SEMANTIC REPRESENTATIONS OF NETWORK FLOW: A PROPOSED STANDARD WITH THE WHAT, THE WHY, AND THE HOW

Eric Dull, Rachel Kartch, Robert Techentin

Agenda

- Background: Eric Dull, Cray
- Key decisions: Rachel Kartch, SEI
- Real-life example: Robert Techentin, Mayo Clinic

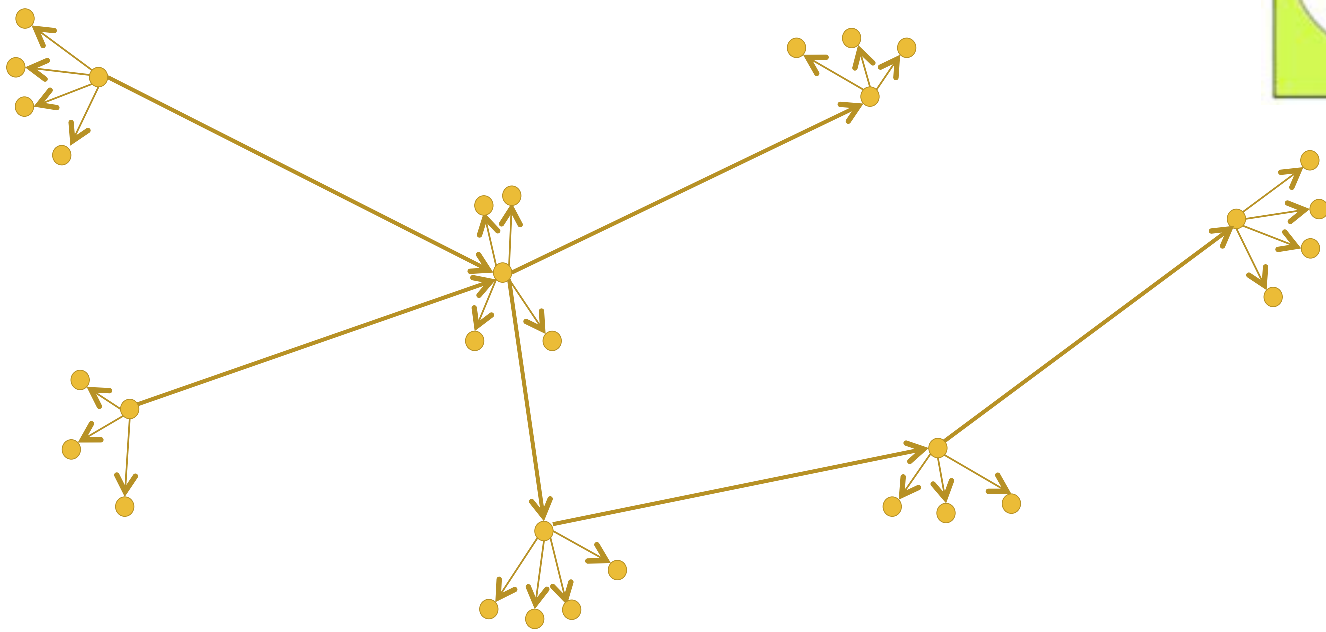
A vendor, an FFRDC, and a medical center walk into a bar…

Background: Why Graphs? Why RDF? Why OCOG?

Eric Dull, Cray

What is a Graph?

- Not a chart
- A fundamental data structure
- A collection of vertices (nodes) and edges (links, relationships, connections)



Why use a graph and graph engines

- **Graphs**
  - Native representation of underlying data
  - Wealth of available algorithms that address analytic questions
  - Data merges for free (more on this on the next slide) [scaled by O(# values present)]
  - Available open source and commercial engines for 10**0 to 10**12 flows
- **Graph engines**
  - Ease of "follow the flow"
  - SPARQL = SQL for graphs
  - Ease of use without professional programmer present

Why use RDF and ontologies

- **RDF**
  - ?s ?p ?o = any structured data is representable in
  - "schema-less" = human readable and definable
  - Representing the entities the same way in multiple data sets makes your data merge
  - Adding new entity- and relationship-types can be done when needed, not only at outset

- **Ontologies**
  - Easily represent the relationships between different types of edges and flows in data -- "find me all servers". Ontologies allow you to represent multiple types of servers and ask about all of them at the same time.
  - Enable standardization of common data types and thus sharing of ETL and analytic software

What is OCOG?

- Cray has observed users reinventing the wheel, each attacking the same problem (how to represent cyber data) before being able to get down to analysis
- In April of 2014, Cray invited representatives of Mayo Clinic, SEI, PSC, and other interested parties to build a standard ontology for cyber-data
- The goal was to eliminate the duplication of effort, and to ease information sharing with a standardized format
- The group first met in Pittsburgh in June of 2014, and chose the name OCOG: Open Cyber Ontology Group

OCOG: Key Decisions

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Rachel Kartch
January 2015

Copyright 2014 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

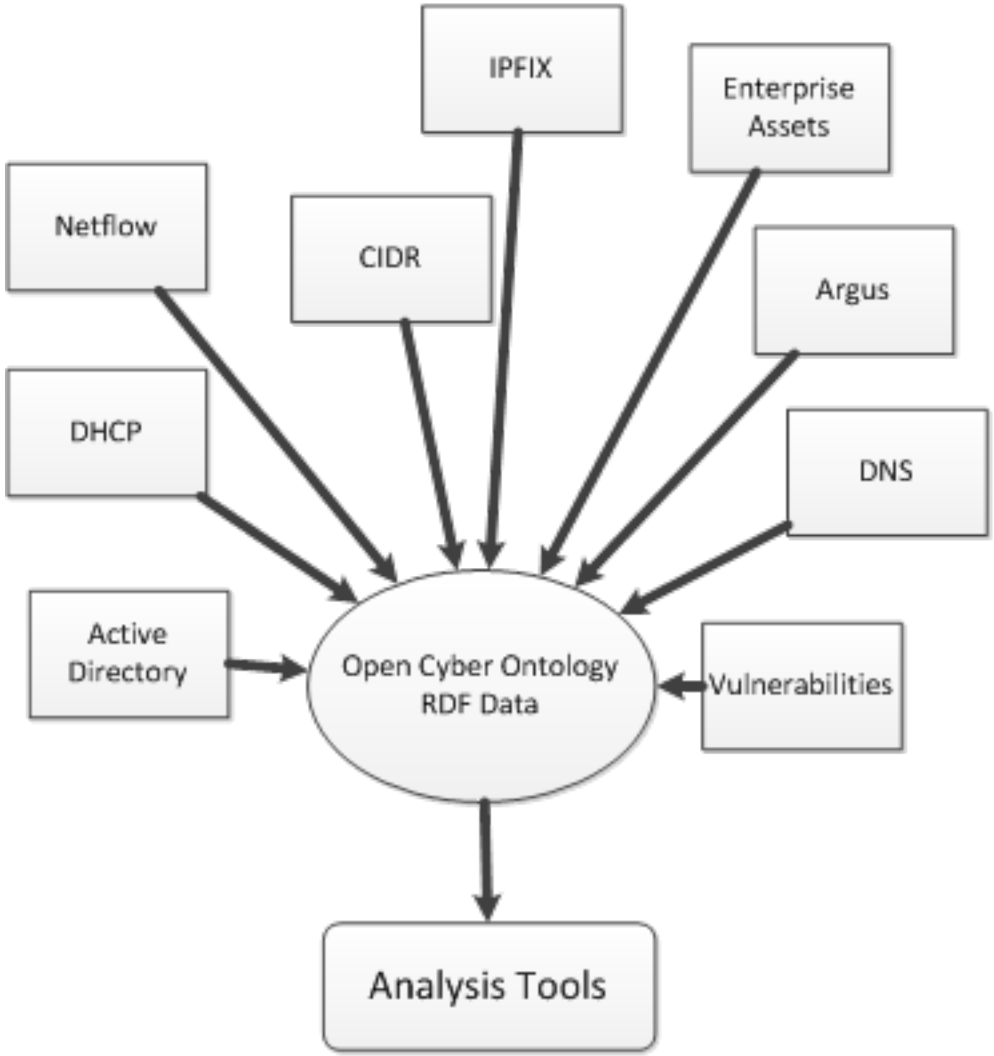
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense.

DM-0001856

Cutting to the chase

http://opencog.net

We have a dream



Key decision 1: Where do we start?

The goal: Rich representation of network objects and events

The challenge:

- Simple transactional data lacks context
- Enrichment data provides context, but is less regular and predictable

The decision: Begin with network flow, which is regular and predictable; build additional ontologies to bolt on and provide context

Key decision 2: What flow standard do we use?

The goal: Base the ontology on a relevant standard already in use, and likely to remain in use for the foreseeable future

The challenge: Multiple standards are in use, most notably NetFlow v5 and v9, Argus, and IPFIX

The decision: Base the ontology primarily on IPFIX, an extensible, open standard that provides support for IPv6, MPLS, multicast, etc. (also mapping to NetFlow v9 field names where possible); this doesn't mean IPFIX will be the only flow standard we can incorporate into the ontology, it's just the first one

Key decision 3: How do we identify a flow?

The goal: Create a flow record identifier that is computable, consistent, and collision-free

The challenge:

- What is the most minimal approach that will result in a sufficiently unique UID?
- How do we balance between readability and uniqueness?

The decision: UID is a 128-bit MD5 hash of 5-tuple plus start time, base64-encoded into an RDF-compatible string

<http://opencog.net/flow#MjA0ODc1MTIzMzA3NTA1MT>

Key decision 4: Which fields are core?

The goal: Identify the set of fields that are required to be present in any RDF representation of network flow

The challenge: There are over 400 IPFIX information elements

- Requiring too many to be present creates storage and verbosity problems
- Too few will limit analysis

The decision: Multiple levels of fields are defined, including a minimum Core level, and a more complete Standard level

Field classes

- Level 1: Flow Identity
  - 5-tuple + start time
- Level 2A: Flow Quantity
  - Level 1 + packet count, byte count, duration
- Level 2B: Protocol-Specific
  - Level 1 + TCP flags, ICMP type code
- Level 2: Flow Detail
  - Level 1 + Level 2A + Level 2B
- ✓ Level 3: Standard
  - Level 2 + exporter/collector, conversion, file, and ontology versioning information
- Level 4-99 (for future work)

Semantics, Ontologies and Graph Analytics for Cyber Defense of a Large Medical Center

Bob Techentin
Mayo Clinic

FloCon
January 13, 2015

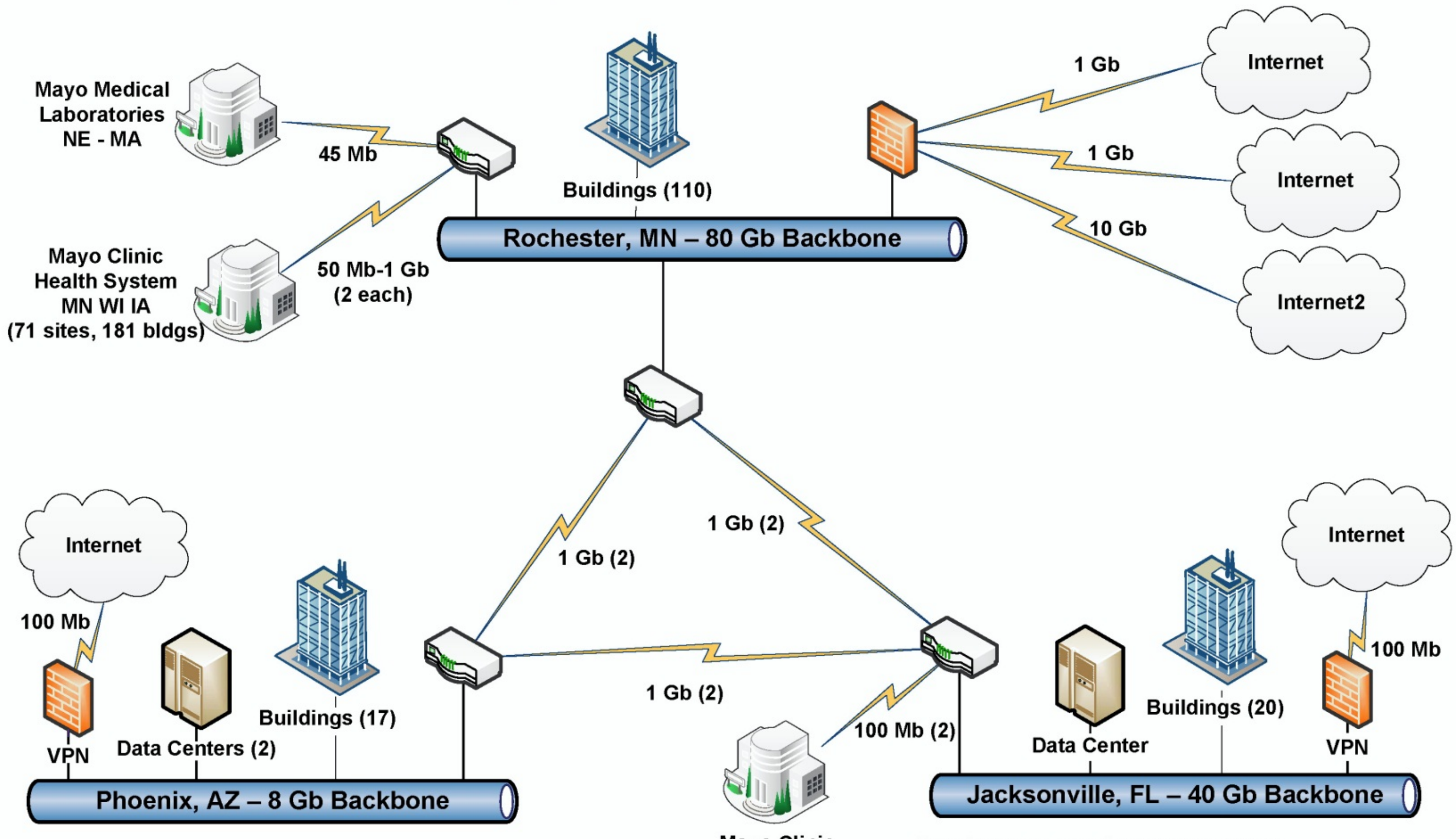
Teamwork

- Special Purpose Processor Development Group
  - Barry Gilbert, Ph.D.
- Biomedical Imaging Resource
  - David R. Holmes, III, Ph.D.
- Office of Information Security
  - Jim Nelms

Will and Charlie Mayo, The Mayo Brothers

MAYO CLINIC NETWORK OVERVIEW

(117,000 Networked Devices; 370 Routers; 4,300 Switches; 5,600 Wireless Access Points; Spanning 330 Buildings in 7 States; Over 55,000 Employees; 1.1 Million Patient Visitors Per Year)



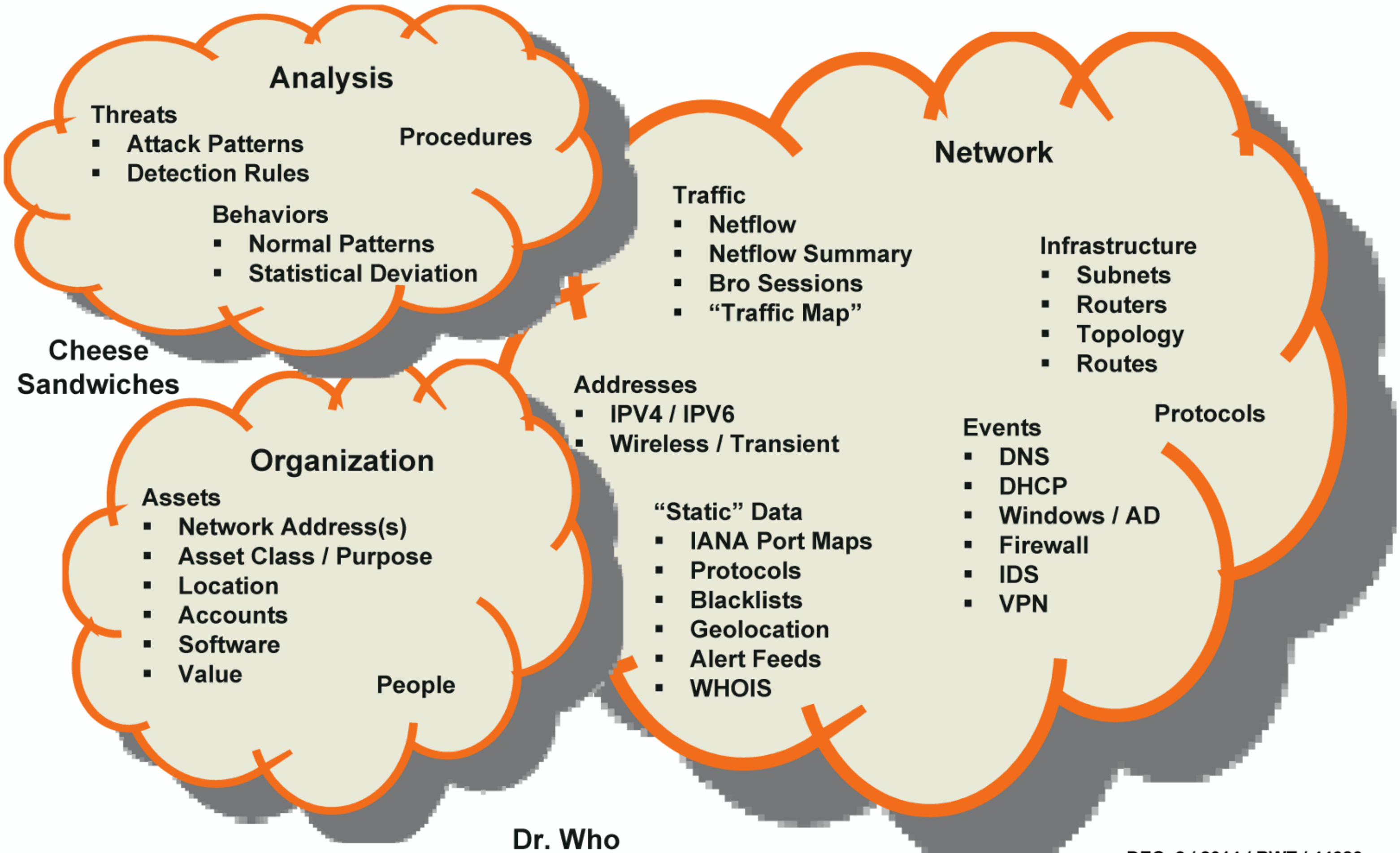
Gb: Gigabits Per Second Network
Mb: Megabits Per Second Network
VPN: Virtual Private Network for remote access

SEP_16 / 2013 / RWT / 44212
MAYO CLINIC
SPPDG

How To Sift Through Terabytes of Data to Find the Needle in the Haystack?

- Graph analytics is one approach
  - Worth investigating when relational databases struggle with massive irregular, noisy datasets
  - Fen-Phen is one example
- Engaged Cray in 2005 regarding XMT, descendent of Tera Computer Company's Multi-Threaded Architecture
  - "Grace", a hybridized 64 Processor XMT-2, now resides at Mayo
  - Now actively applying "graph analytics" to clinical data
- Cyber traffic analysis has similar characteristics
  - Huge irregular datasets, often incomplete, from disparate sources

POTENTIAL DOMAIN ELEMENTS FOR CYBER ONTOLOGY
(Lists Are Notional; Some Elements Are Essential; Others Must Be Excluded)

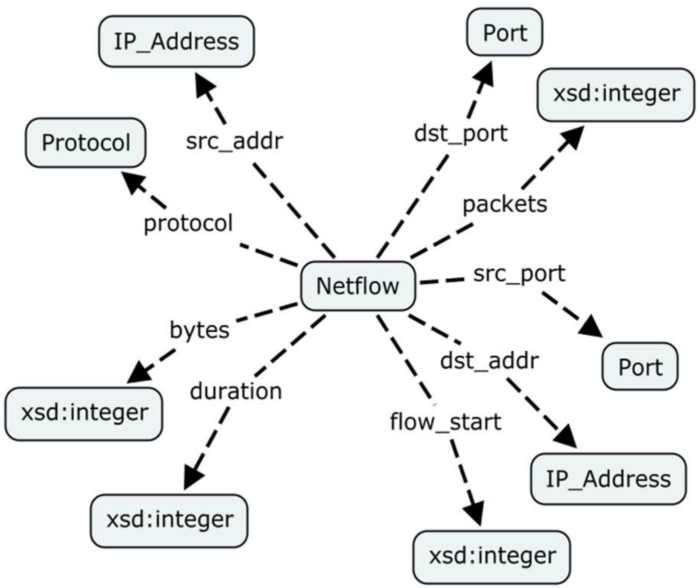


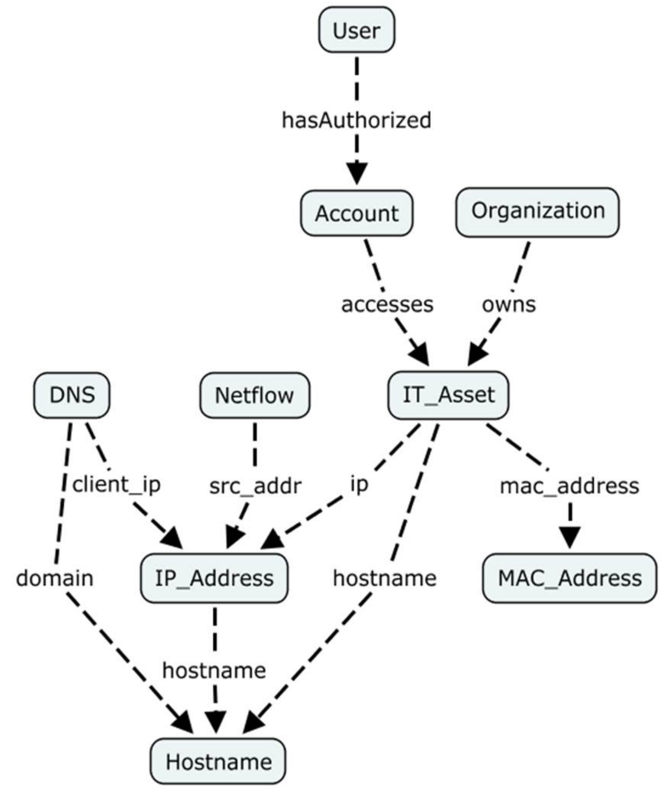
Mayo Clinic Cyber Model (MCCM) and Open Cyber Ontology Group (OCOG)

- Mayo began developing MCCM in 2013
  - Initially captured Netflow, DNS, DHCP plus network structure and enterprise data
  - Defines relationships between (for example) IANA port numbers and assets owned by different business units
- However, Mayo and Cray (and others) had different approaches and naming conventions, even for simple things like port numbers
- OCOG allows sharing a common ontology for common concepts: i.e., don’t reinvent the wheel

SOME COMPONENTS OF MAYO CLINIC CYBER MODEL (MCCM)
(Resource Description Format (RDF) Ontology Links Transactional and Enterprise Data Sources)

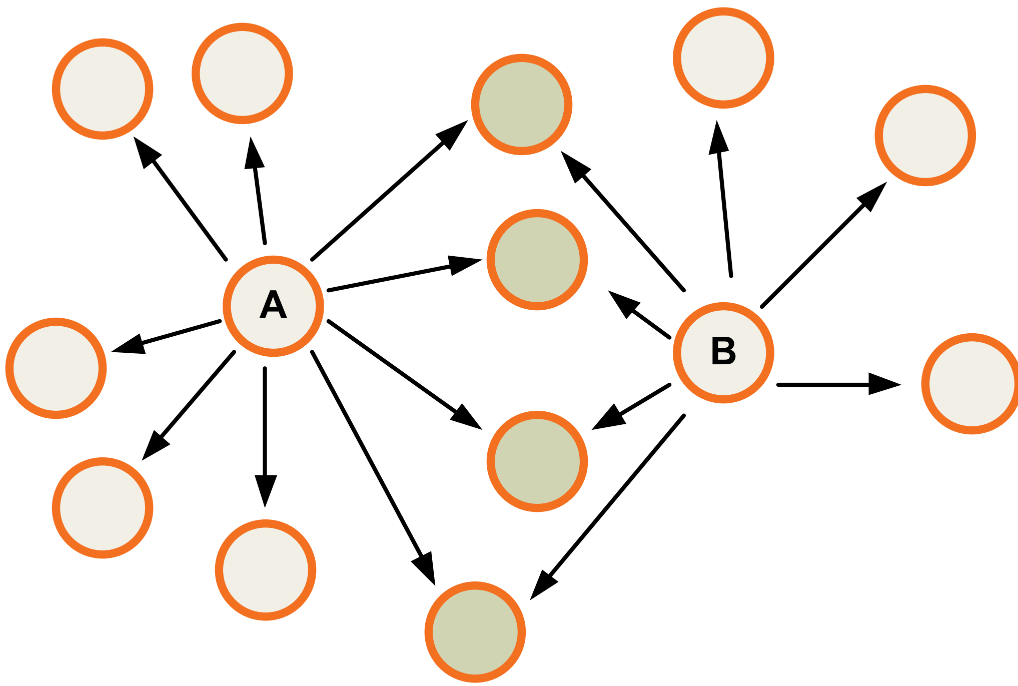
Netflow records are translated into RDF graph of relationships between addresses, ports, protocols, and traffic measures.





Integrating Transactional Data (e.g., Netflow, DNS, AD Logs) With Enterprise Data Enables Rapid Analysis Across Many Variables

SUBGRAPH SIMILARITY MEASUREMENT BY JACCARD INDEX
(Similarity of Graph Vertices and Edges Based on Set Theory)



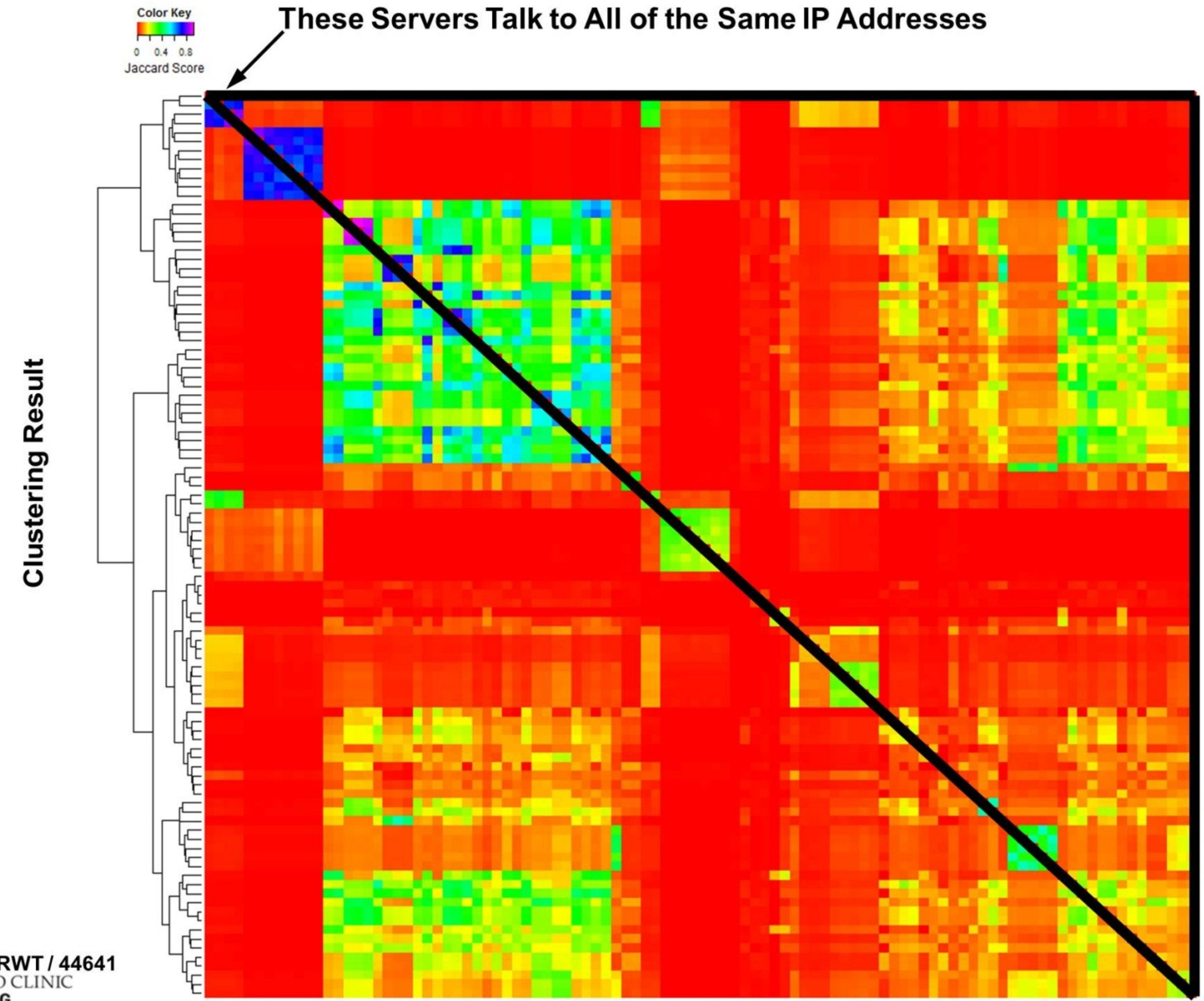
The Jaccard Index Measures Subset Similarity as the Ratio of the Number of Elements in the Intersection and the Union

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

There are Several Options For Semantic Graphs

- **Count Typed Edges**
- **Count Unique Edge Types**
- **Count Incoming vs. Outgoing Edges**
- **Count Vertices**
- **Count Vertex Types**

NETWORK TRAFFIC SIMILARITY ANALYSIS FOR 100 IP ADDRESSES WITH HIGHEST TRAFFIC
(Jaccard Similarity Score Computed on IP Traffic Destinations for 100 Servers on MTA Side of Hybrid XMT-2 Supercomputer at Mayo Clinic; Hierarchical Clustering of Similarity Scores Computed on XT5 Blades in Hybrid XMT-2; Clustering Reveals Similar Behavior; Intensity Indicates Similarity Score)



OCT_07 / 2014 / RWT / 44641
MAYO CLINIC
SPPDG

Conclusion

- Graph analytics can find meaningful relationships in huge and complex datasets in the cyber realm
  - Successful searches and analyses at Mayo Clinic (and elsewhere) demonstrate utility of the approach
- A standard ontology can describe many aspects of computer networks and behaviors
  - Transactional sources (e.g., Netflow, DNS, AD, Firewall)
  - Static and structural sources (e.g., port assignments, CIDR blocks)
  - Common patterns for behaviors of interest
- Mayo Clinic has already performed some real-world analysis of network activity using RDF and graph analytics