High-Throughput Real-Time Network Flow Visualization

Daniel Best

Research Scientist | Information Analytics

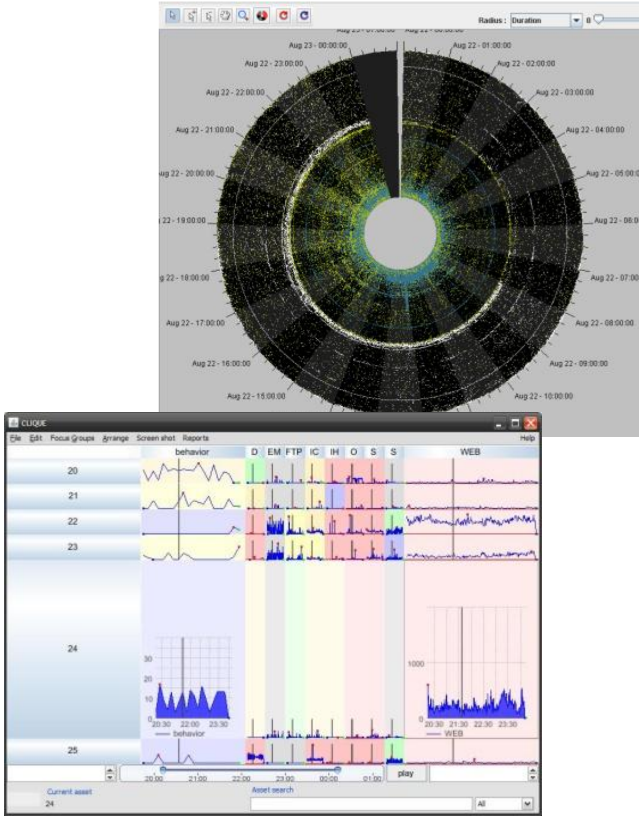
daniel.best@pnl.gov

Douglas Love, Shawn Bohn, William Pike

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Tools and a Pipeline to Provide Defense in Depth



- Traffic Circle
  - Visualization for situational awareness
- Correlation Layers for Information Query and Exploration (CLIQUE)
  - Network behavior visualization using LiveRac interface
- Middleware for Data-Intensive Computing (MeDiCi)
  - Data pipeline

Motivation

- Improve upon current analysis capabilities
  - Provide a mechanism for multiple tools to feed off the same data
  - Move away from batch processing of flow data
  - Support both forensic and real-time monitoring capabilities
- Implement a streaming analytics tool set
  - Handle large volumes of flow data (millions to billions) per view
  - Help analysts gain situational understanding of current state of a network
- Provide engaging flow visualizations
  - Visualization of both raw data and aggregates
  - Automated identification of off normal conditions

What our users want

- They want to know what normal looks like – from an individual host, to a group, to an enterprise
- They want ways to overcome limitations in analyzing raw transactions
  - Lots of data (billions of transactions/day)
  - Lots of unique actors (IPv6: 6.67 * 1027 IP addresses per square meter on Earth)
  - Lots of noise
- If they know what they're looking for, they can build a signature to detect it. **But what's in the data that they don't already know to look for?**
  - Need to link data reduction techniques with exploratory analysis interfaces
- They want to know where to focus!

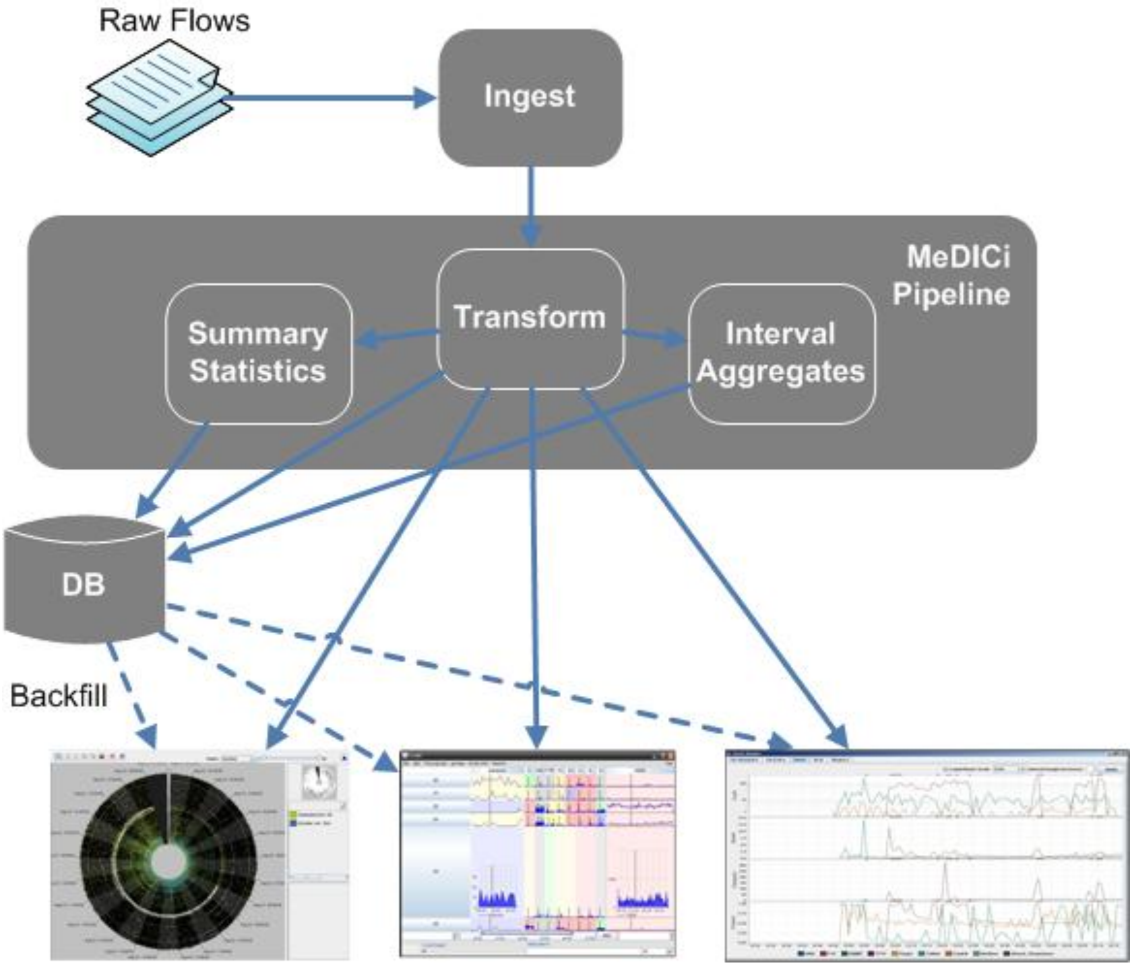
Middleware for Data-Intensive Computing (MeDICi)

- Provides a high-throughput data communication and processing pipeline
  - Creates the substrate for real-time information sharing
- Mechanism to hand information to multiple tools
  - Multiple tools “subscribe” to MeDICi data, so that tools can be combined for defense in depth
- Capable of streaming data transformations
  - Handles data changes needed by a client prior to being transmitted to that client
    - Offloads computational cycles to the MeDICi server

Using MeDICi for real-time analytics

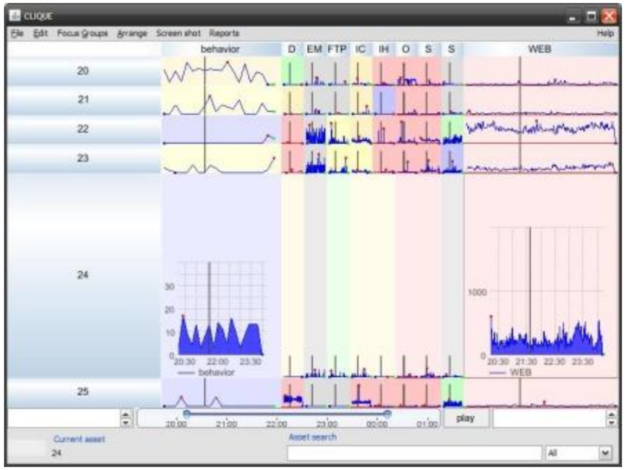
- Components can be created by different developers using various languages
  - Traffic Circle and CLIQUE use Java, but components from other languages can be wrapped easily
- Information is passed between components using a producer / listener mechanism
  - Apache ActiveMQ message broker
  - JMS messaging standard
- Components are chained together to create a pipeline
  - Aggregates
  - Summary Statistics
  - Others

A sample MeDICi pipeline



CLIQUE

- Produces behavioral summaries based on a live sensor feed
  - A behavior is a model of predicted activity based on past activity
  - CLIQUE helps visualize the deviation of an actor from its predicted activity
  - Working at the level of behaviors helps cope with large data volumes
  - Display “walks” along with incoming flow information to show current state
- Helps highlight trends and patterns in high-volume flows
  - Capability to compare behavior deviations with category activity

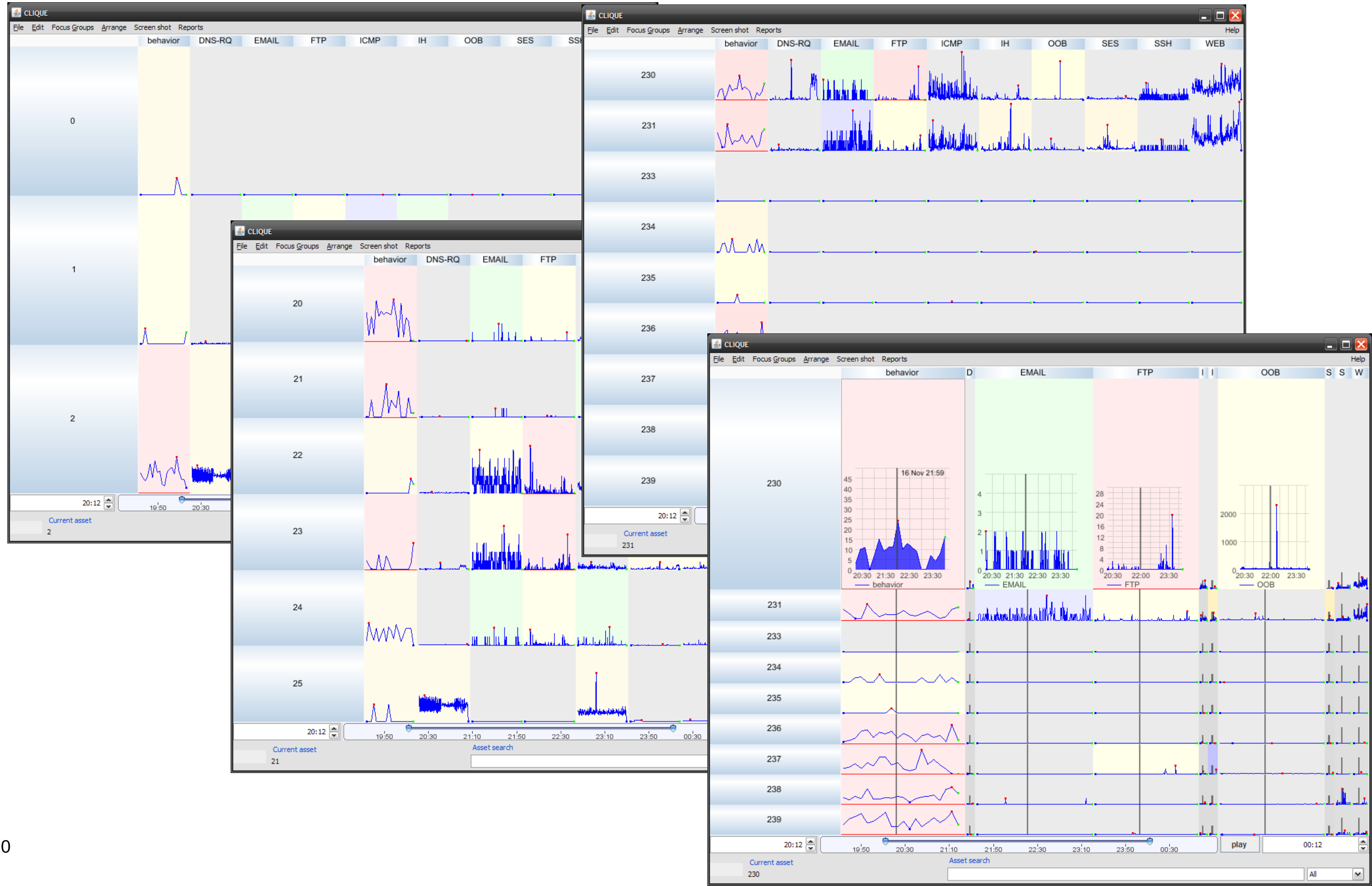


Current asset
24

CLIQUE behavior analysis

- Utilizes Symbolic Aggregate approXimation (SAX)
  - To deal with the scaling issue in the temporal dimension (scales by summarization/aggregation)
- Creates a SAX representation across all categories for a given actor
- Converts SAX representation to a glyph
  - Produces a language
- Creates matrix of glyphs and temporal sub-segments
- Compares current matrix and historic matrix to yield behavior deviation plot

CLIQUE Walkthrough

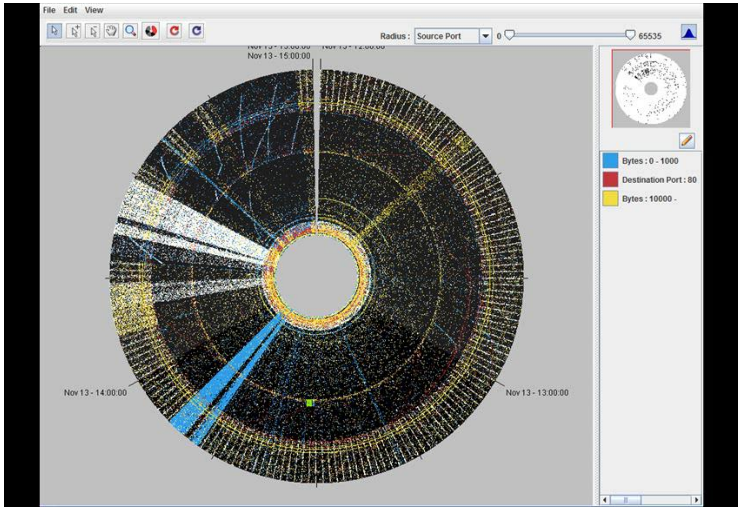


Traffic Circle

- Interactive visualization of patterns in high volume flow data
  - When you can see more data you can see patterns previously hidden when examining by subsets
- Visualizes raw flow information
  - Drill-through from CLIQUE
  - Manages throughput by utilizing a threaded architecture that distributes query and rendering
- Can operate in forensic mode or real-time animation mode

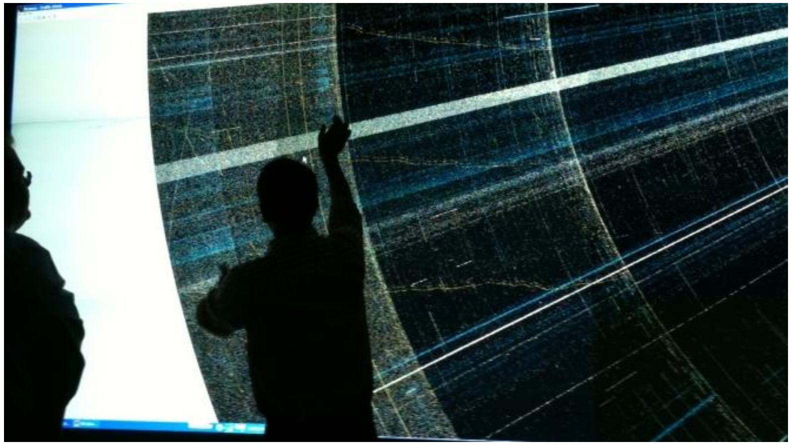
Traffic Circle

- Circular “time wheel” metaphor
  - Flows ordered by start time
  - Arc length corresponds to duration
  - Spinnable interface
- Filters can be added
  - Color coding
  - Hide / show capability
- Operationally demonstrated at data volumes upward of 125 million flows
  - Using high-performance backfill database

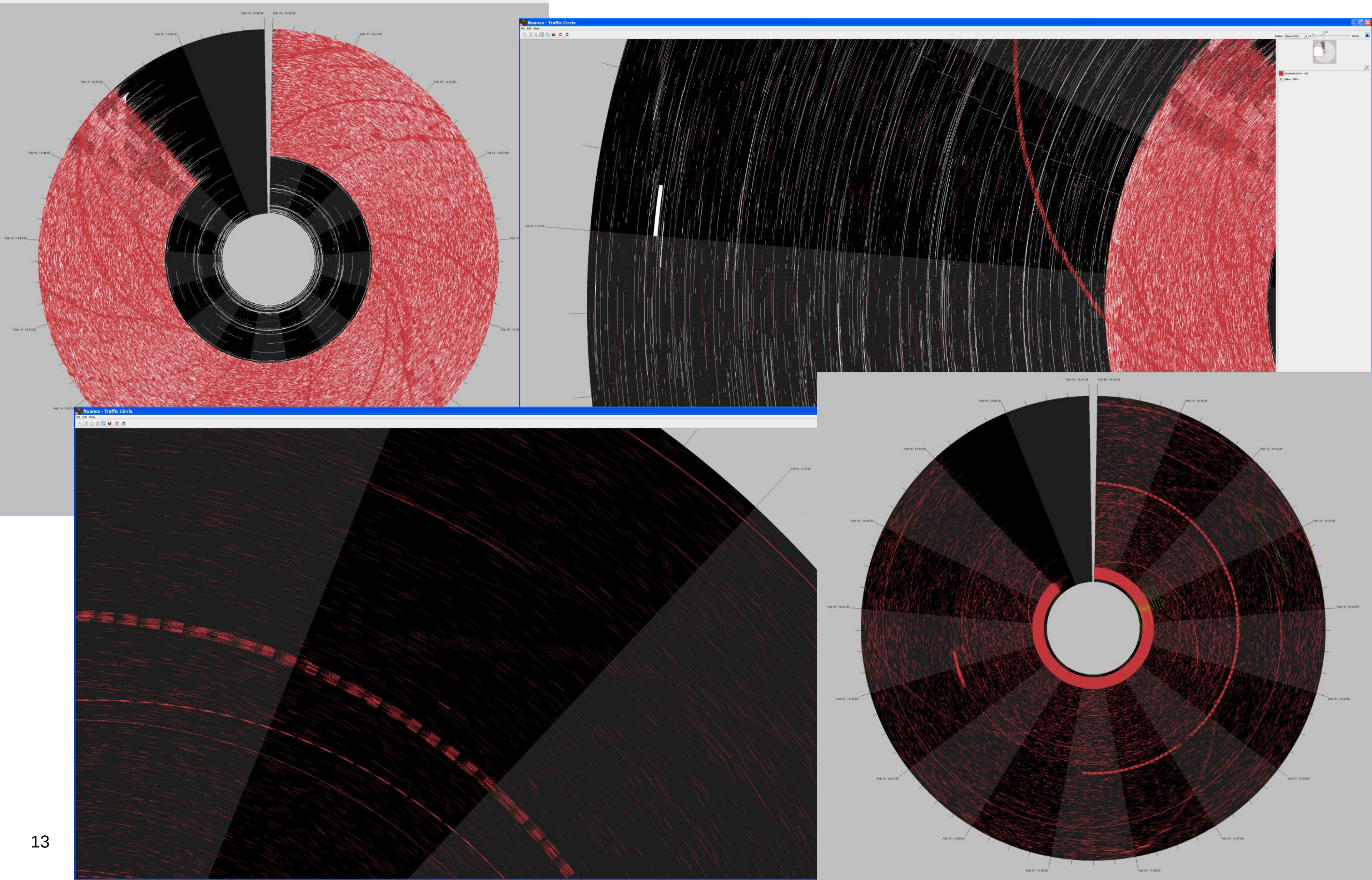


Radius : Source Port 0 65535

🟦 Bytes : 0 - 1000
🟥 Destination Port : 80
🟨 Bytes : 10000 -



Traffic Circle Walkthrough



Conclusions

- Visualization at different levels of abstraction supports situational awareness in large data sets

- High-throughput pipelines are necessary to scale visual analysis to operational data volumes

- Modeling helps analysts baseline normal activities and quickly identify off-normal conditions

Future Directions

- Develop a predictive capability
  - Nascent behavioral changes can be detected visually in real-time in Traffic Circle and CLIQUE
  - CLIQUE classifier enables **sequence detection** for proactive threat identification

- Explore extensions to other domains
  - Financial fraud detection
  - SCADA system reliability and security