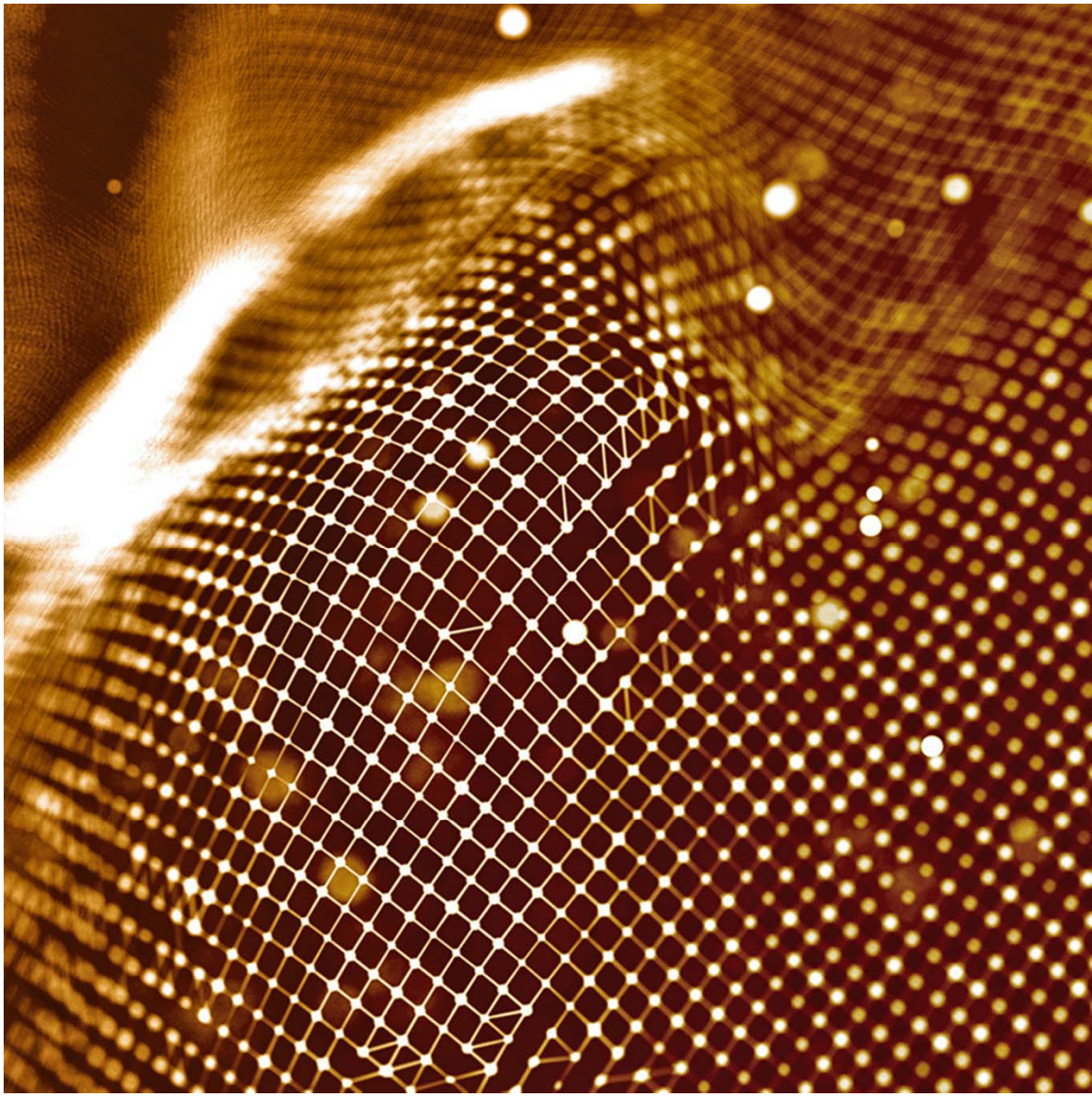
High Performance Hypergraph Analytics of Domain Name System Relationships

Cliff Joslyn

Sinan Aksoy, Dustin Arendt,
Louis Jenkins (U Rochester), Brenda Praggastis,
Emilie Purvine, Marcin Zalewski

HICSS Symposium on Cybersecurity Big Data Analytics
PNNL-SA-140454
January 8, 2019

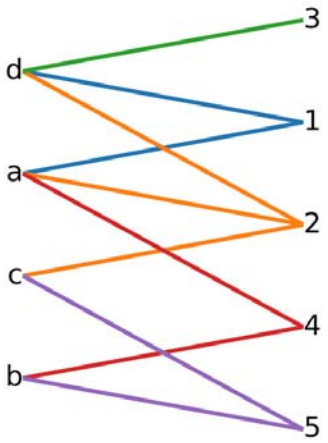


High Performance Hypergraph Analytics of Domain Name System Relationships

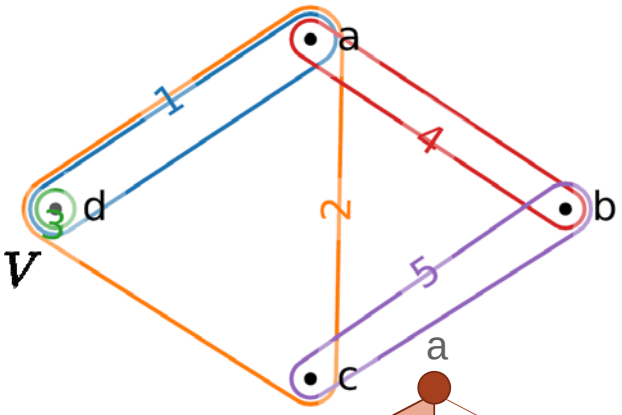
- **Hypergraph Analytics:** Towards a hypernetwork science
  - Multidimensional graph analytics for cyber data
  - Graphs vs. hypergraphs
- **DNS Data As a Hypergraph**
- **Chapel Hypergraph Library (CHGL)**
  - High performance hypergraph analytics
- **Initial Results**
  - Load scaling
  - Basic hypergraph analysis
  - Distributional statistics
  - Hypercomponent analysis
  - Segments and some motifs

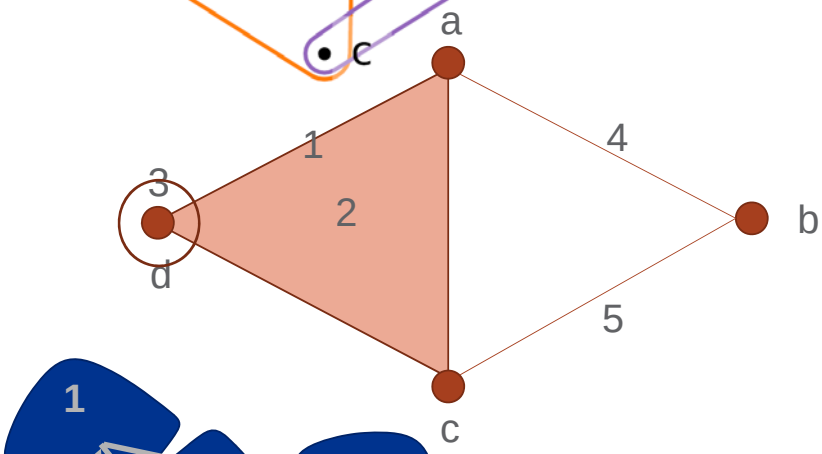
Hypergraphs

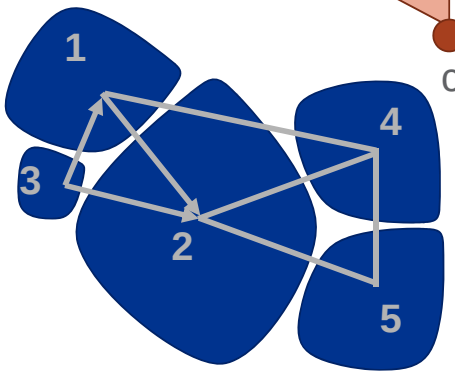
- **Hypergraph:** $\mathcal{H} = \langle V, \mathcal{E} \rangle$ Multiset $\mathcal{E} = \{e\}, e \subseteq V$
- **Multiple forms:**
  - **Set multi-system:** $\mathcal{E} \subseteq 2^V$
    - ✓ Euler diagram
    - ✓ Simplicial diagram
    - ✓ “Pebble diagram” (line graph)
  - **Incidence matrix:** $V \times \mathcal{E}$
  - **Bipartite graph:**



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



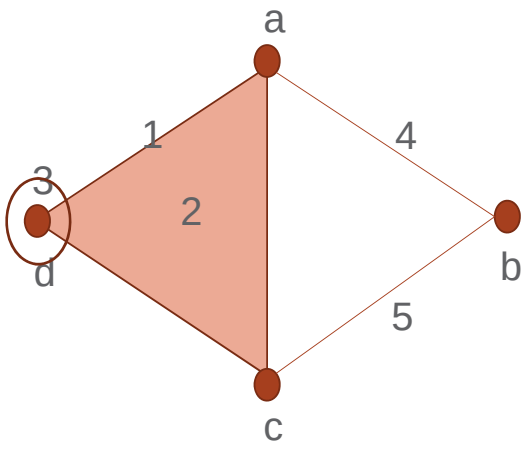




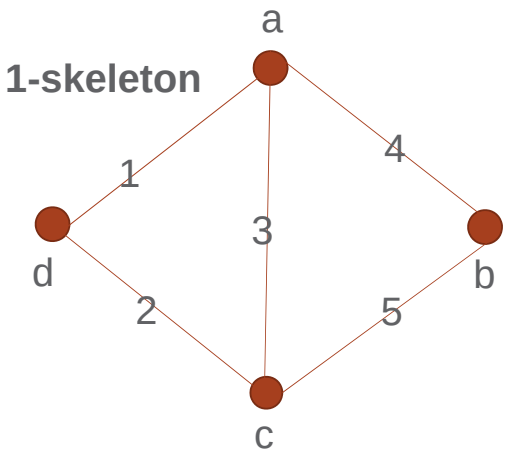
Towards Hypernetwork Science

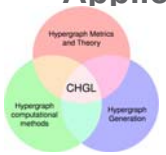
- **Hypergraphs explicitly generalize graphs:**
  - Graphs are 2-uniform hypergraphs
- **Many/most graph models of complex data "squeeze" data down to 1-skeletons**
  - Lose critical information about multi-way relationships
- **Towards a "hypernetwork science"**
  - Hypergraphs well known in math and CS
  - Very few data science applications
  - *Reluctance to approach hypergraphs' higher complexity: potentially exponential vs. quadratic*
- **PNNL is pioneering**
  - **Mathematics:** Hyperpath analysis, spectral approaches
  - **Methodologies:** Hypernetwork science
  - **Algorithms:** Shortest hyperpaths, hyper-components, generation
  - **HNX Software:** Exploratory analytics, visualization
  - **CHGL Software:** High performance scaling
  - **Applications:** Cyber analytics, information integration

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



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





https://github.com/pnnl/chgl

https://github.com/pnnl/HyperNetX

Jenkins, LP; Bhuiyan, T; Harun, S; *et al.*: (2018) "Chapel Hypergraph Library (CHGL)", 2018 IEEE High Performance Extreme Computing Conf. (HPEC 2018)

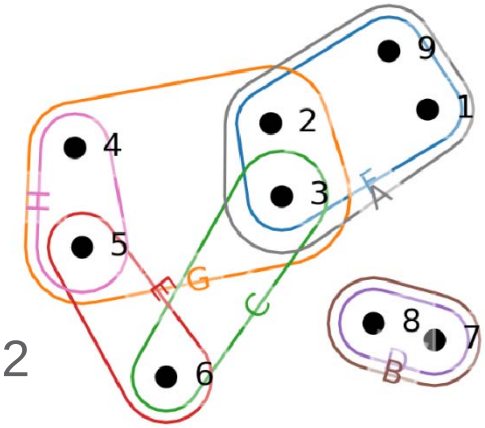
Purvine, EAH; Aksoy, S; Joslyn, CA; Nowak, K; B Praggastis, M Robinson : (2018) "A Topological Approach to Representational Data Models", 20th Int. Conf. on Human-Computer Interaction (HCI International), LNCS 10904, pp. 90-109

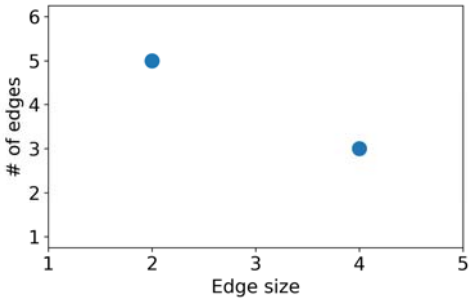
Hypergraph Attributes

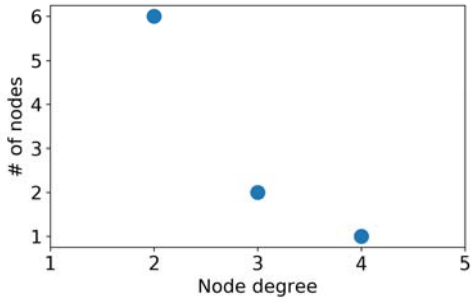
- **Multiple edges = duplicate columns:** $A = F, B = D$
- **Equivalent vertices = duplicate rows:** $1 = 9, 7 = 8$
- **Singleton edges**
- **Edge size:** # vertices/edge
- **Vertex degree:** # edges/vertex
- **Included edges:** $H \subset G$
- **Hypernetwork properties:** e.g. s=2
  - **s-adjacency:** $|A \cap G| = 2$
  - **s-path:** $H \mapsto G \mapsto A$
  - **s-components:** $\{A, F, G, H\}, \{B, D\}, \{C\}, \{E\}$
  - **s-diameter:** 2

The longest shortest path in the largest s-component is 2 steps.

	A	B	C	D	E	F	G	H
1	X					X		
2	X					X	X	
3	X		X			X	X	
4							X	X
5					X		X	X
6			X		X			
7		X		X				
8		X		X				
9	X					X		

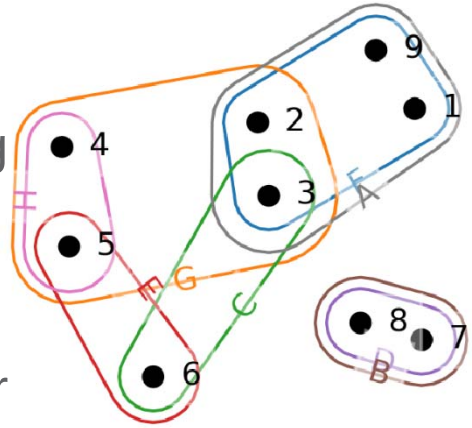




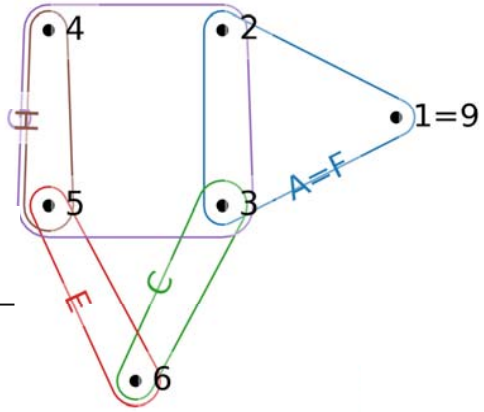


Collapsing

- **Multiplicity:** Source of weighting
  - **Multiple Edges:** Multiset to set
  - **Equivalent Vertices:** Characteristic vertex per wedge
- **Included Edges:** Not needed for s-components
- **Isolated Singletons:**
  - Count and discard
  - Represent default DNS activity



	A	B	C	D	E	F	G	H
1	X					X		
2	X					X	X	
3	X		X			X	X	
4							X	X
5					X		X	X
6			X		X			
7		X		X				
8		X		X				
9	X					X		

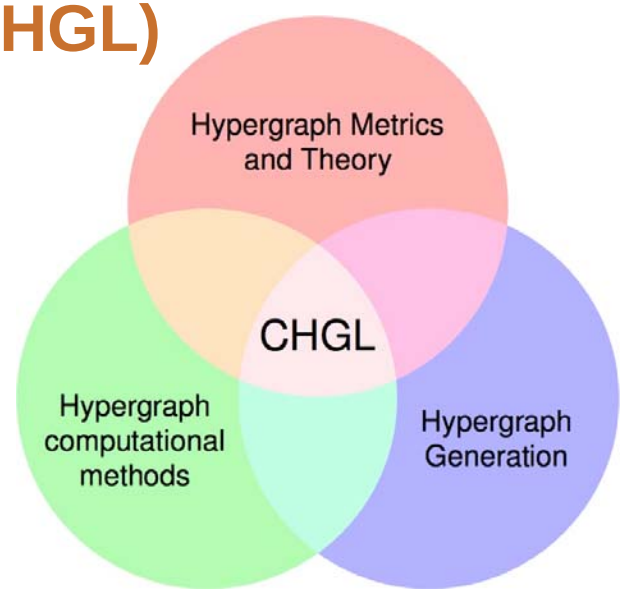


		2	2	1	1	1	1
		A=F	B=D	C	E	G	H
2	1=9	X					
1	2	X				X	
1	3	X		X		X	
1	4					X	X
1	5				X	X	X
1	6			X	X		
2	7=8		X				

	Initial	Collapsed	Non-Singleton Components
$\|V\|$	9	7	6
$\|E\|$	8	6	5
Aspect ratio	1.125	1.167	1.200
# Cells	23	14	13
Density	0.319	0.333	0.433

Chapel Hypergraph Library (CHGL)

- **Hypergraph Computation for HPC**
- **Chapel Programming Language:** Cray Inc.
  - Parallelism, HPC built-in
  - Modern look and feel
  - Strong typing, generics, etc.
- **CHGL Design Goals**
  - **Genericity**
    - ✓ Abstract interfaces describing classes of data structures
    - ✓ Reusable algorithms
  - **Performance**
    - ✓ Distributed-memory scalable performance
    - ✓ Rely on Chapel for the basics
    - ✓ Design efficient data structures and algorithms
  - **Usability**
    - ✓ Interface levels (e.g., simple for most tasks, advanced for customization)
    - ✓ Modern feel
    - ✓ User-centric: prioritize user experience over developer convenience



https://github.com/pnnl/chgl

DNS use case

- **Hypergraph:** IP X Domain
  - Nodes = IP addresses
  - Hyperedges = domain names
- **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
- **Extracted from records:**
  - `qname` = domain name requested
  - `rdata` = IP addresses resolved
- **Analytical Questions:**
  - (*general exploration*) Can we find IPs or domains that are used abnormally?
  - (*targeted exploration*) Given known bad IPs or domains what does their local hypergraph look like?
  - (*future*) Can we identify where there might be missing data?


```
{
    "date": "20161001",
    "qname": "gatech.edu.",
    "qtype": 1,
    "rdata": "130.207.160.173",
    "ttl": 300,
    "authority_ips": "128.61.244.253,168.24.2.35",
    "count": 80,
    "hours": 16710647,
    "source": "gt",
    "sensor": "active-dns"
}
```

ActiveDNS Data Format and Cleaning

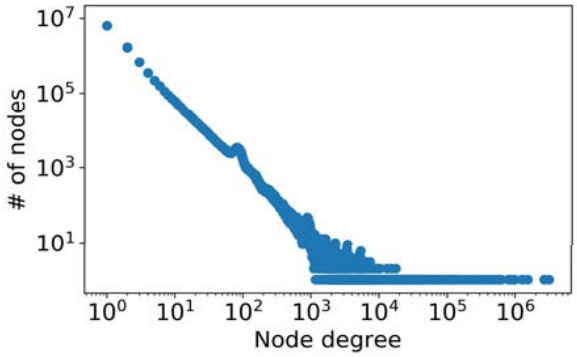
- 1200 files/day, Avro format, parsed into CSV
- We remove records with…
  - Empty `qname` or `rdata`
  - Domain names for the `rdata` (should be IPs)
  - IPs as `qname` (should be domains)
- **Analysis of Full Day:** April 26, 2018 (arbitrary)
  - Each file on this date has on average 900K records total
  - After cleaning each file has on average 180K valid records
- **Cluster:** 16 nodes, 20 processors/node, 768 GB/node
  - Currently single node operation
  - Future move to distributed memory

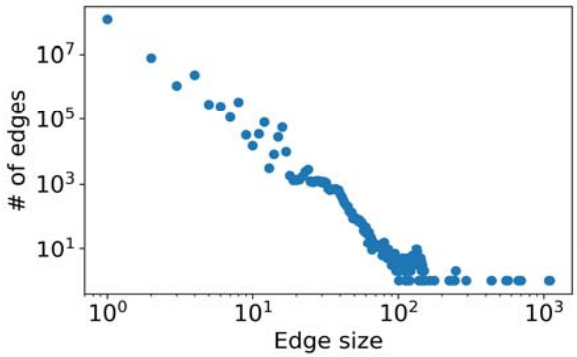
One Day's Data

	Initial	Collapsed	Non-Singleton Components
$\|V\|$	10.6M	10.3M	557K
$\|E\|$	131.2M	11.0M	1.2M
Aspect ratio	0.081	0.941	0.460
# Cells	157.4M	25.7M	15.9M
Density	1.14 E-7	2.26 E-7	2.35 E-5

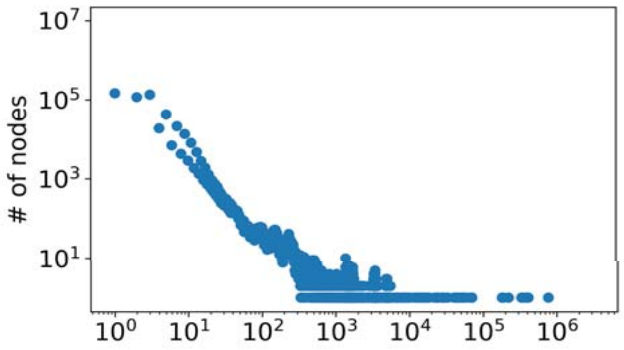
- 9.8M isolated singletons

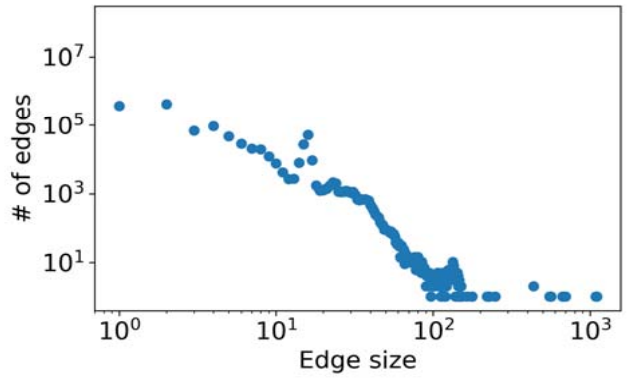
Original





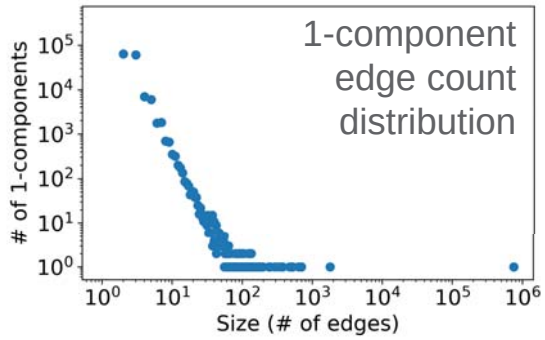
Fully Reduced

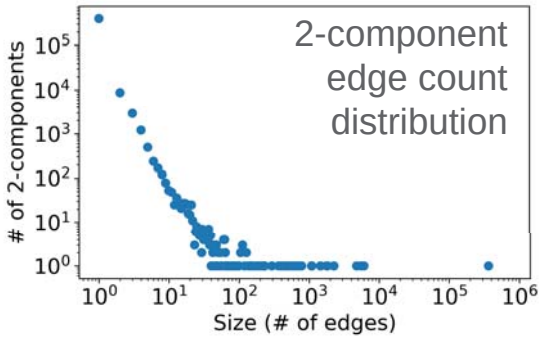


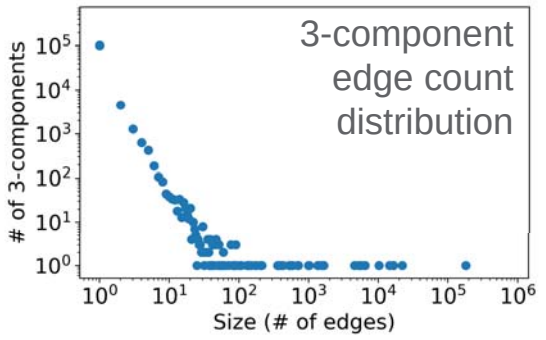


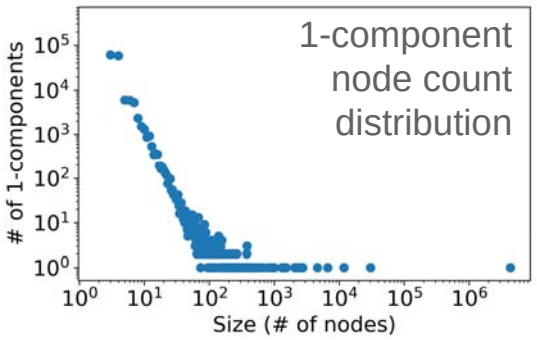
s-Component computation and size distributions

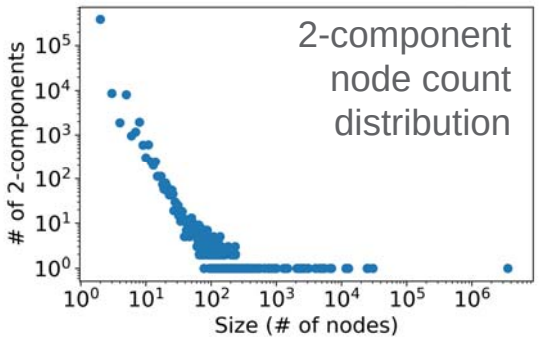
- 9.8M isolated singletons already removed
- *s*-Connected components (*s* = 1, 2, 3)
  - Dominates overall execution time (~60%)

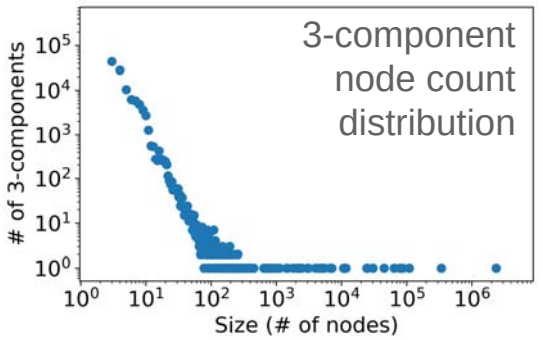






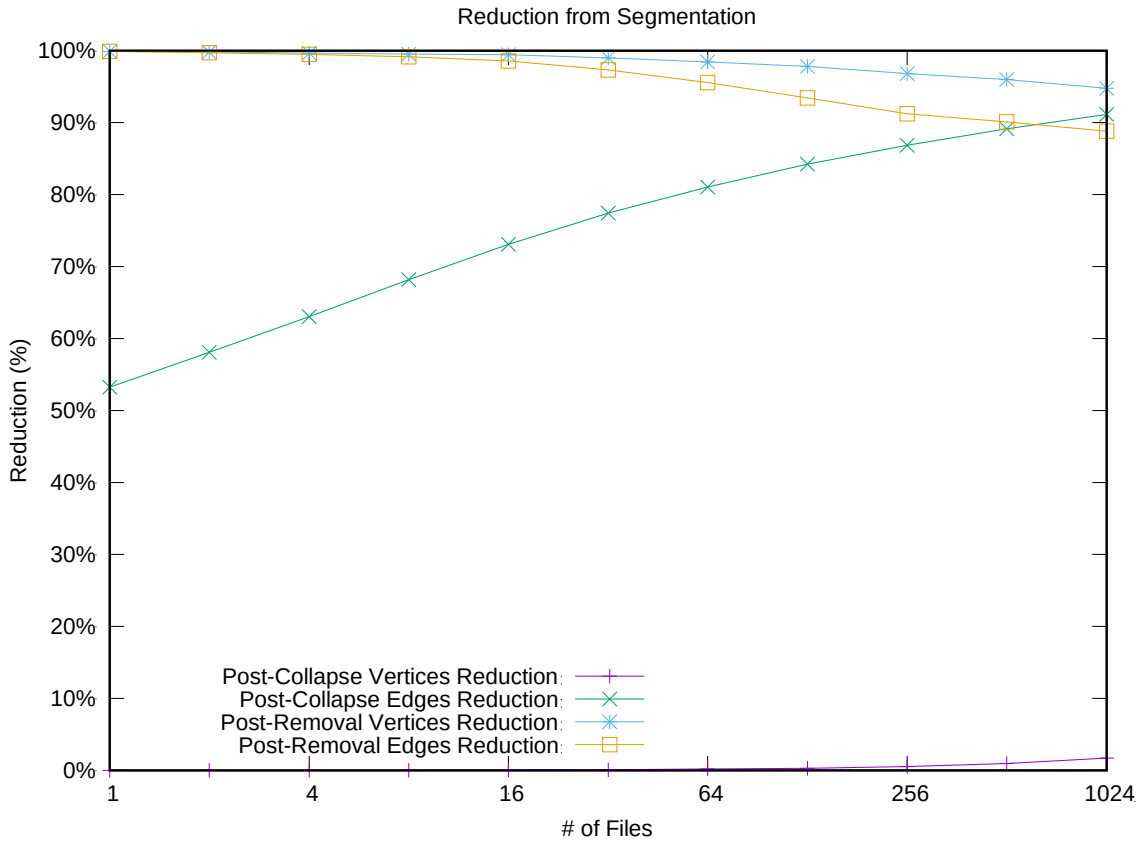






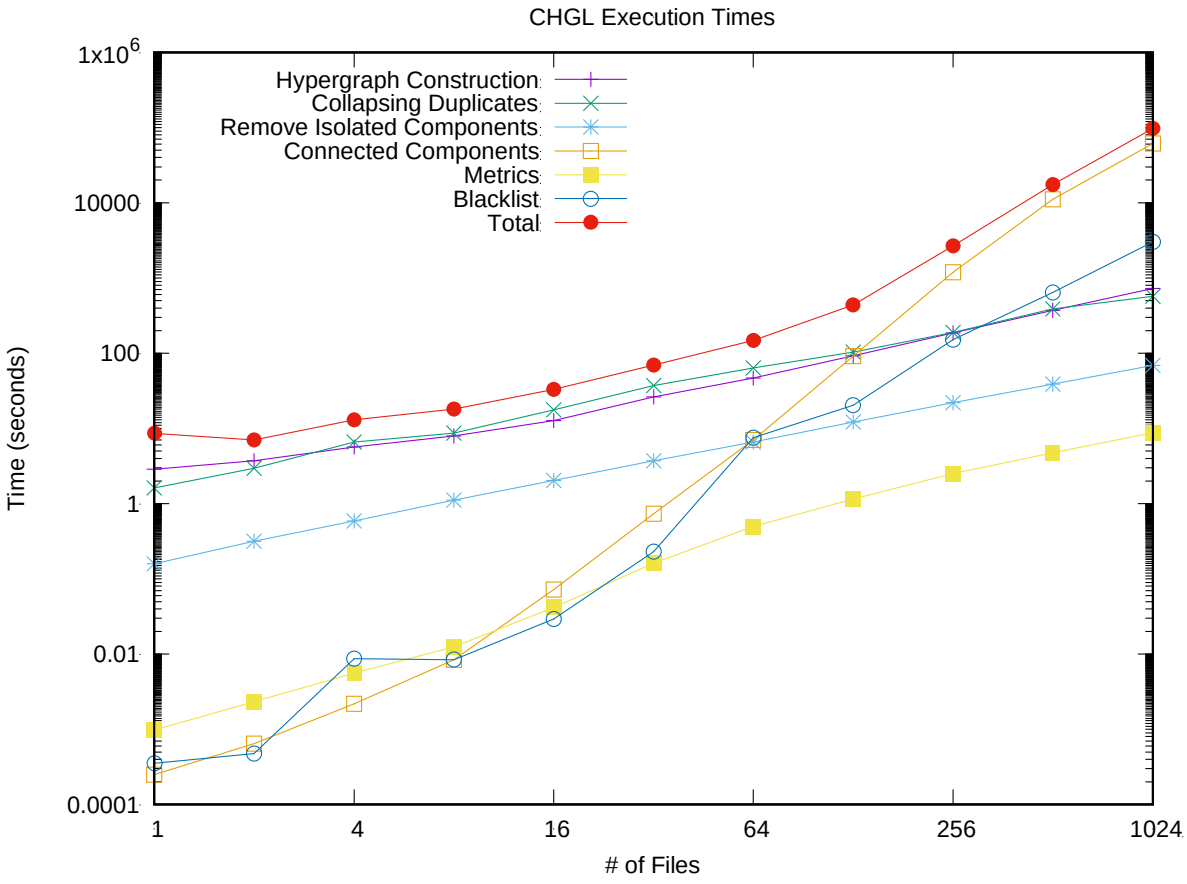
Analysis of Hypergraph Reduction

- Collapsing scales with size of graph
  - Not worth collapsing for nodes on small graphs
- Removing isolated singleton components degrades
  - Converges to true # of isolated singleton components



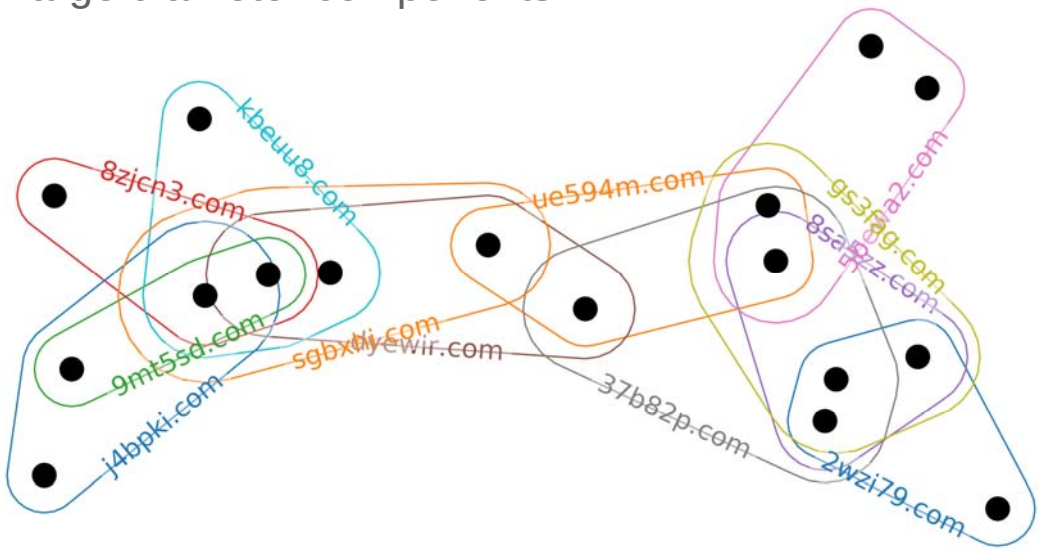
Performance Analysis: Initial Implementation

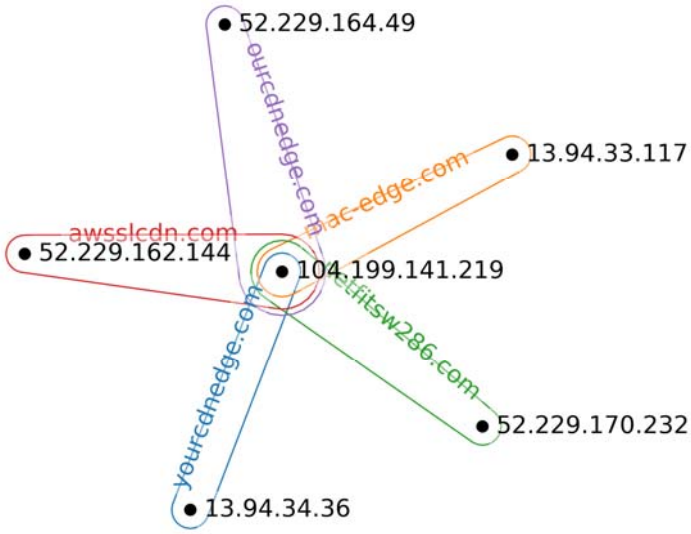
- Loading Hypergraph is only ~1% of execution time
- Connected Components is ~60% of execution time
- ~1 day compute time to process 1 day of data.
- Processed on one compute node
  - Distributed coming soon…
  - GPU support possible…



General exploration of *s*-components using HNX

- Resulting *s*-components are analyzed to find motifs of known behavior and outliers in certain hypergraph metrics
  - Star motif
  - Large diameter components

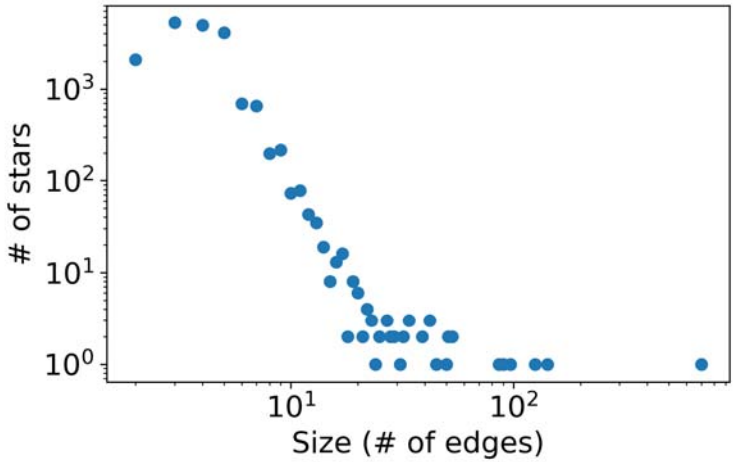


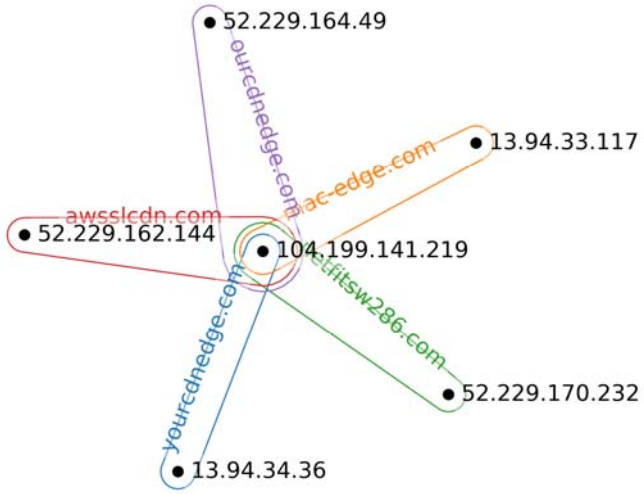


Star Motifs

All observations about DNS records and IP or domain registration were found using publically available services like WHOIS and BGP routing.

- Searched all 1-components for stars and computed their sizes (# edges)
- Largest star is outlier with 642 leaves, consistent with DNS sinkhole behavior
  - Central node 17.17.17.17 with start of authority (SOA) record proclaiming “sinkhole root@sinkhole”
  - Leaf nodes come from 640 distinct /16 (first two octets) IP ranges
- Smaller stars more consistent with content delivery networks (CDNs)
  - All IPs and domains within the same, or a relatively small set of, ranges and organizations
  - Example: Central IP address registered to Google Cloud, leaves registered to Microsoft Corporation. All five domains are registered through the hosting site GoDaddy.com.

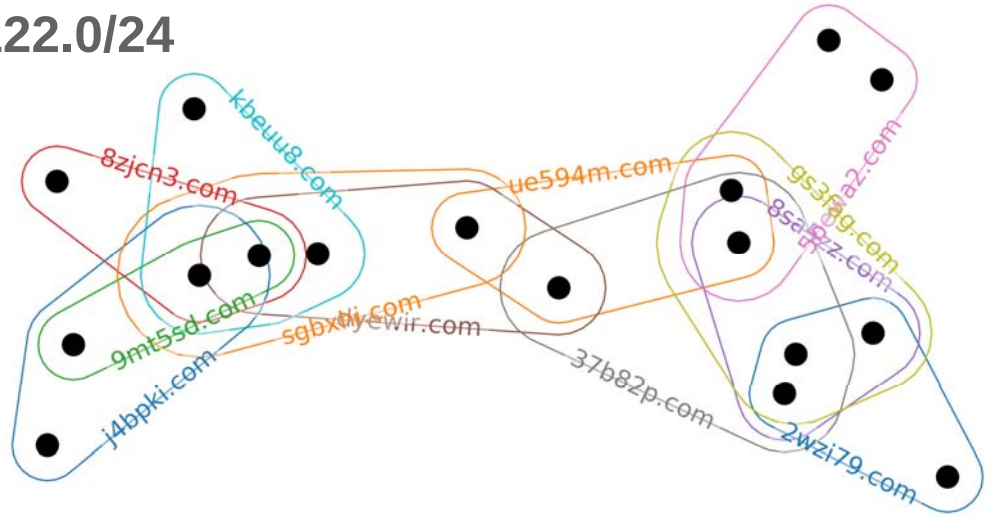




All observations about DNS records and IP or domain registration were found using publically available services like WHOIS and BGP routing.

Large diameter components

- **Computed diameter of all 2- and 3-components**
- **Max diameter (6) 2-Component:** Consistent with *fast flux* behavior
  - Relationship between IP and domain is very short-lived
  - Used by botnets to hide malicious content delivery sites and make malware networks more difficult to discover
- **All domains with IPs in 103.86.122.0/24**
  - In late NOV 2018 the IPs were 103.86.123.0/24 with time to live (TTL) of 120 seconds
  - Now many of the domains have no associated IP addresses



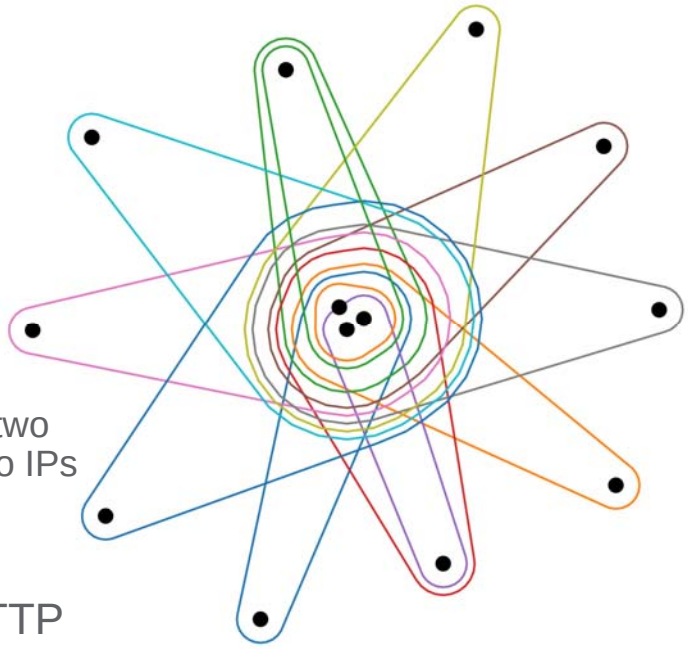
Targeted exploration starting with CHGL

- FireEye Threat Research Blog
  - “On the Hunt for FIN7: Pursuing an Enigmatic and Evasive Global Criminal Operation”
  - Contains list of ‘blacklisted’ IP addresses and DNS names
    - ✓ Regex for DNS Names: [a-zA-Z]{4,5}\.[pw|us|club|info|site|top], e.g., pvze.club
- Use a PropertyMap to map DNS names to edges and IP addresses to nodes
  - Constructed with initial hypergraph for entire day of data
  - Updated appropriately with each phase of segmentation
- After segmenting, scan PropertyMap for blacklisted IP or DNS Names
  - If found, print out the IP or DNS Name, its s-neighborhood and the connected component it is in for s = 1, 2, 3
  - Passed to HNX for processing

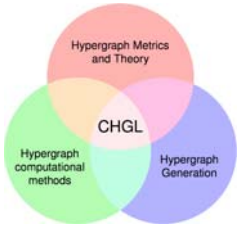
All observations about DNS records and IP or domain registration were found using publically available services like WHOIS and BGP routing.

Targeted components in HNX

- Many of the domains found by CHGL are in the largest *s*-component and likely are not connected to one another within that component
- Found a set of ten domains that follow the known regex pattern and are all contained within the same small 2-component (16 edges) and 3-component (13 edges)
  - The 3-component is nearly a star
  - No common intersection among all domains although there are two central IPs with each domain containing at least one of these two IPs
  - All domains in this component are registered by “Chengdu west dimension digital”
- Targeted analysis could be used to discover how known TTP signatures may be present within a data set.



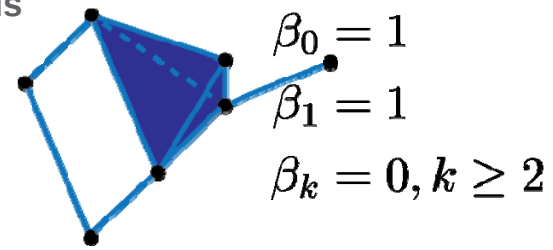
Future Work



https://github.com/pnnl/chgl

https://github.com/pnnl/HyperNetX

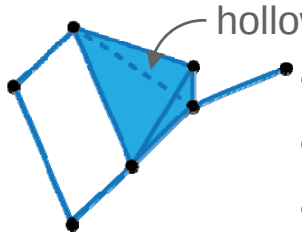
- **Theory**
  - **Hypernetwork paper pending!**
  - **Multiplicity weightings**
  - **Spectral approaches**
- **Methodology**
  - **Application of hypernetwork science methods:** Centrality, connectivity, clustering coefficients
  - **Topology and homology:** "Gap identification"
  - **"Pivot" ability to more data fields**
- **HNX**
  - **Released, testing underway**
  - **Topology and homology**
  - **Interactive visualization**
- **CHGL**
  - **Distributed memory hash table (property map)**
  - **Topology and homology**
  - **Chapel graph library also in development**
- **Applications**
  - **Ground truth V+V**
  - **Expand to other cyber data sets**
  - **Non-cyber applications:** Computational biology, social hypernetworks



$\beta_0 = 1$

$\beta_1 = 1$

$\beta_k = 0, k \geq 2$



$\beta_0 = 1$

$\beta_1 = 1$

$\beta_2 = 1$

$\beta_k = 0, k \geq 3$

- $\beta_0 = 2$

$\beta_k = 0, k \geq 1$

Pacific Northwest
NATIONAL LABORATORY

THANK YOU!