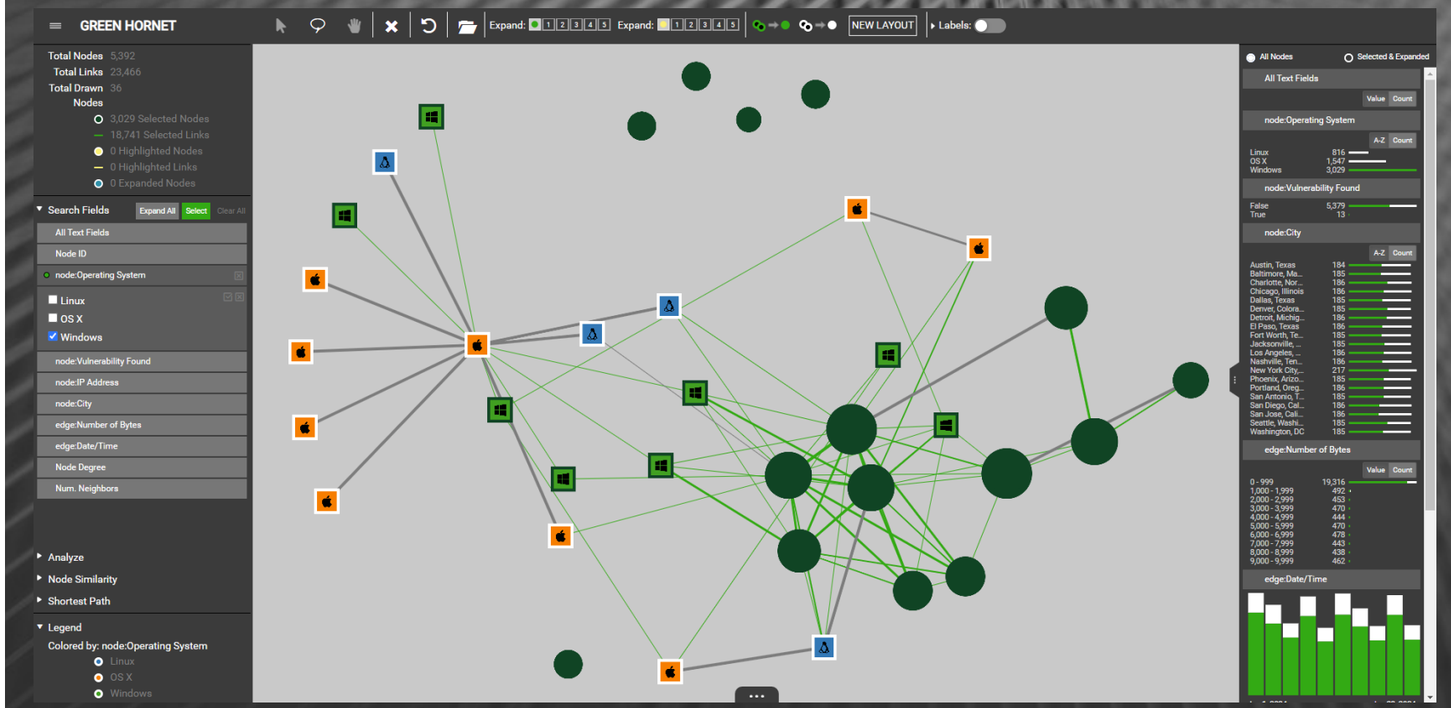
Pacific Northwest
NATIONAL LABORATORY

Green Hornet: Visual Analytics for Large Graphs

Patrick Mackey, Nicholas Cramer, Dave Gillen, Doug Love, Liz Faultersack



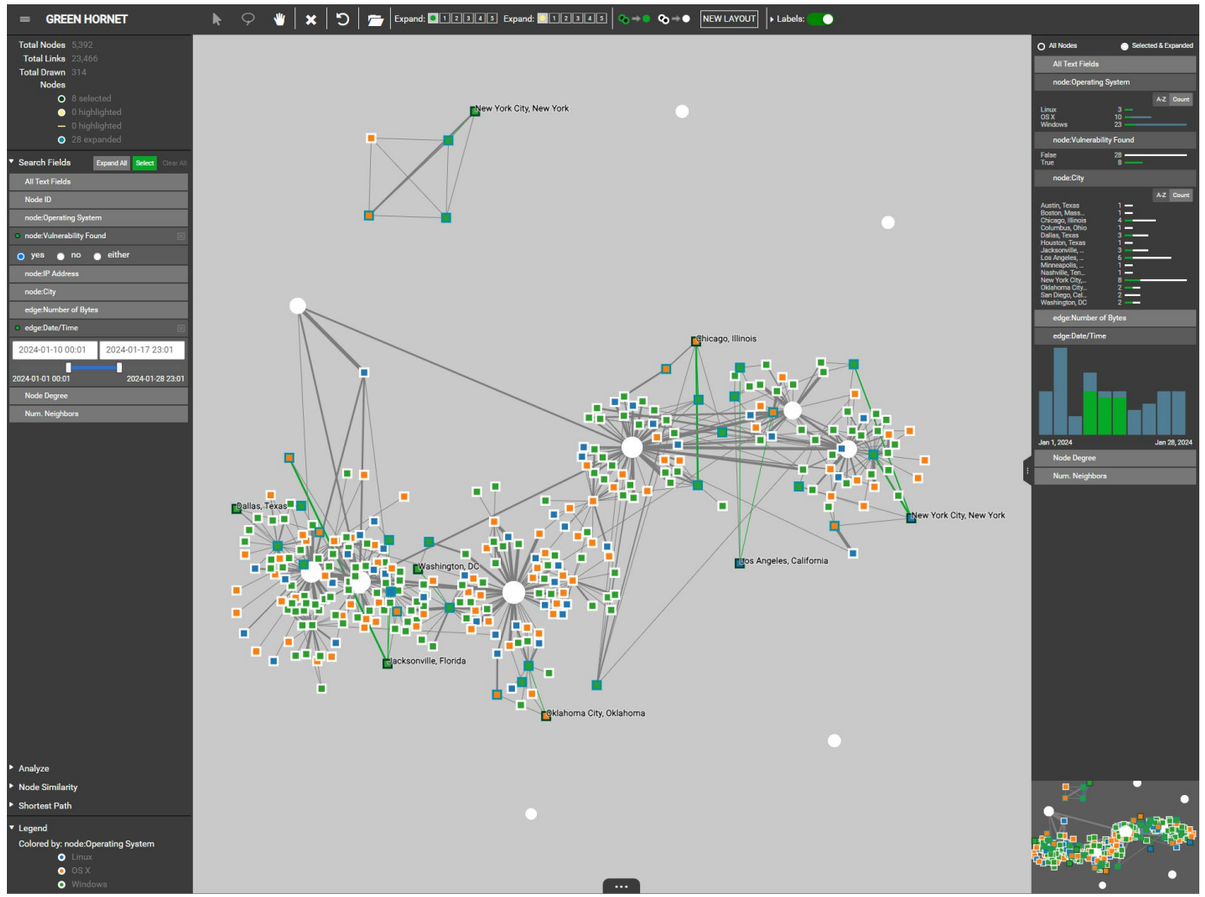
U.S. DEPARTMENT of ENERGY BATTELLE

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

Pacific Northwest NATIONAL LABORATORY

What *is* Green Hornet?

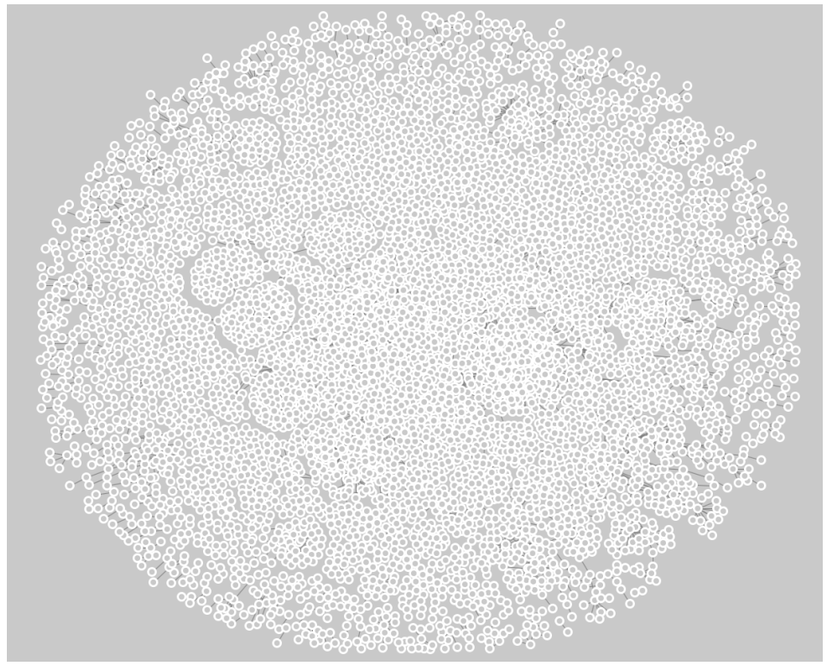
- Green Hornet is a graph visual analytics tool for enabling analysts to explore graphs with up to **1 million** nodes and edges.
- It was designed to be a powerful graph analytics tool *without* requiring expertise in mathematics or data science.
- It makes such large graphs understandable through use of “**supernodes**”.



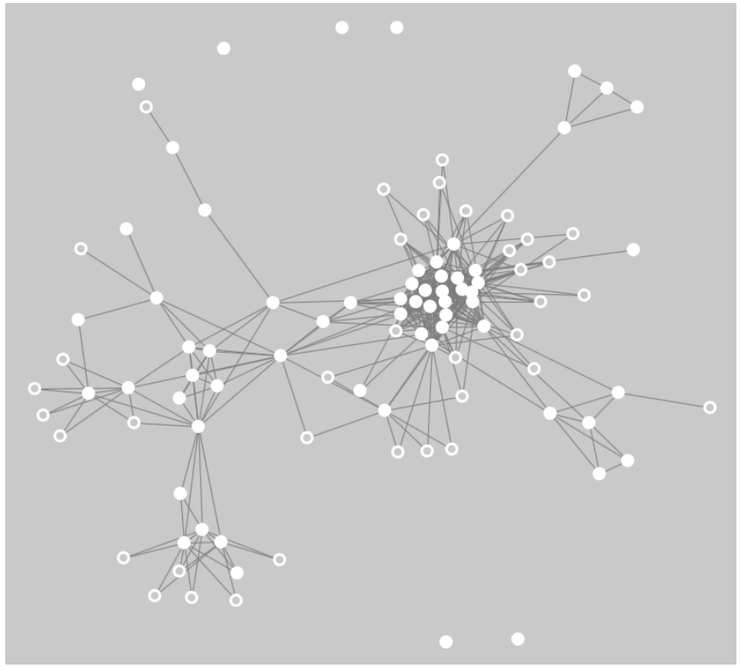
Supernode Based Approach

- Large graphs quickly become “hairballs” as the number of nodes and links increases
- Green Hornet uses **supernodes** to cluster groups of nodes
- The resulting graph has much less clutter, and enables users to see the underlying structure better

Full Graph



Supernodes



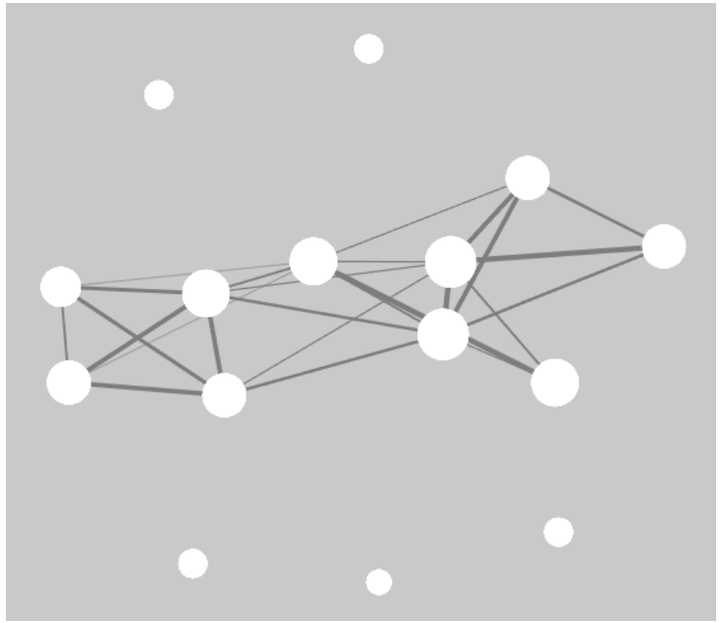
The same graph, before and after clustering.

The structure becomes much more apparent with the use of supernodes.

Graph size: 6,396 nodes, 9,368 links

Pacific Northwest NATIONAL LABORATORY

Clustering Options



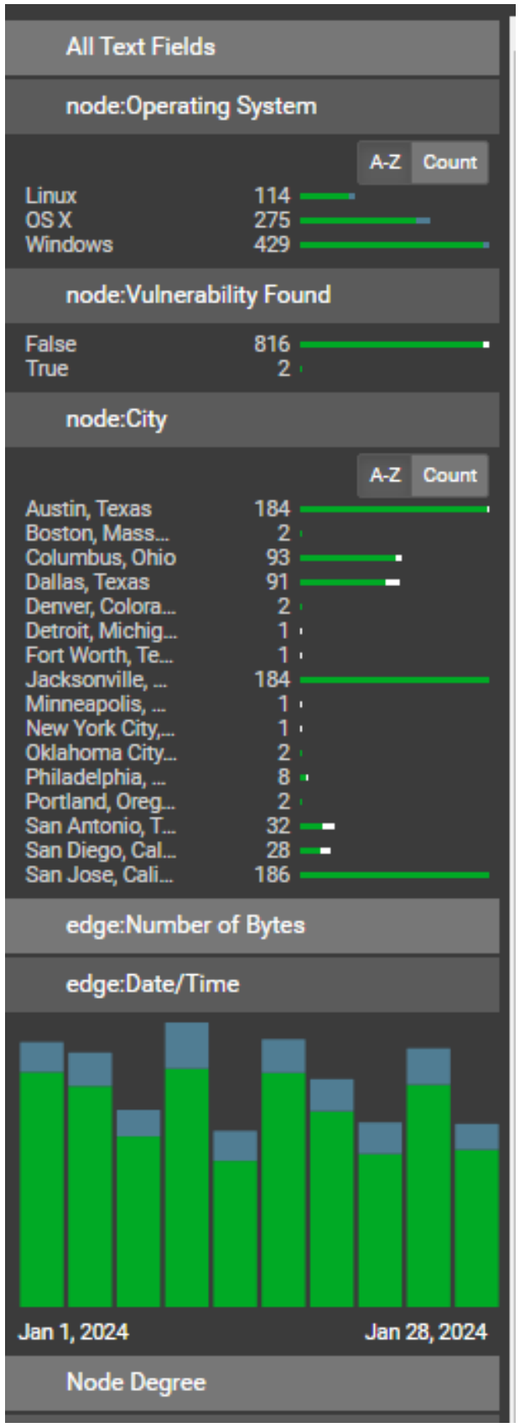
Different options are available for clustering, including:

- Louvain community detection
  - Clusters nodes that are more connected to each other than the rest of the graph (e.g., tightly connected network resources)
- Structural equivalence
  - All nodes that share the same neighbors become clustered together (e.g., all accounts that interact with the same IP addresses)
- Node property-based approach
  - Any user-defined node property can be used to create the clusters (e.g., city associated with an IP address)
- Manually – via lassoing groups of nodes together
  - Allows users to create supernodes interactively as needed

Pacific Northwest NATIONAL LABORATORY

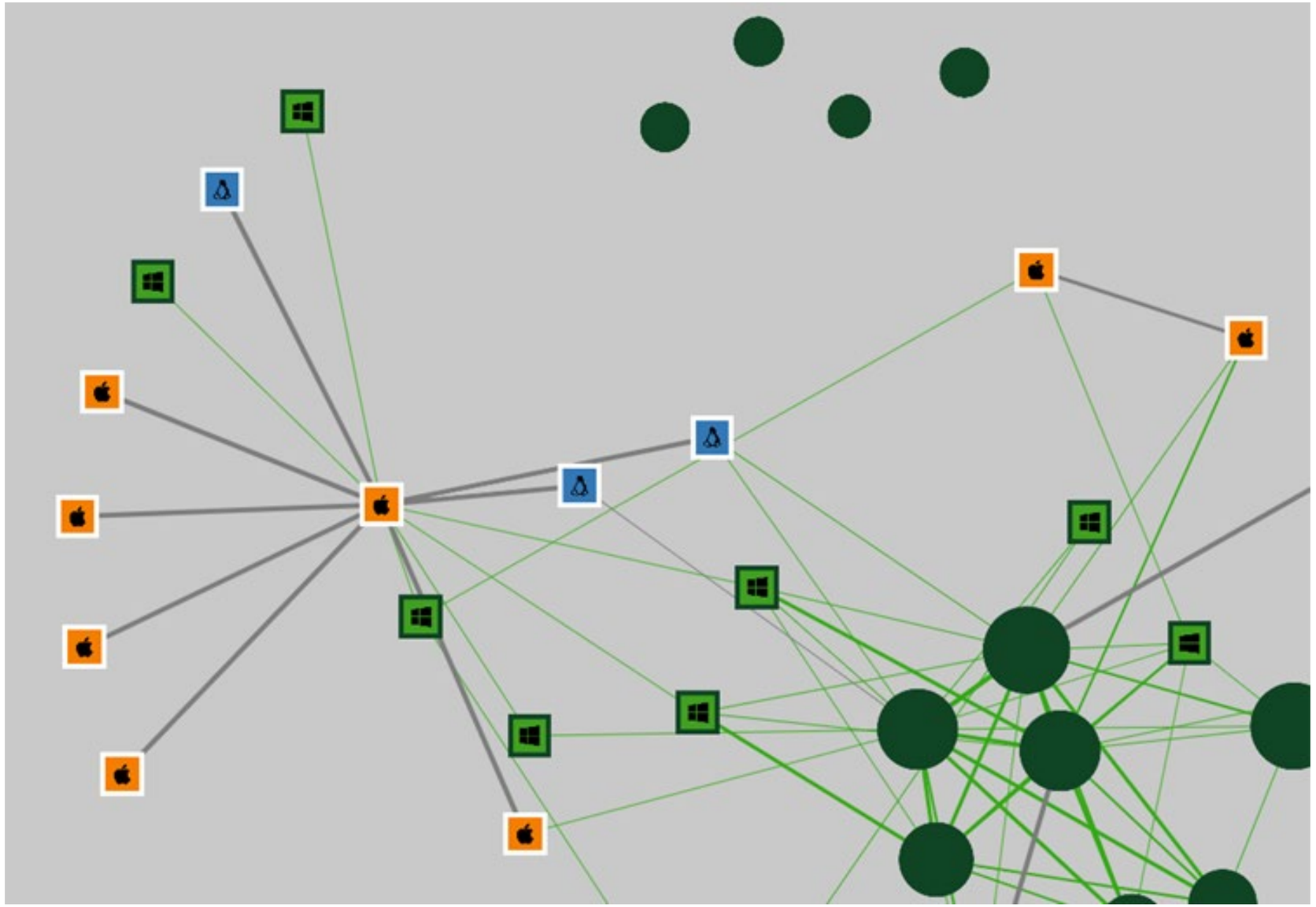
Node and Edge Properties

- Green Hornet is specifically designed to work for **property graphs**
- Many kinds of properties can be represented on nodes or edges:
  - Free text (e.g., error message additional details)
  - Categorical text (e.g., operating system, device browser)
  - Integers (e.g., number of log-in attempts, port address)
  - Floating point (e.g., importance scores, edge weights)
  - Booleans (e.g., true or false status)
  - IP addresses (including CIDR search capability)
  - Date/time
- Search capabilities provided for all properties
- Includes summary of graph properties
  - Available for whole graph or sub-selection



Iconography

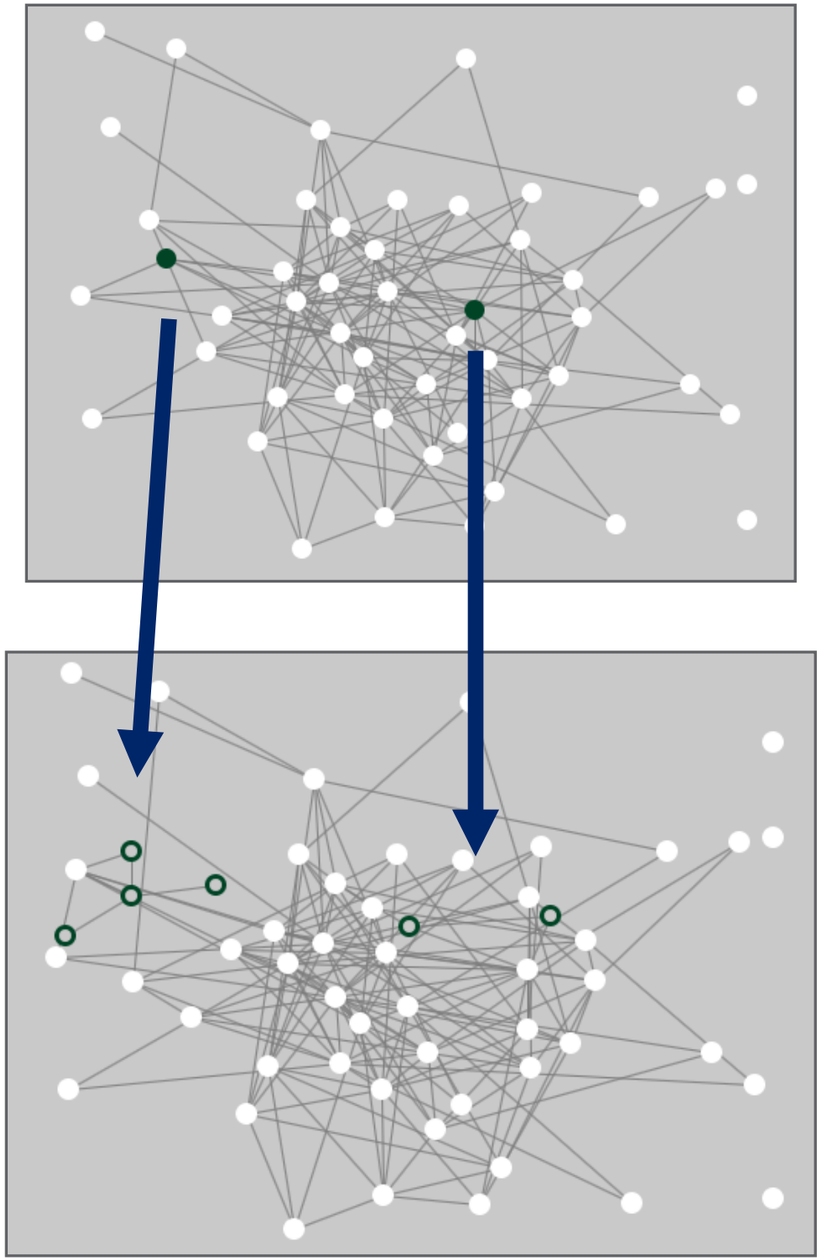
- Node type and properties can be represented through use of **color** and **icons.**
- These can be **customized** by the user to more intuitively represent important concepts in their data.



Pacific Northwest NATIONAL LABORATORY

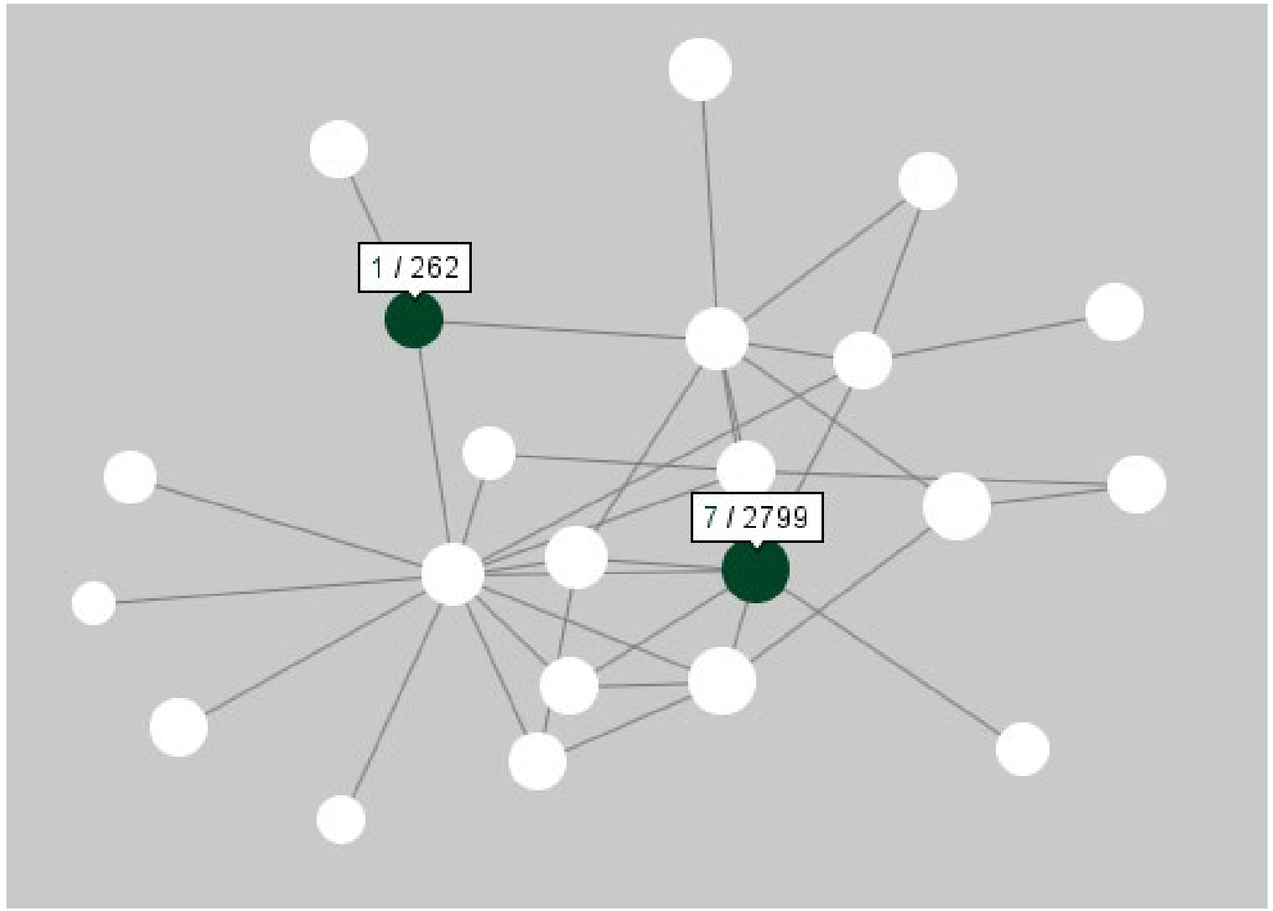
Interactive Exploration

- While supernodes allow you to see the high-level structure, users also want to investigate specific nodes and edges.
- This is typically done via a process of:
  - **Selection** (either manually, or by searching for specific node/edge properties, or other graph characteristics)
  - **Expansion** of specific nodes from their supernodes
- The goal is to enable users to simultaneously zoom-in to nodes of interest while maintaining high-level global view of the graph.



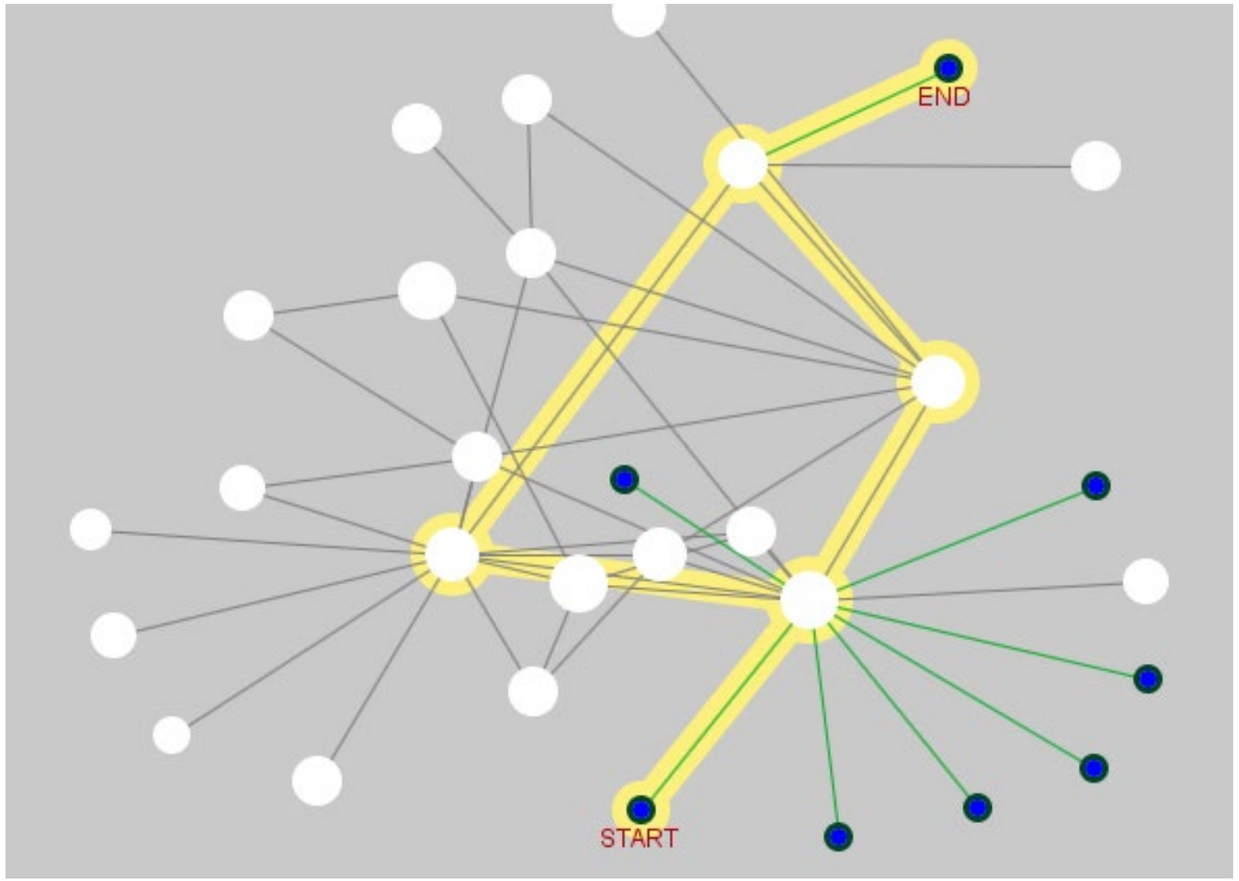
Graph Selection

- Node/edge selection is one of the primary ways users explore a graph
- Selections are made in one of the following ways:
  - Searching by node/link properties
  - Manual selection of nodes/links
  - Searching by graph patterns like shortest path
- Any supernode with at least one individual node/link is considered “selected” and shown in green
- Many of the features for graph exploration and visualization are based around which nodes and links you currently have selected
- Supernode labels show the number of selected nodes as well as the total number of nodes

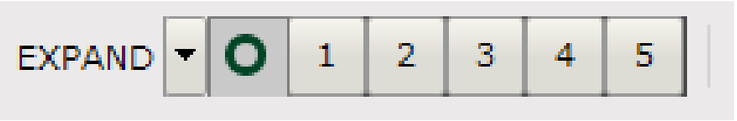


Highlighted Nodes/Links

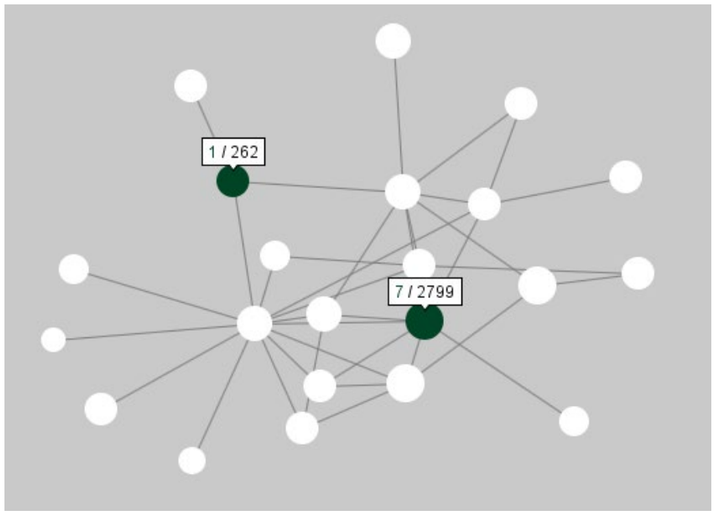
- In addition to the “selection”, users can also highlight nodes/links of interest
- This is primarily done through graph algorithms like:
  - Shortest paths
  - k-cores
  - Vertex similarity
- It can also be done through node and link properties
- Highlighted nodes/links are marked in yellow
- Highlighted nodes/links can be converted into the selection

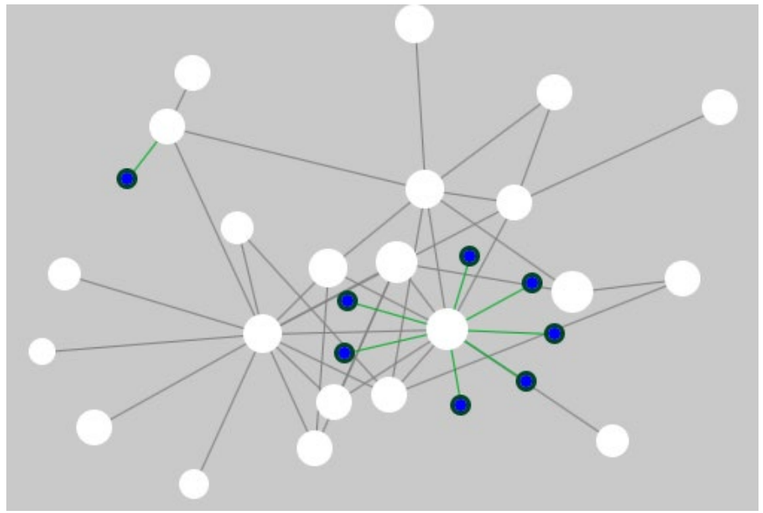


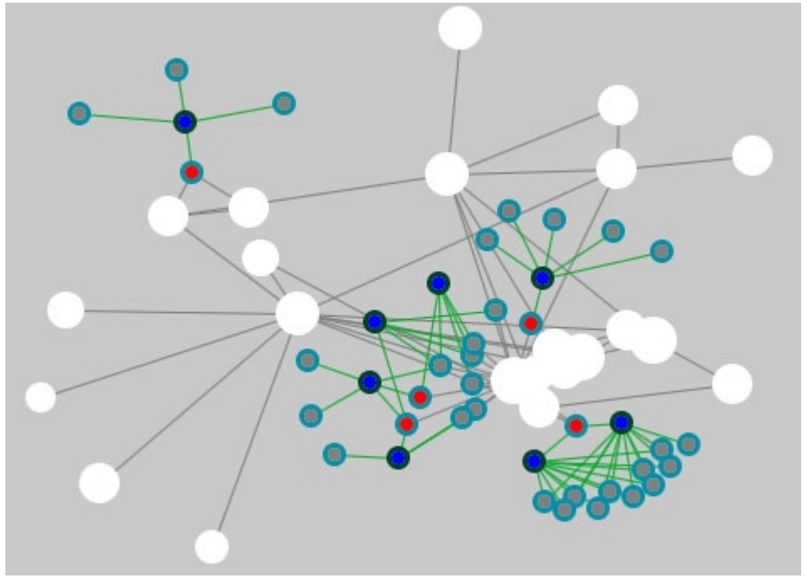
Expanding Nodes of Interest



- Both selected and highlighted nodes can be extracted from their supernodes
- This allows the user to see individual nodes of interest while keeping the rest of the graph clustered
- The neighbors of the selection/highlight can also be expanded up to 5 hops away

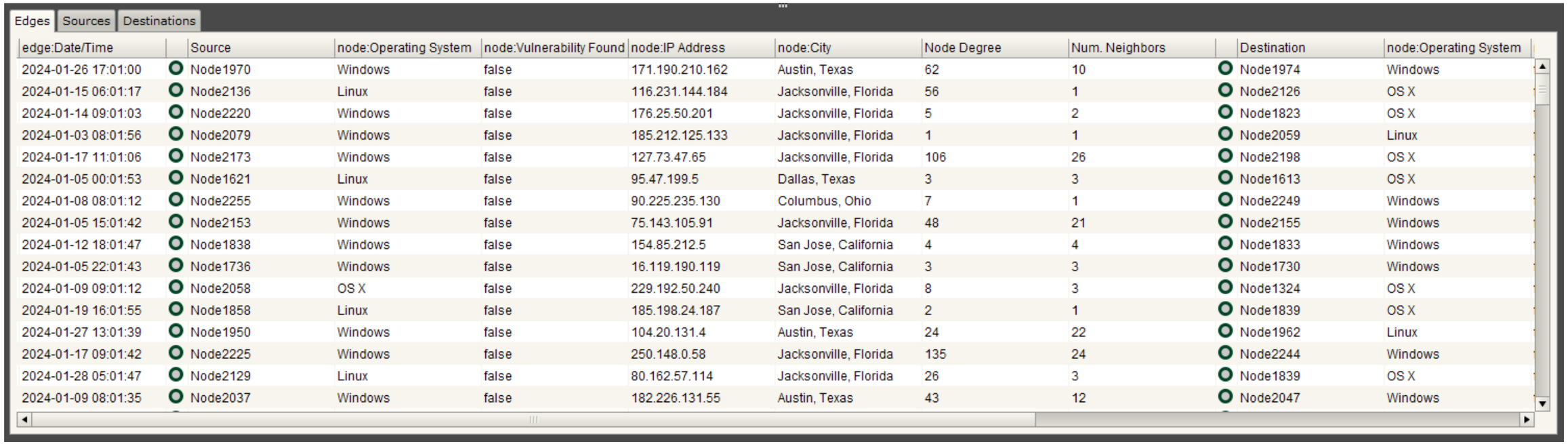






List Panel (Spreadsheet View)

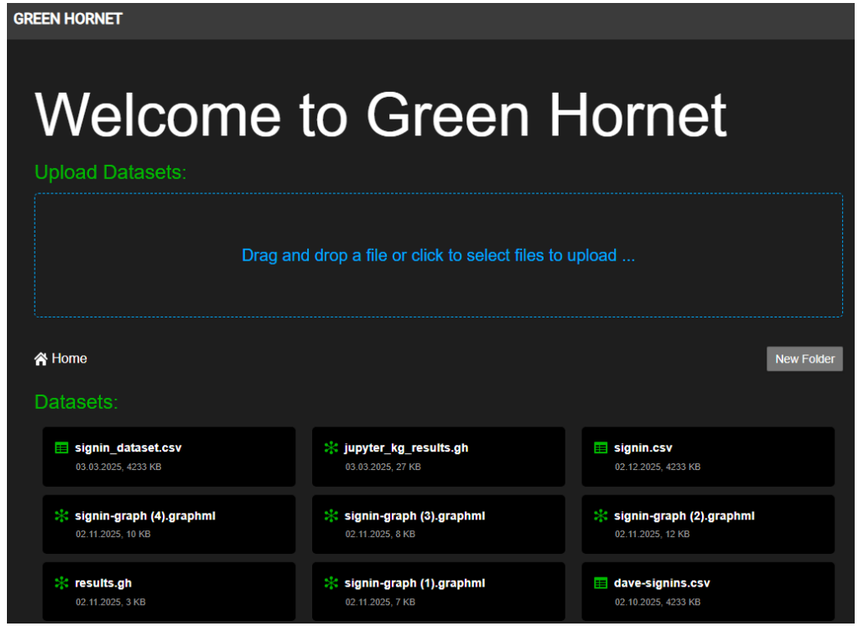
- Users can view the raw data related to nodes/edges and their properties
- This reflects the current selection and expansion
- Clicking on rows highlights the specific nodes/links representing the data
- Selections can be exported for use in other tools / reports



Edges | Sources | Destinations

edge:Date/Time	Source	node:Operating System	node:Vulnerability Found	node:IP Address	node:City	Node Degree	Num. Neighbors	Destination	node:Operating System
2024-01-26 17:01:00	Node1970	Windows	false	171.190.210.162	Austin, Texas	62	10	Node1974	Windows
2024-01-15 06:01:17	Node2136	Linux	false	116.231.144.184	Jacksonville, Florida	56	1	Node2126	OS X
2024-01-14 09:01:03	Node2220	Windows	false	176.25.50.201	Jacksonville, Florida	5	2	Node1823	OS X
2024-01-03 08:01:56	Node2079	Windows	false	185.212.125.133	Jacksonville, Florida	1	1	Node2059	Linux
2024-01-17 11:01:06	Node2173	Windows	false	127.73.47.65	Jacksonville, Florida	106	26	Node2198	OS X
2024-01-05 00:01:53	Node1621	Linux	false	95.47.199.5	Dallas, Texas	3	3	Node1613	OS X
2024-01-08 08:01:12	Node2255	Windows	false	90.225.235.130	Columbus, Ohio	7	1	Node2249	Windows
2024-01-05 15:01:42	Node2153	Windows	false	75.143.105.91	Jacksonville, Florida	48	21	Node2155	Windows
2024-01-12 18:01:47	Node1838	Windows	false	154.85.212.5	San Jose, California	4	4	Node1833	Windows
2024-01-05 22:01:43	Node1736	Windows	false	16.119.190.119	San Jose, California	3	3	Node1730	Windows
2024-01-09 09:01:12	Node2058	OS X	false	229.192.50.240	Jacksonville, Florida	8	3	Node1324	OS X
2024-01-19 16:01:55	Node1858	Linux	false	185.198.24.187	San Jose, California	2	1	Node1839	OS X
2024-01-27 13:01:39	Node1950	Windows	false	104.20.131.4	Austin, Texas	24	22	Node1962	Linux
2024-01-17 09:01:42	Node2225	Windows	false	250.148.0.58	Jacksonville, Florida	135	24	Node2244	Windows
2024-01-28 05:01:47	Node2129	Linux	false	80.162.57.114	Jacksonville, Florida	26	3	Node1839	OS X
2024-01-09 08:01:35	Node2037	Windows	false	182.226.131.55	Austin, Texas	43	12	Node2047	Windows

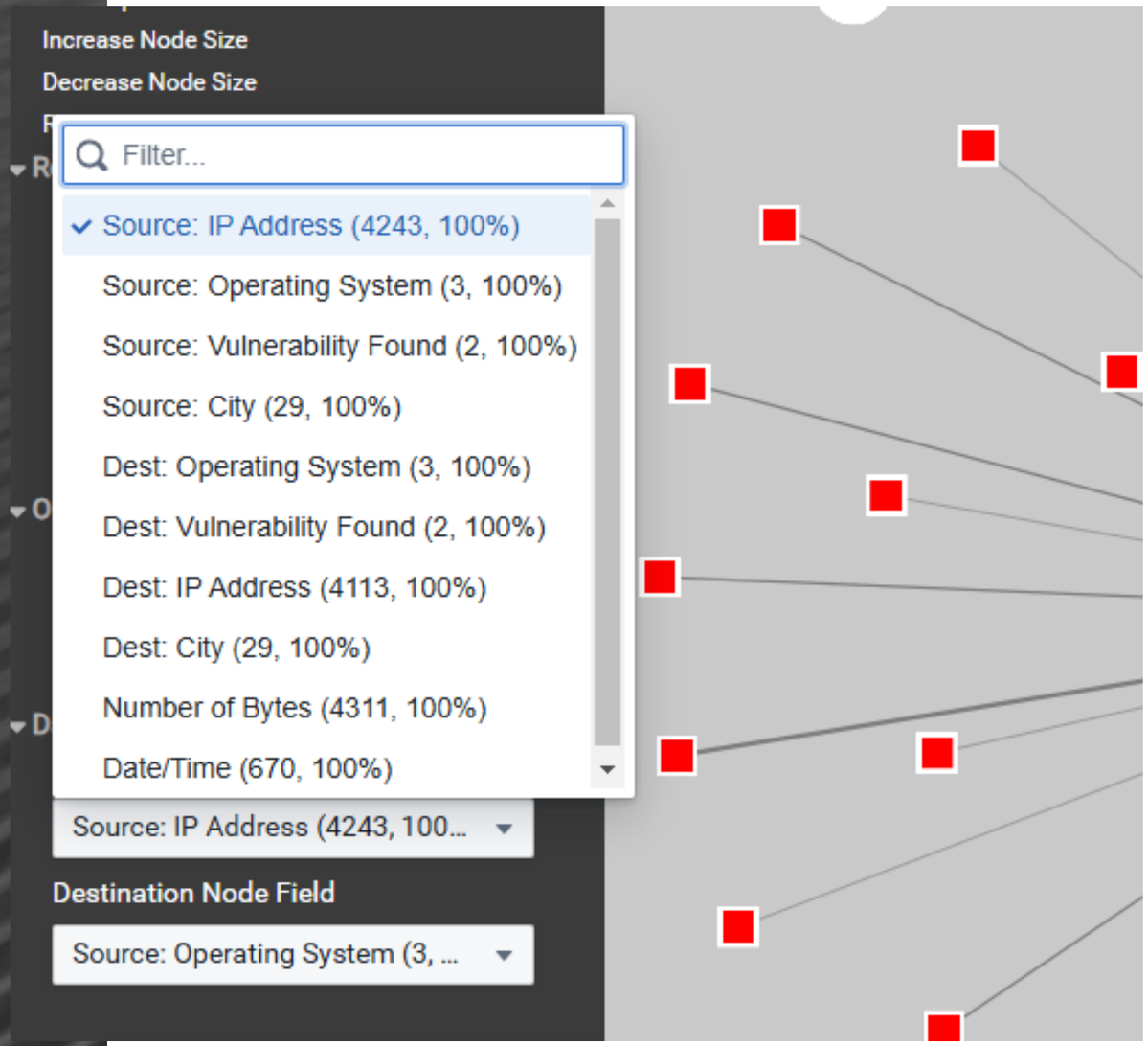
Data Formats



Green Hornet supports the following file formats:

- GDF files:
  - An extension of CSV data specifically for graphs
- GraphML
  - A common standard for graph modeling
- Green Hornet's proprietary *.gh format
  - Enables specialized Green Hornet features
- CSV files
  - Any CSV can be loaded as a graph
  - Includes special graph modeling capabilities, enabling users to interactively re-define how the CSV represents the graph

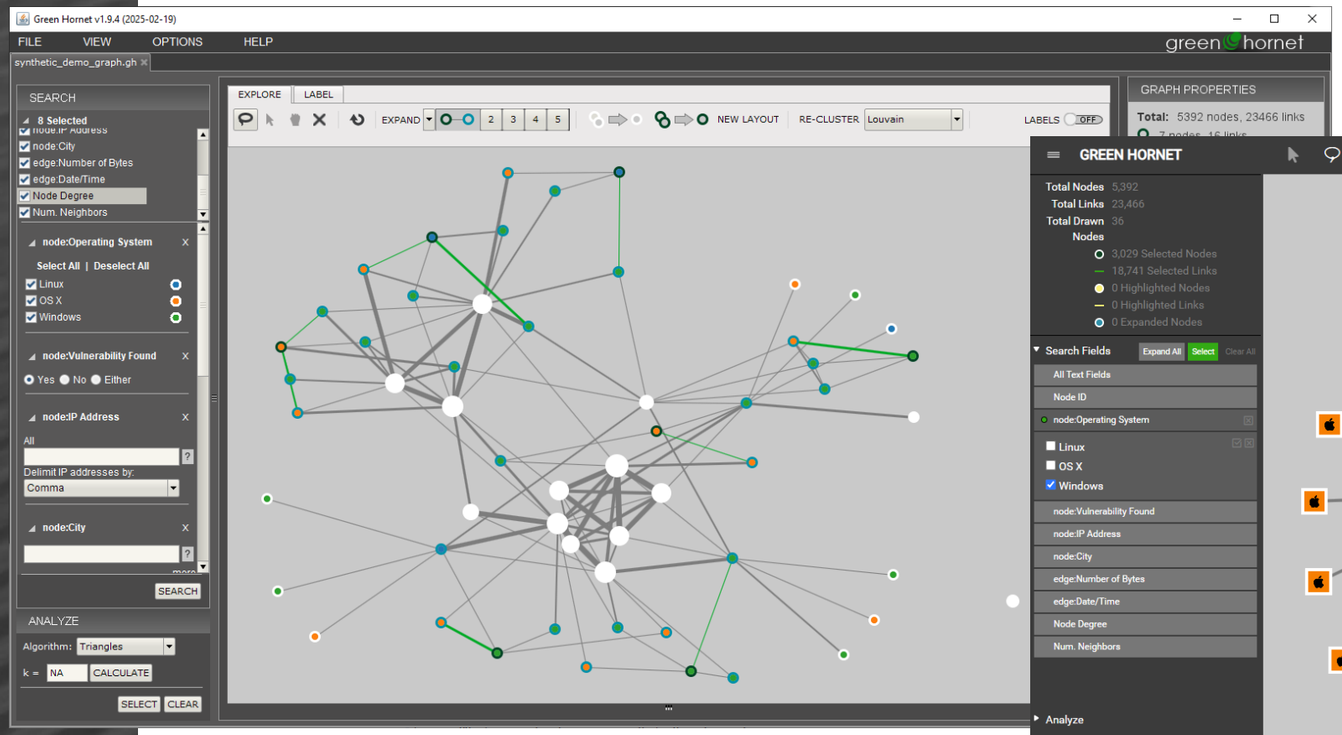
Integrated Graph Modeling

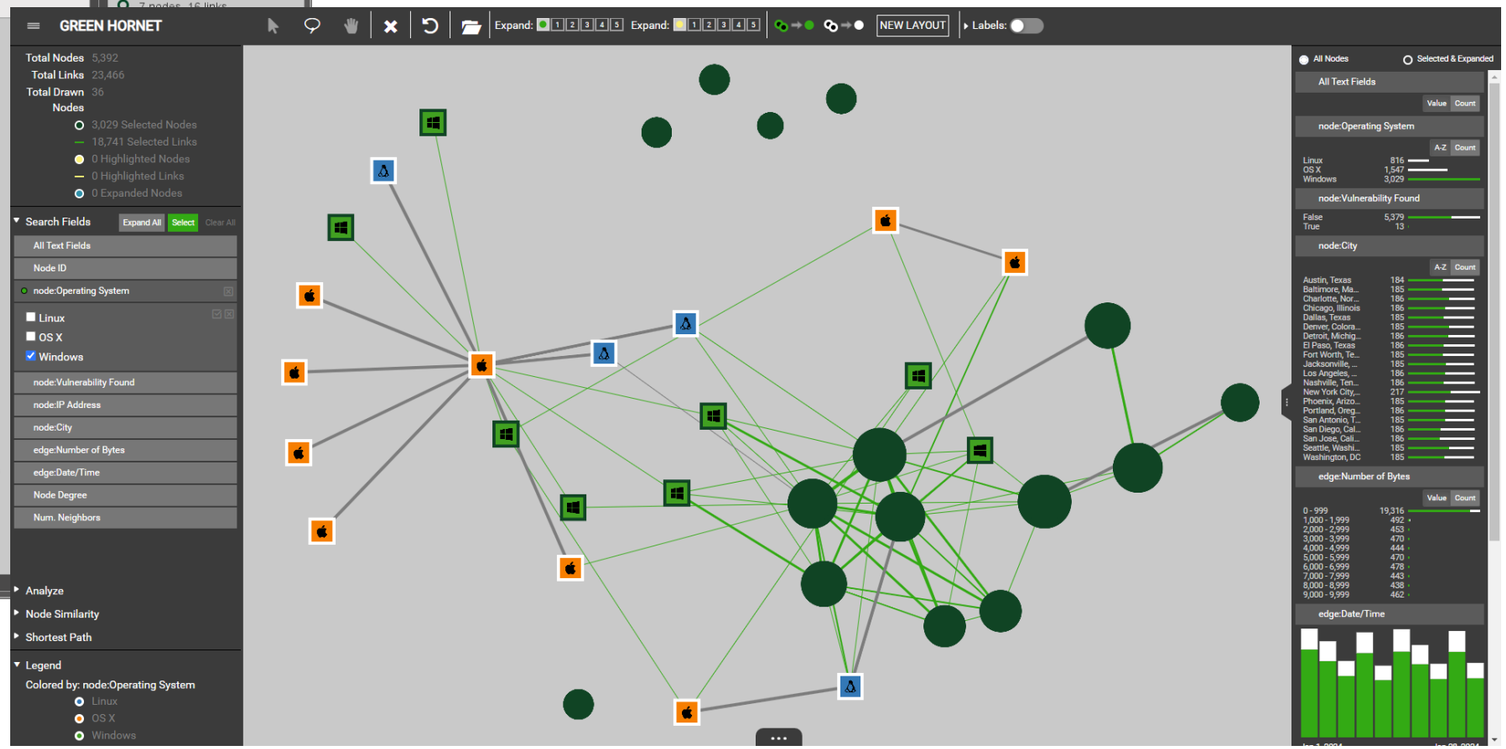


- Any pair of columns can be selected as the source and destination, for example:
  - **User Account** to **Error Message**
  - **IP Address** to **Role**
  - Etc.
- Users can interactively change the columns to create different **“perspectives”** into the same data
- The graph **selection** persists during changes
  - Enables users to see how the same data looks from different graph “perspectives”
  - For example, different types of relationships associated with a particular error message

Pacific Northwest NATIONAL LABORATORY

Desktop and Web Based Options

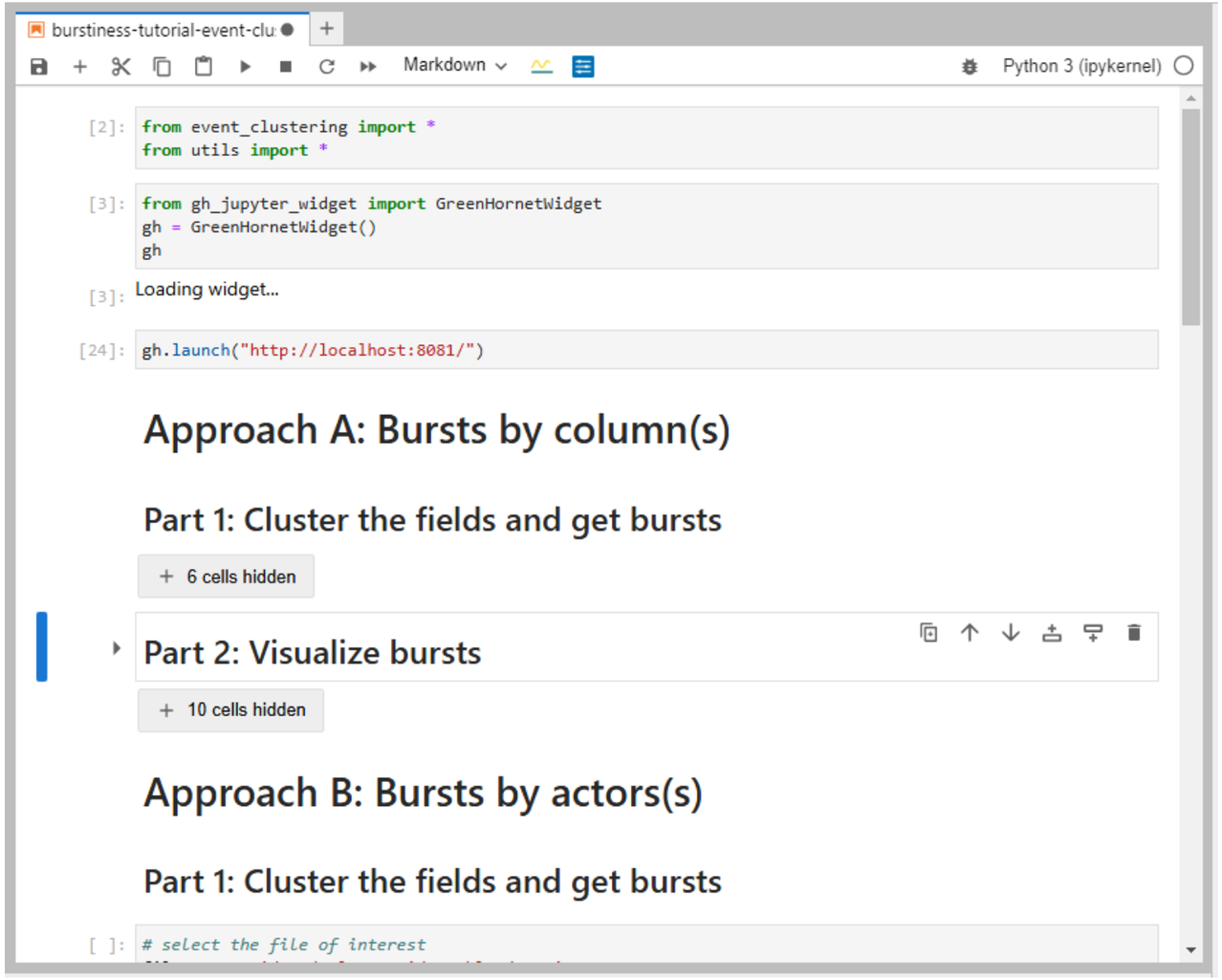




Both pure desktop and thin-client web-based versions available for deployment

Jupyter Notebook Integration

- Green Hornet Web also supports **Jupyter Notebook** integration
- Notebooks can be used to:
  - Create graphs
  - Modify an existing dataset
  - Perform graph analytics
  - Machine learning
- Results can be viewed automatically in Green Hornet
- Green Hornet data and interactions (e.g., selection and highlight) can be loaded into your notebook



Pacific Northwest
NATIONAL LABORATORY

Thank you