphs and computational tools — A survey for users. In Journal of Knot Theory and Its Ramifications (Vol. 29, Issue 13, p. 2043007). World Scientific Pub Co Pte Lt.

Zigzag Persistence of Temporal Hypergraphs: *Associated Simplicial Complex*





Pacific Northwest NATIONAL LABORATORY



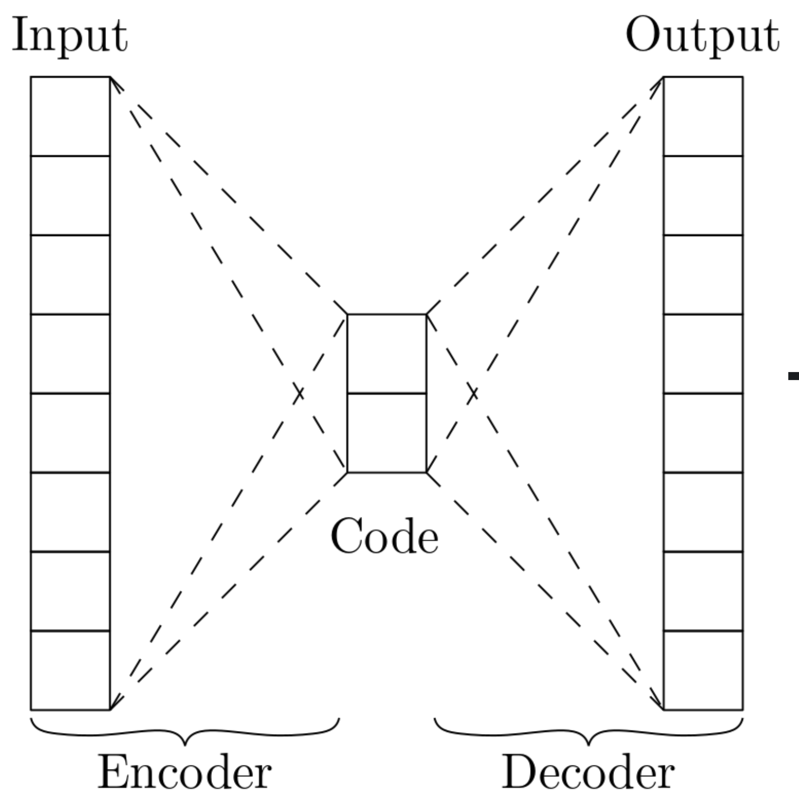
Benign (Training) Hosts:	0005, 0006, 0010, 0012, 0071, 0162 0213, 0222, 0274, 0304, 0461, 0906
Malicious (Testing) Hosts:	0201, 0402, 0660

Sub-window of Barcode

Vectorization: Adcock-Carlsson Coordinates

$$ACC(D_p) = \Big[\sum_i b_i(d_i - b_i), \\ \sum_i (d_{\max} - d_i)(d_i - b_i), \\ \sum_i b_i^2(d_i - b_i)^4, \\ \sum_i (d_{\max} - d_i)^2(d_i - b_i)^4\Big]$$

Autoencoder



Loss Between Input and Output

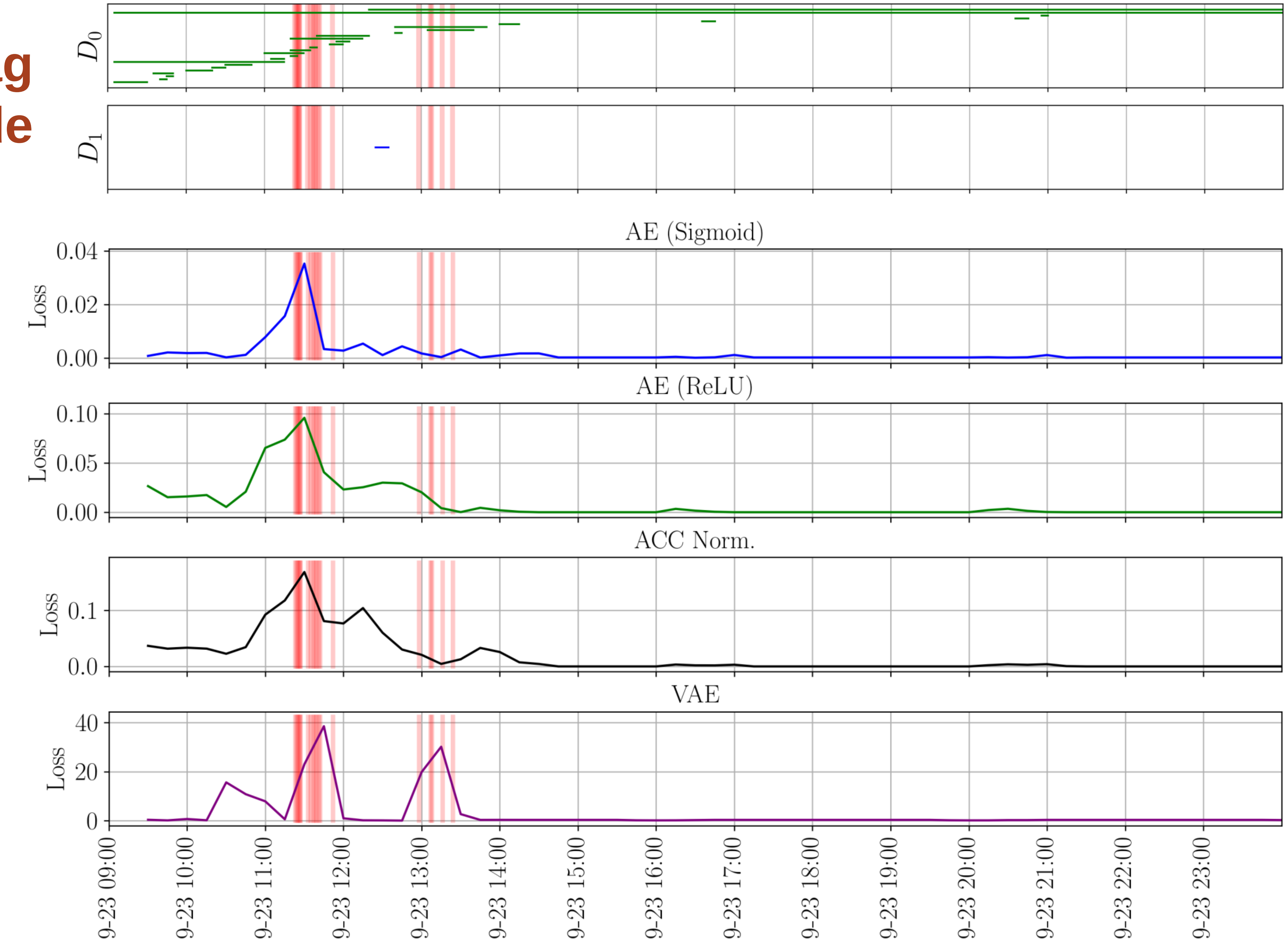
Results

Powershell Empire, Mimikatz, and injection attacks

Ping Sweep and ARP Scan attack

Zigzag barcode

Autoencoder Loss



Summary Statistics of Results

Host(s)	ACC ($\times 10^{-3}$)			Summary Statistics		
	25%	**50%**	**75%**	**25%**	**50%**	**75%**
201 (Benign IPs)	0.04	0.11	0.19	0.68	0.92	1.19
201 (Malicious IPs)	1.21	3.93	7.96	5.31	6.96	8.81
Training Hosts (24th)	0.07	0.14	0.26	0.76	1.03	1.34
Training Hosts (23rd)	0.06	0.12	0.26	0.72	1.01	1.39

- Median **ACC** loss ratio of malicious and benign is **35.7**
- Median **summary statistics** loss ratio of malicious and benign is **7.6**

Take-aways

- The sea of cyber log data allows for real time and forensic analysis by cyber defenders
- Adversaries are constantly innovating so we can’t just look for signatures of known behavior
- Mathematicians can lend a hand by providing insight into complex data structures, statistical anomaly detection, and principled decision making
  - Graphs
  - Hypergraphs
  - Topology
  - Machine learning

Pacific Northwest NATIONAL LABORATORY

Acknowledgements – this is a team effort!

- Sinan Aksoy
- Molly Baird
- Dan Best
- Alyson Bittner
- Paul Bruillard
- Clara Buck
- Gregory Henselman-Petrusek
- Helen Jenne
- Cliff Joslyn
- Bill Kay
- Audun Myers
- Katy Nowak
- Christopher Potvin
- Brenda Praggastis
- Garret Seppala
- Jackson Warley
- Stephen Young

This work was partially supported by the Topological Data Analysis for Cyber (TDAC) program at Pacific Northwest National Laboratory.

Pacific Northwest
NATIONAL LABORATORY

Thank you