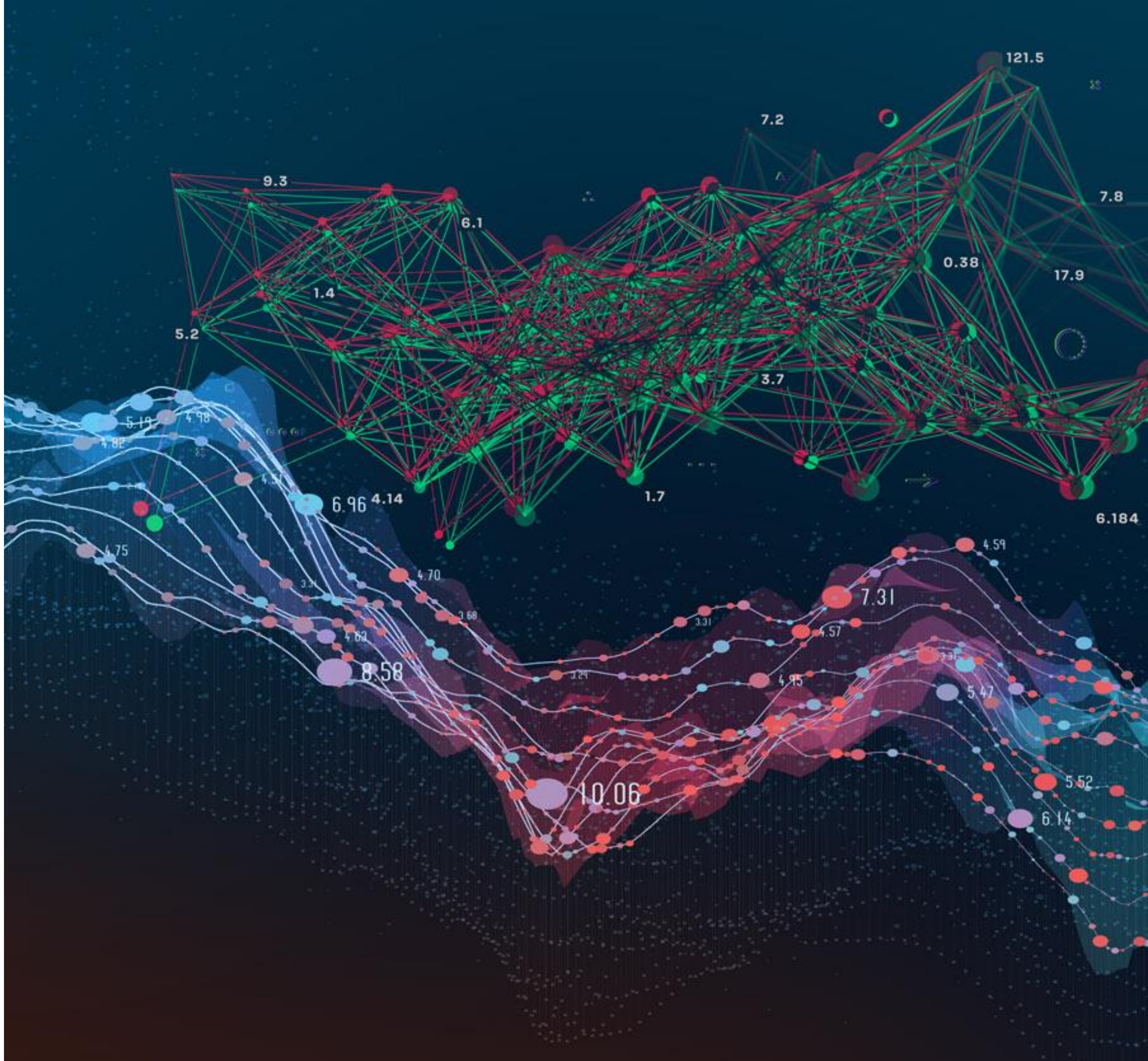
Mathematics for Cybersecurity

April 26, 2025
Metro NY MAA Section Meeting

Emilie Purvine
Chief Data Scientist

…and a ***ton*** of collaborators!

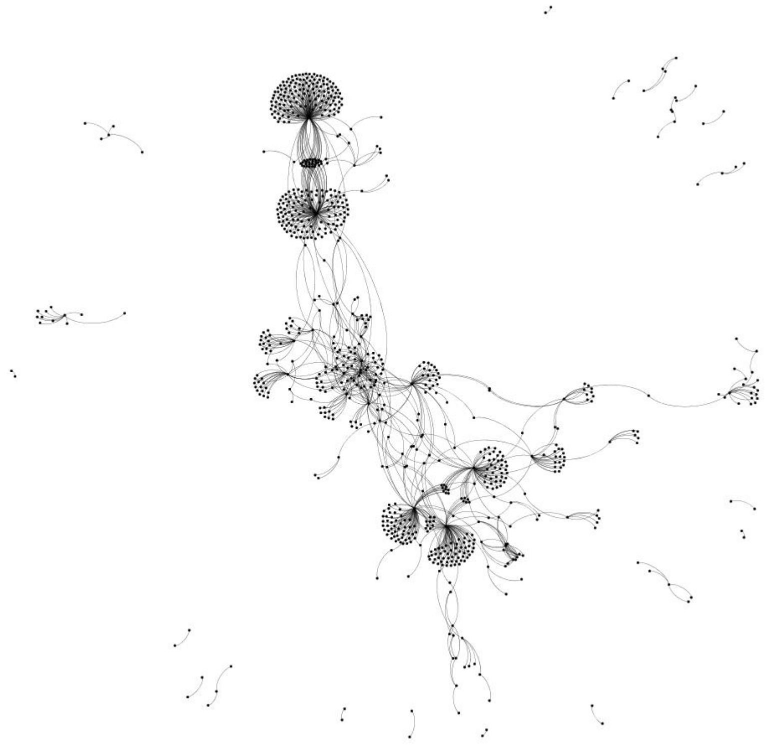
PNNL is operated by Battelle for the U.S. Department of Energy

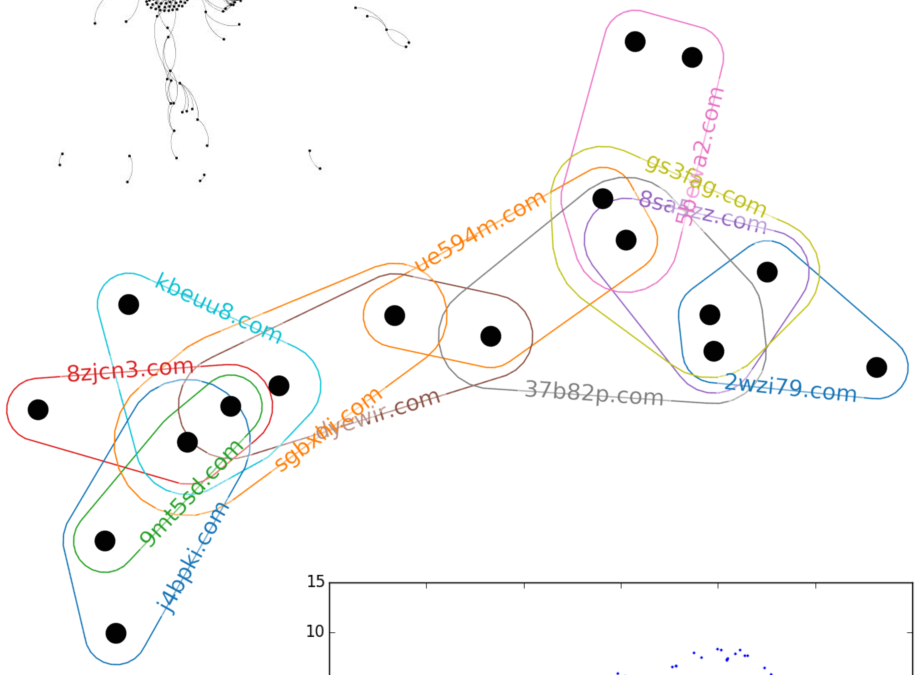


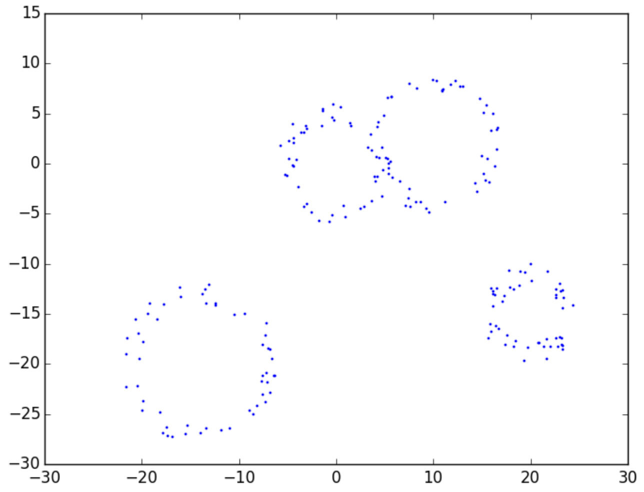
PNNL-SA-210124

Plan of the talk

- My path to a nonacademic career
- Cybersecurity 101 (accelerated version!)
- Graphs and hypergraphs via network flow
- Topology via high-dimensional data







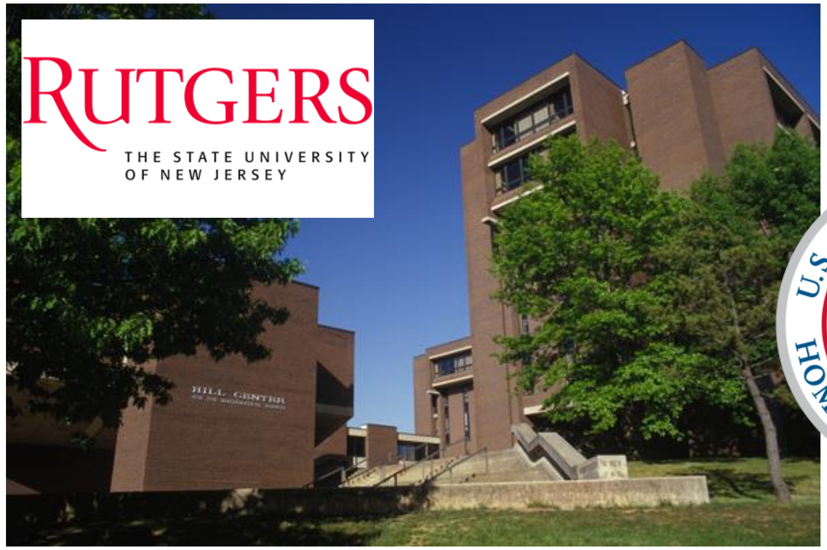
Pacific Northwest NATIONAL LABORATORY

My path to PNNL

- My path was quite linear. Perhaps too linear…
- Undergrad @ University of Wisconsin
  - Math education → Math
  - Summer math programs for women
  - Study abroad in math program
  - Undergrad research
  - Internship with small gov't contractor
- Grad @ Rutgers
  - Planned to NOT work in academia after graduation
  - Pure math, not applied. That was a choice.
  - Fellowship with DHS → internships at PNNL
- Postdoc @ PNNL started summer 2011

BSM
BUDAPEST SEMESTERS IN MATHEMATICS

Paul Heideman and I doing undergrad research at UW



U.S. DEPARTMENT OF HOMELAND SECURITY

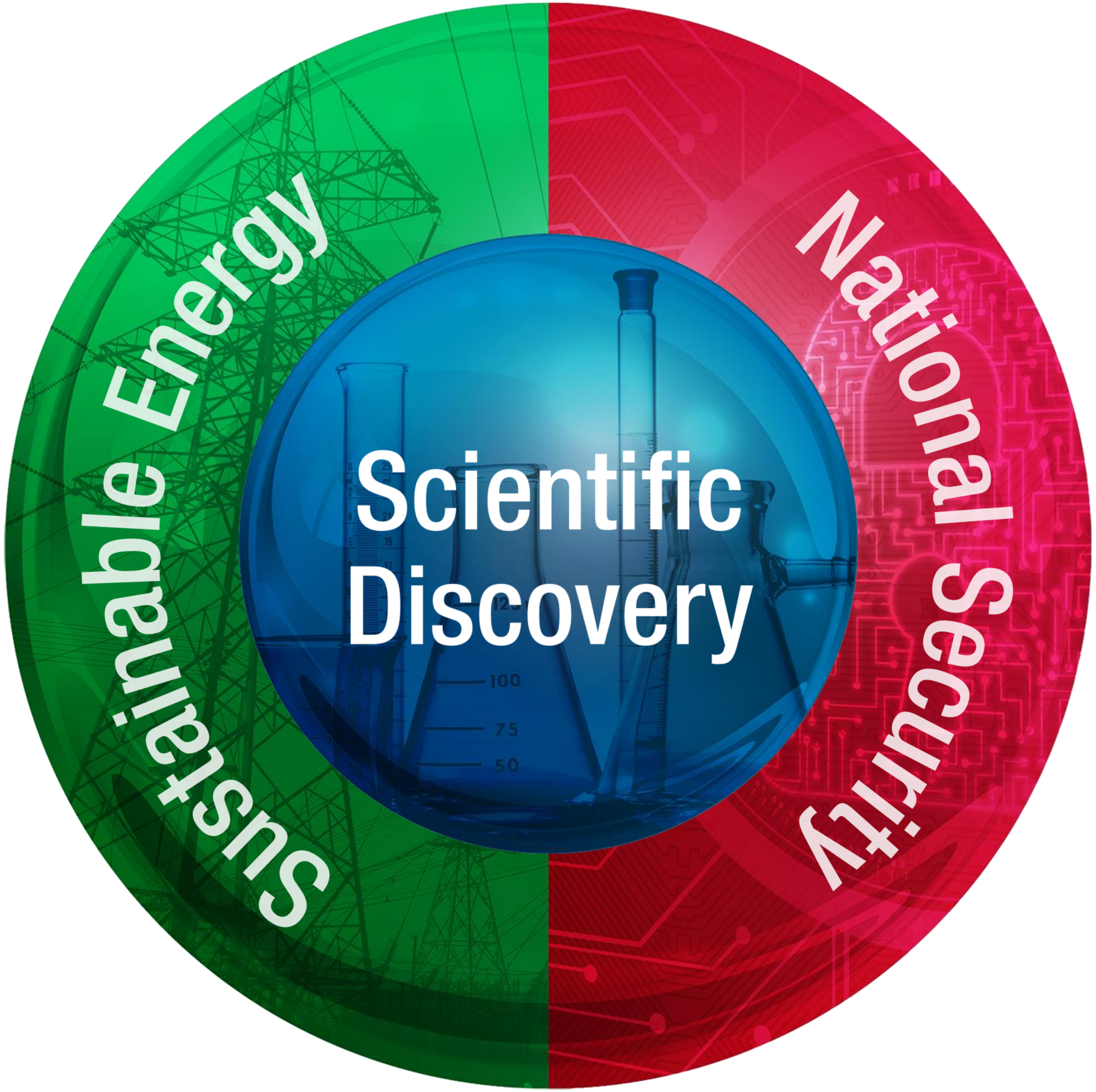
DIMACS
Center for Discrete Mathematics and Theoretical Computer Science
Founded as a National Science Foundation Science and Technology Center

Pacific Northwest
NATIONAL LABORATORY

DOE's 17 **national laboratories** tackle critical scientific challenges

Pacific Northwest
NATIONAL LABORATORY

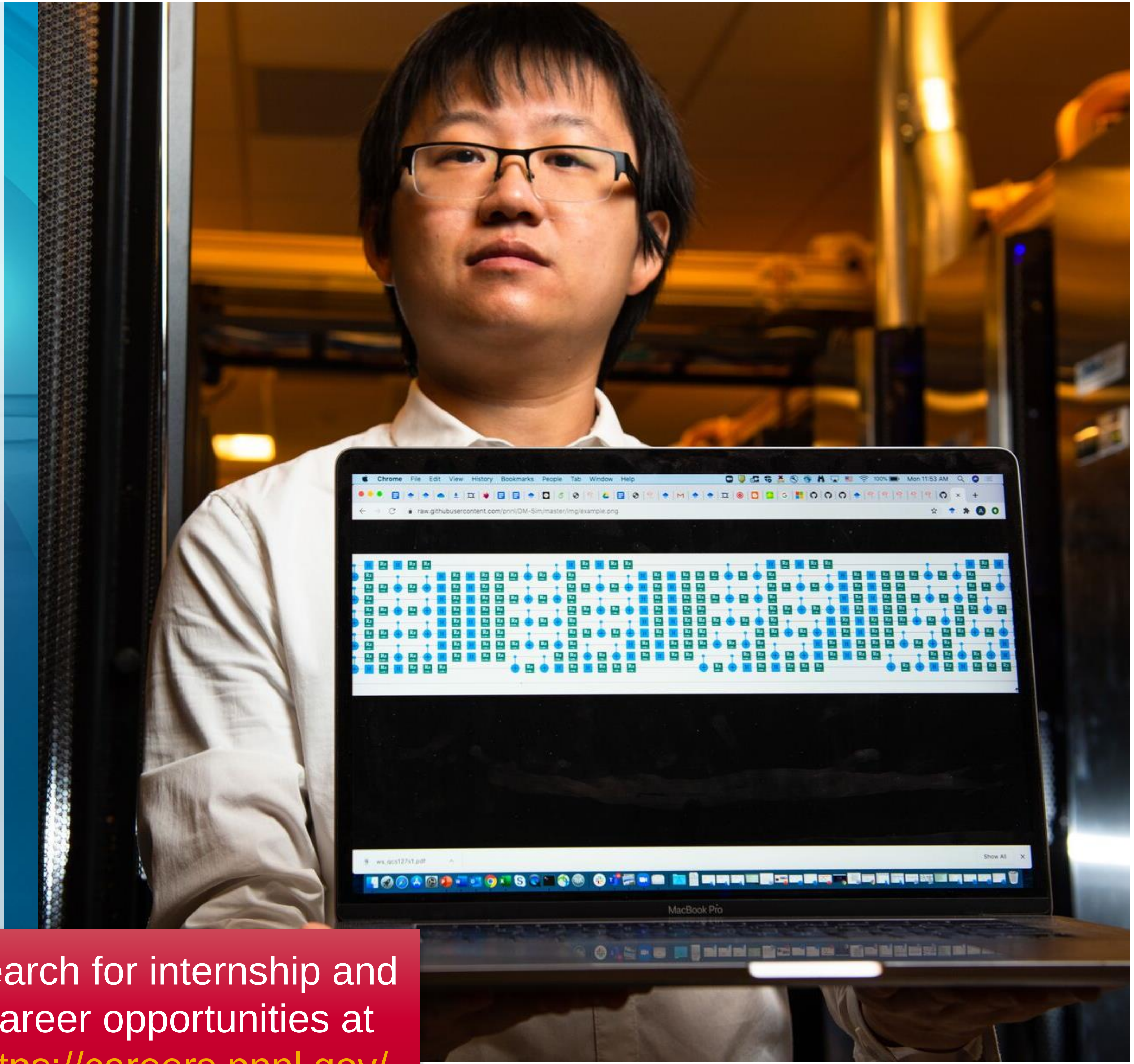
PNNL is **advancing scientific frontiers** and **providing solutions** to critical national needs



Scientific Discovery

DATA SCIENCE

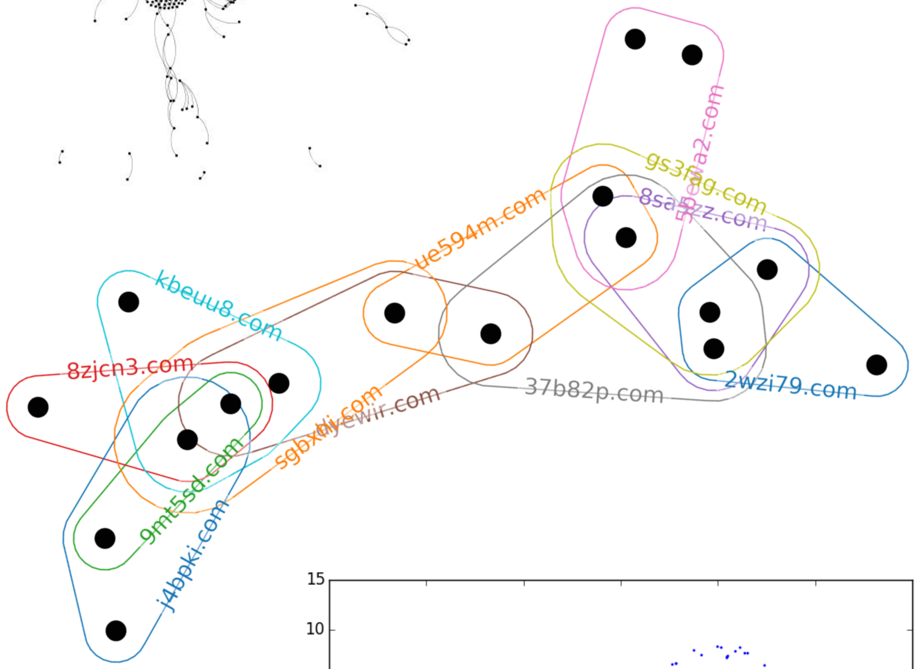
- Join **extreme scale computing** and **big data**
- Deliver advanced **visualization** technologies and novel **algorithms**
- Apply **artificial intelligence** and **machine learning** to complex computational problems

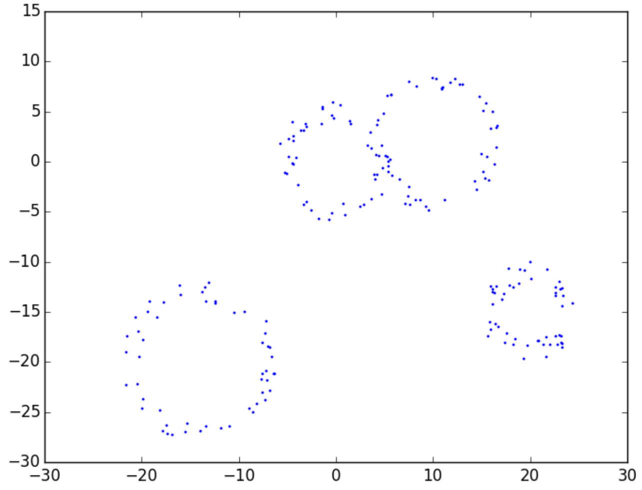


Search for internship and career opportunities at https://careers.pnnl.gov/

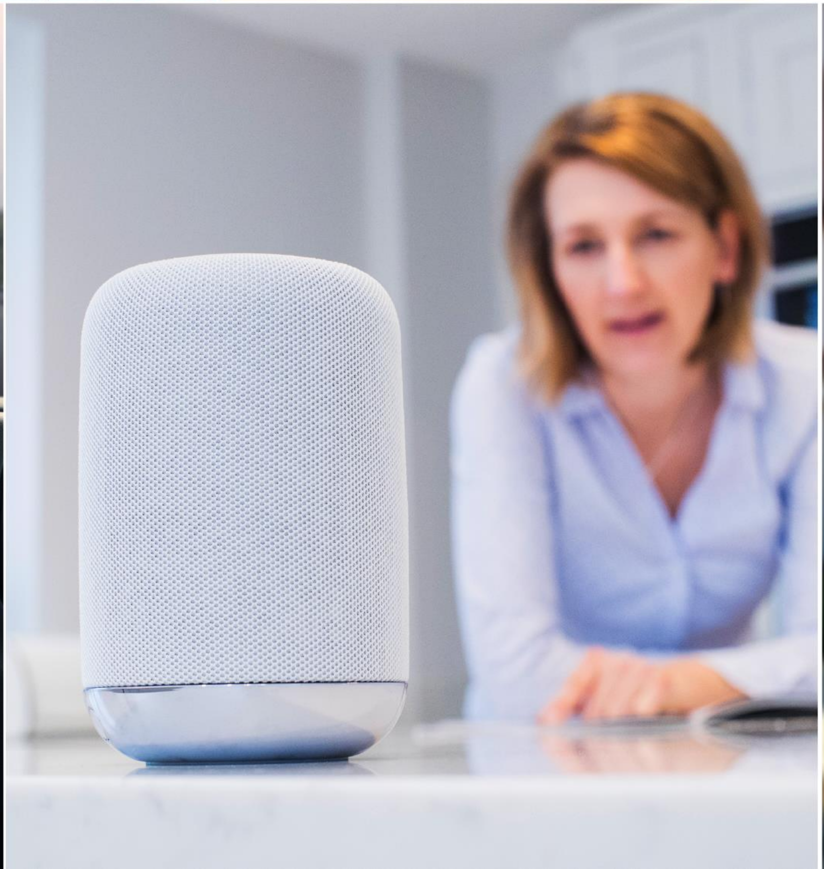
Plan of the talk

- My path to a nonacademic career
- Cybersecurity 101 (accelerated version!)
- Graphs and hypergraphs via network flow
- Topology via high-dimensional data

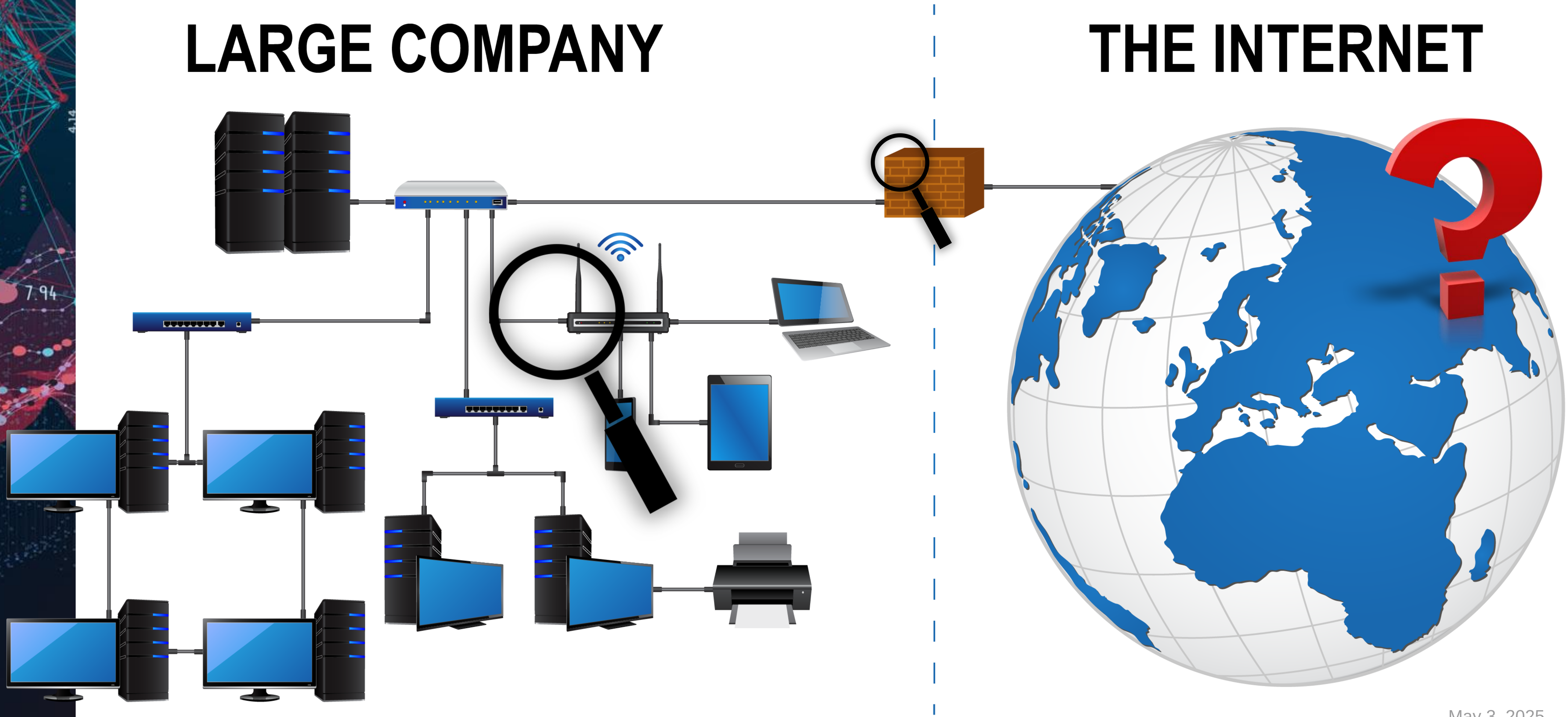




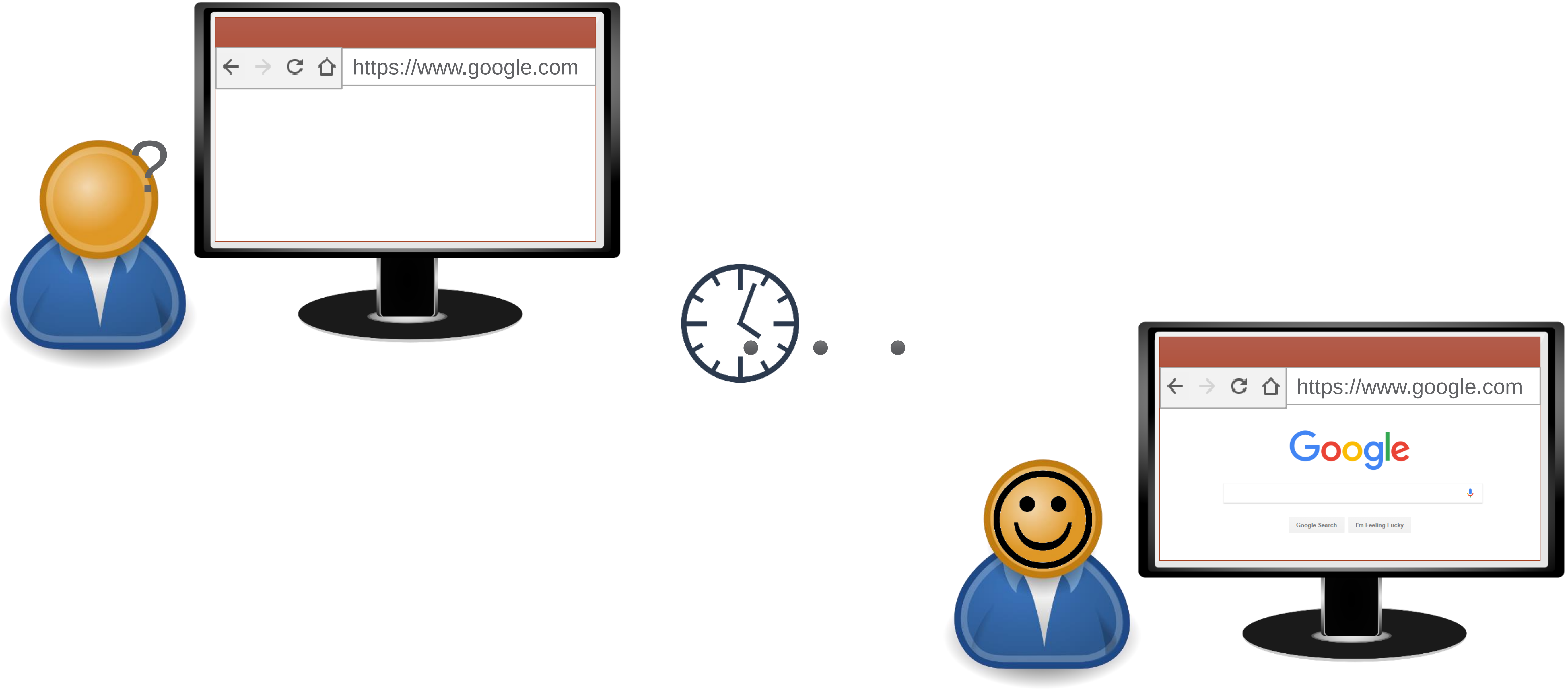
Internet of things – how many do you have?



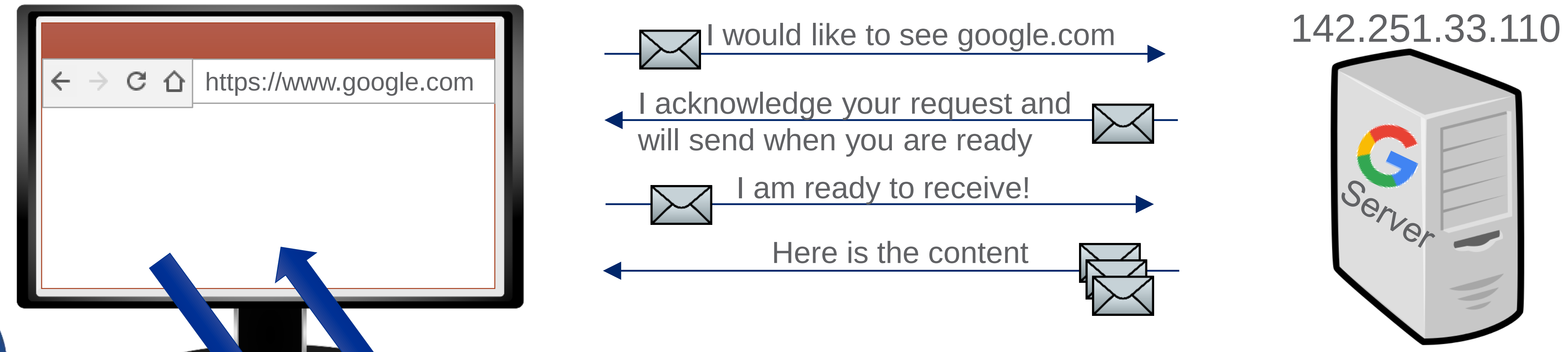
Where can defenders "see"?

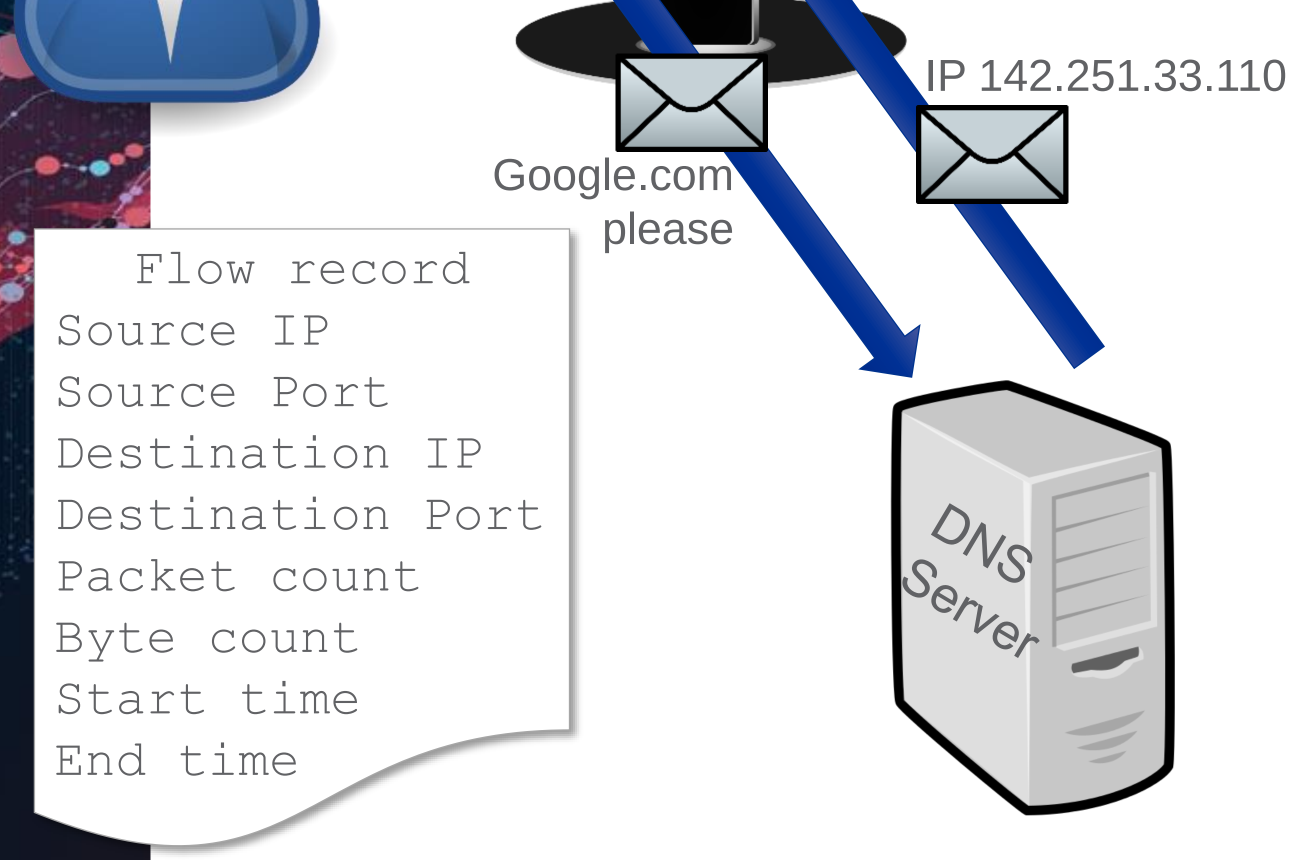


Internet communication – what you experience



Internet communication – behind the scenes



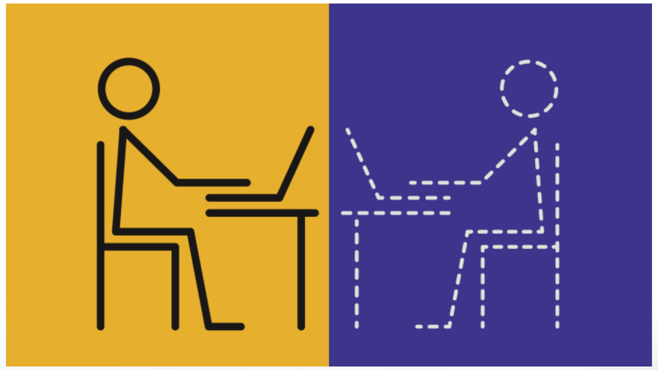


- First your browser must find the IP address for google.com via a DNS server
- Then your browser establishes a connection with the google.com server via a TCP 3-way handshake
- Each message is broken up into potentially multiple packets and reassembled at the destination
- Packets can be aggregated into conversations called flow (e.g., IPFlow, NetFlow) ***DATA!!***

Pacific Northwest
NATIONAL LABORATORY

A Sea of Data

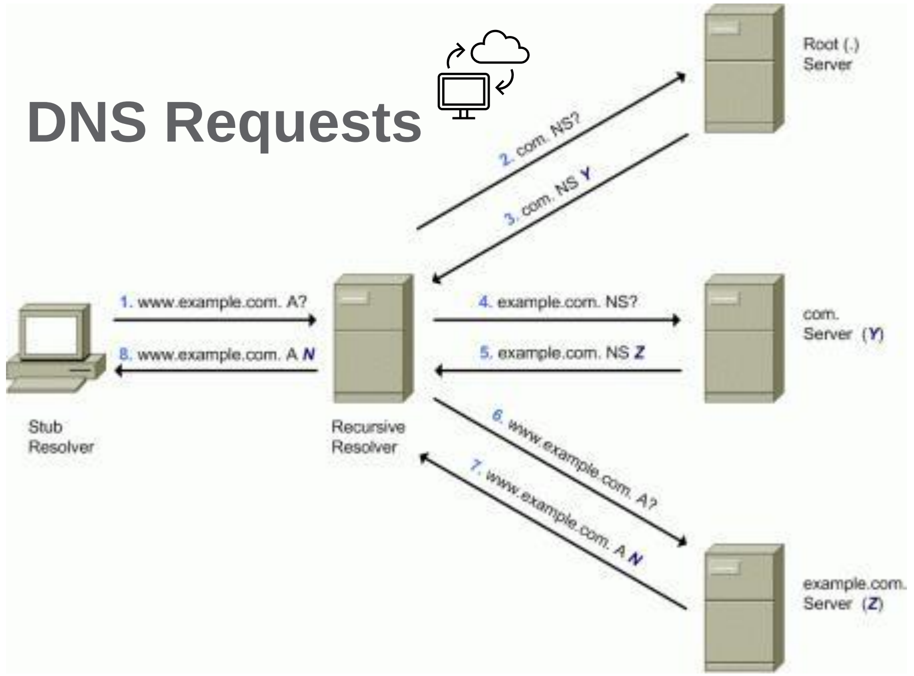
Authentication: SSH, Kerberos, …



Flow record

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

DNS Requests



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
```

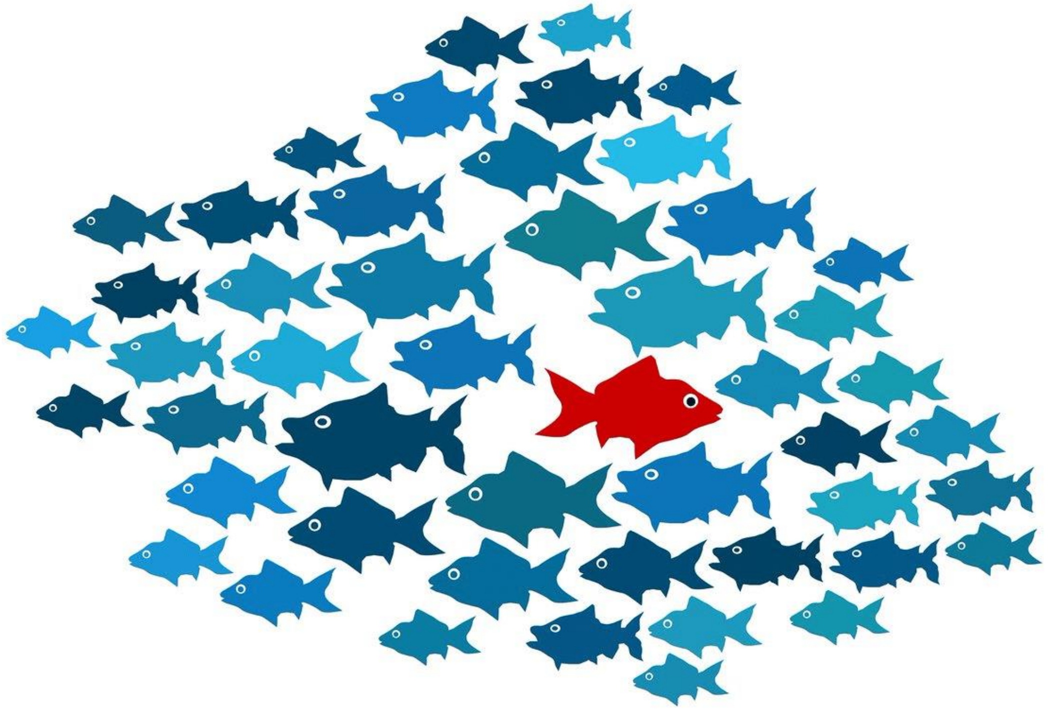
Aligning data with the cyber kill chain

- 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 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)



Pacific Northwest
NATIONAL LABORATORY

"OODA loop" – where can mathematicians fit?

Observe
- Collect data

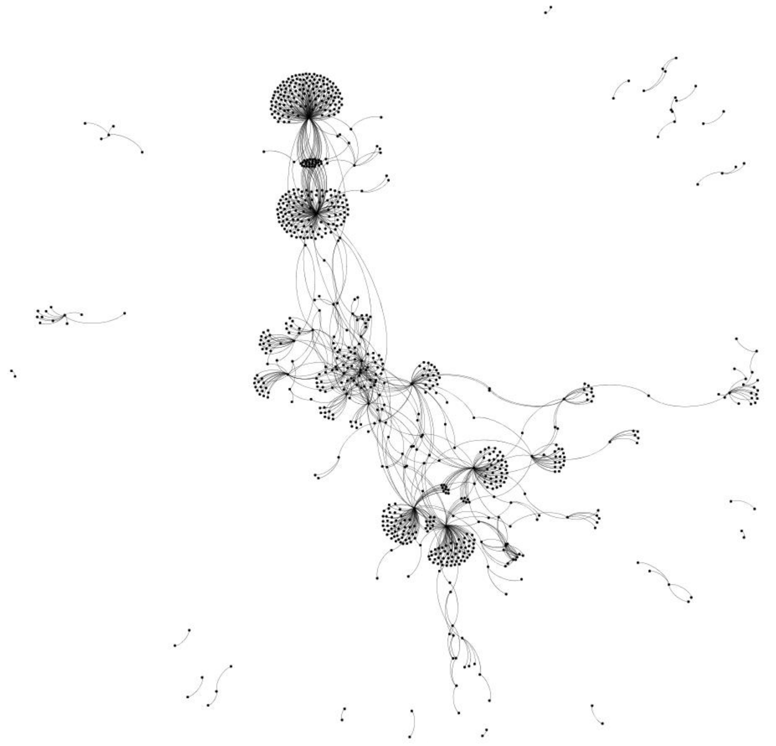
Orient
- Develop situational awareness
- Run "analytics" on collected data

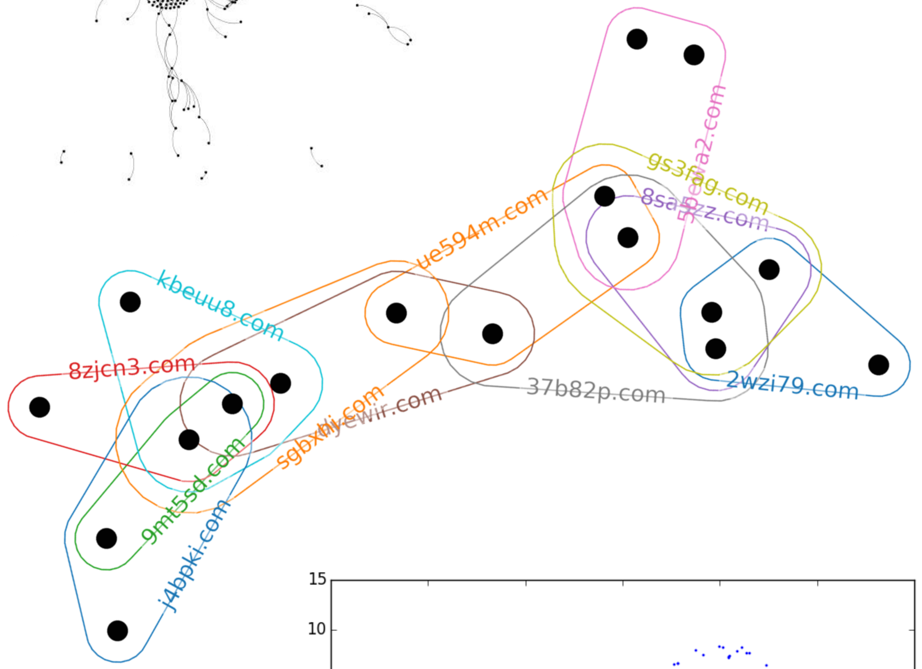
Decide
- What should we do based on observed data
- E.g., patch, segment, target additional data collection

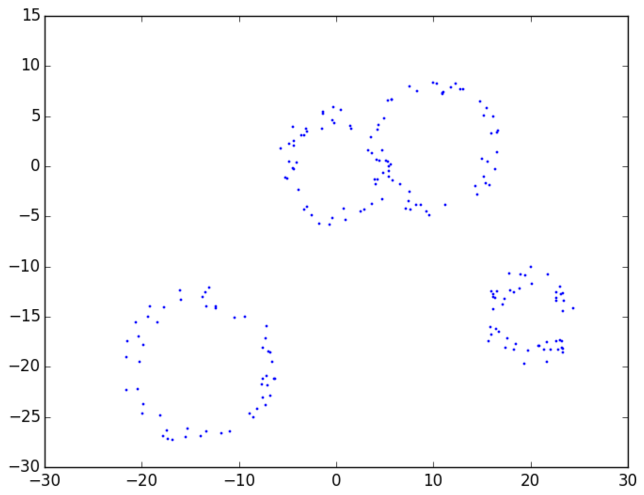
Act
- Carry out the actions

Plan of the talk

- My path to a nonacademic career
- Cybersecurity 101 (accelerated version!)
- Graphs and hypergraphs via network flow
- Topology via high-dimensional data







Application #1: Host and network data

time	action-object	host	principal	pid	source IP	dest IP	dest port	protocol	image path
9/24 10:45:00	MESSAGE-FLOW	SysClient0501	bantonio	2192	10.20.5.191	10.20.2.66	5999	UDP	python.exe
9/24 10:45:02	START-FLOW	SysClient0501	bantonio	836	132.197.158.98	202.6.172.98	80	TCP	powershell.exe
9/24 10:45:25	MESSAGE-FLOW	SysClient0501	bantonio	5100	142.20.57.246	142.20.61.132	80	TCP	outlook.exe
9/24 10:45:29	START-FLOW	SysClient0501	bantonio	648	142.20.57.246	202.6.172.98	443	TCP	powershell.exe

- Snapshot of data from Operationally Transparent Cyber (OpTC) data set which includes both *host* and *network* events
  - Network data: communications between computers (recorded as “IP addresses”). Records the two computers and metadata about the connection. (See table above.)
  - Host data: processes occurring on individual computers.
- **Questions:** What connection patterns exist? How do they change over time? Can we find unusual patterns or connections? What do they mean?

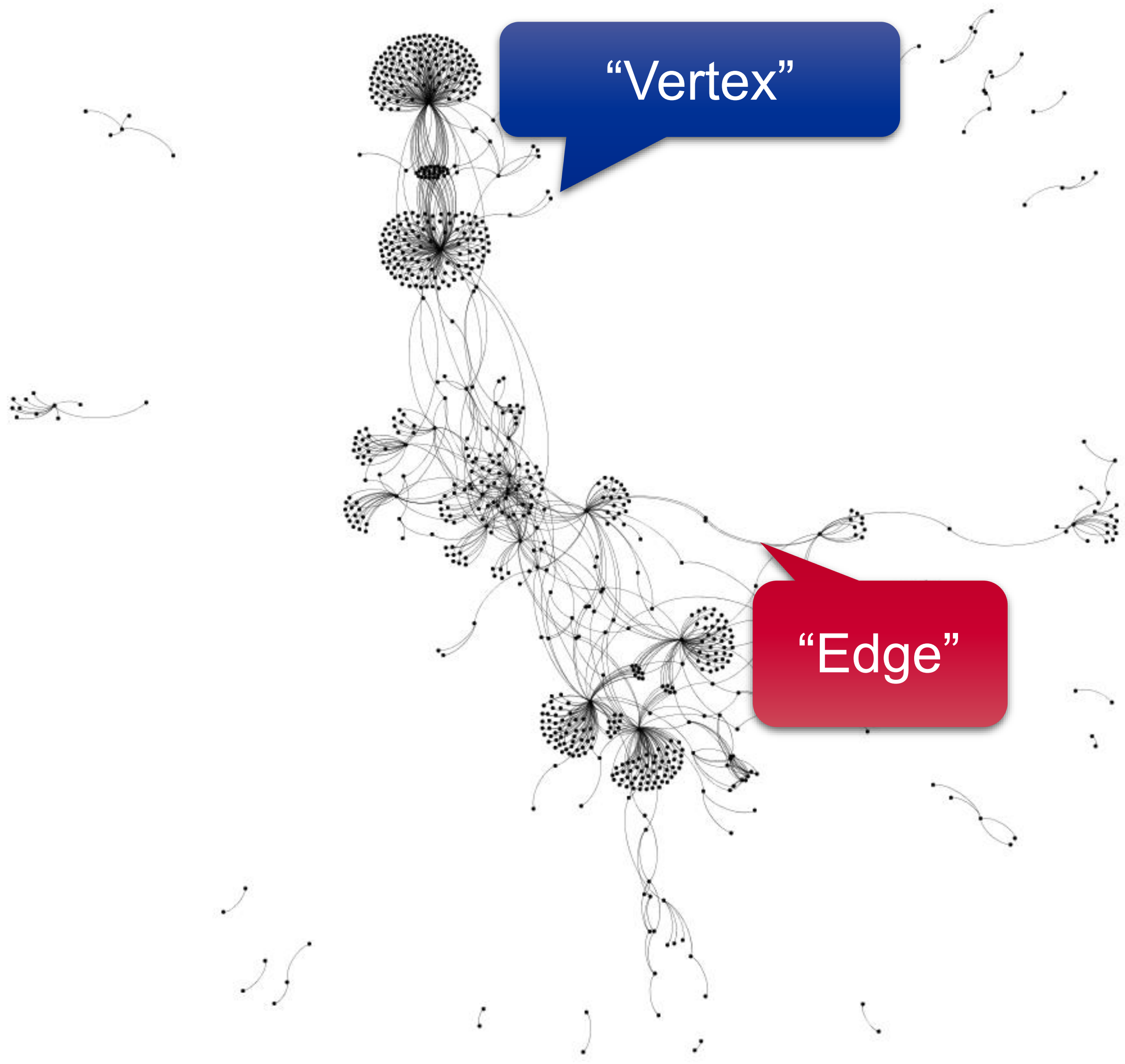
https://github.com/FiveDirections/OpTC-data/

Mathematical model for communications: Graph

- **Graphs provide a mathematical model of data focused on 2-way relationships**
  - To ***ask*** certain kinds of questions
    - ✓Connectivity of entities
    - ✓Clustering structure
  - To ***model*** certain kinds of interactions
    - ✓Pairwise relationships

$$G = (V, E), E \subseteq \binom{V}{2}$$

  - Network flow graph:
    - ✓ V = IP addresses / hosts
    - ✓ E = communications



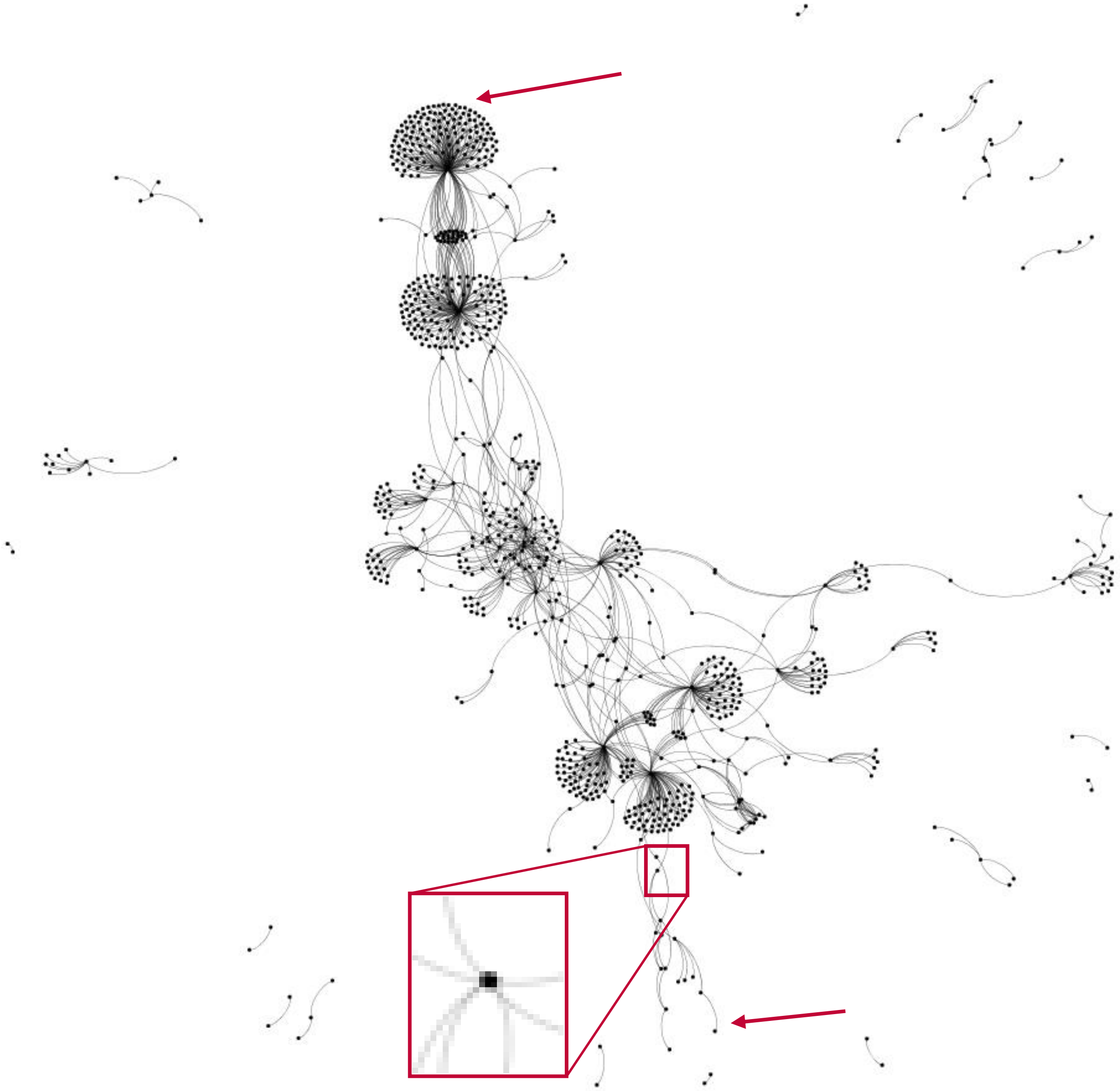
Data from http://csr.lanl.gov/data/cyber1/

2 minutes of flow in LANL system: |V|=842, |E|=1038

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

Graph properties

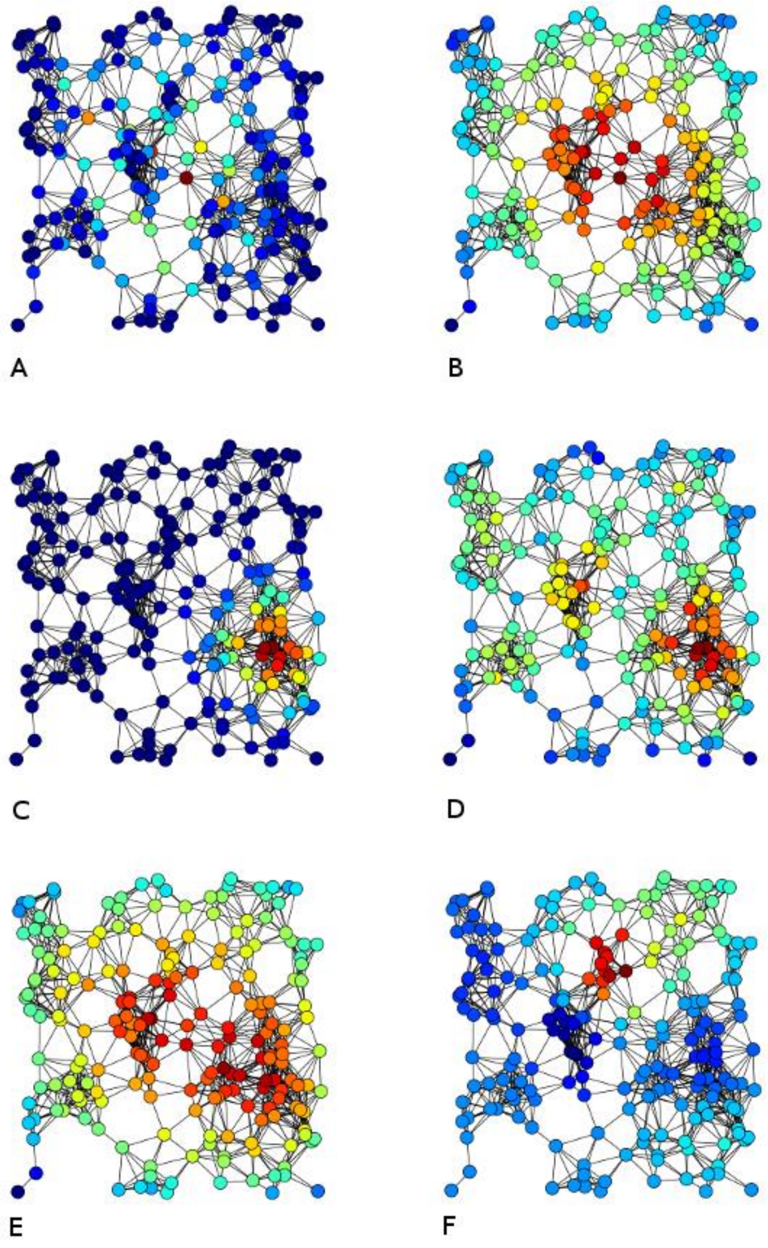
- Degree (distribution)
- Walk, Path, Diameter
- Connected components
- Centrality
- Clustering coefficient
- Triangle counting
- …



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

Graph properties

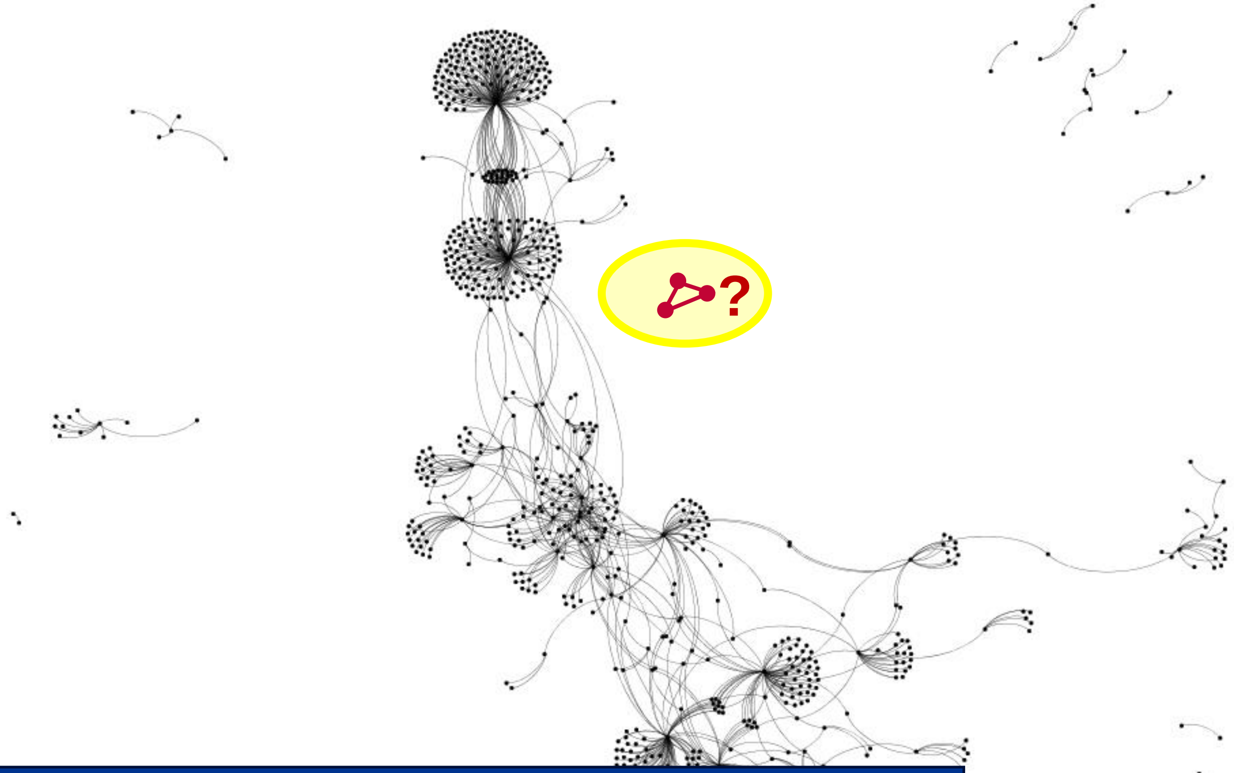
- Degree (distribution)
- Walk, Path, Diameter
- Connected components
- Centrality – measured for each vertex
  - Betweenness: measure of belonging to shortest paths
  - Closeness: measure of average distance to other vertices
  - Eigenvector: Solution to $Ax = \lambda x$
  - Degree: degree of vertex
  - 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
- …*



Recall our questions: What connection patterns exist? How do they change over time? Can we find unusual patterns or connections? What do they mean?

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

Generative graph models – what and why

- What are they?
  - Models that create graphs possessing properties we are interested in
- What are they good for?
  - Null model for algorithm testing and experiments
  - Create synthetic graphs on different scales
  - Create surrogate graph to protect anonymity of data
  - Graph generation process may give insight into properties being matched

Our focus

- What makes them good?
  - Inputs are compact & few
  - Easily measured from real data or generated artificially
  - Generalized and formalized generation process
  - Avoid ad hoc methods/restrictive assumptions on inputs

Classic simple generative models

Erdős-Rényi

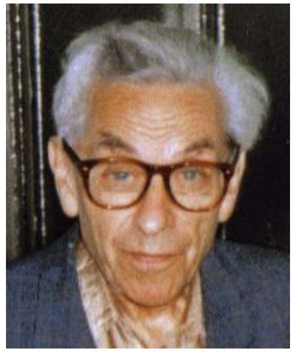
- **Matches:** average degree / density
- **Inputs:** number of vertices (n), edge probability (p)

u ? v

$P(u,v) = p$

All edges independent

- Connected if $p > \log(n)/n$



Paul Erdős

Alfred Rényi

Chung-Lu

- **Matches:** degree distribution
- **Inputs:** degree sequence $\{d_i\}$ where d_i is desired degree of v_i

v_i ? v_j

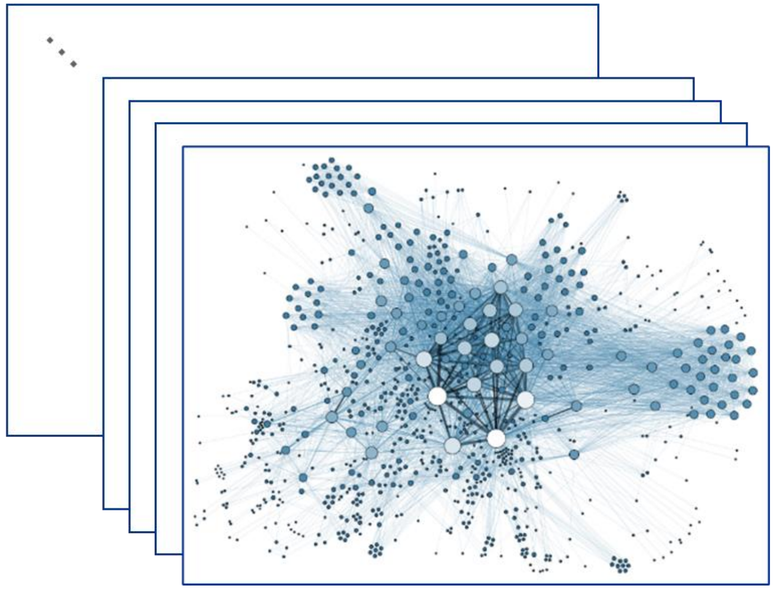
$$P(v_i, v_j) = \frac{d_i d_j}{\sum d_k}$$

- Typically small-world

Fan Chung

Linyuan Lu

Dynamic graphs – background



- **Graph:** $G = (V, E)$ with vertices V and edges E
  Dynamic graph: $\{G_t\}_{t \in T}$ where $G_t = (V_t, E_t)$ is a graph and T is a set of times
- Dynamic graph considered as a 3-tensor with dimensions $T \times V \times V$
  - Entry at index (t, v, w) if $(v, w) \in E_t$
- (Static) Random graph models often used as null models – Erdos-Renyi, Chung-Lu, other specialized models
- Random *dynamic* graphs:
  - Dynamic Erdos-Renyi [1] – missing edges appear with probability α, existing edges disappear with probability β
  - Dynamic Chung-Lu [1] – Poisson process, edges added at rate λ, removed at rate μ
  - Dynamic block model [1] – Poisson process, rate of addition and removal of edges depends on group membership
  - Additional survey of methods [2]

1 Xiao Zhang, Cristopher Moore, and Mark EJ Newman. "Random graph models for dynamic networks." *The European Physical Journal B* 90.10 (2017): 200.
2 Holme, Petter, and Jari Saramäki. "Temporal networks." Physics reports 519.3 (2012): 97-125.

Hagberg-Lemons-Misra (HLM)[3] Model

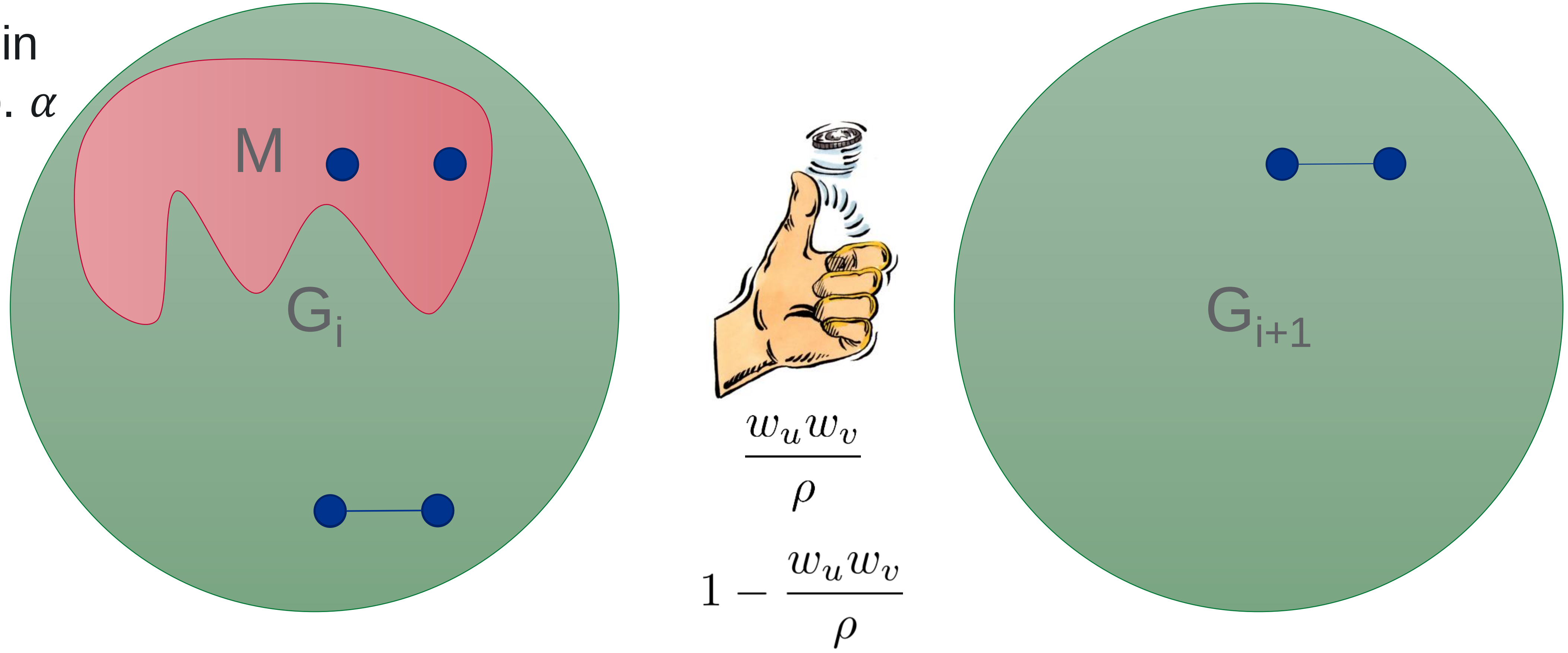
[3] A. Hagberg, N. Lemons, and S. Misra, Temporal reachability in dynamic networks, in Dynamic Networks and Cyber-Security, WORLD SCIENTIFIC (EUROPE), Mar 2016, pp. 181–208.

Desired degree sequence: $w_1, w_2, \ldots, w_n$
G_1 is Chung-Lu graph with this deg. sequence

Parameter $\alpha \in [0,1]$ controls extent to which G_{i+1} depends on G_i

- HLM was created to mimic network traffic
- Each G_i is Chung-Lu
- Cannot capture density changes over time.

Pairs (u, v) in M with prob. α



Note: can generalize to arbitrary edge probability matrix P and alphas for each edge

Pacific Northwest
NATIONAL LABORATORY

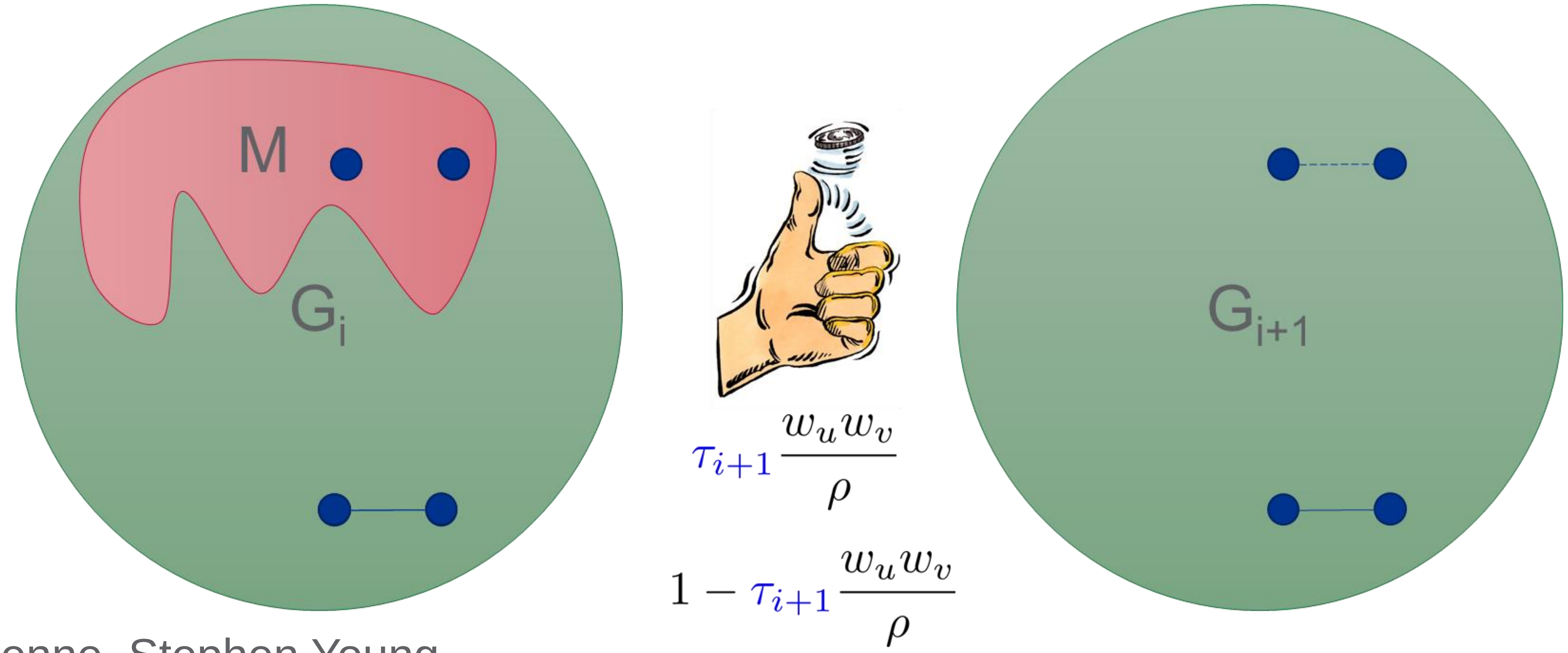
Temporal HLM = THeLMa Model

Desired degree sequence: $w_1, w_2, \ldots, w_n$
Temporal parameter: $\tau_1, \tau_2, \ldots, \tau_n$
G_1 is Chung-Lu graph with this deg. sequence times τ_1

- THeLMa: Include density parameter in evolution
- G_i no longer Chung-Lu because of presence of τ parameters

Parameter $\alpha \in [0,1]$ controls extent to which G_{i+1} depends on G_i

Pairs (u,v) in M with prob. α



Joint with Sinan Aksoy, Helen Jenne, Stephen Young

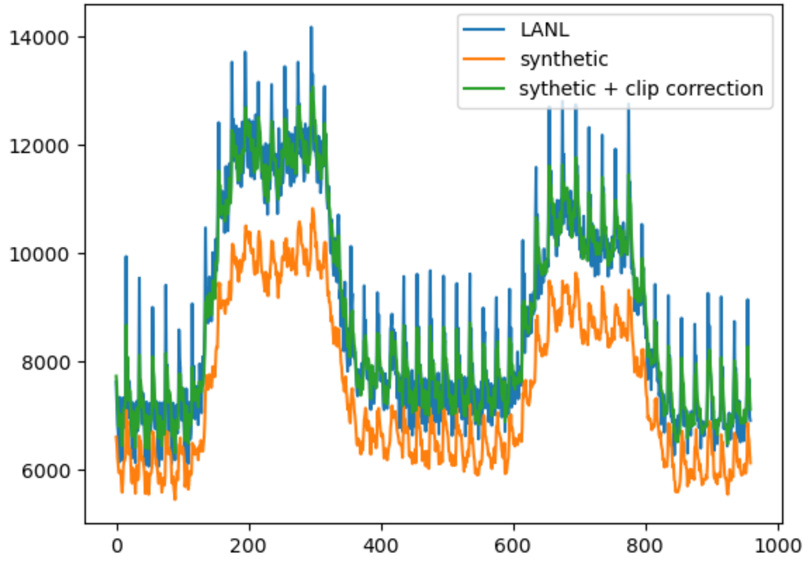
THeLMa as a flexible network baseline model

- **Assumption:** anomalies are sparse in network data
- *Measure* simple parameters from observed data
  - Average degree sequence across time = average degree for each vertex
  - τ = Number of edges for each time step
  - α… it's not so simple, but it's possible (MLE estimator) [paper in progress]
- *Generate* a THeLMa sequence using the measured parameters
- Use the generated dynamic graph as a baseline for:
  - Anomaly detection
  - Background identification/subtraction
  - Anomaly injection for algorithm testing

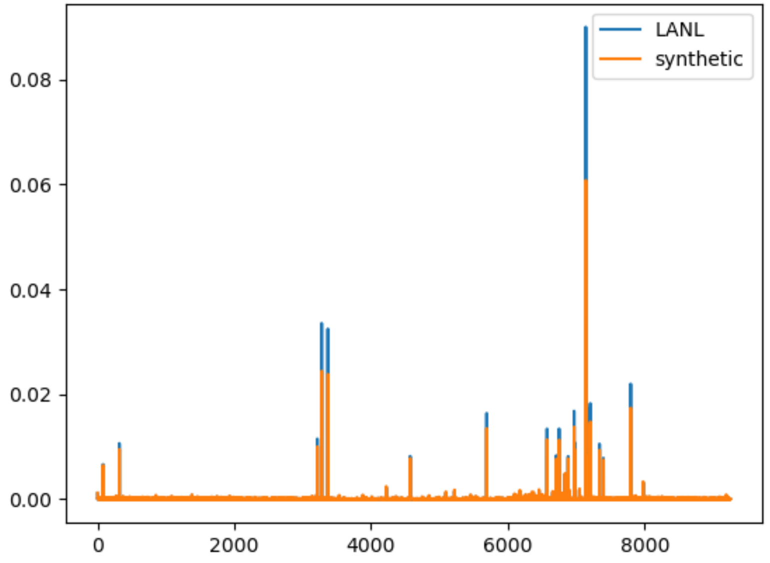
Case study: synthetic LANL data

Two days of network flow[4] in 3-minute time windows

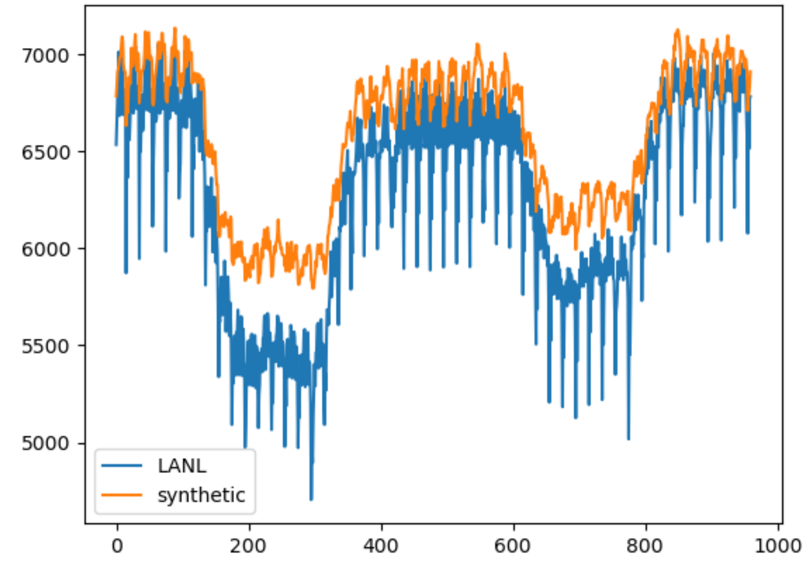
Number of edges (tau)



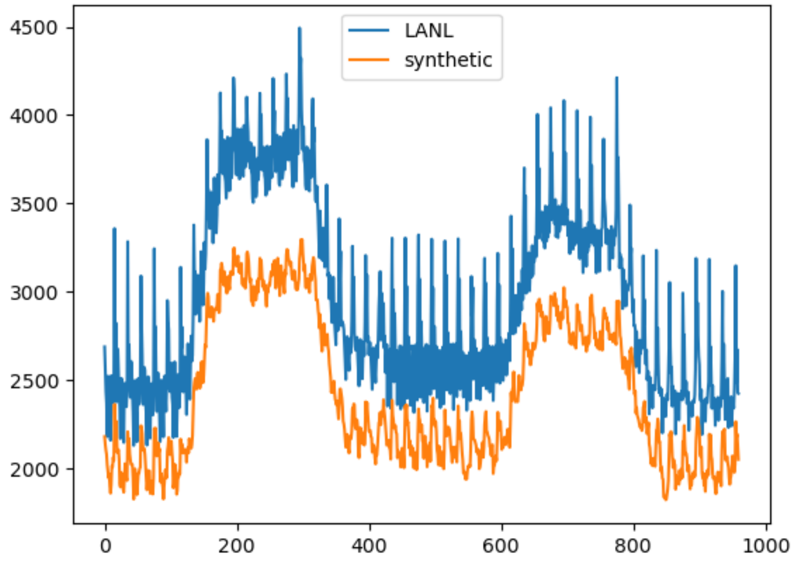
Normalized degree sequence (W)



Number of components



Largest component

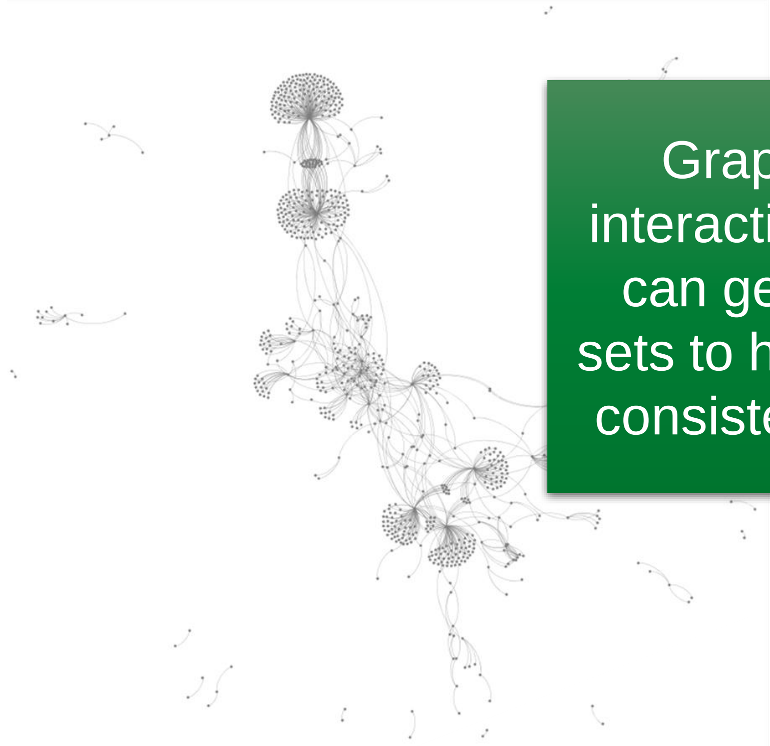


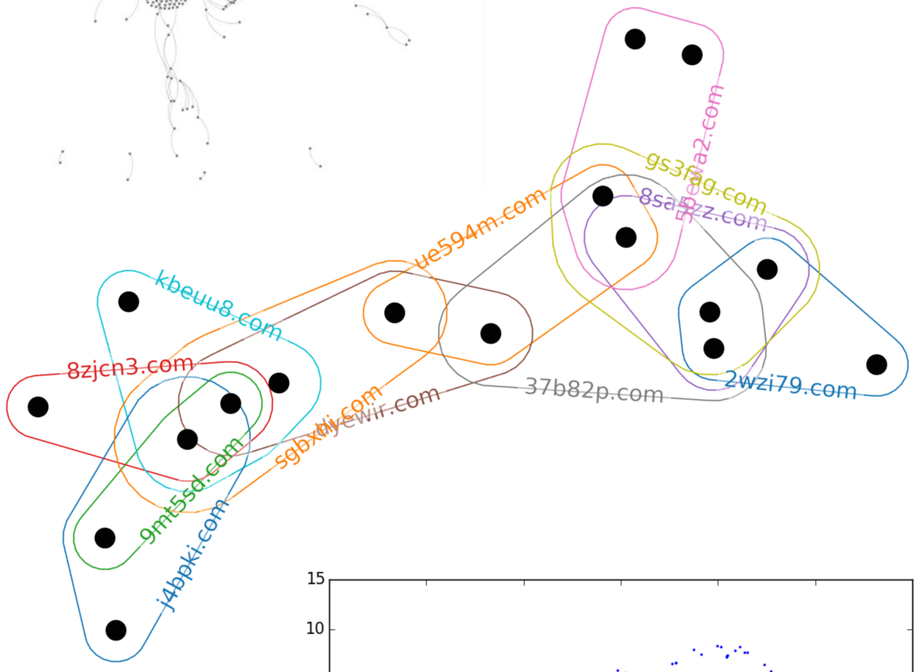
[4] https://csr.lanl.gov/data/cyber1/

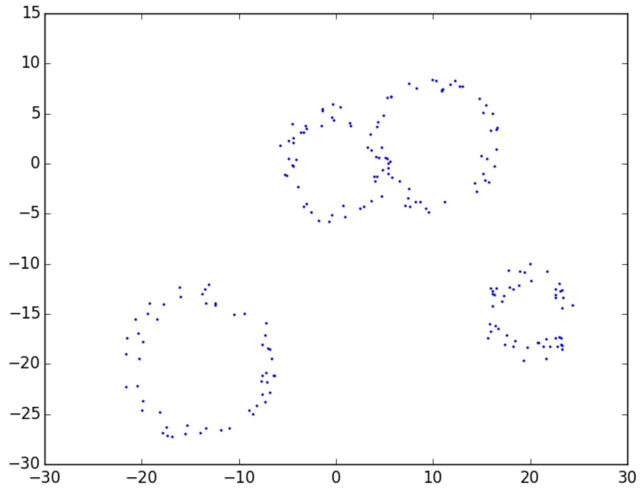
Plan of the talk

- My path to a nonacademic career
- Cybersecurity 101 (accelerated version!)
- Graphs and hypergraphs via network flow
- Topology via high-dimensional data

Graphs model pairwise interactions. Random models can generate realistic data sets to help explore properties consistent with real systems.







Graph structure is only part of the story…

time	action-object	host	principal	pid	source IP	dest IP	dest port	protocol	image path
9/24 10:45:00	MESSAGE-FLOW	SysClient0501	bantonio	2192	10.20.5.191	10.20.2.66	5999	UDP	python.exe
9/24 10:45:02	START-FLOW	SysClient0501	bantonio	836	132.197.158.98	202.6.172.98	80	TCP	powershell.exe
9/24 10:45:25	MESSAGE-FLOW	SysClient0501	bantonio	5100	142.20.57.246	142.20.61.132	80	TCP	outlook.exe
9/24 10:45:29	START-FLOW	SysClient0501	bantonio	648	142.20.57.246	202.6.172.98	443	TCP	powershell.exe

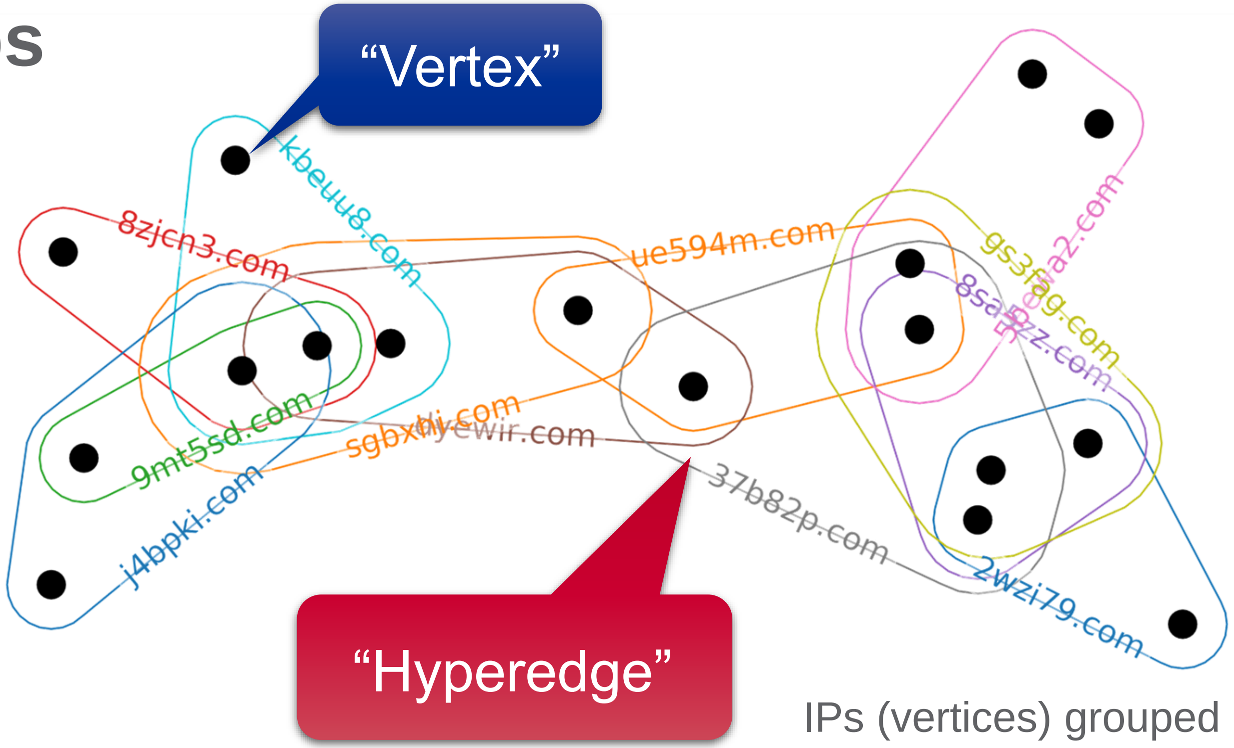
- Network flow has so much more information than just pairs of IPs!
  - Ports can be surrogates for type of communication
  - Protocols dictate the format of communication
  - Host, principal (not always present in network flow) indicate who on the computer is connected to the communication
  - Image path and PID (also not always present) tie the communication to a specific process
- How do we incorporate this information in a *structural* way?

Mathematical model for group relationships: Hypergraph

- **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

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

  - Cyber hypergraph:
    - ✓ Vertices = IPs, ports, users, executables, …
    - ✓ Hyperedges = "behaviors"



IPs (vertices) grouped into website domains (hyperedges)

What kind of data generate Hypergraphs?

Imagine your tabular data:

- **Attributes:** Entities (rows) are indicated as having specific attributes or properties (columns)

	svchost.exe	powershell.exe	lsass.exe	firefox.exe
Host01	0	1	0	0
Host02	0	0	1	1
Host03	1	1	1	1
Host04	0	0	1	0
Host05	1	0	1	1

- **Joint relationships:** Entities jointly participate in some relationship or activity

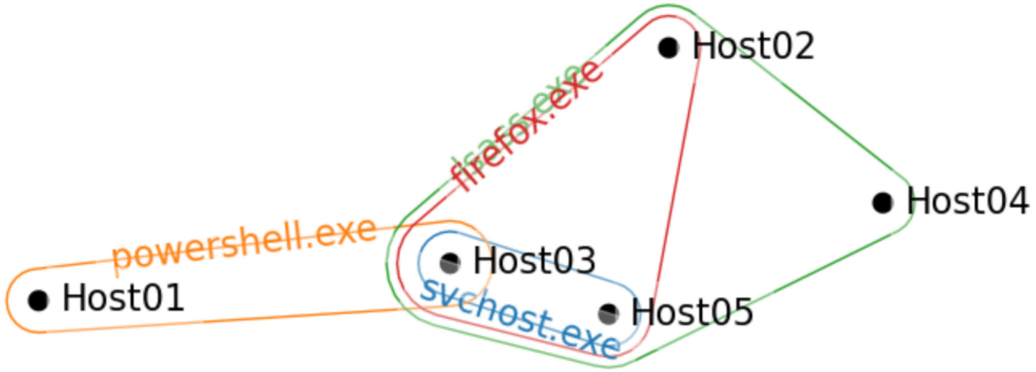
	camcountry.com	crowmedicine.com	sonymusicnashville.com	elvisthemusic.com
10.16.236.100	1	1	1	0
10.16.237.100	1	1	0	1
10.16.238.100	1	1	1	1
10.16.235.100	0	1	1	1

- **Numeric data:** consider thresholding the data

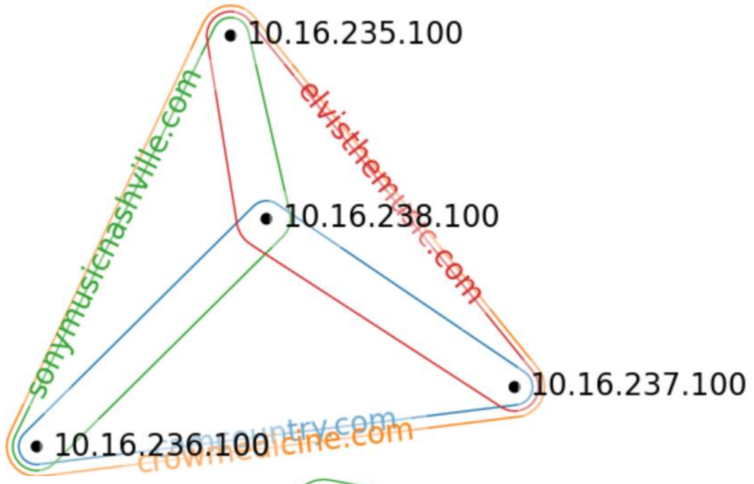
  e.g., cell value > 1

	Port 80	Port 443	Port 22	Port 3389
10.20.30.40	10.262	0.869	0.619	0.989
10.20.30.41	0.312	14.609	0.106	17.427
10.20.30.42	7.401	0.674	4.977	0.831
10.20.30.43	0.282	17.785	8.053	0.195
10.20.30.44	18.484	14.705	0.028	16.451

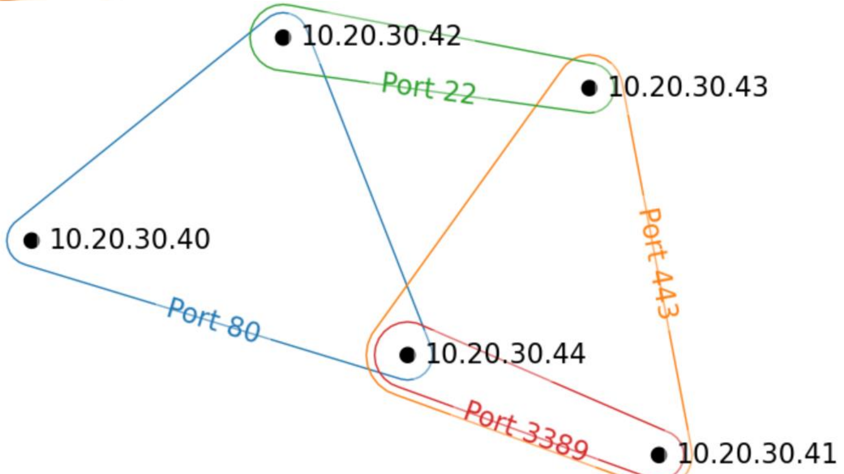
What kind of data generate Hypergraphs?



	svchost.exe	powershell.exe	lsass.exe	firefox.exe
Host01	0	1	0	0
Host02	0	0	1	1
Host03	1	1	1	1
Host04	0	0	1	0
Host05	1	0	1	1

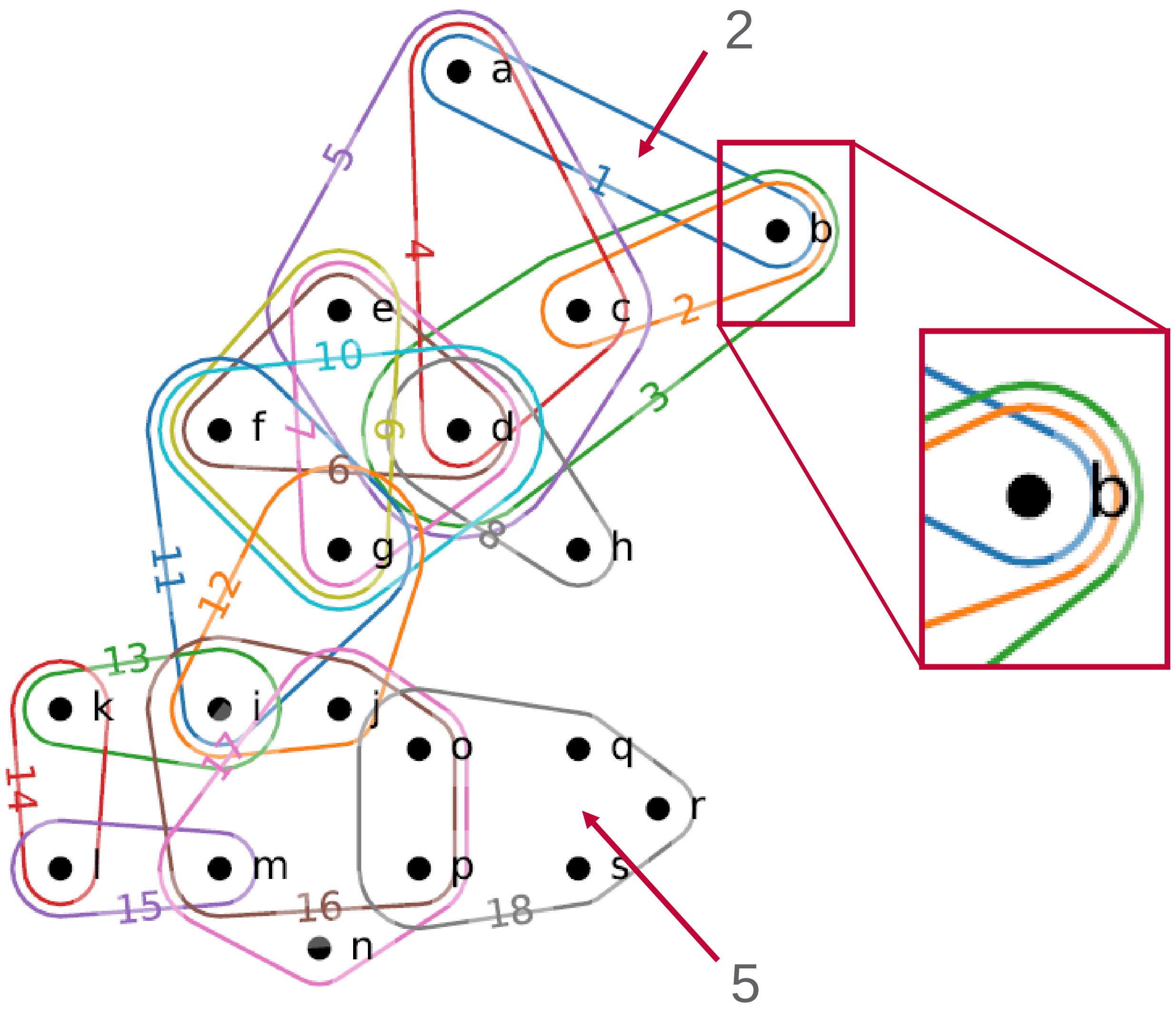


	camcountry.com	crowmedicine.com	sonymusicnashville.com	elvisthemusic.com
10.16.236.100	1	1	1	0
10.16.237.100	1	1	0	1
10.16.238.100	1	1	1	1
10.16.235.100	0	1	1	1



	Port 80	Port 443	Port 22	Port 3389
10.20.30.40	1	0	0	0
10.20.30.41	0	1	0	1
10.20.30.42	1	0	1	0
10.20.30.43	0	1	1	0
10.20.30.44	1	1	0	1

Hypernetwork science



Hypergraph properties

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

Vertex or edge?

Walks on edges or vertices?

– **For graphs, essentially the same.**

- Each pair of vertices in G belong to at most 1 edge, so:

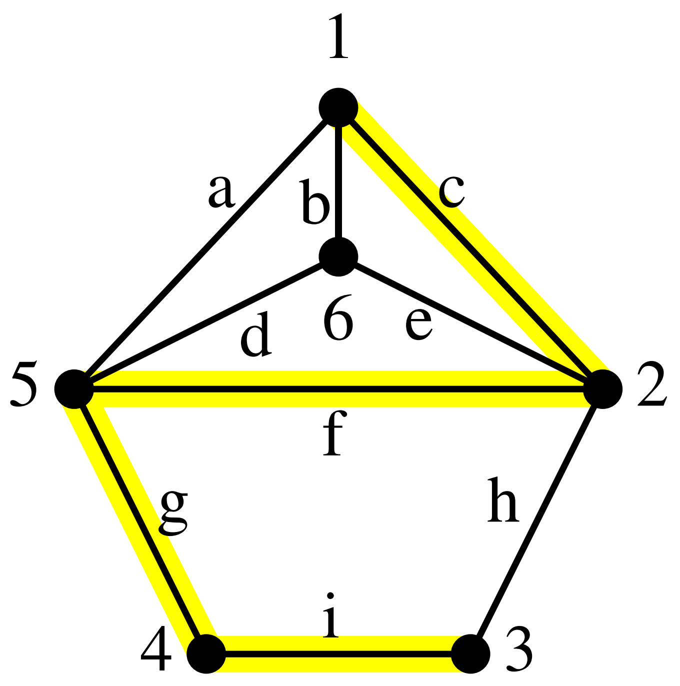
$$\underbrace{v_0, v_1}_{\text{adjacent}}, \ldots, \underbrace{v_{k-1}, v_k}_{\text{adjacent}} \rightarrow \underbrace{e_1}_{\{v_0, v_1\}}, \ldots, \underbrace{e_k}_{\{v_{k-1}, v_k\}}$$

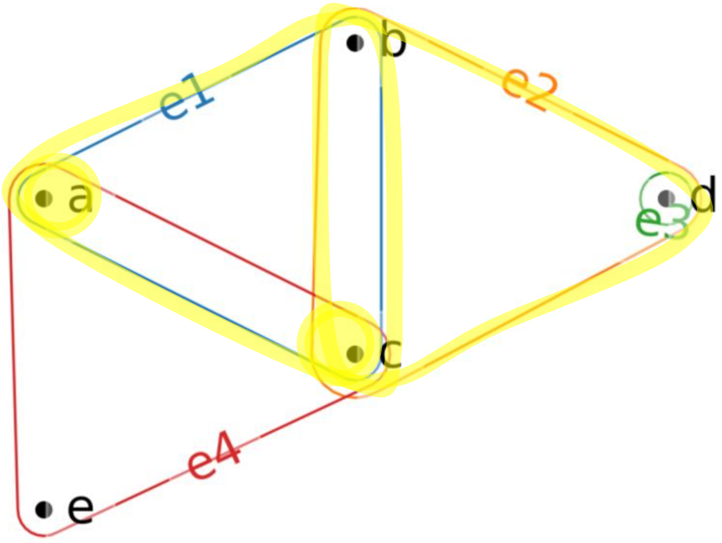
- Each pair of edges in G intersect in at most 1 vertex, so:

$$\underbrace{e_1, e_2}_{\text{incident}}, \ldots, \underbrace{e_{k-1}, e_k}_{\text{incident}} \rightarrow \underbrace{v_0}_{e_1 \setminus e_2}, \underbrace{v_1}_{e_1 \cap e_2}, \ldots, \underbrace{v_{k-1}}_{e_{k-1} \cap e_k}, \underbrace{v_k}_{e_k \setminus e_{k-1}}$$

– **For hypergraphs, not the same.**

- Each pair of vertices can belong to many edges.
- Each pair of edges can intersect at many vertices.





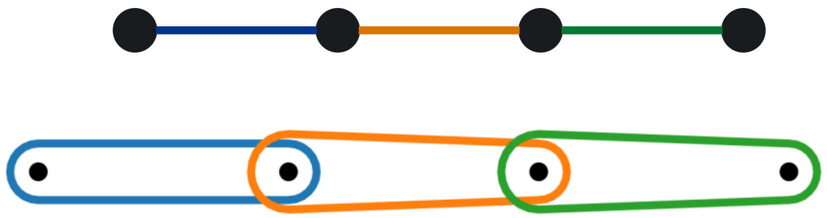
***Walks between edges*:** sequence of successively intersecting edges ← Our focus

***Walks between vertices*:** sequence of successively adjacent vertices

Hypergraph walks have width

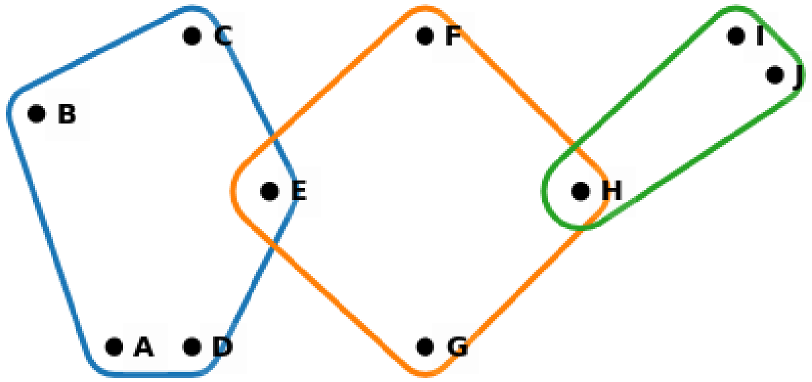
- **s-Walk:** sequence of edges $e_1, \ldots, e_k$ such that $|e_i \cap e_{i+1}| \geq s$
- Walks/paths in hypergraphs have width in addition to length:

A 2-Uniform Hypergraph Path:
(Edgewise) Length = 2
Width = 1

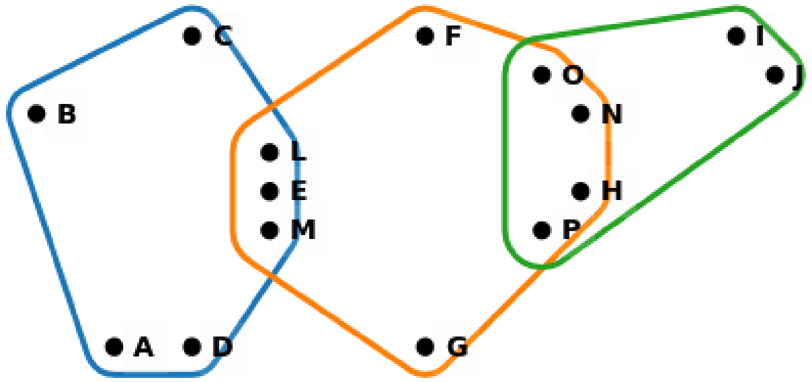


As a 2-uniform HG

Two Hypergraph Paths:
Same length = 2



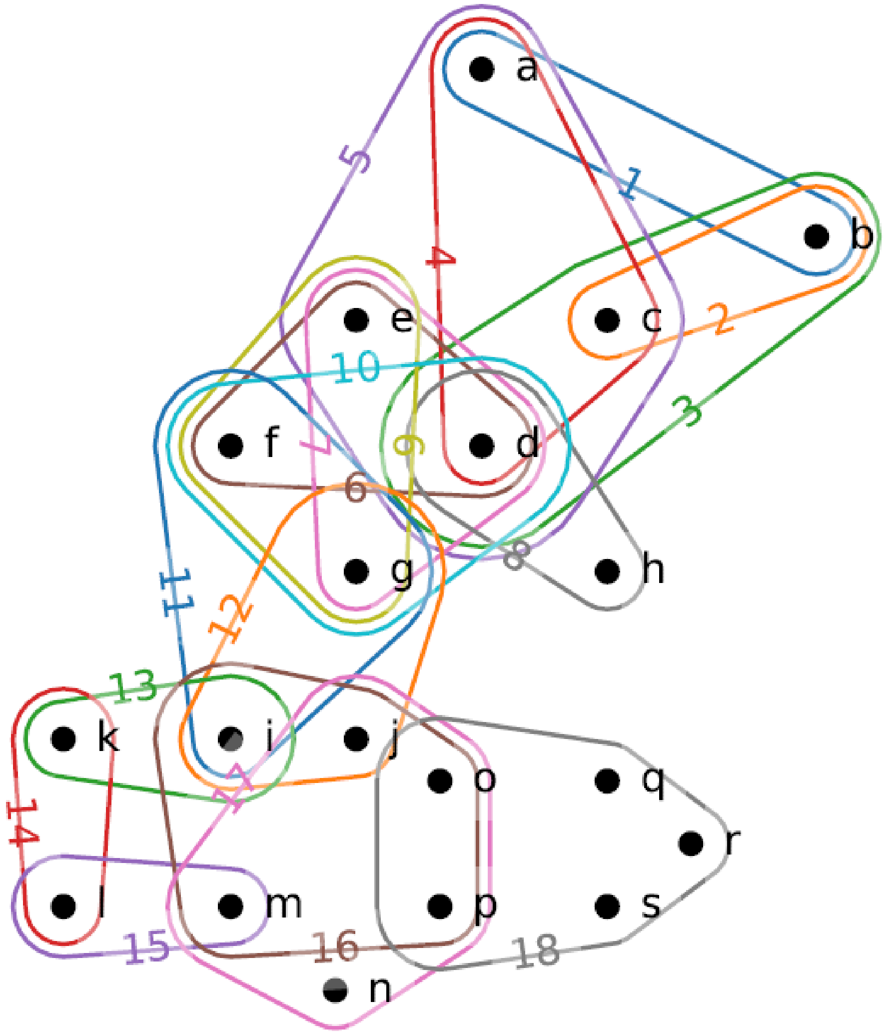
***Weak* interactions:** Width=1



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

- **s-Path** = s-Walk where edges are not repeated

Hypernetwork science



Hypergraph properties

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

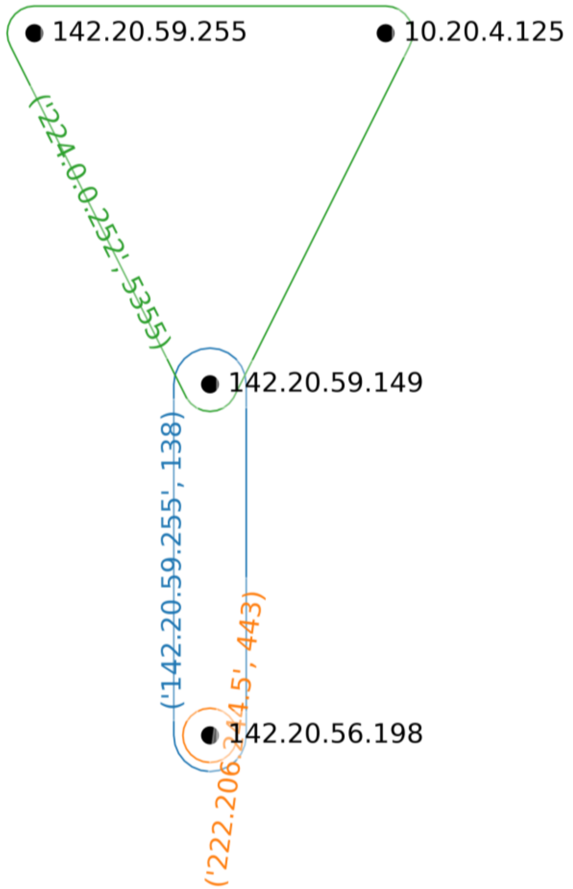
Vertex or edge?

Pacific Northwest
NATIONAL LABORATORY

Hypergraph construction from multi-column data

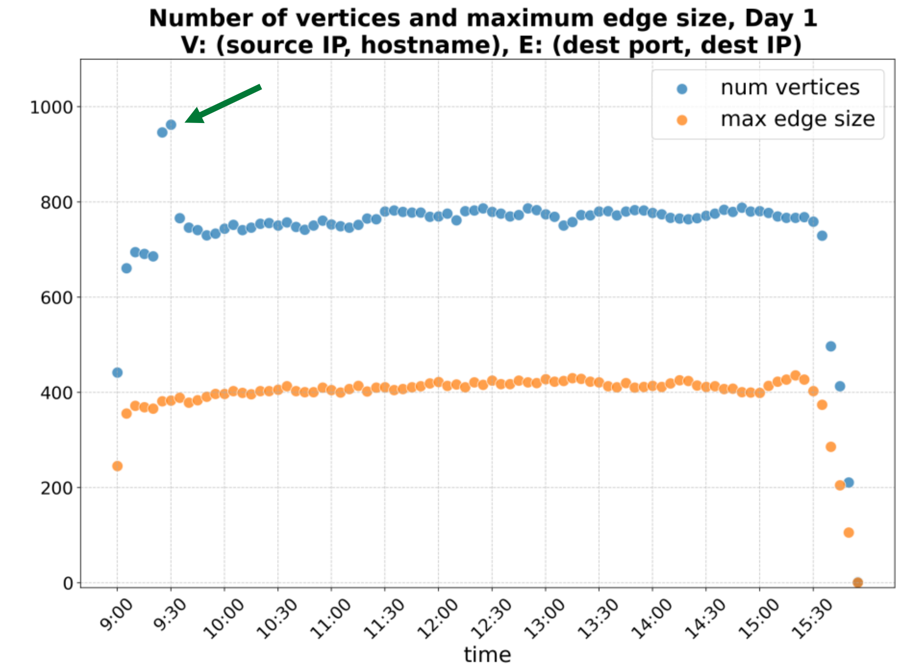
hostname	principal	pid	src_ip	dest_ip	dest_port	l4protocol	image_path
SysClient0201.systemia.com	NT AUTHORITY\SYSTEM	4	142.20.56.198	142.20.59.255	138	UDP	System
SysClient0201.systemia.com	NT AUTHORITY\NETWORK SERVICE	864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
SysClient0201.systemia.com	NT AUTHORITY\NETWORK 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.exe
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.exe

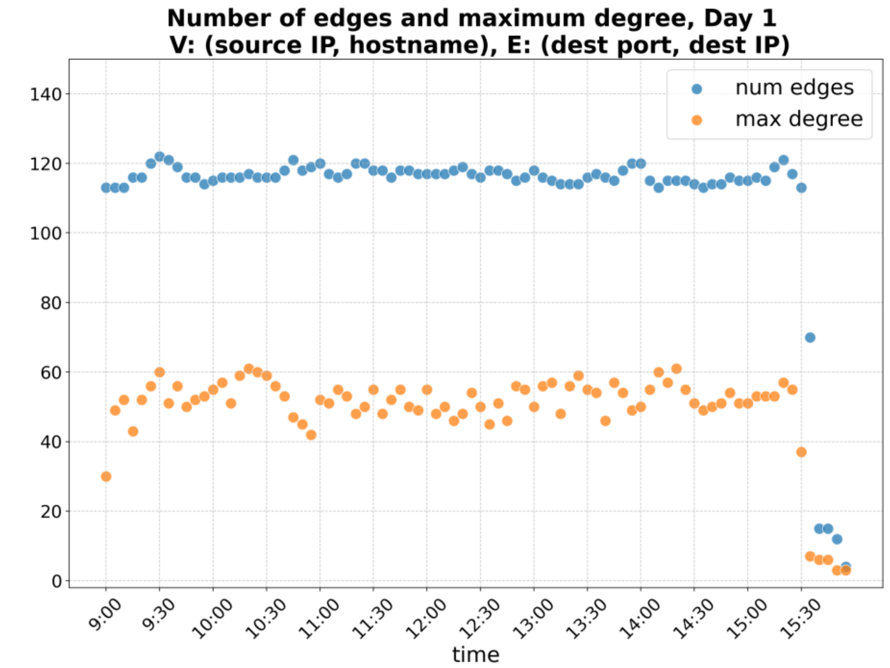
- Multi-dimensional data set: *n*D-array, *n*-column data frame
- Specify column set for hyperedges (yellow)
  - **Unique combinations:**
    (142.20.59.255, 138), (224.0.0.252, 5355), (222.206.244.5, 443)
- Specify disjoint column set for vertices (blue)
  - **Unique vertices:**
    142.20.59.149, 10.20.4.125, 142.20.59.255, 142.20.56.198
- A vertex is contained in a hyperedge if there is a record with that combination in the data. Think “hyperedges = common behaviors”

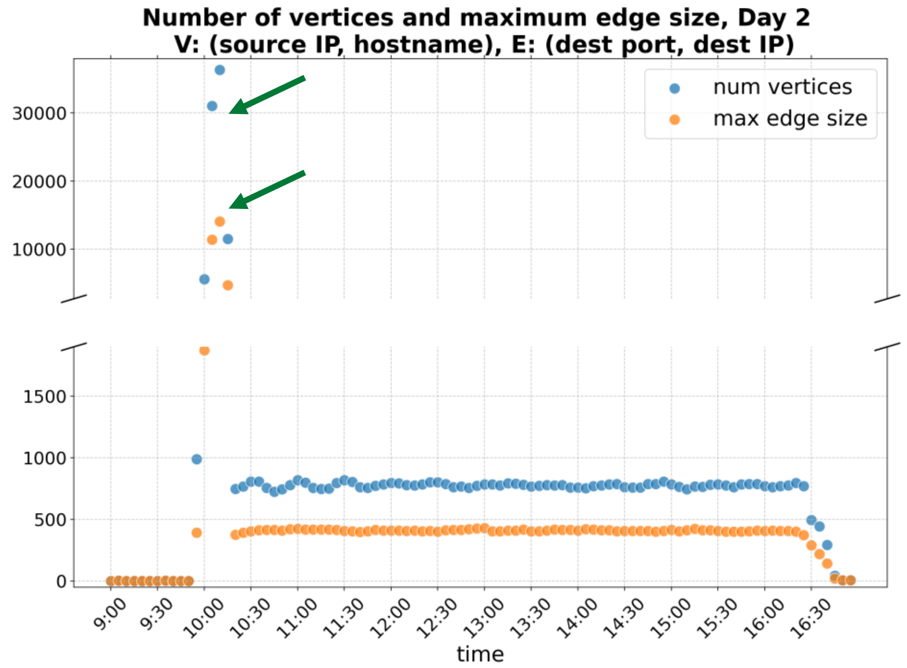


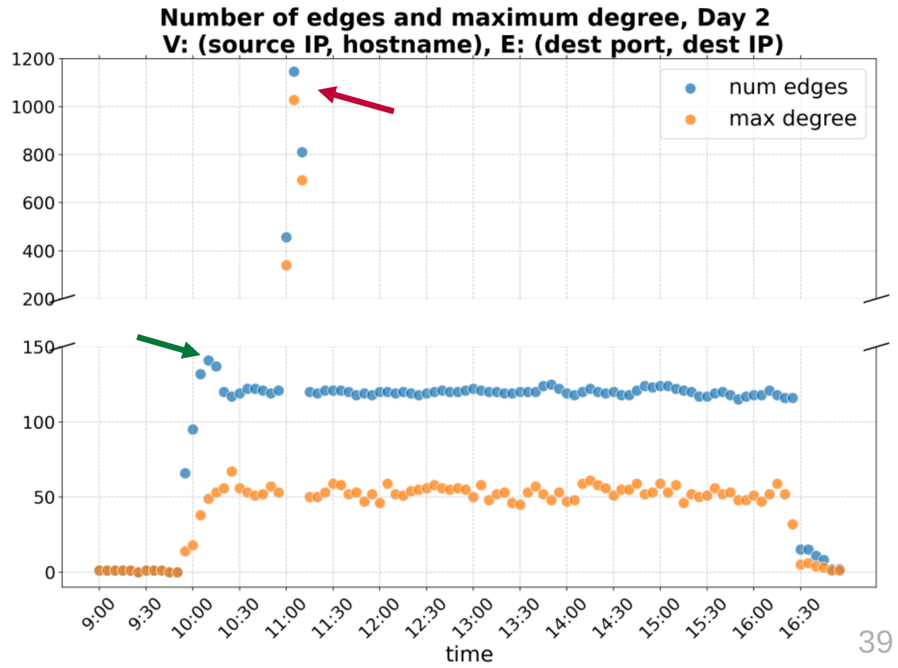
Identifying anomalies via simple dynamic hypergraph measures

- “Do the simple thing first”
- Sometimes just counting things (vertices, edges, degrees, edge sizes) and looking for temporal changes gives you insight.
  - A. Network simulation startup activity
  - B. Actual red team activity – “Deathstar” to scan domain
- But there’s much more to find in this data…







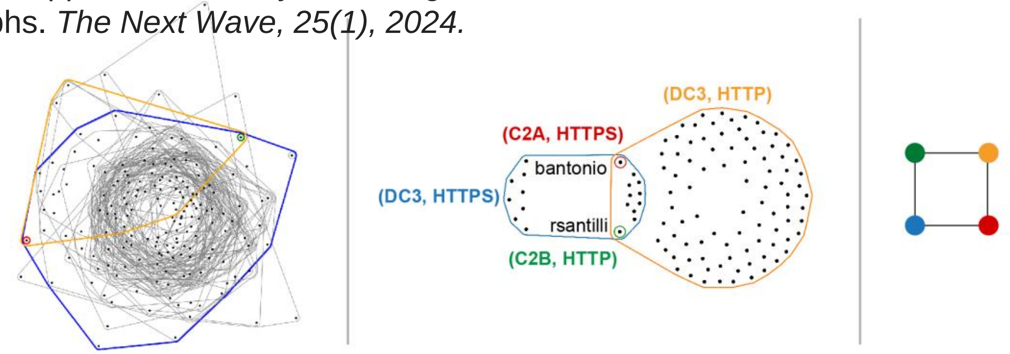


Joint with Helen Jenne

Jenne, H., Aksoy, S.G., Best, D., Bittner, A., Henselman-Petrusek, G., Joslyn, C., Kay, B., Myers, A., Seppala, G., Warley, J., Young, S.J., and Purvine, E. Stepping Out of Flatland: Discovering behavior patterns as topological structures in Cyber hypergraphs. *The Next Wave, 25(1), 2024.*

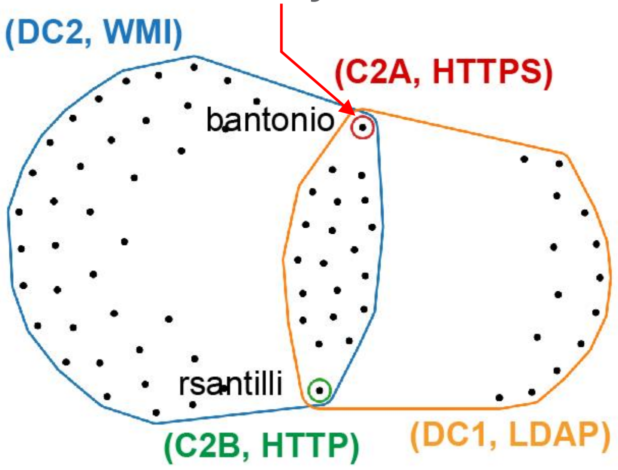
Pacific Northwest
NATIONAL LABORATORY

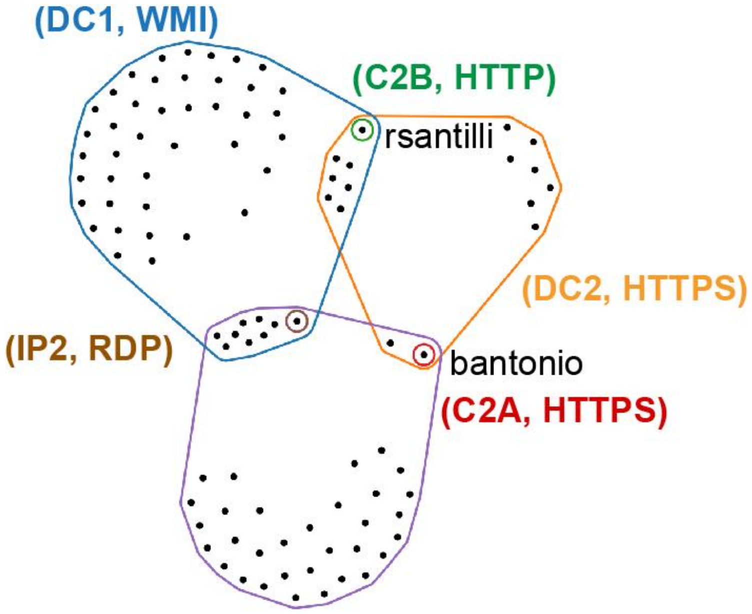
Finding complex patterns of connectivity

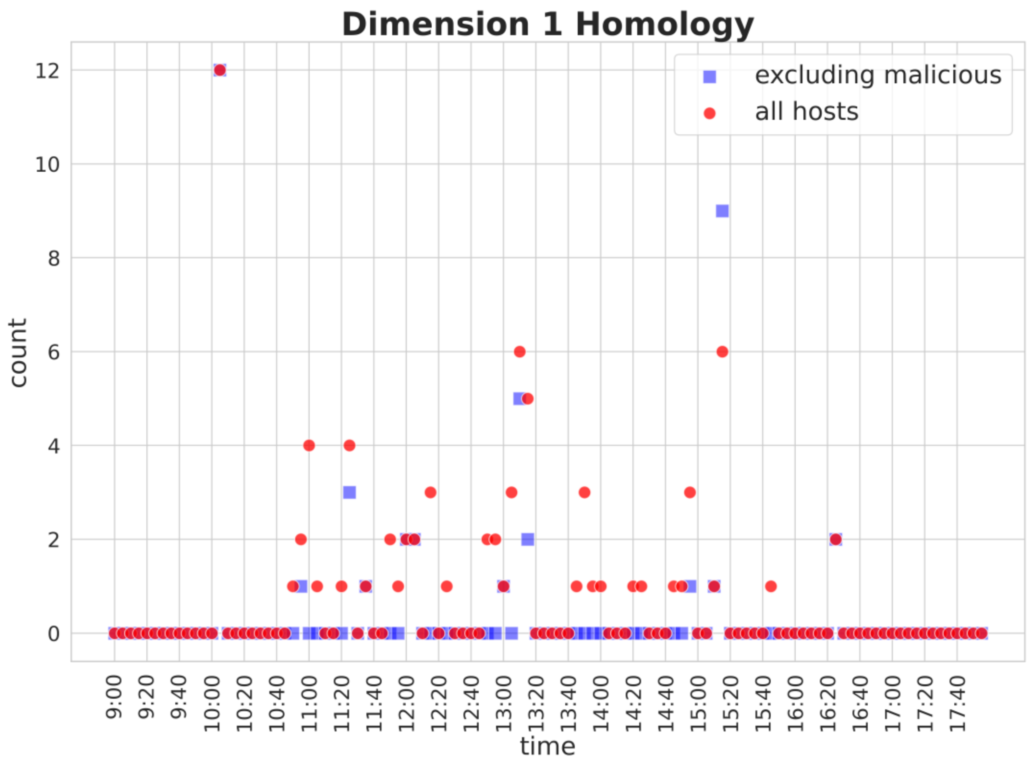


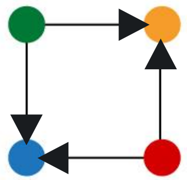
- While adversaries try to fly below the radar they still operate within the network and likely do things that are abnormal. Their activities may create unusual patterns of connectivity.

Red singleton edge represents 335 identical edges containing vertex for Sysclient0501

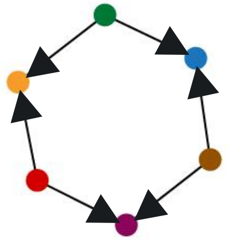








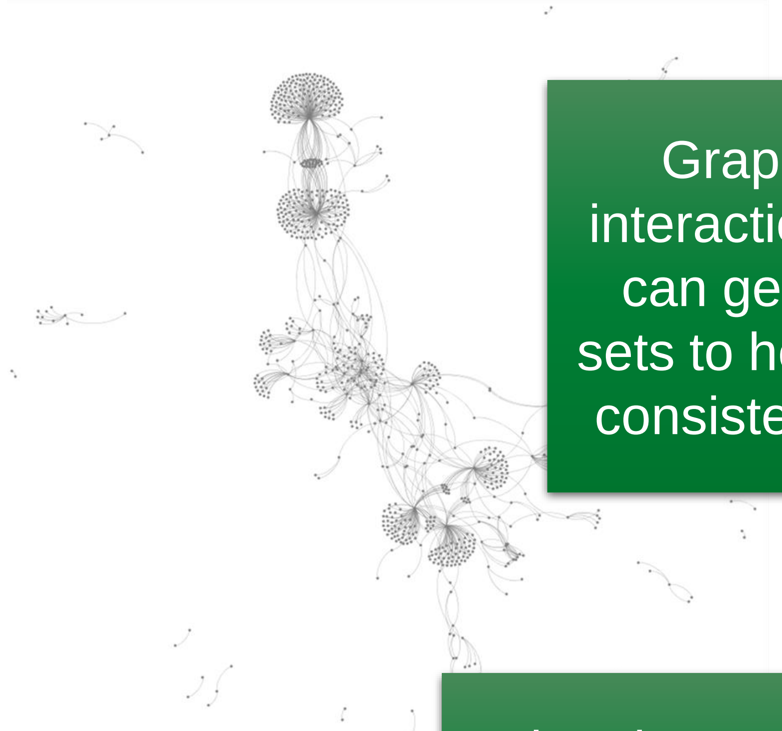
Edge containment structures



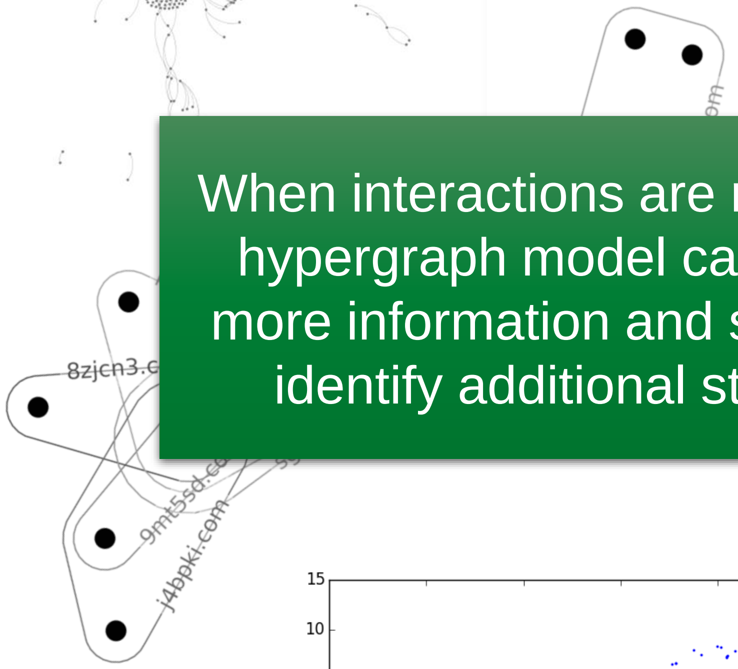
This structure is not always tied to malicious activity, but it is rare in this data and thus potentially of interest.

Plan of the talk

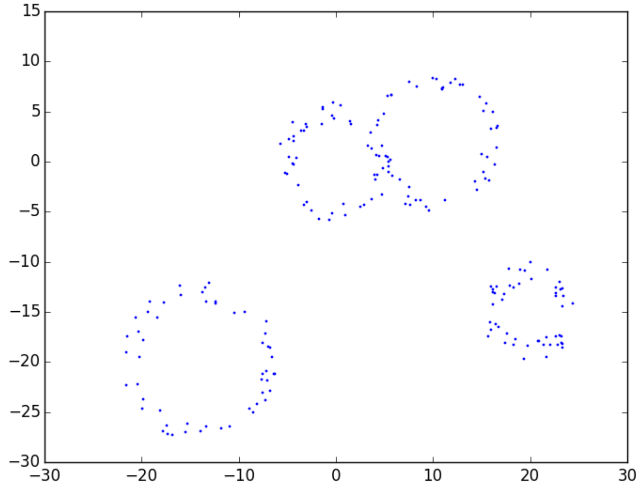
- My path to a nonacademic career
- Cybersecurity 101 (accelerated version!)
- Graphs and hypergraphs via network flow
- Topology via high-dimensional data



Graphs model pairwise interactions. Random models can generate realistic data sets to help explore properties consistent with real systems.

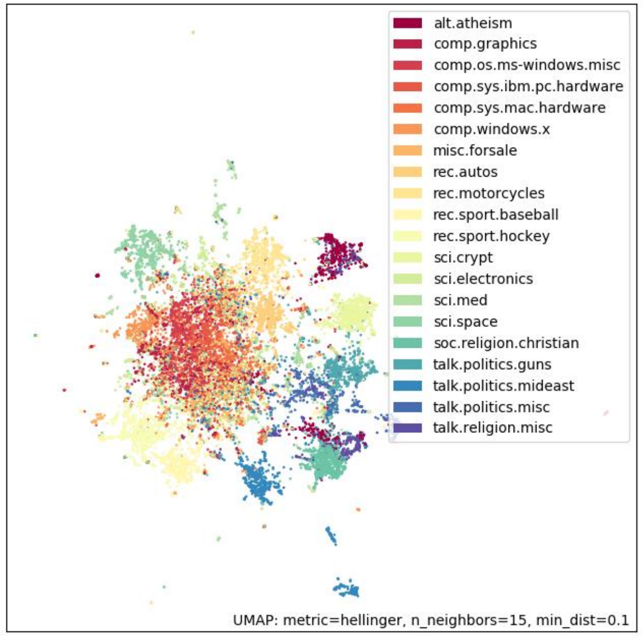


When interactions are multi-way a hypergraph model can capture more information and sometimes identify additional structure.

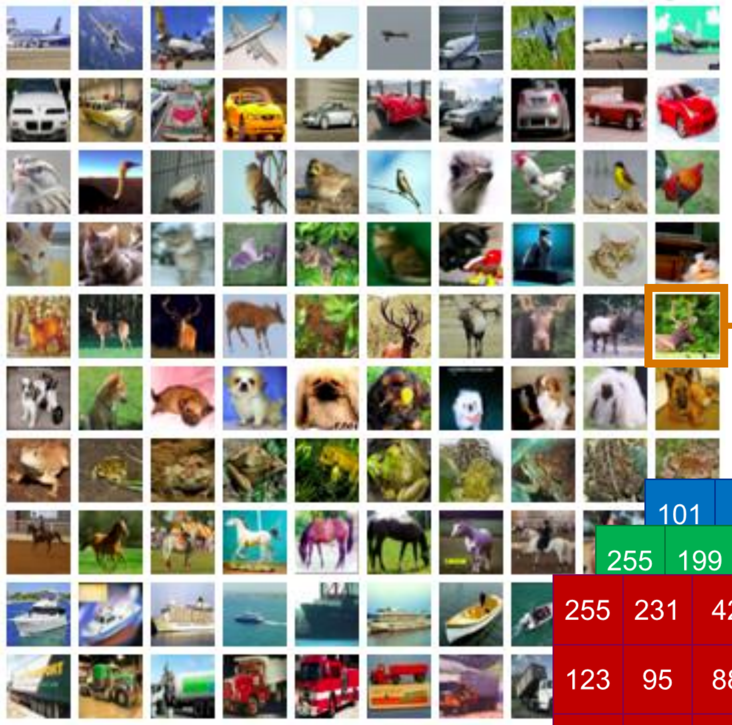


Application #2: High-dimensional data (generally)

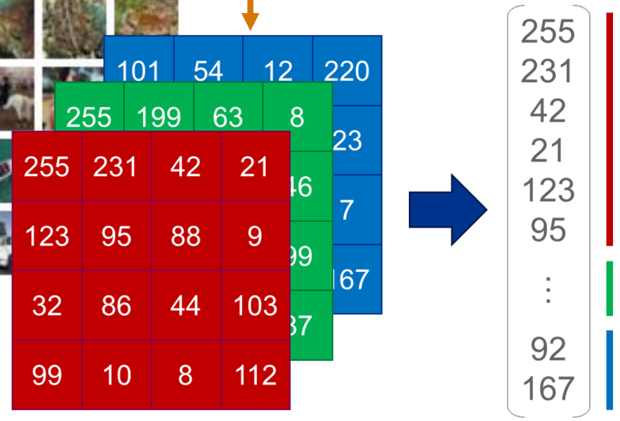
[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

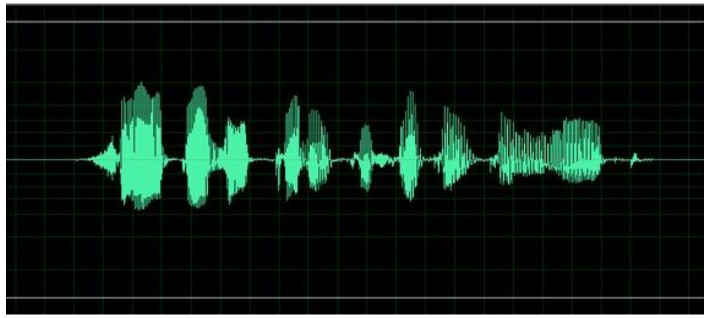


https://umap-learn.readthedocs.io/en/latest/document_embedding.html

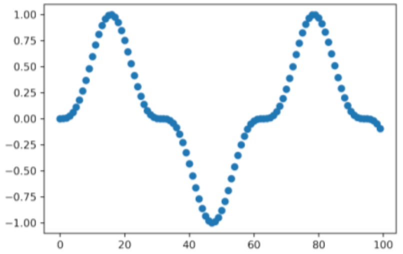


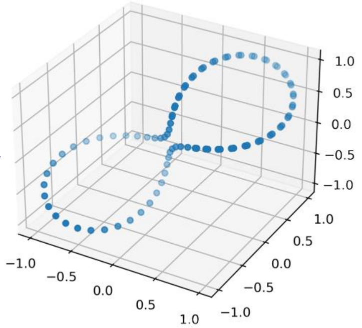
https://www.kaggle.com/c/cifar-10





[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)





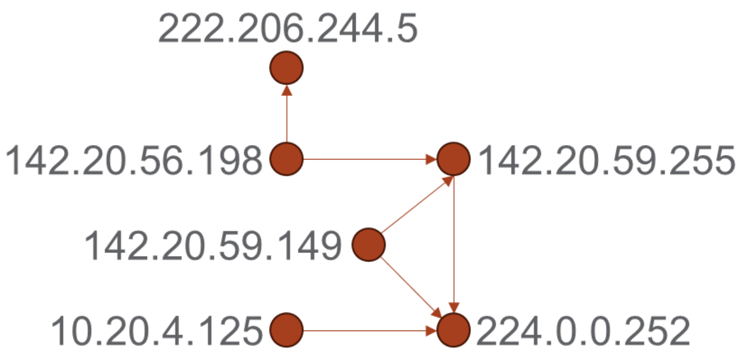
PID	Src IP	Dst IP	Dst Port	Protocol	Image path
4	142.20.56.198	142.20.59.255	138	UDP	System
864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
864	142.20.59.255	224.0.0.252	5355	UDP	svchost.exe
636	142.20.56.198	222.206.244.5	443	TCP	firefox.exe
4	142.20.59.149	142.20.59.255	138	UDP	System
864	142.20.59.149	224.0.0.252	5355	UDP	svchost.exe

Creating high dimensional data from cyber logs: Feature engineering

- For a given time-window of data we can create a feature vector
- Hand-crafted features:
  - Count of unique values in a column
  - Count occurrences of specific values in a column
  - Numerical aggregations – min, max, mean, median, sum
  - Max degree of a graph of the data
  - …
- Machine learned features
  - Train an autoencoder or LLM on log lines, aggregate all encoded lines in a window

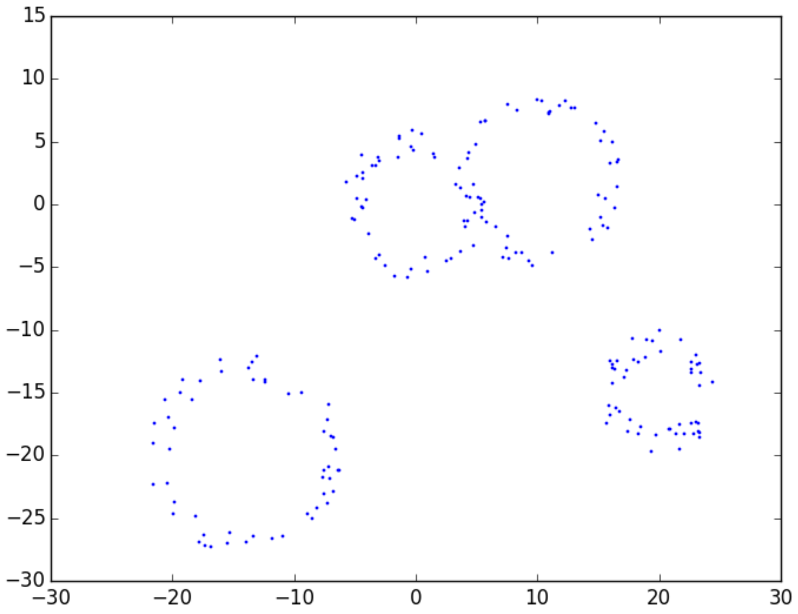
PID	Src IP	Dst IP	Dst Port	Protocol	Image path
4	142.20.56.198	142.20.59.255	138	UDP	System
864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
864	142.20.59.255	224.0.0.252	5355	UDP	svchost.exe
636	142.20.56.198	222.206.244.5	443	TCP	firefox.exe
4	142.20.59.149	142.20.59.255	138	UDP	System
864	142.20.59.149	224.0.0.252	5355	UDP	svchost.exe

Feature	Value
Count of Src IPs	4
Count of Dst Port = 443	1
Count of Image path = svchost.exe	3
Max in-degree in Src IP -> Dst IP graph	3
Max out-degree in Src IP -> Dst IP graph	2

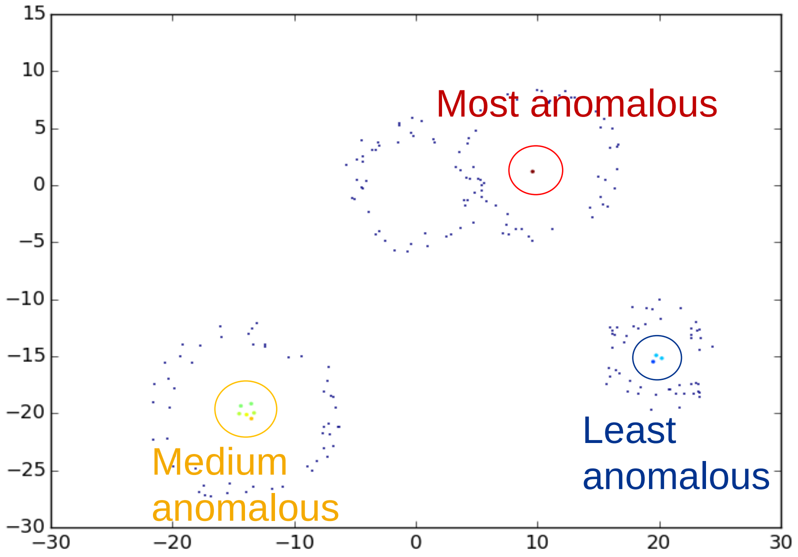


Temporal anomaly detection from feature point clouds

- **Main assumption:** Behavior varies smoothly from set of recent small time windows to the next window
- **Method:**
  - Partition data into time intervals, create a single vector for each
  - Baseline contains many time intervals – many vectors – current time interval is single vector
  - How, and how much, does adding the single vector change the structure of the collection of baseline vectors?



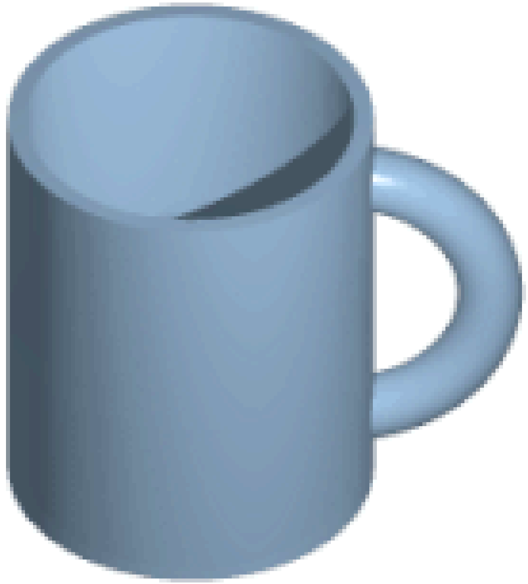
Baseline data



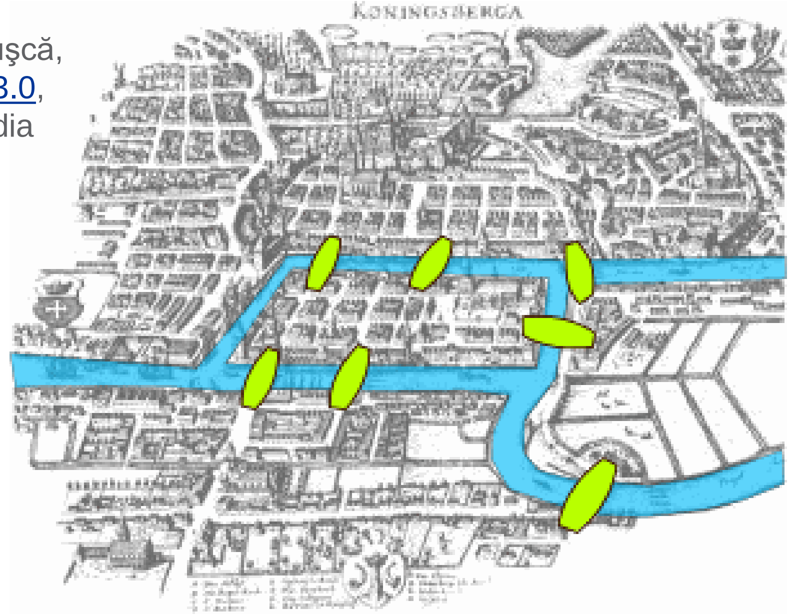
Baseline with additional points added

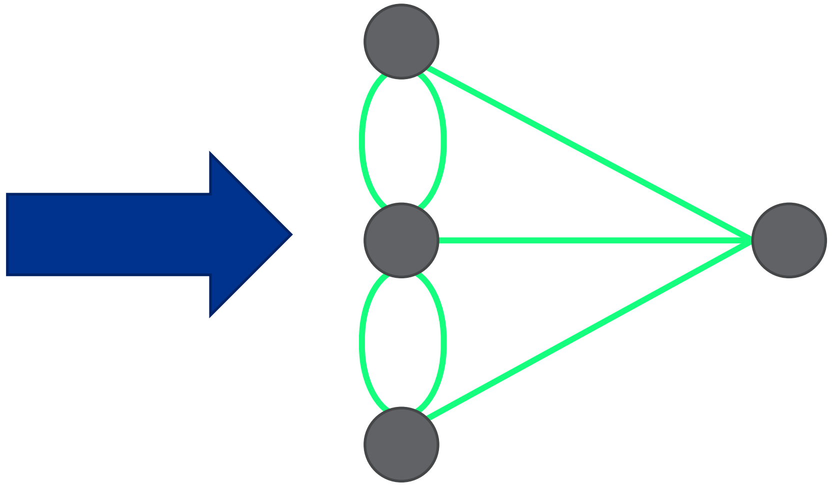
Structure = Topology

- Geometry without distance; stretchy geometry
- Properties (= holes) of geometric objects preserved under "continuous deformation" – stretching and twisting are ok but tearing and gluing are not
- Abstract an object into a simpler version that preserves certain properties – "topological invariants"



Donut? Coffee cup?





1736 Leonhard Euler, Seven Bridges of Königsberg

Persistent Homology

- Given a point cloud we want to understand its coarse topological structure
- Connect points at increasing distance thresholds
- Track birth and death threshold for topological features ("holes")

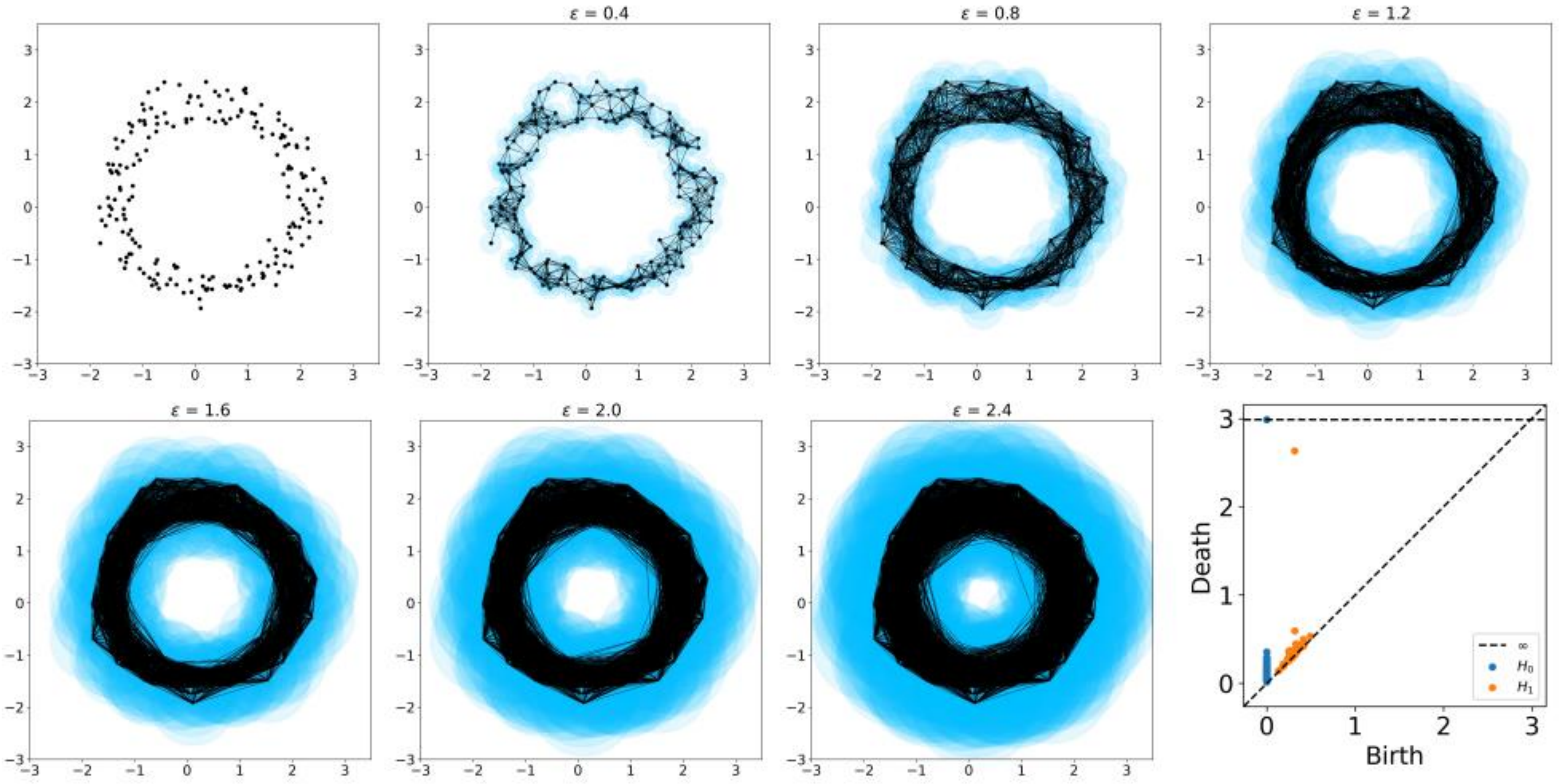
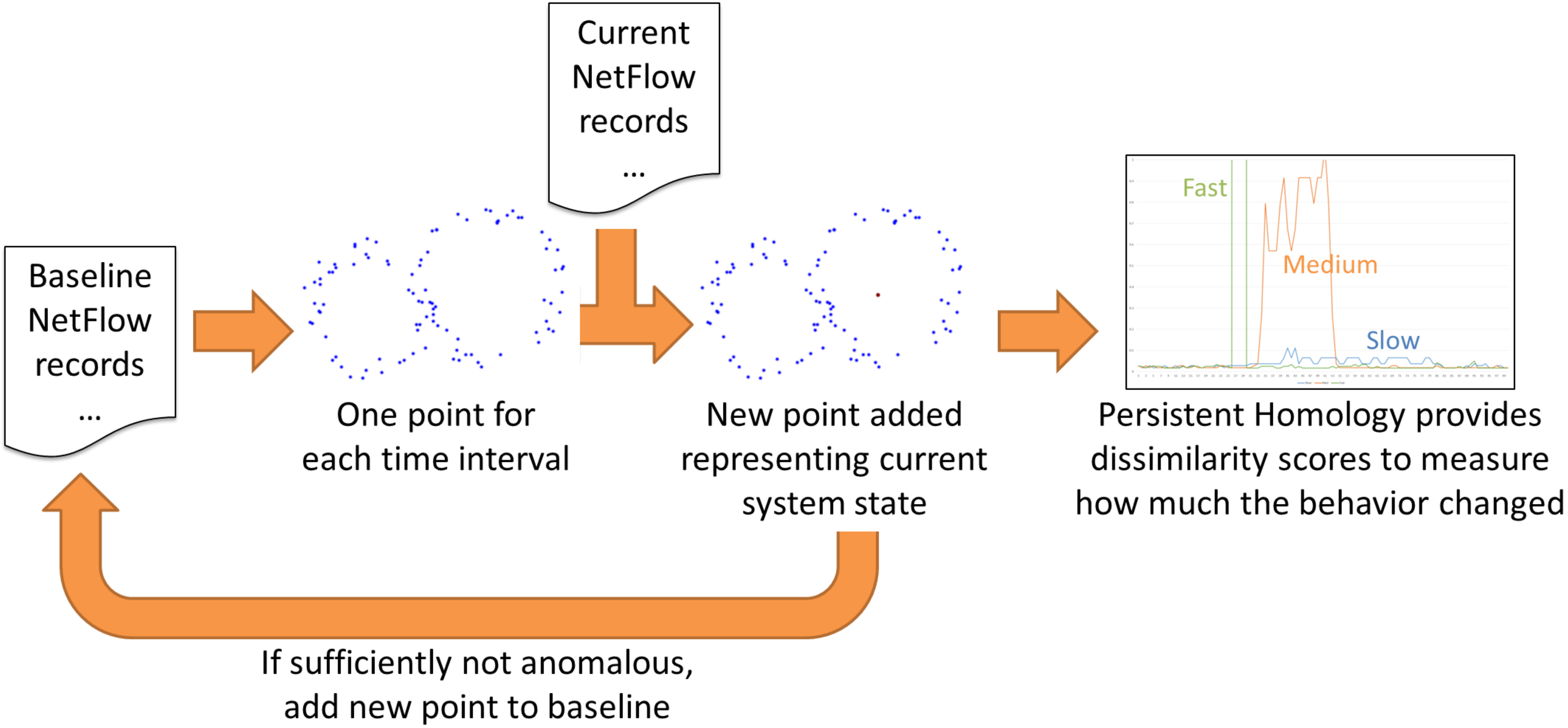


Image credit: Sarah Tymochko

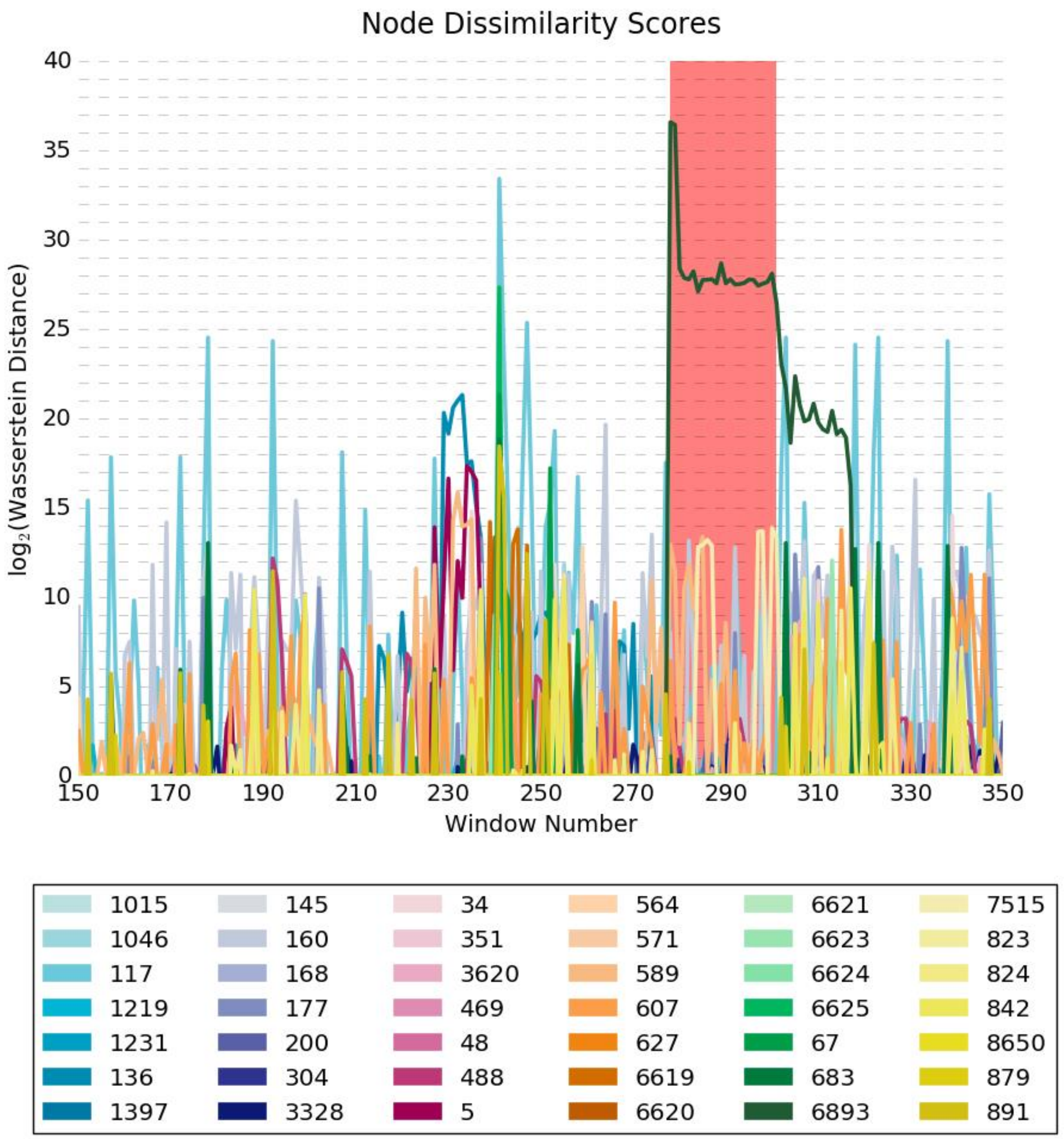
Anomaly detection pipeline



Joint with Helen Jenne, Dan Best, Paul Bruillard, Alyson Gauthier, Greg Henselman-Petrusek, Cliff Joslyn, Bill Kay, Audun Myers, Kathleen Nowak, Garret Seppala, Stephen J. Young

Use case example: BitTorrent Detection

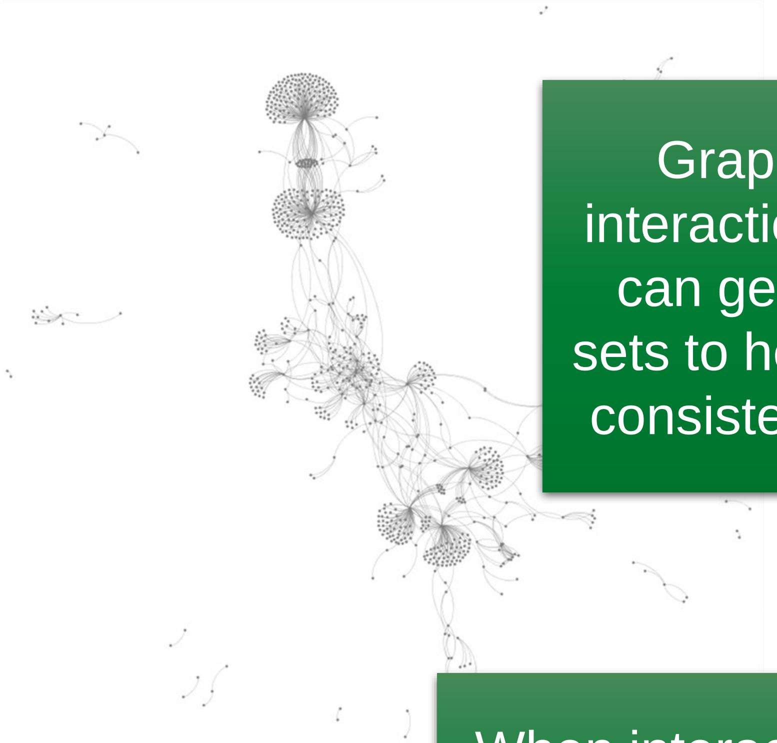
- Network flow for a single building was captured
  - BitTorrent traffic added after the fact by node 6893 during windows 278-301
- Feature vectors came from counts of small graph patterns
- Our pipeline was able to detect an anomaly from 6893 during the correct time windows



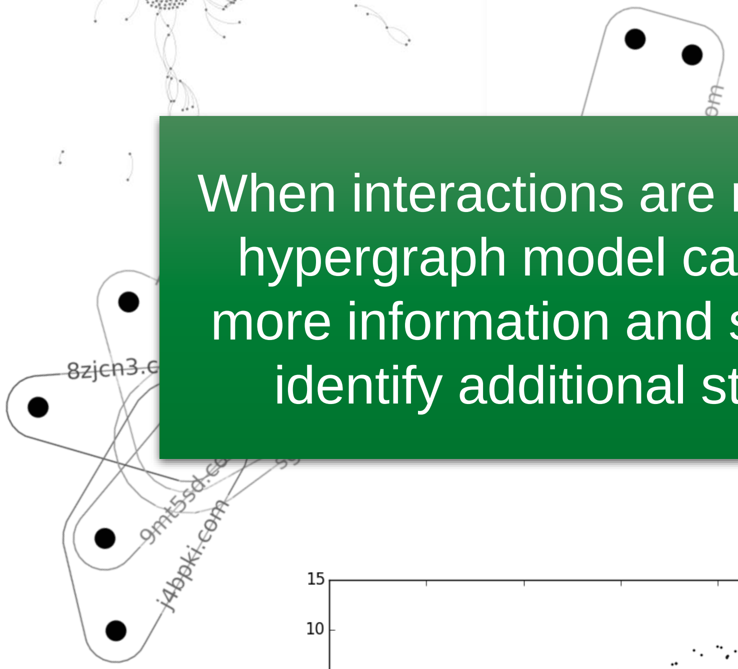
Christopher R. Harshaw, Robert A. Bridges, Michael D. Iannacone, Joel W. Reed, and John R. Goodall. 2016. GraphPrints: Towards a Graph Analytic Method for Network Anomaly Detection. In Proceedings of the 11th Annual Cyber and Information Security Research Conference (CISRC '16).

~~Plan~~ End of the talk

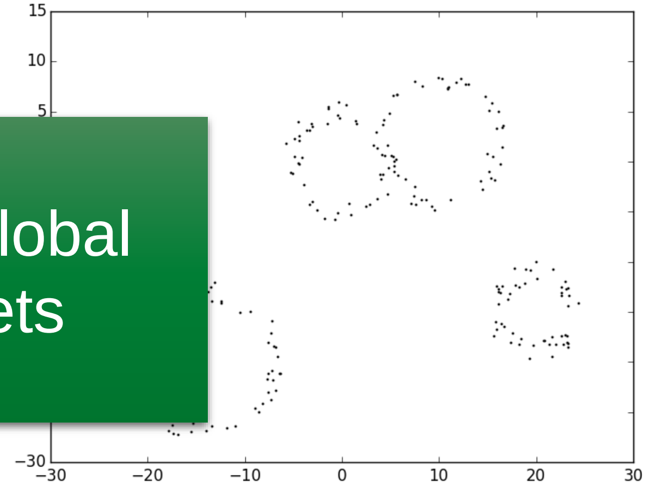
- My path to a nonacademic career
- Cybersecurity 101 (accelerated version!)
- Graphs and hypergraphs via network flow
- Topology via high-dimensional data



Graphs model pairwise interactions. Random models can generate realistic data sets to help explore properties consistent with real systems.



When interactions are multi-way a hypergraph model can capture more information and sometimes identify additional structure.



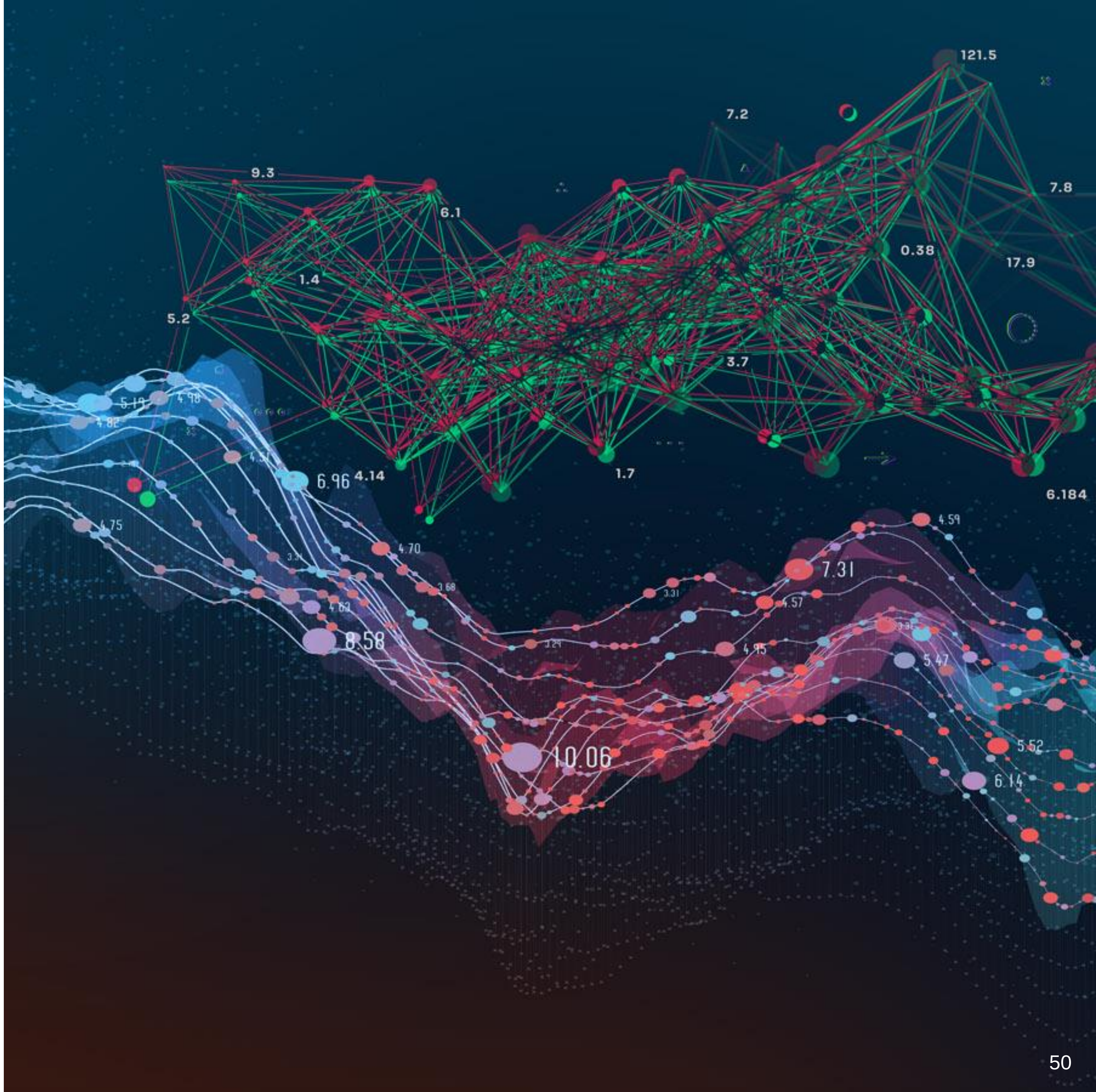
Topology captures global features in data sets

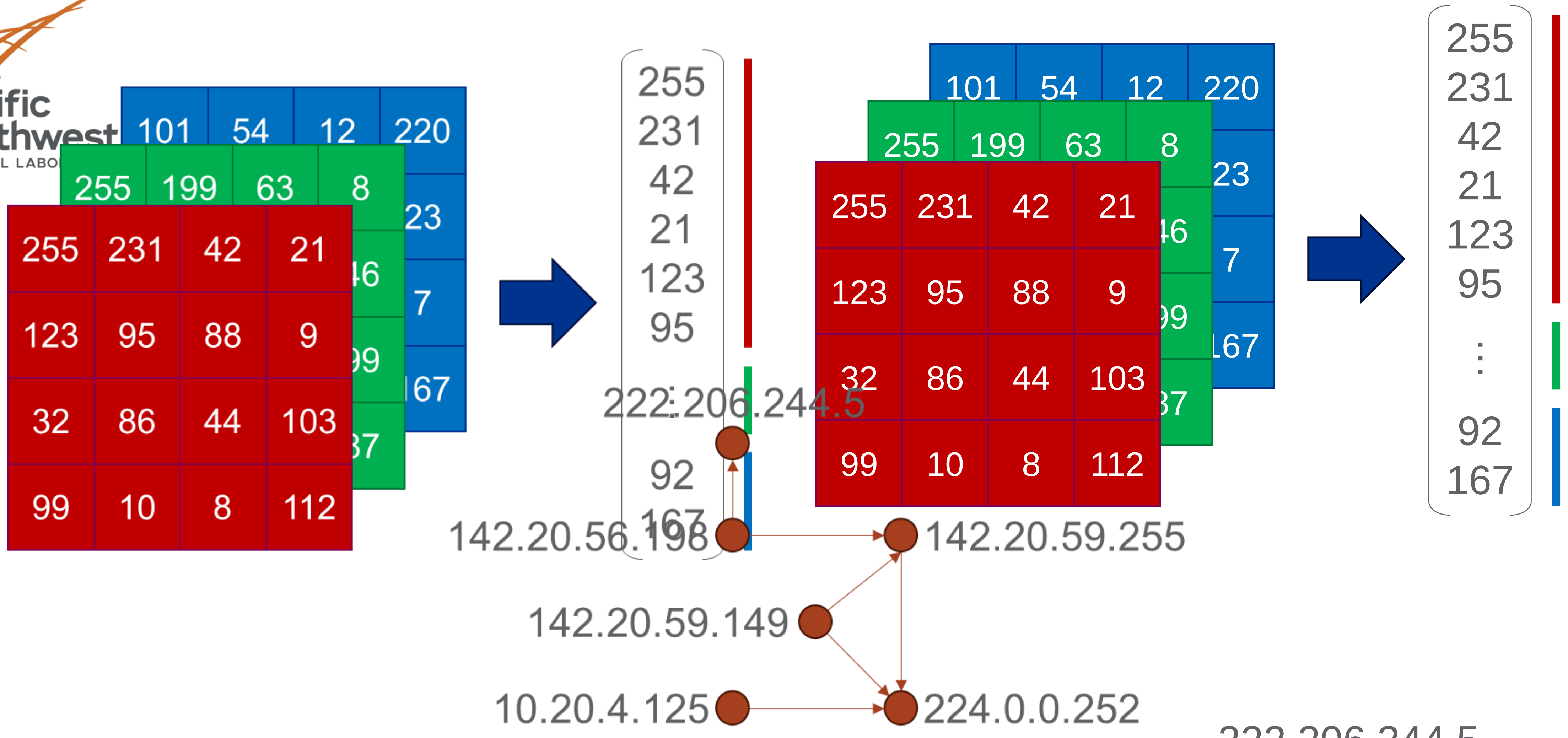
Pacific Northwest
NATIONAL LABORATORY

Thank you

Check out our internships and jobs!
https://careers.pnnl.gov/

Contact me with questions!
Emilie.Purvine@pnnl.gov





PID	Src IP	Dst IP	Dst Port	Protocol	Image path
4	142.20.56.198	142.20.59.255	138	UDP	System
864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
864	142.20.59.255	224.0.0.252	5355	UDP	svchost.exe
636	142.20.56.198	222.206.244.5	443	TCP	firefox.exe
4	142.20.59.149	142.20.59.255	138	UDP	System
864	142.20.59.149	224.0.0.252	5355	UDP	svchost.exe