Anomaly Detection in Bipartite Networks

Mohammed Eslami, Ph.D
George Zheng, Hamed Eramian, Georgiy Levchuk

Disclaimers

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).

Outline

netrias®

Goal: Formulate cyber logs as bipartite graphs and an an analytical workflow that use graph features to highlight events of interest to a cyber analyst.

- Overview of Network Defense
- Cyber Data Represented as Bipartite Graphs
- Graph Analytical Components, Features, and Workflow for Cyber Security
- Scalability and Examples
- Conclusion/Next Steps

The Challenge of Network Defense

netrias

Rapid identification of network anomalies in billions of records across a heterogeneous logs.

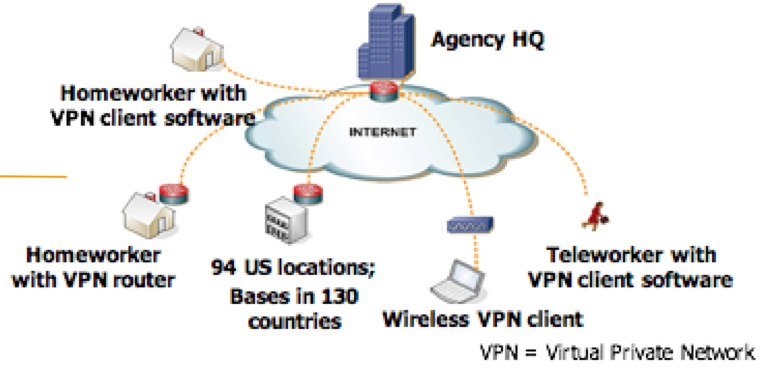
Enterprise Netflow and log data:

12 billion events per day,

1 TB per day of communications

>60,000 employees,

>570,000 users



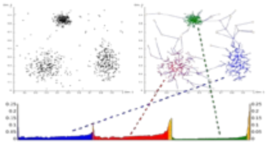
Moving beyond State of the Art:

Rule-based signatures → Adaptive behavior detection

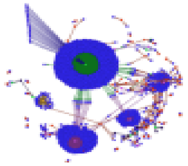
Stateless single IP analyses → Context based decisions

Manual analysis → Guided automation

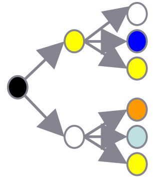
Source: Deason, L. et. al. Scalable Temporal Analytics to Detect Automation and Coordination. Flocon 2017



Clustering



Graphs



Sequence

The Challenge of Network Defense

Rapid identification of network anomalies in billions of records across a heterogeneous logs.

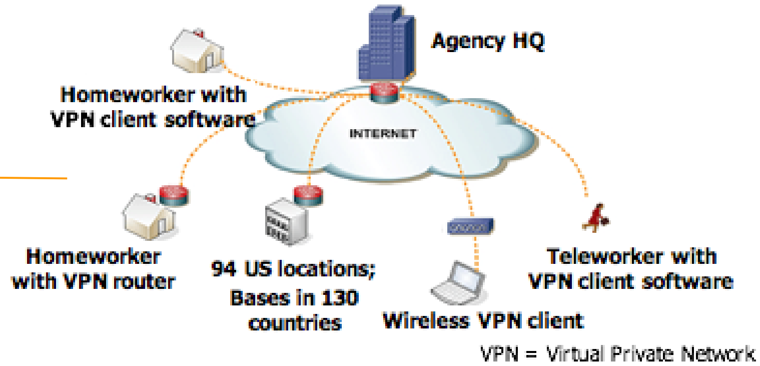
Enterprise Netflow and log data:

12 billion events per day,

1 TB per day of communications

>60,000 employees,

>570,000 users



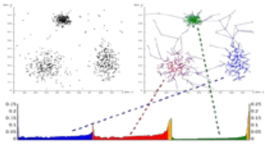
Moving beyond State of the Art:

Rule-based signatures → Adaptive behavior de

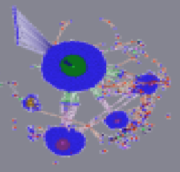
Stateless single IP analyses → Context-based decisi

Manual analysis → Guided automation

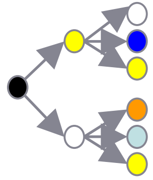
Source: Deason, L. et. al. Scalable Temporal Analytics to Detect Automation and Coordination. Flocon 2017



Clustering



Graphs



Sequence

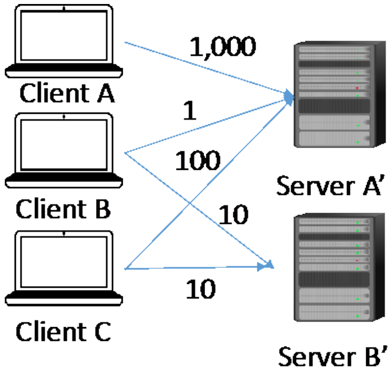
Types of Bipartite Graphs from Enterprise Networks

Bipartite graphs, graphs that have edges only between two distinct entity types, provide an opportunity to capture the relationships between entities within and across types but pose a unique set of challenges in their storage, scalable analysis, and interpretability.

IP-IP Graphs

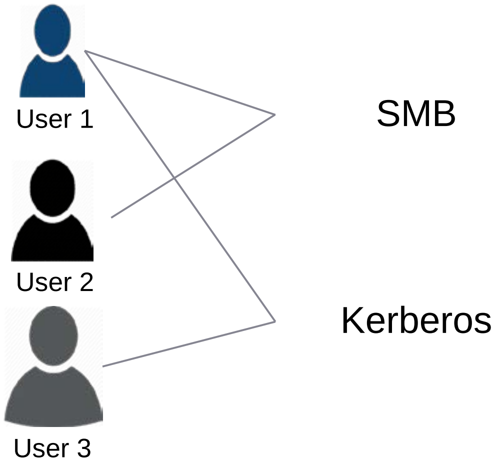
Netflow records – edges only between internal/external IPs

Client-Server Graphs



DNS, HTTP, SMTP, etc. logs – edges only between client IP and server IP

User-Service Graph

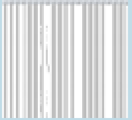


RDP, SMB, Kerberos, etc. logs- edges only between users and services used

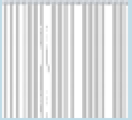
Bipartite Graph Analysis for Enterprise Scale Network Defense

Analytical suite infers relationships between similar entities, scales to billions of records, and provides rapid situational awareness to SOC analyst.

Data

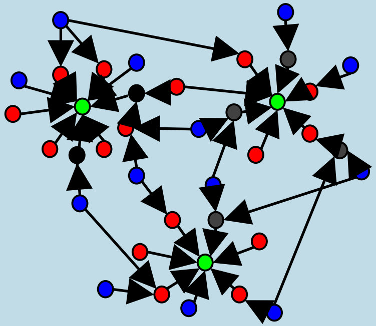


…

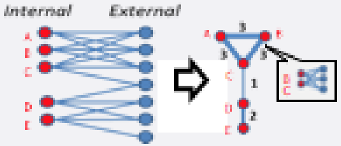


Sensor 1

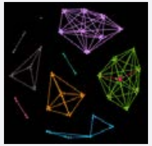
Sensor N



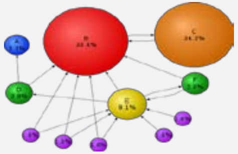
Analytics



Directional Graph Projections



Community Detection



Source: Page Rank. Wikipedia

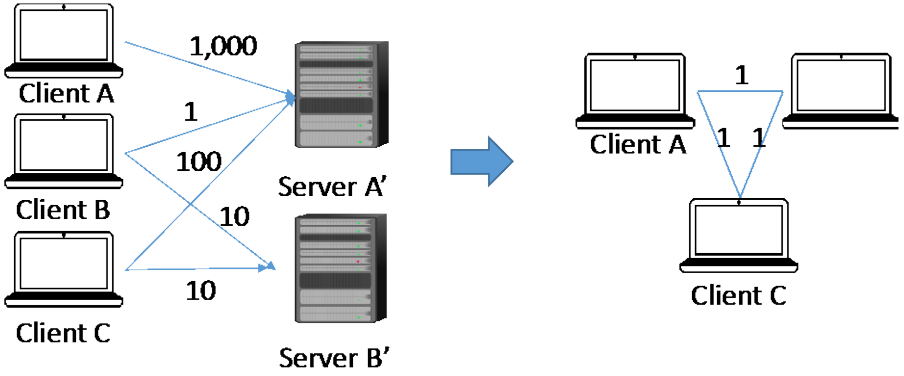
First Order Graph Features

Analytical Contributions

1. A novel method to capture and represent similarity between network entities
2. A scalable method to compute directional graph projections for enterprise scale networks
3. A mapping between extracted graph features and relevant cyber use cases to an analyst

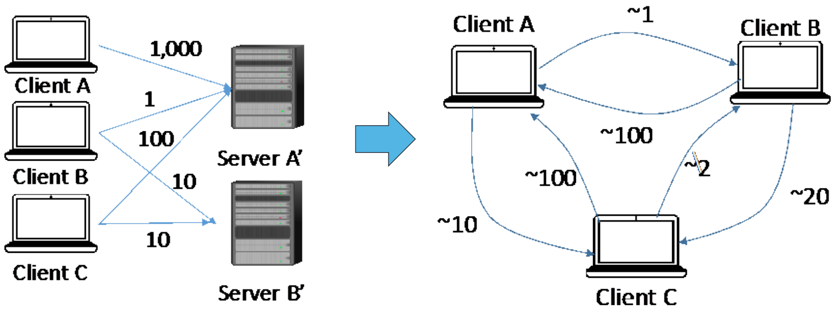
Analytics: Directional Graph Projections

Traditional Graph Projections



Nuances introduced by different graph weights and different destination nodes are ignored

Directed Graph Projections



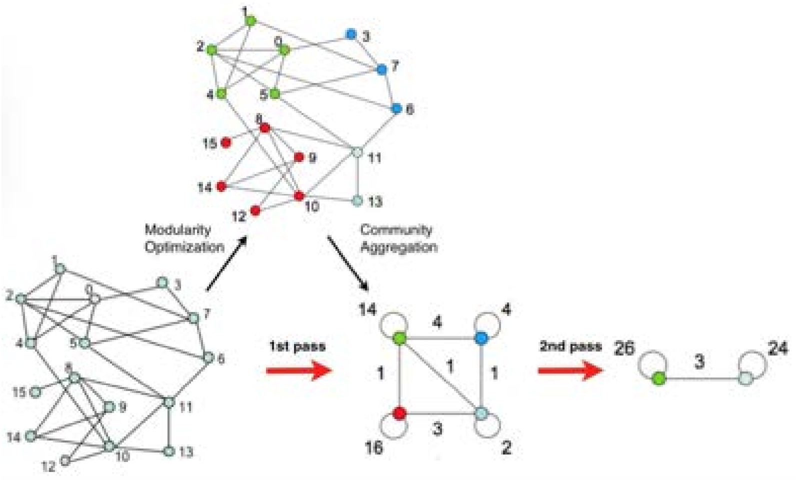
Asymmetric similarity measure can capture difference in usage of uncommon servers between clients

Analytics: Community Detection

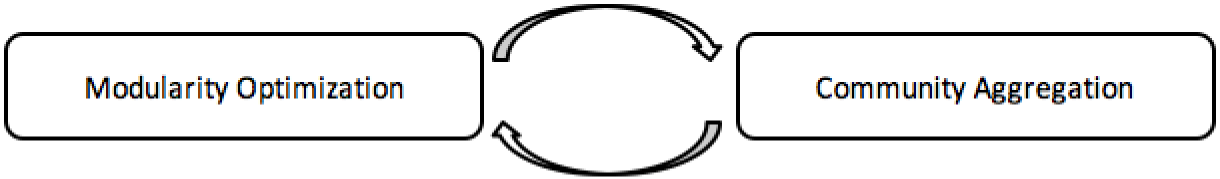
Identify communities within a network that are more connected to each other than other parts of the network.

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

- Σ_{in} is the sum of the weights of the links inside C
- Σ_{tot} is the sum of the weights of the links incident to nodes in C
- k_i is the sum of the weights of the links incident to node i
- $k_{i,in}$ is the sum of the weights of the links from i to nodes in C
- m is the sum of the weights of all the links in the network.



Reference: Blondel, V. et al. *Fast unfolding of communities in large networks, 2008*



Analytics: Interpretable First Order Graph Features

Graph Feature	Cyber Story
Raw Degree	# of requests made, # of services used, …
Raw Weighted Degree	Amount I’m using a specific service
Projected Degree	# of entities that I think I am similar to because we use a common service
Projected Community Size	# of entities I am actually similar to
Projected Page Rank	My “significance” as compared to other entities (ex. Admins will use more services than clients)

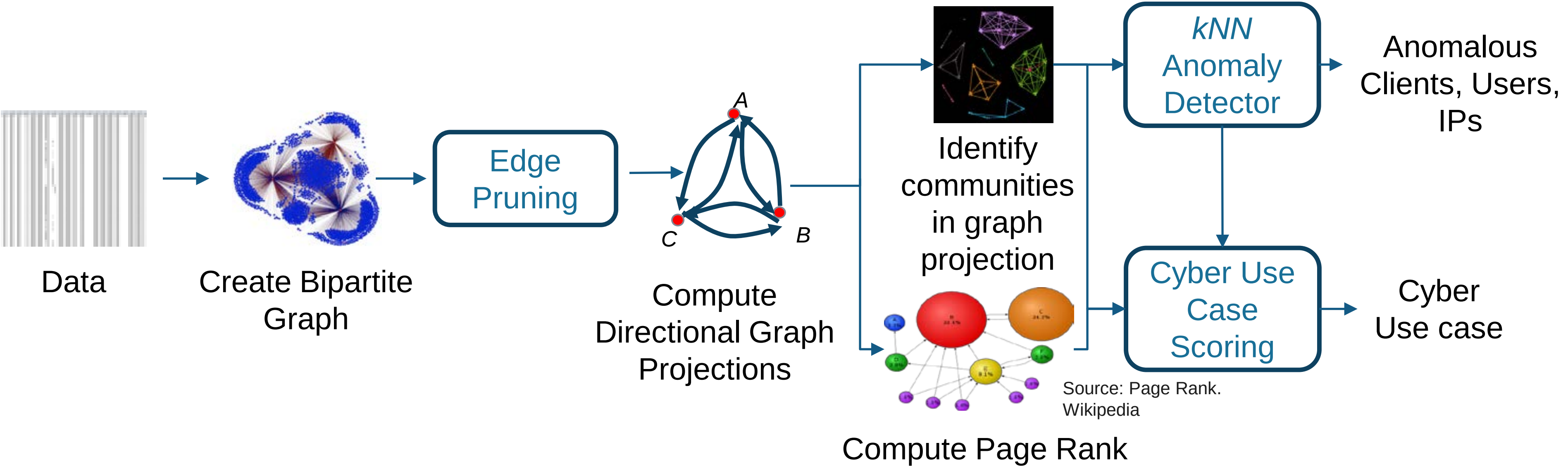
Cyber Use Case to Graph Feature Mapping

netrias®

Data Type	Cyber Use Case Description	Features
User → Service	Infer user roles	**Admins** = High projected degree, community size, and page rank **Non-admins** = High projected degree but small community size and page-rank
Client → Server	Infer similarities between groups of clients	**Typical Client Systems** = High community size, projected degree, and low page rank
Internal IP → External IP	Identifying firewalls, VPNs, or other network access points from flow data	**Firewalls** = High raw degree, weighted degree, projected degree, community size, and page rank **VPNs** = High raw degree, weighted degree, projected degree, but small community size, and page rank

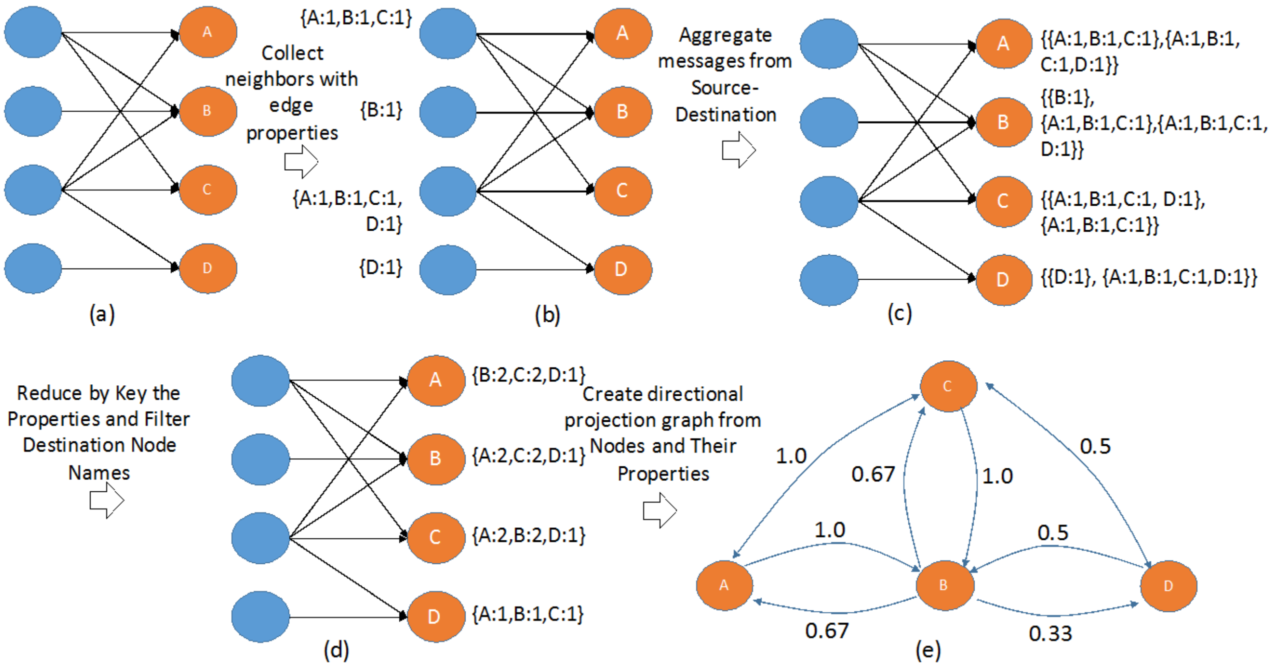
Graph Analytic Workflow

Modularization and integration identifies cyber use cases from graph feature mappings and also provides flexibility to identify anomalies within and across derived communities.



Scaling Directional Graph Projections

Message passing algorithms on graph data structure allows for custom asymmetric similarity measure and scales to O(*e*).



Technology Base

Graph Analytic Workflow

GraphX

mllib

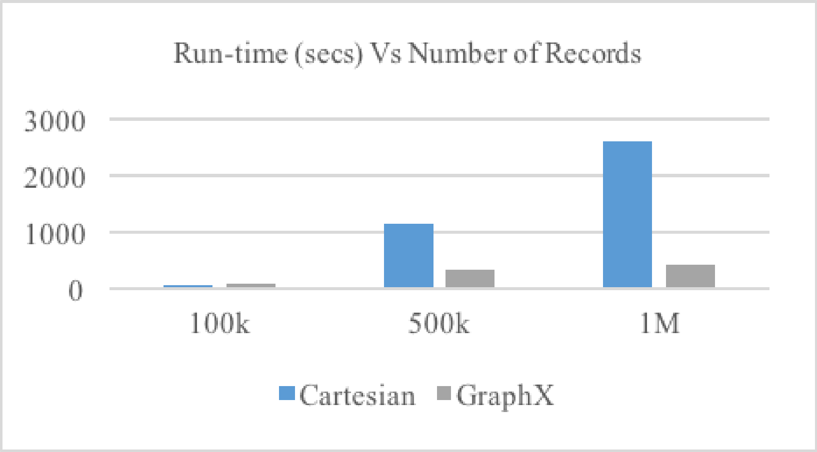
Breeze

Spark

10 executors
4GB driver memory
3GB executor memory

cloudera

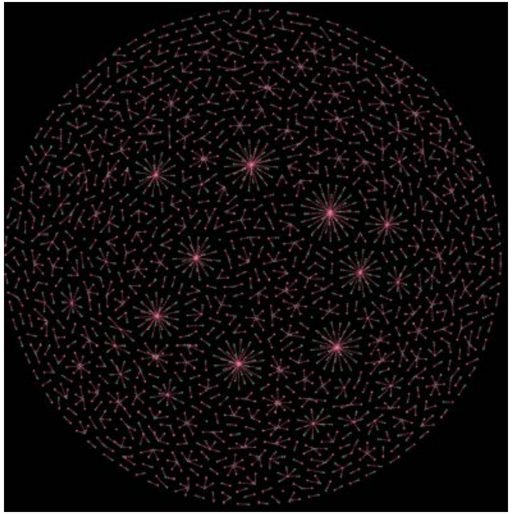
hadoop



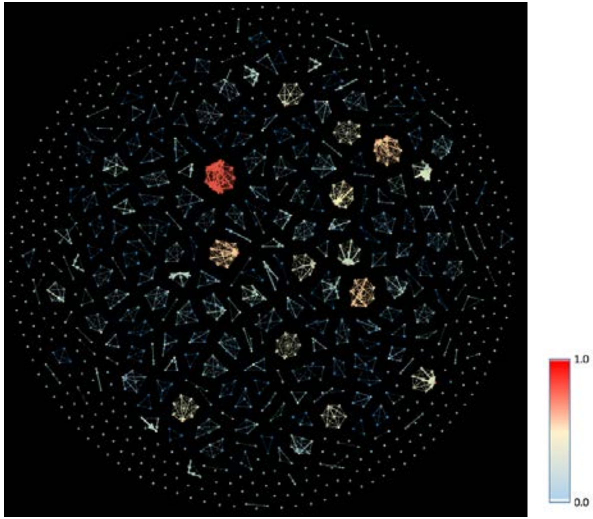
45x Dell servers, 17.28 TB RAM, and 2.304 PB HDFS Storage

Use Case: Netflow from Edge of Network

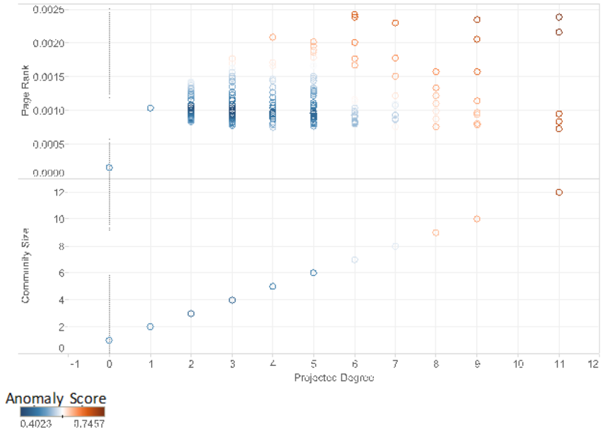
k-Nearest Neighbor anomaly detection on graph projection features highlighted the single client in the largest community that made communication with a particular DNS server an anomalous number of times.



Client-Server DNS Graph



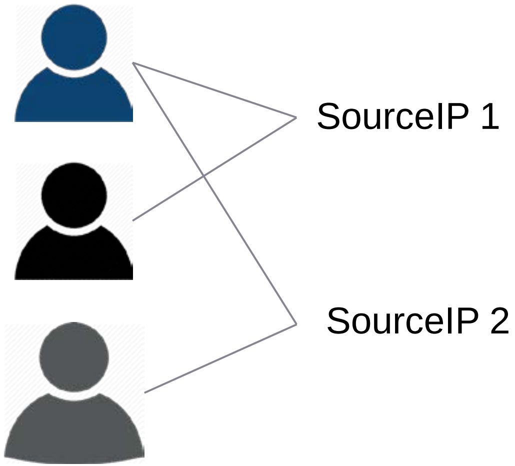
Client Graph Projection colored by anomaly score



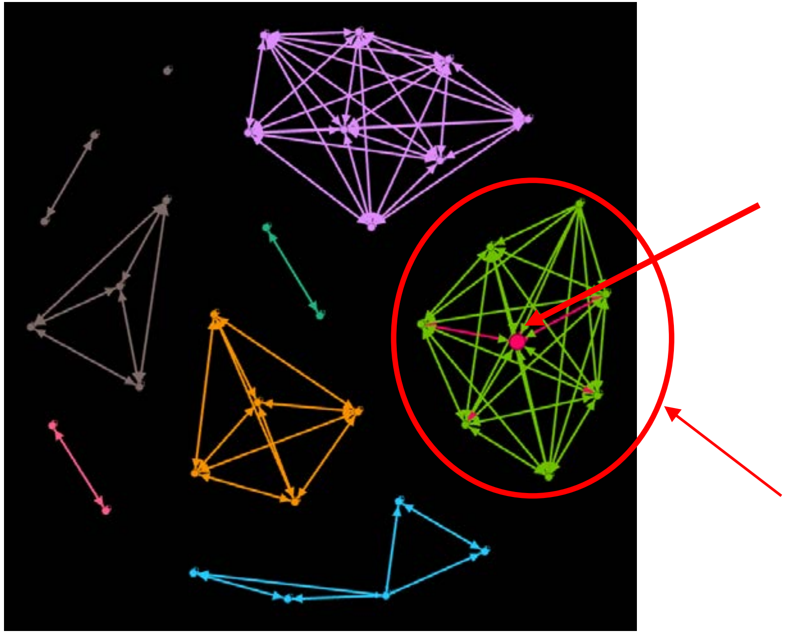
Explainable features highlight most anomalous client

Use Case: RDP Logs

Graph projections onto username from Remote Desktop Protocol (RDP) logs highlights that communities of users that login from the same IP have multiple aliases.



Bipartite graph:
username → source ip



Single user identified within a community with an alias of no resemblance to other users.

{jsmith, johns, johnsmith, …, adamb}

Conclusions and Next Steps

Conclusions

1. A novel method to capture and represent similarity between network entities

2. A scalable method to compute directional graph projections for enterprise scale networks

3. A method to rapidly visualize, identify, and interpret anomalies from cyber logs using graph features

Next Steps:

1. Identify more relevant and concrete cyber use cases for improvements and expansions on various similarity metrics and graph features.

2. We would like to extend our work to better account for temporally evolving graphs to identify significant events that occur on a network at a particular time.

netrias®

reveal the hidden state of the system™