Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Data Fusion

Enhancing NetFlow Graph Analytics

EMILIE PURVINE, BRYAN OLSEN, CLIFF JOSLYN

Pacific Northwest National Laboratory

FloCon 2016

Outline

- Introduction
- NetFlow
  - Windows Event Log data
  - Remote Desktop Protocol (RDP) sessions
- Approach to fusion of NetFlow and Windows Event Log data
- Exploratory data analysis of fused data
- Topological analysis
  - Spectral methods
  - Persistent Homology

Introduction

- *Remote Desktop Sessions*
  - *Important to analyze in the context of NetFlow*
- *Data Sources*
  - *NetFlow (using cisco NetFlow v5)*
  - *Windows Event Logs*
    - *Windows Logging Service (WLS)*
      - *Developed by the Department of Energy's Kansas City Plant*
      - *Enhance and standardize information coming from Windows logging*
      - *Incorporated network interface information to create a hybrid data set enabling more accuracy in NetFlow/event log fusion at the enterprise level*
- *We will describe our lessons learned when fusing WLS and NetFlow sessions*

The Challenge

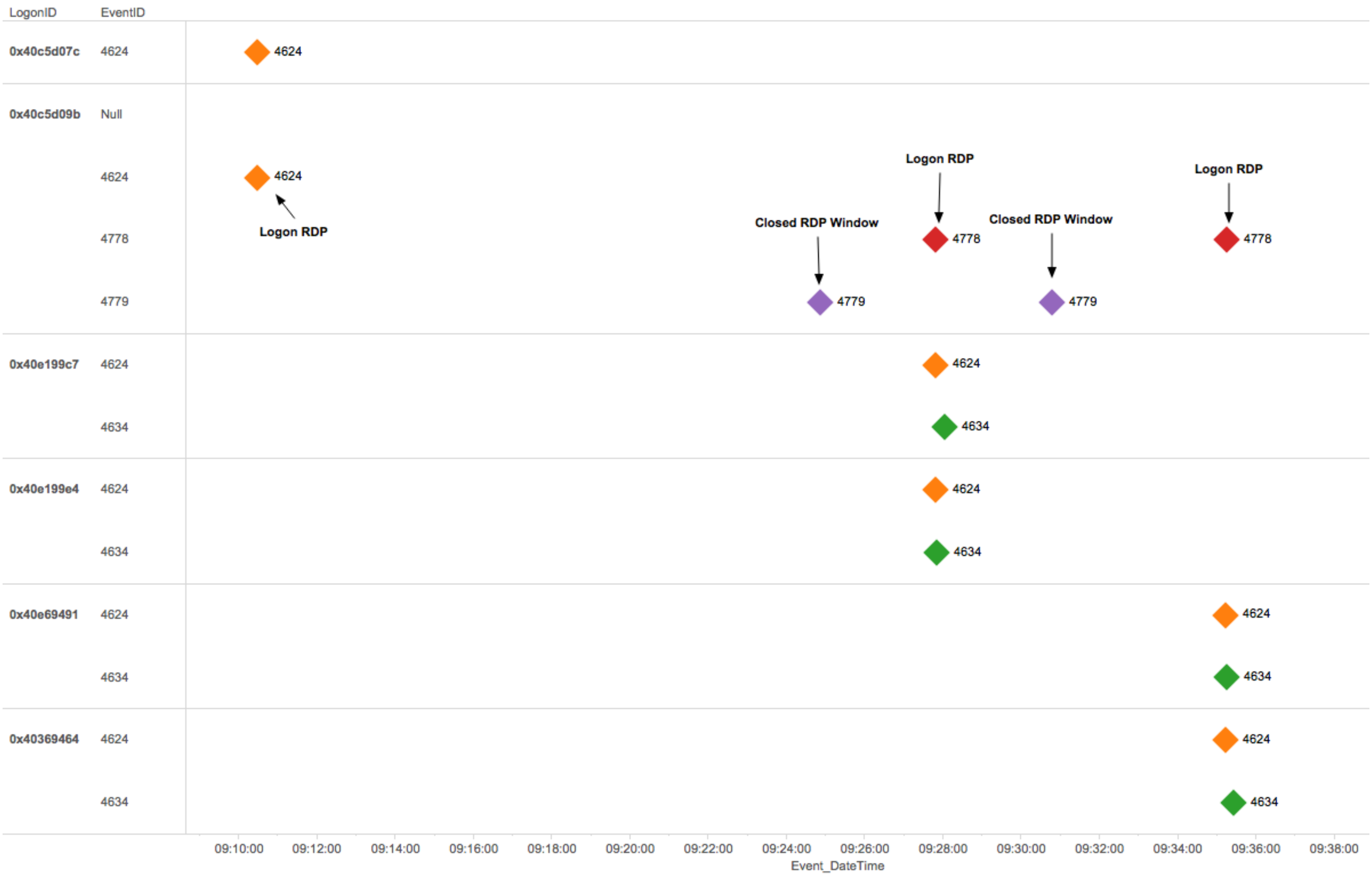
- Research needs a way to “map” remote logins as the are represented in Windows event logs to the associated NetFlow records

- The mapping will highlight the relationship and fidelity of both datasets as representatives for remote login behavior

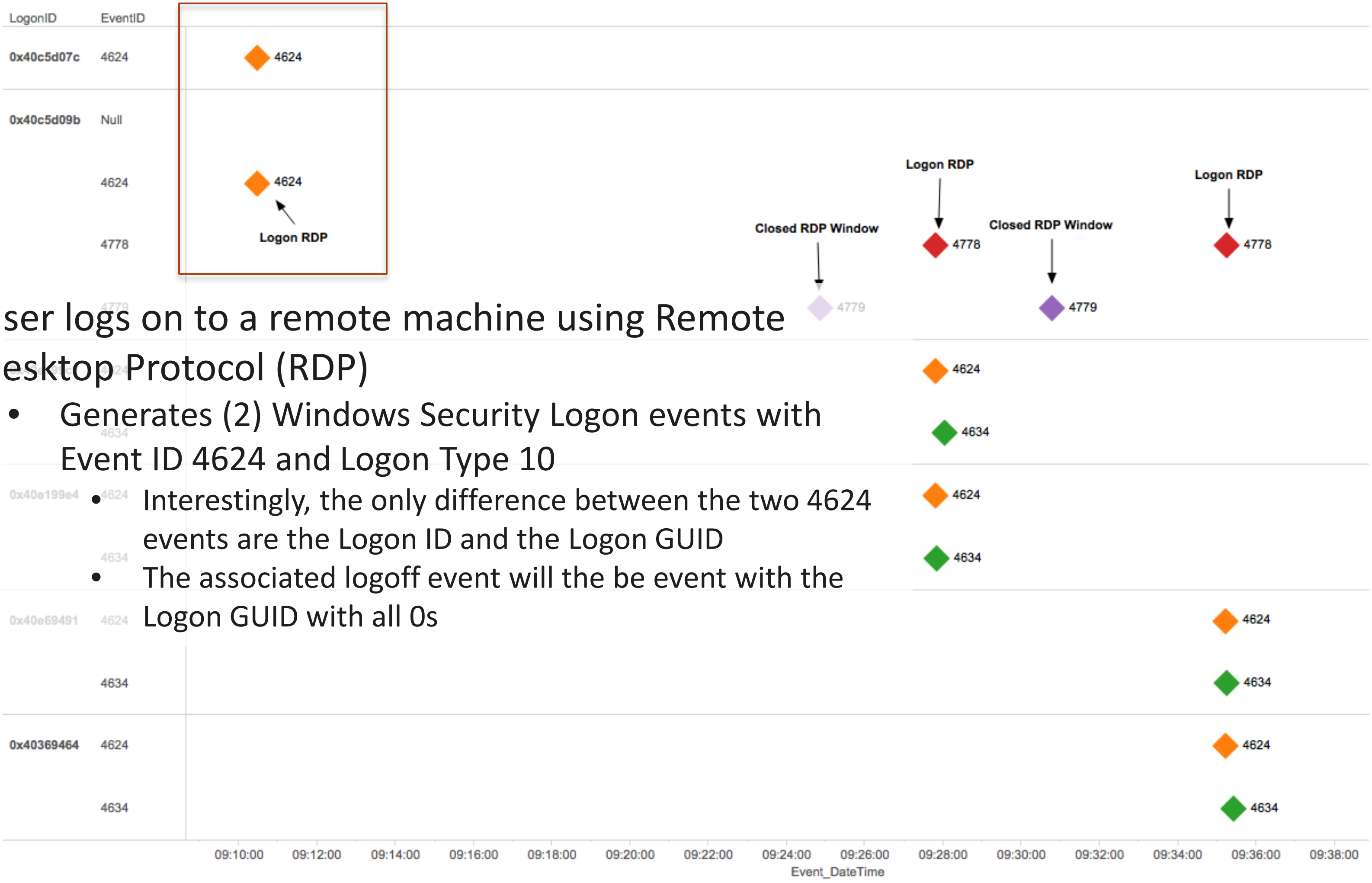
- Provide understanding for how each source may be used for topological and graph based approaches

Windows Event Illustrated - *Remote Desktop Sessions*

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*



Windows Event Illustrated - Remote Desktop Sessions

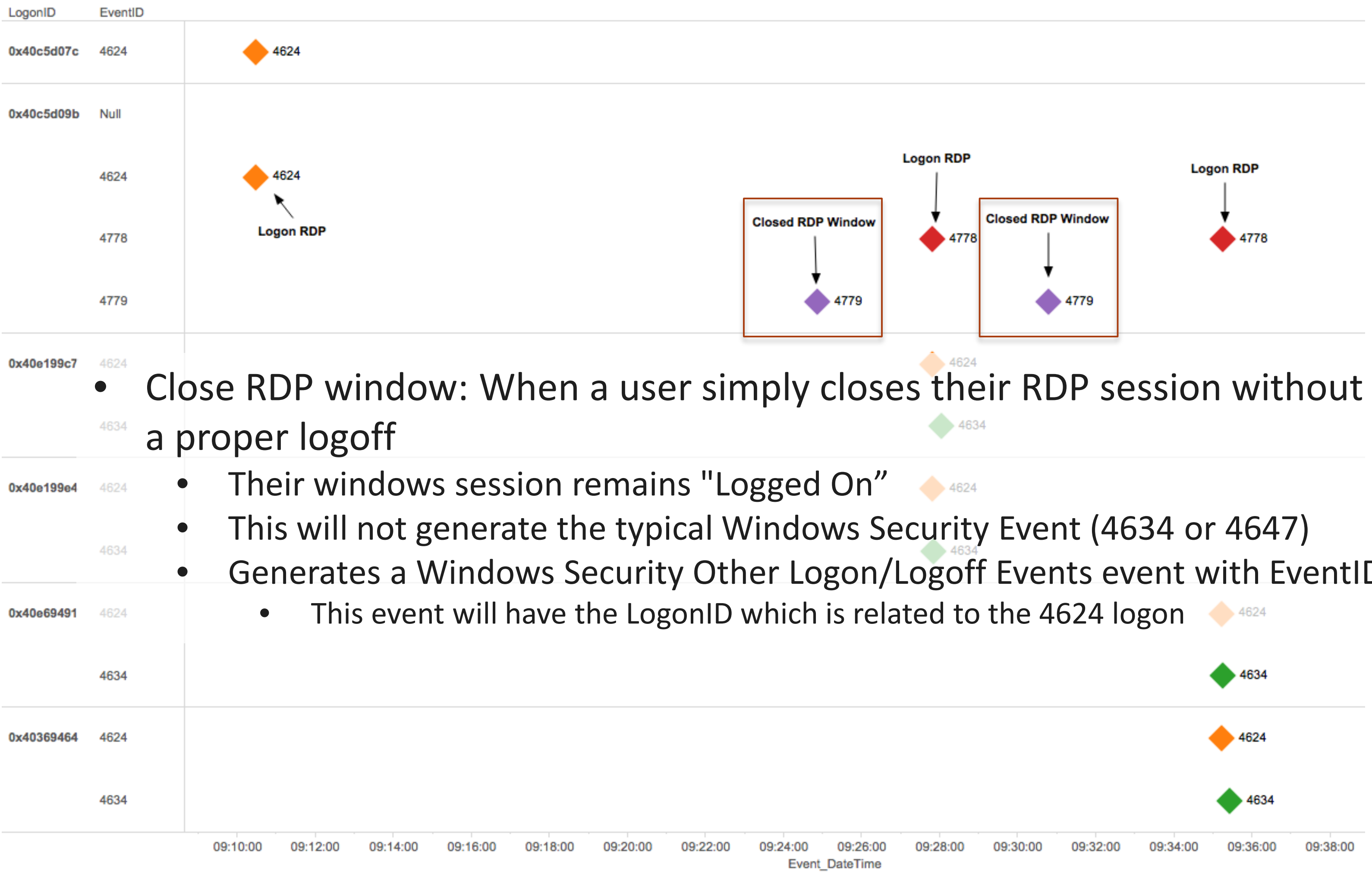


- User logs on to a remote machine using Remote Desktop Protocol (RDP)
  - Generates (2) Windows Security Logon events with Event ID 4624 and Logon Type 10
    - Interestingly, the only difference between the two 4624 events are the Logon ID and the Logon GUID
    - The associated logoff event will the be event with the Logon GUID with all 0s

Windows Event Illustrated - Remote Desktop Sessions

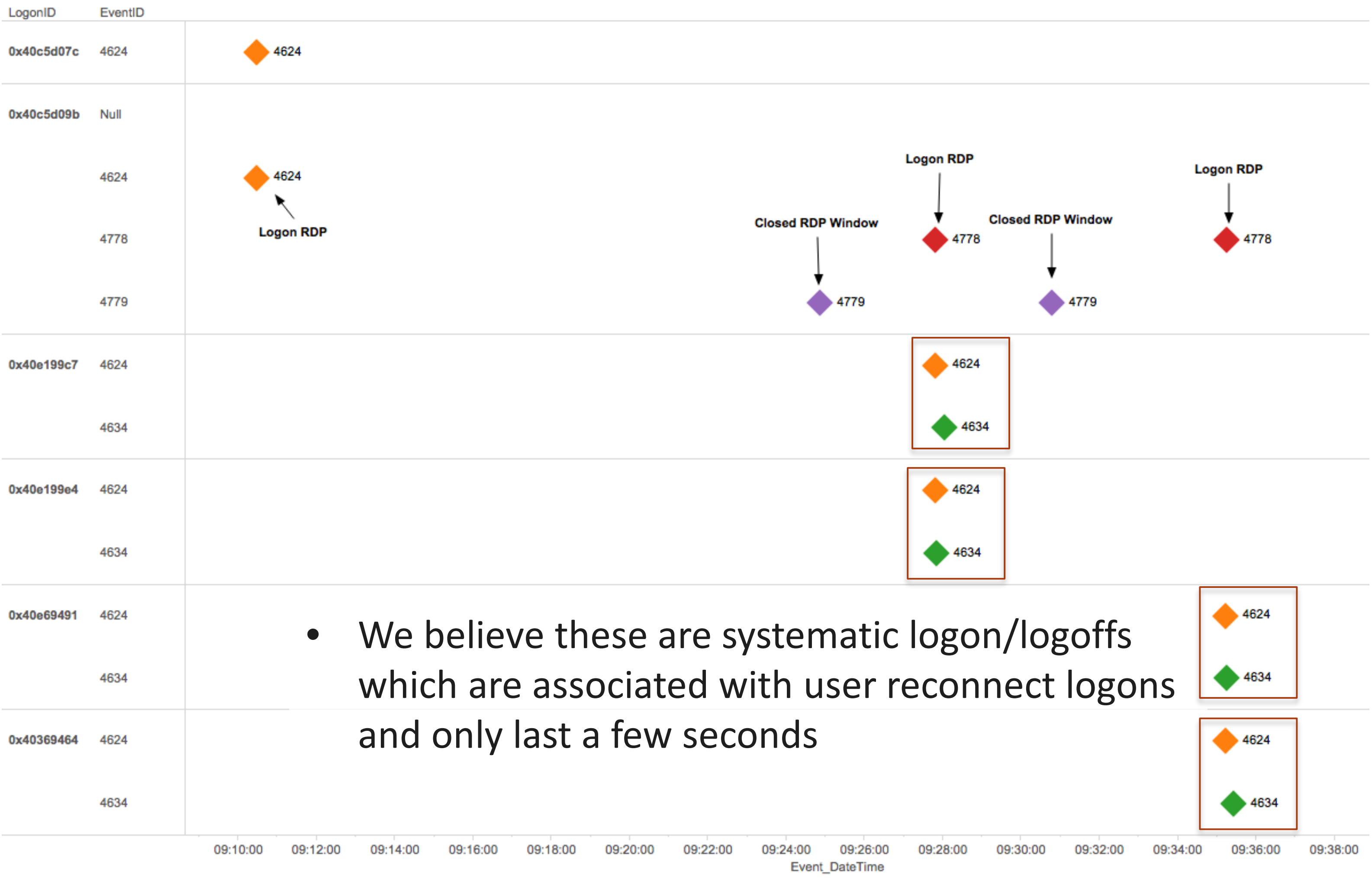
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*



- Close RDP window: When a user simply closes their RDP session without doing a proper logoff
  - Their windows session remains "Logged On"
  - This will not generate the typical Windows Security Event (4634 or 4647)
  - Generates a Windows Security Other Logon/Logoff Events event with EventID 4779
    - This event will have the LogonID which is related to the 4624 logon

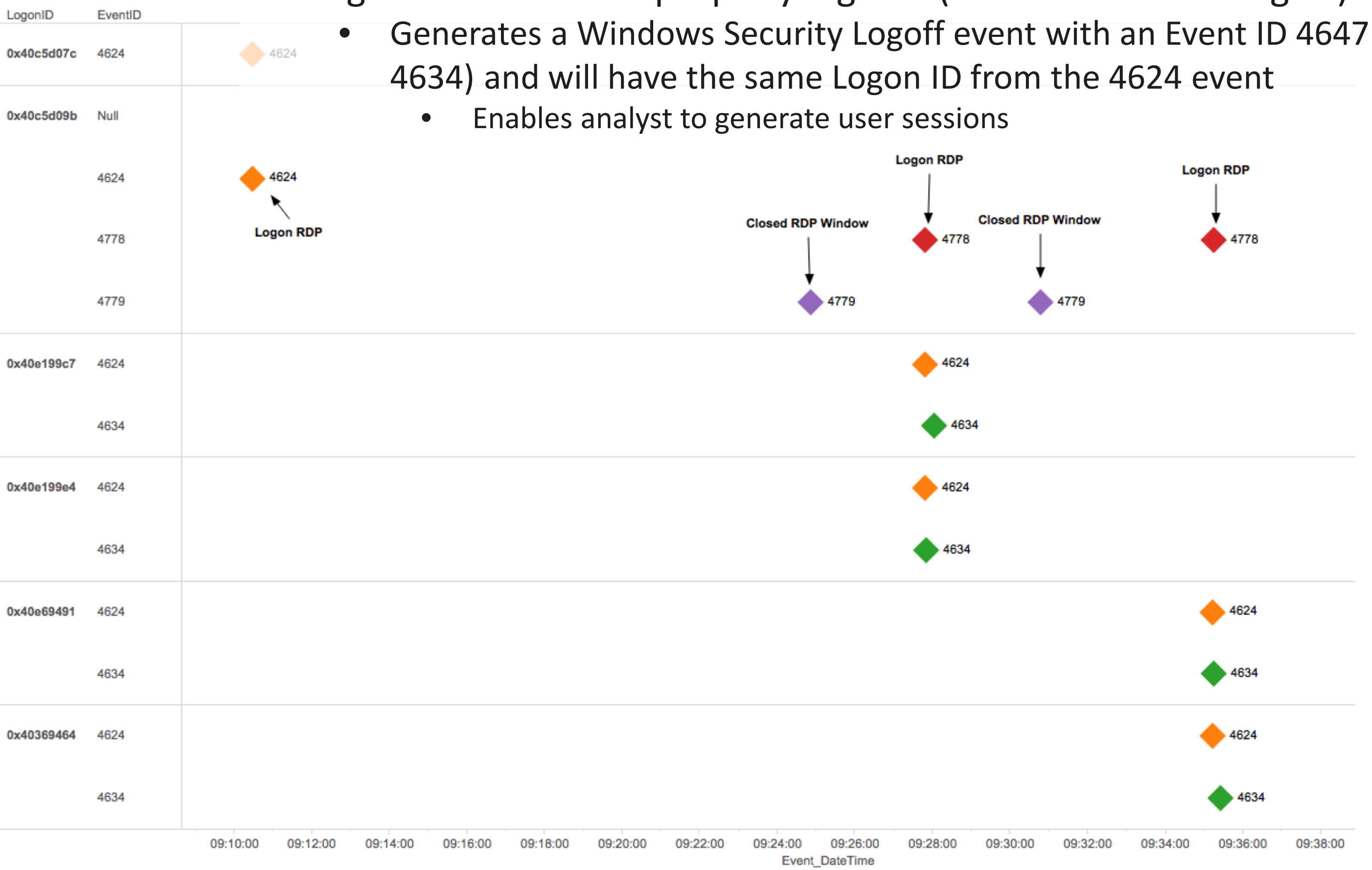
Windows Event Illustrated - *Remote Desktop Sessions*



- We believe these are systematic logon/logoffs which are associated with user reconnect logons and only last a few seconds

Windows Event Illustrated - Remote Desktop Sessions

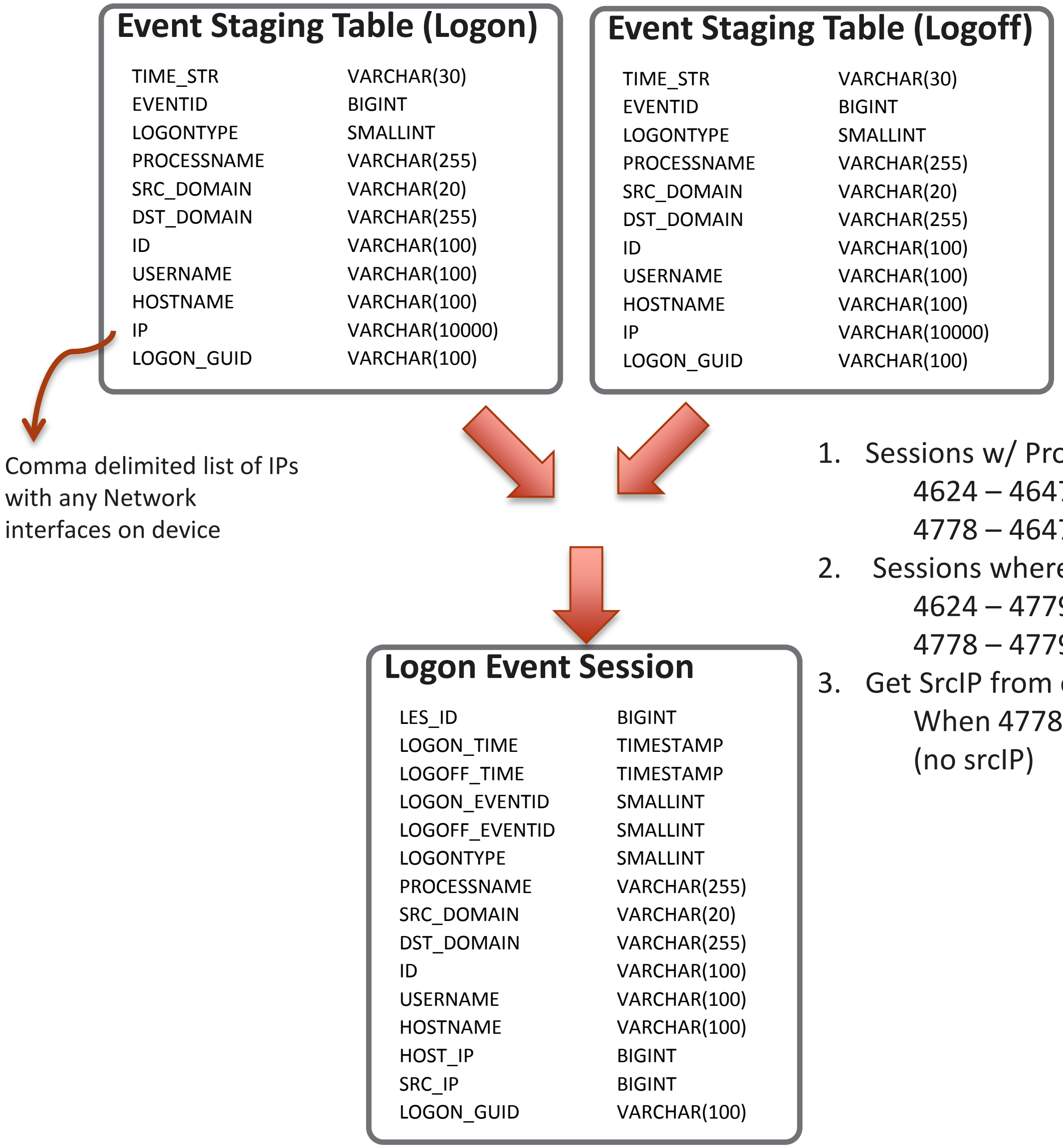
- Logoff: When a user properly logs off (user clicks start->logoff) RDP
  - Generates a Windows Security Logoff event with an Event ID 4647 (or 4634) and will have the same Logon ID from the 4624 event
    - Enables analyst to generate user sessions



Supporting Database Tables

Flow Table

FLOW_ID	BIGINT
SIP	BIGINT
DIP	BIGINT
SPORT	INTEGER
DPORT	INTEGER
PROTOCOL	SMALLINT
PACKETS	BIGINT
BYTES	BIGINT
FLAGS	VARCHAR(100)
STIME	NUMERIC
DURATION	NUMERIC
ETIME	NUMERIC
SENSOR	VARCHAR(100)
DIRECTION_IN	SMALLINT
DIRECTION_OUT	SMALLINT
STIME_MSEC	NUMERIC
ETIME_MSEC	NUMERIC
DUR_MSEC	NUMERIC
ITYPE	VARCHAR(10)
ICODE	VARCHAR(10)
INITIALFLAGS	VARCHAR(100)
SESSIONFLAGS	VARCHAR(100)
ATTRIBUTES	VARCHAR(100)
APPLICATION	VARCHAR(100)



Event Staging Table (Logon)

TIME_STR	VARCHAR(30)
EVENTID	BIGINT
LOGONTYPE	SMALLINT
PROCESSNAME	VARCHAR(255)
SRC_DOMAIN	VARCHAR(20)
DST_DOMAIN	VARCHAR(255)
ID	VARCHAR(100)
USERNAME	VARCHAR(100)
HOSTNAME	VARCHAR(100)
IP	VARCHAR(10000)
LOGON_GUID	VARCHAR(100)

Comma delimited list of IPs with any Network interfaces on device

Event Staging Table (Logoff)

TIME_STR	VARCHAR(30)
EVENTID	BIGINT
LOGONTYPE	SMALLINT
PROCESSNAME	VARCHAR(255)
SRC_DOMAIN	VARCHAR(20)
DST_DOMAIN	VARCHAR(255)
ID	VARCHAR(100)
USERNAME	VARCHAR(100)
HOSTNAME	VARCHAR(100)
IP	VARCHAR(10000)
LOGON_GUID	VARCHAR(100)

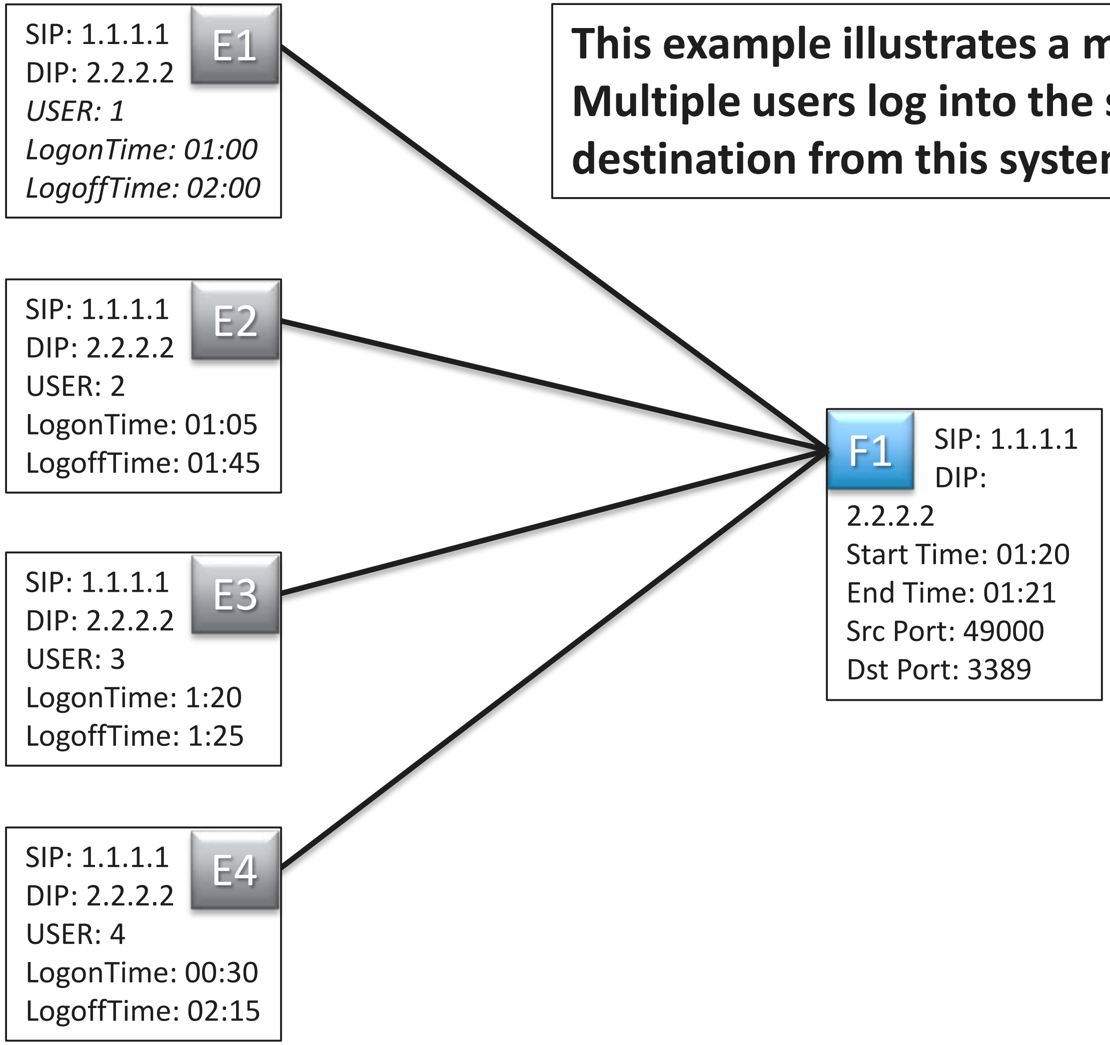
Logon Event Session

LES_ID	BIGINT
LOGON_TIME	TIMESTAMP
LOGOFF_TIME	TIMESTAMP
LOGON_EVENTID	SMALLINT
LOGOFF_EVENTID	SMALLINT
LOGONTYPE	SMALLINT
PROCESSNAME	VARCHAR(255)
SRC_DOMAIN	VARCHAR(20)
DST_DOMAIN	VARCHAR(255)
ID	VARCHAR(100)
USERNAME	VARCHAR(100)
HOSTNAME	VARCHAR(100)
HOST_IP	BIGINT
SRC_IP	BIGINT
LOGON_GUID	VARCHAR(100)

1. Sessions w/ Proper Logon and Logoff
   4624 – 4647
   4778 – 4647
2. Sessions where closed window
   4624 – 4779
   4778 – 4779
3. Get SrcIP from event 4624
   When 4778 is logon event
   (no srcIP)

Findings: Many Sessions → 1 Flow

This example illustrates a multi-user machine: Multiple users log into the same remote destination from this system



E1
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 1
LogonTime: 01:00
LogoffTime: 02:00

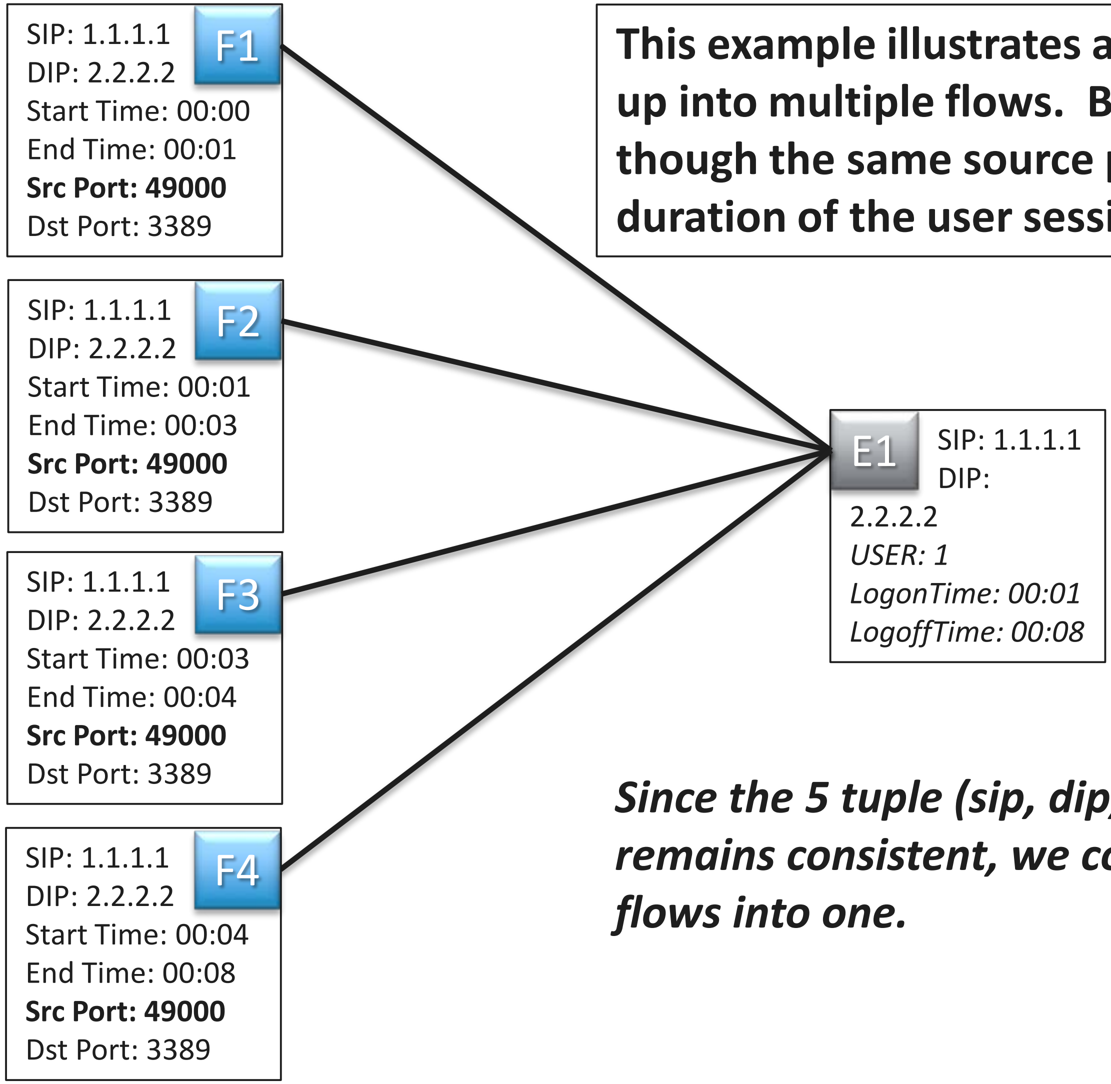
E2
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 2
LogonTime: 01:05
LogoffTime: 01:45

E3
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 3
LogonTime: 1:20
LogoffTime: 1:25

E4
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 4
LogonTime: 00:30
LogoffTime: 02:15

F1
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 01:20
End Time: 01:21
Src Port: 49000
Dst Port: 3389

Findings: Many Flows → 1 Event



F1
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 00:00
End Time: 00:01
Src Port: 49000
Dst Port: 3389

F2
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 00:01
End Time: 00:03
Src Port: 49000
Dst Port: 3389

F3
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 00:03
End Time: 00:04
Src Port: 49000
Dst Port: 3389

F4
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 00:04
End Time: 00:08
Src Port: 49000
Dst Port: 3389

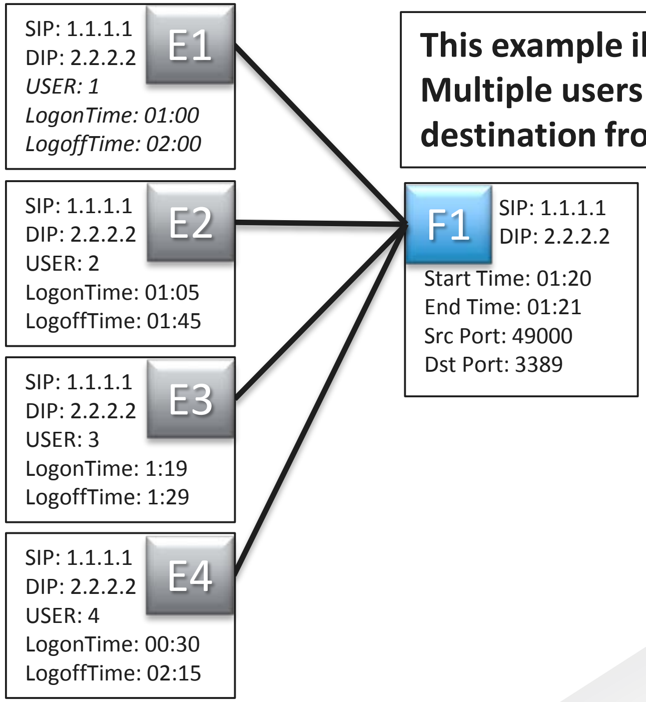
E1
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 1
LogonTime: 00:01
LogoffTime: 00:08

This example illustrates a user session broken up into multiple flows. But….It appears as though the same source port is used for the duration of the user session

Since the 5 tuple (sip, dip, sport, dport, prot) remains consistent, we could aggregate these flows into one.

Findings: Aggregation can help

This example illustrates a multi-user machine: Multiple users log into the same remote destination from this system



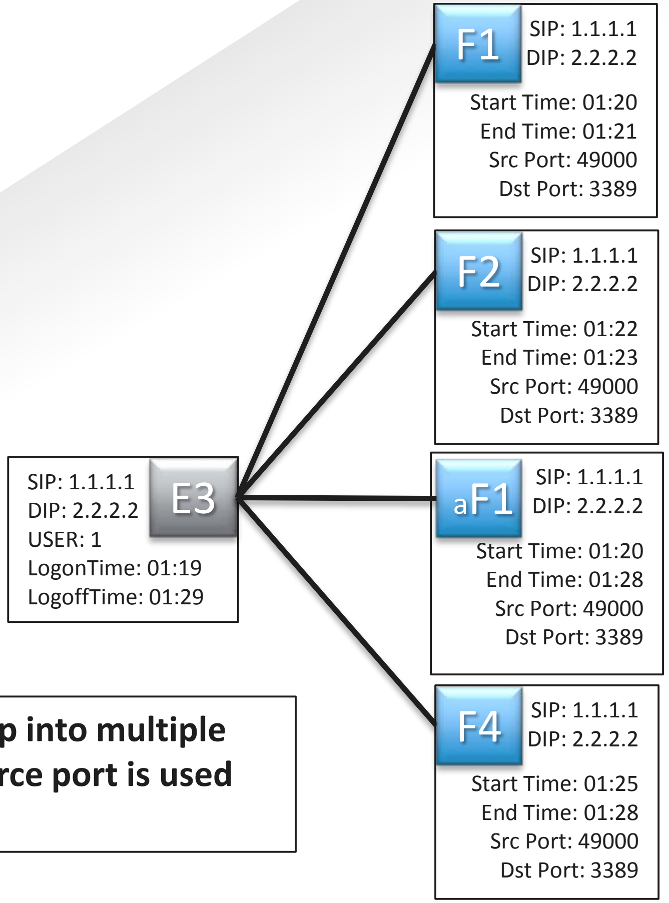
E1
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 1
LogonTime: 01:00
LogoffTime: 02:00

E2
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 2
LogonTime: 01:05
LogoffTime: 01:45

E3
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 3
LogonTime: 1:19
LogoffTime: 1:29

E4
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 4
LogonTime: 00:30
LogoffTime: 02:15

F1
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 01:20
End Time: 01:21
Src Port: 49000
Dst Port: 3389



E3
SIP: 1.1.1.1
DIP: 2.2.2.2
USER: 1
LogonTime: 01:19
LogoffTime: 01:29

F1
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 01:20
End Time: 01:21
Src Port: 49000
Dst Port: 3389

F2
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 01:22
End Time: 01:23
Src Port: 49000
Dst Port: 3389

aF1
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 01:20
End Time: 01:28
Src Port: 49000
Dst Port: 3389

F4
SIP: 1.1.1.1
DIP: 2.2.2.2
Start Time: 01:25
End Time: 01:28
Src Port: 49000
Dst Port: 3389

This example illustrates a user session broken up into multiple flows. But….It appears as though the same source port is used for the duration of the user session

What we learned trying to join session

- “Join” remote login events to NetFlow records using the following conditions
  - Flow records must have a Duration > 0
  - Flow records must have a Destination Port of 3389
  - Event sessions must NOT have a logoff Event ID of 4634.
    - Automatic/systematic logoffs which only last a few seconds
  - Flow Source IP = Event session Source IP
  - Flow Destination IP = Event session Host IP
  - Flow Start Time >= Event Session Start Time (- 1 minute)
  - Flow End Time <= Event Session Stop Time (+ 1 minute)

Mapping Flow to RDP Sessions

- Learned that our NetFlow data had to be aggregated.
  - Many flows for an actual “session”
  - Enabled more accurate joins between RDP session table and Flows
- Joined on…
  - Source and Destination IP
  - Flow start time between event start time +/- 1min
  - Flow end time between event end time +/- 1min
- Created a Mapping table that includes
  - Aggregated FlowID and Logon Event Session ID (LES_ID)
- Created views to represent flow / session data

Fusion enables graph comparisons

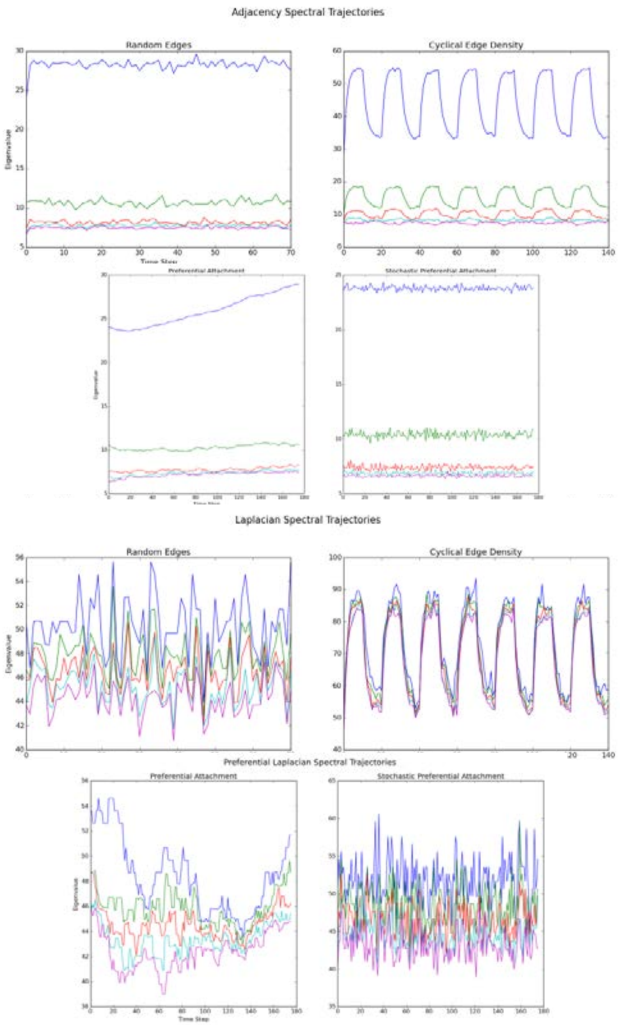
- *Compare a NetFlow graph with the login graph*
- *Enables…*
  - *Higher level understanding of linked events*
  - *Deviations within session behavior*

- *Initial work focused on understanding of RDP sessions and how those would represent themselves in both NetFlow and windows event log data*

Spectral and topological methods applied to both Flow and Login graphs

Dimensionality Reduction for Graphs

- Graphs are complex objects, |V|+|E| pieces of information needed to describe
- ***Aim:*** map a graph into a lower dimensional space, study a dynamic graph sequence by following a trajectory through the lower dimensional space
- Questions
  - What should the mapping be?
  - How do dynamics depend on the mapping?
- Possible mappings
  - **Graph spectrum** – top eigenvalues of an adjacency or Laplacian matrix
  - Degree distribution
  - Information measures on and label distributions
  - Combination of graph measures



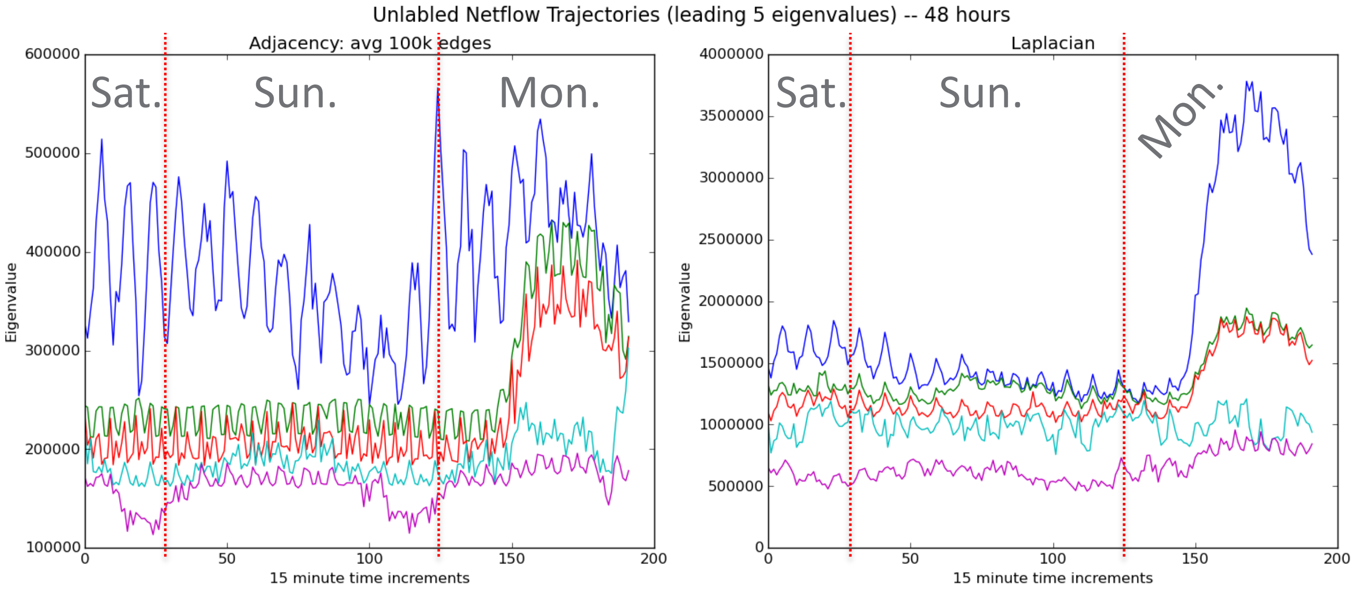
Dynamics of random graph evolution using spectrum of adjacency matrix (top 4 images) and Laplacian matrix (bottom)

Spectral Methods

- For graph $G = (V,E)$ create adjacency and Laplacian matrices
  - Adjacency: $A = \{a_{ij}\}$ where $a_{ij} = 1$ if (v_i, v_j) is an edge, $a_{ij}=0$ otherwise
  - Diagonal degree: $D = \{d_{ij}\}$ where $d_{ii}=deg(v_i)$ and $d_{ij}=0$ if $i \neq j$
  - Laplacian: $L = D - A$
- Graph spectrum is the set of eigenvalues for A or L
- Things we know about the eigenvalues:
  - Laplacian:
    - Eigenvalues are all non-negative
    - Multiplicity of zero eigenvalue is number of connected components
    - Second smallest eigenvalue related to connectivity of graph
  - Adjacency:
    - Largest eigenvalue related to max and average degree
    - Sum of all eigenvalues is zero
- Goal – watch evolution of largest eigenvalues in both graphs to monitor behavior of cyber system

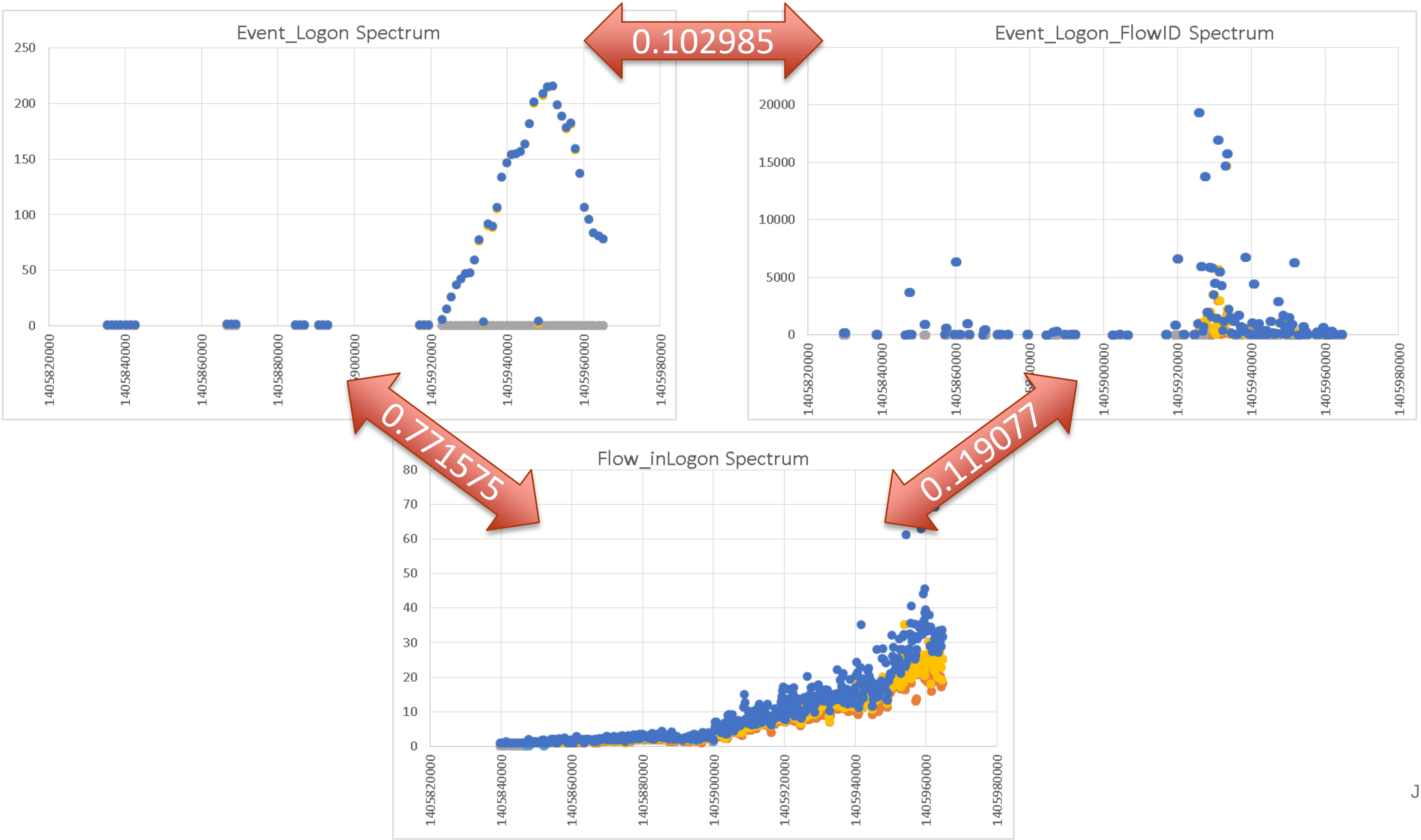
PNNL NetFlow Graphs

- 48 hours of data (5pm Saturday 7/19/14 – 5pm Monday 7/21/14)
  - Each graph spans 60 minutes with 45 minute overlap between consecutive graphs
- Regular cyclic behavior on weekend, ramp up in behavior Monday morning
- Problem: We have no ground truth about events in this data
  - We have talked with our cyber team to confirm that these regular-looking events are expected



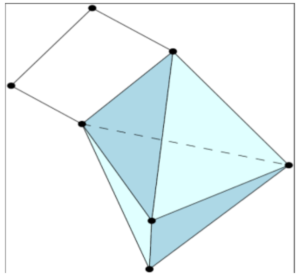
Comparison of Flow and Login Spectrum

- Start time = 7/19/2014, 6:33:20 PM
- End time = 7/21/2014, 3:00:00 PM

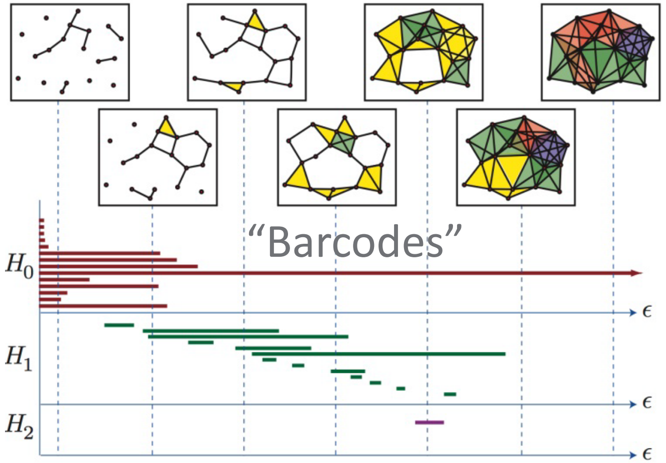


Finding the Shape of Data

- **Homology:** a characterization of the "holes" in a *single topological object* across different dimensions
  - Not-filled-in 4-cycle attached to hollow double tetrahedron
  - Has one hole in one dimension (the not-filled-in 4-cycle) and one hole in two dimensions (the hollow double-tetrahedron)



- **Persistent Homology (PH):** Given a *single data set* (as a point cloud or points in a metric space), what is its most prevalent underlying topological space?
  - Sweep through different distance thresholds and characterize space's shape (homology) at each
  - Most "persistent" features indicate most likely shape of data sample space

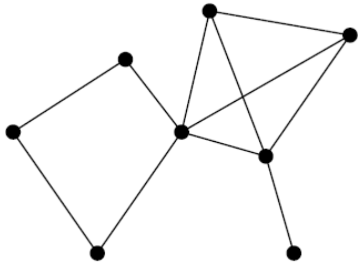


Application to Cyber Systems

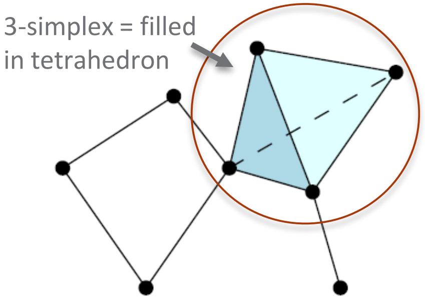
- Cyber system modeled as a dynamic graph – sequence of graphs corresponding to rolling time intervals
- PH on each graph in the sequence
  - A single graph thought of as a metric space with the *shortest path metric*
    - Also investigating other metric spaces and point clouds from each graph
  - Resulting *Betti numbers* provides a signature of the underlying shape of the graph when considered as this metric space
  - Evolution of this shape gives characterization of system behavior
- For neighboring graphs (in time) compare their Betti number vectors and plot distance as it changes over time

Topological spaces from a single graph

- For graph G = (V,E) create *filtration* of *simplicial complexes (SC)* based on shortest path distance:
  - d=0 – all vertices isolated (every vertex is distance zero only to itself)
  - d=1 – connect vertices at distance 1 (add all edges) and create *simplicies* for all completely connected subgraphs
  - d=2 – connect vertices at distance 2 and create *simplices* for all completely connected subgraphs
  - …
- SC for distance d is always contained in SC for distance d+1



Original graph



3-simplex = filled
in tetrahedron

Distance 1

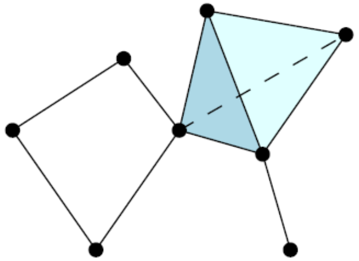
Filtration = sequence of objects with d^{th} object contained in $d+1^{st}$ object for all d

k-simplex = convex hull of $k+1$ independent points in dimension k

e.g., 0-simplex is a point, 1-simplex an edge, 2-simplex a triangle, 3-simplex a tetrahedron

Comparing two graphs topological structures

- **Definition:** The nth *Betti number* is the rank of the nth homology group
  - b_0 = # of connected components
  - b_1 = # of 1 dimensional loops
  - b_2 = # of 2 dimensional voids or cavities
- PH gives a sequence of Betti numbers for each dimension



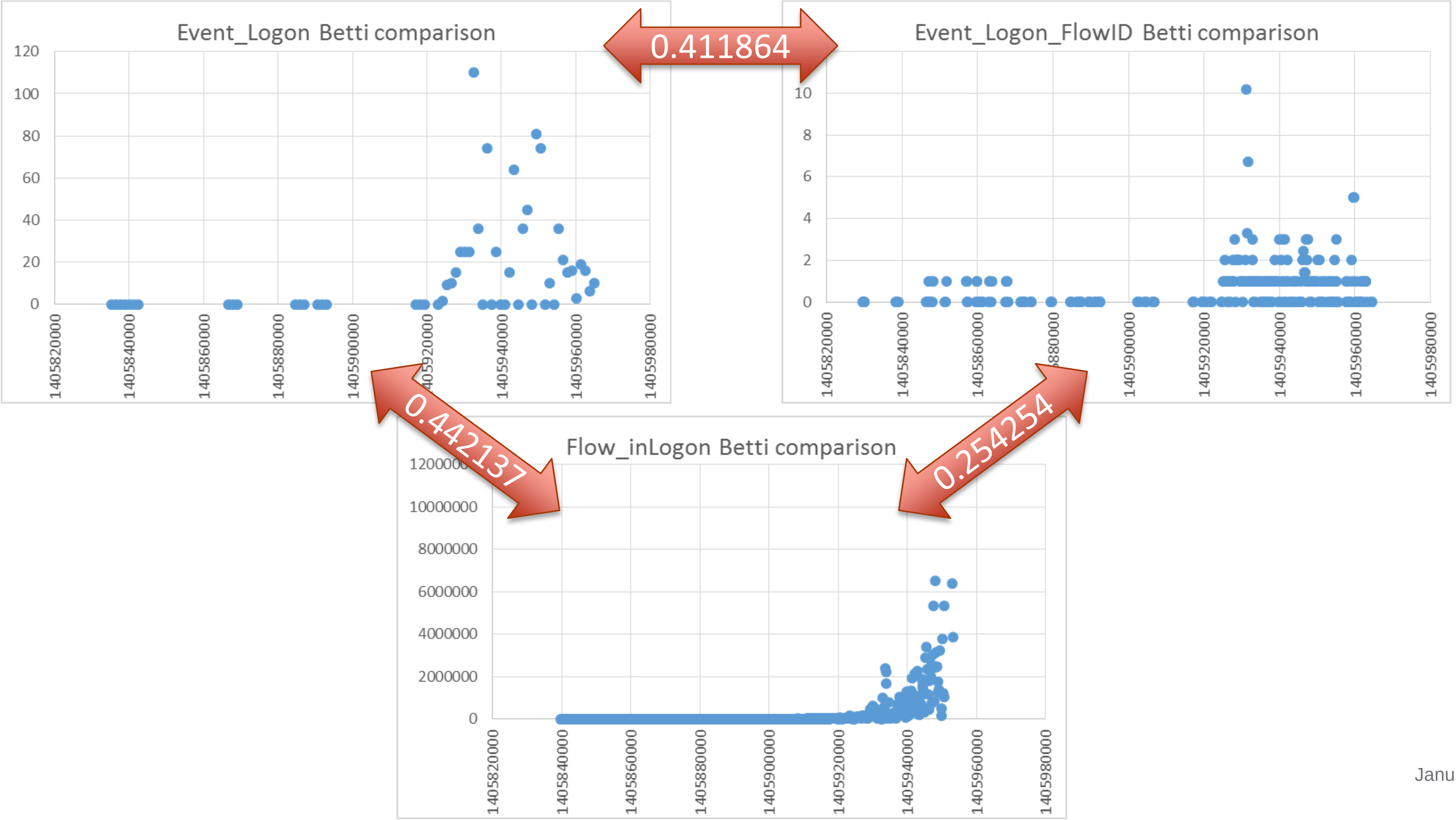
b_0=1; b_1=1; b_2=0

	Dimension		
Distance ≤	0	1	2
0	163	0	0
1	58	0	0
2	58	0	228
3	58	0	1082
4	58	0	2438

- Comparing two of these Betti number sets
  - Vectorize each and calculate Euclidean distance between them
  - E.g., < 163, 0, 0 | 58, 0, 0 | 58, 0, 228 | 58, 0, 1082 | 58, 0, 2438 >

Flow vs. Login Betti Numbers

- Start time = 7/19/2014, 6:33:20 PM
- End time = 7/21/2014, 3:00:00 PM



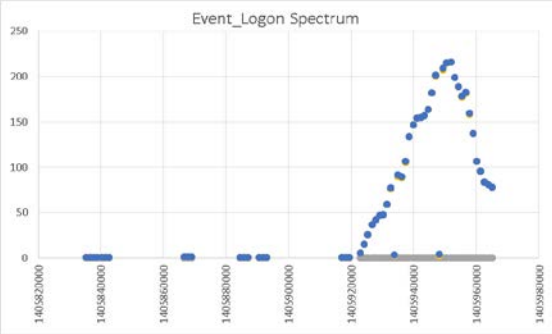
Comparison of Spectrum and Betti numbers

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

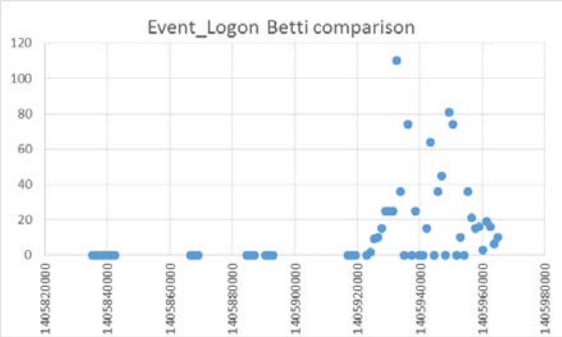
Spectrum values

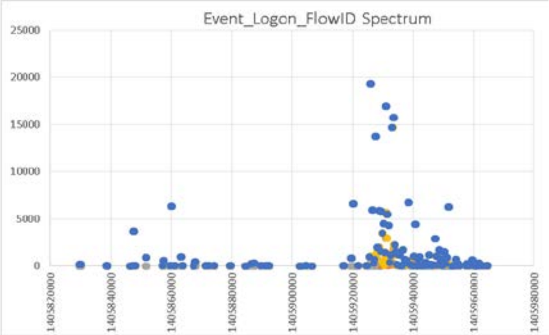
Correlation values

Betti comparison

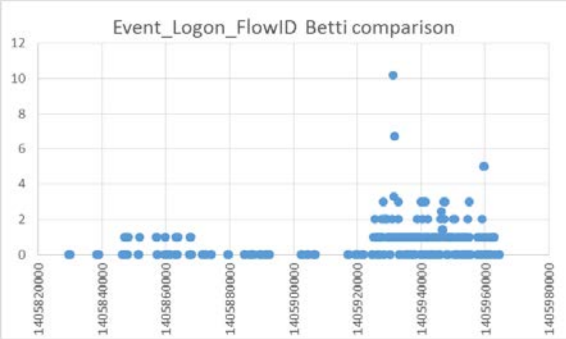


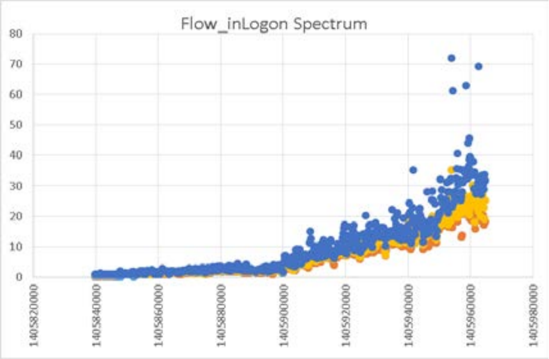
0.137568



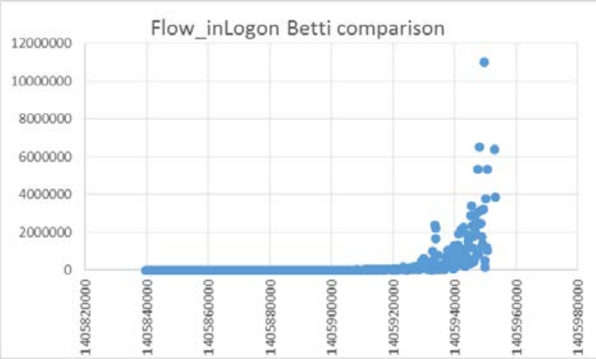


0.328083





0.494285



Summary & Future Work

- Automation of data ingest and sessionization of flow and login records
- Initial topological analysis of NetFlow and login data shows
  - PH and Betti number analysis is similar to graph spectrum with some weak correlation between the two
  - Login and Flow record data (both spectrum and Betti number comparison) show some correlation as well
- Current work in developing methods to draw cyber-relevant conclusions from the results of our topological analysis methods
- Future work will refine algorithms and further investigate the link between analyses on NetFlow and login data

Acknowledgements

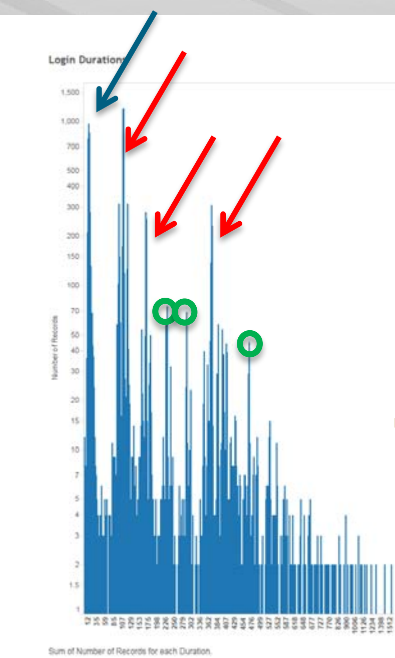
- The research described in this presentation is part of the Asymmetric Resilient Cybersecurity Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.
  - ARC leadership: Nick Multari, Chris Oehmen

- Topological Analysis of Graphs (TAGs) additional team members
  - Paul Bruillard
  - Chase Dowling
  - Katy Nowak

Backup Slides

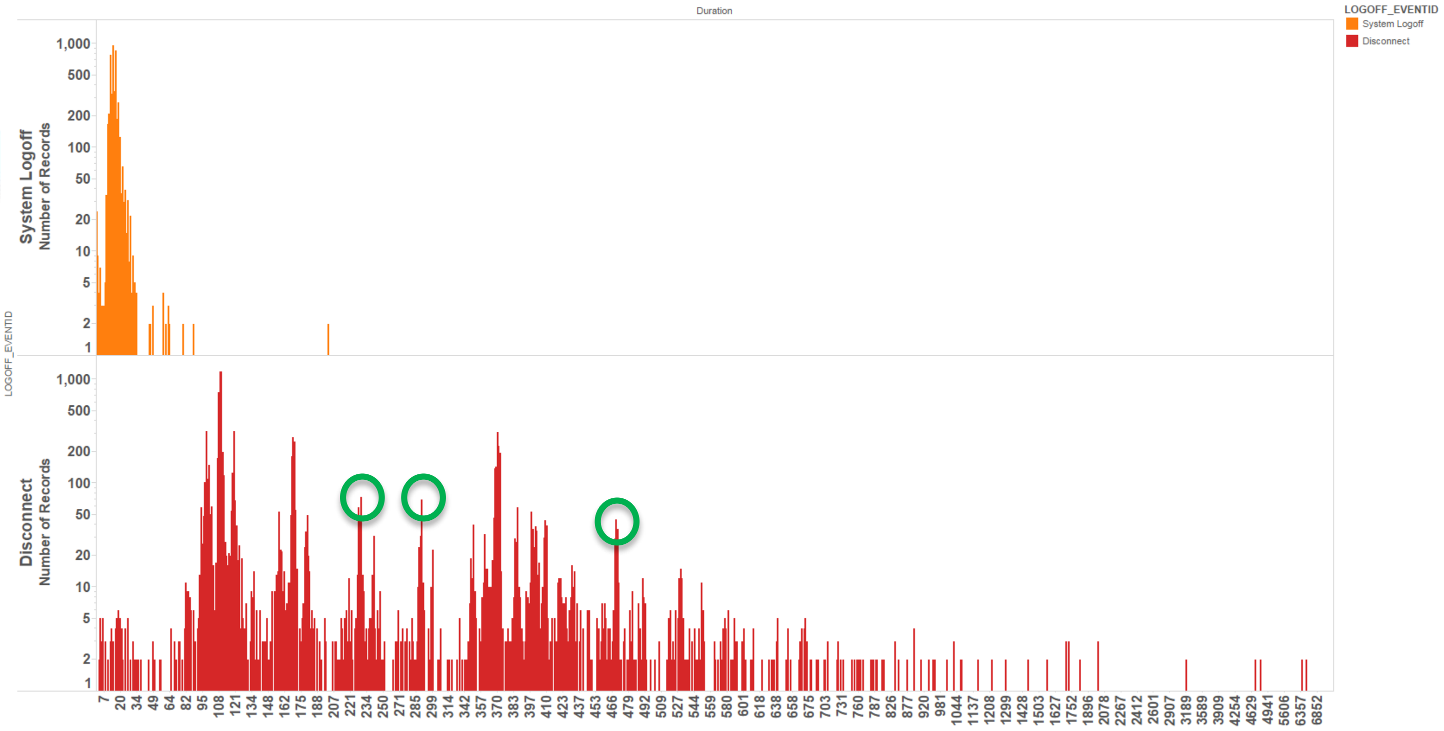
Login duration

- Notice the multiple modality of the login durations
- Systematic logoff events explain first mode
- Other modes are in disconnect logoff type



Sum of Number of Records for each Duration.

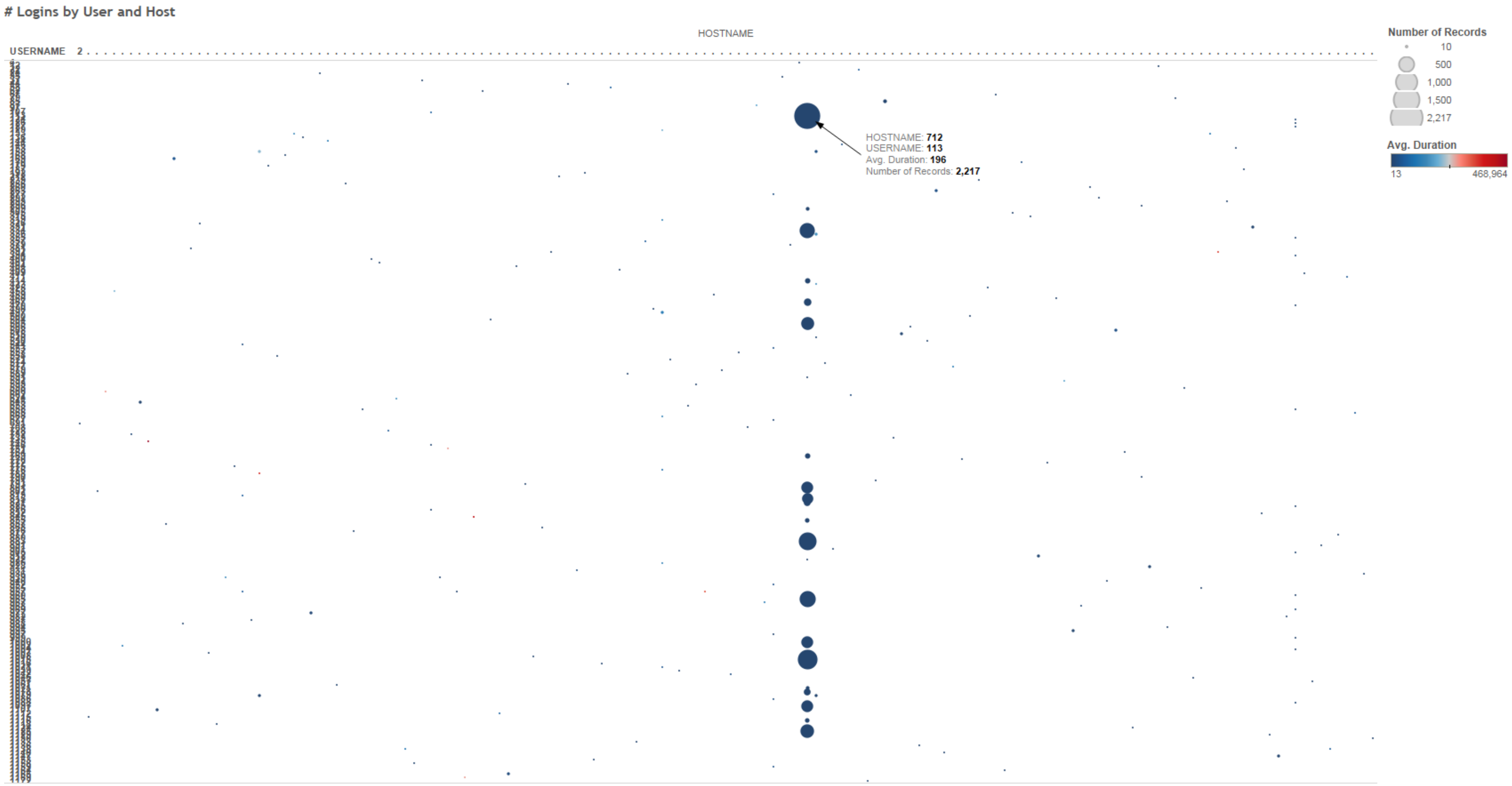
Login Durations by Login Type



Sum of Number of Records for each Duration broken down by LOGOFF_EVENTID. Color shows details about LOGOFF_EVENTID. The view is filtered on LOGOFF_EVENTID and Duration. The LOGOFF_EVENTID filter excludes User Logoff and 4647. The Duration filter keeps 2,186 of 2,554 members.

# Logins by User and Host

- Host 712 is heavily used by many users, much more than any other host



# Logins by User and Host

Average of Duration (color) and sum of Number of Records (size) broken down by HOSTNAME vs. USERNAME. The view is filtered on sum of Number of Records, which ranges from 10 to 2,217.