

The background of the slide features a large, faint watermark of the University of Delaware seal. The seal is circular and contains the text "UNIVERSITY OF DELAWARE" around the perimeter. In the center, there is a shield with the words "GRAMM", "PHILOL", "RHETOR", "ETHIC" on the left and "METAPH", "LOGICA", "MATHE" on the right. Below the shield, the year "1743" is visible.

# Visualization of Shared System Call Sequence Relationships in Large Malware Corpora

Authors:

Josh Saxe   David Mentis   Chris Greamo

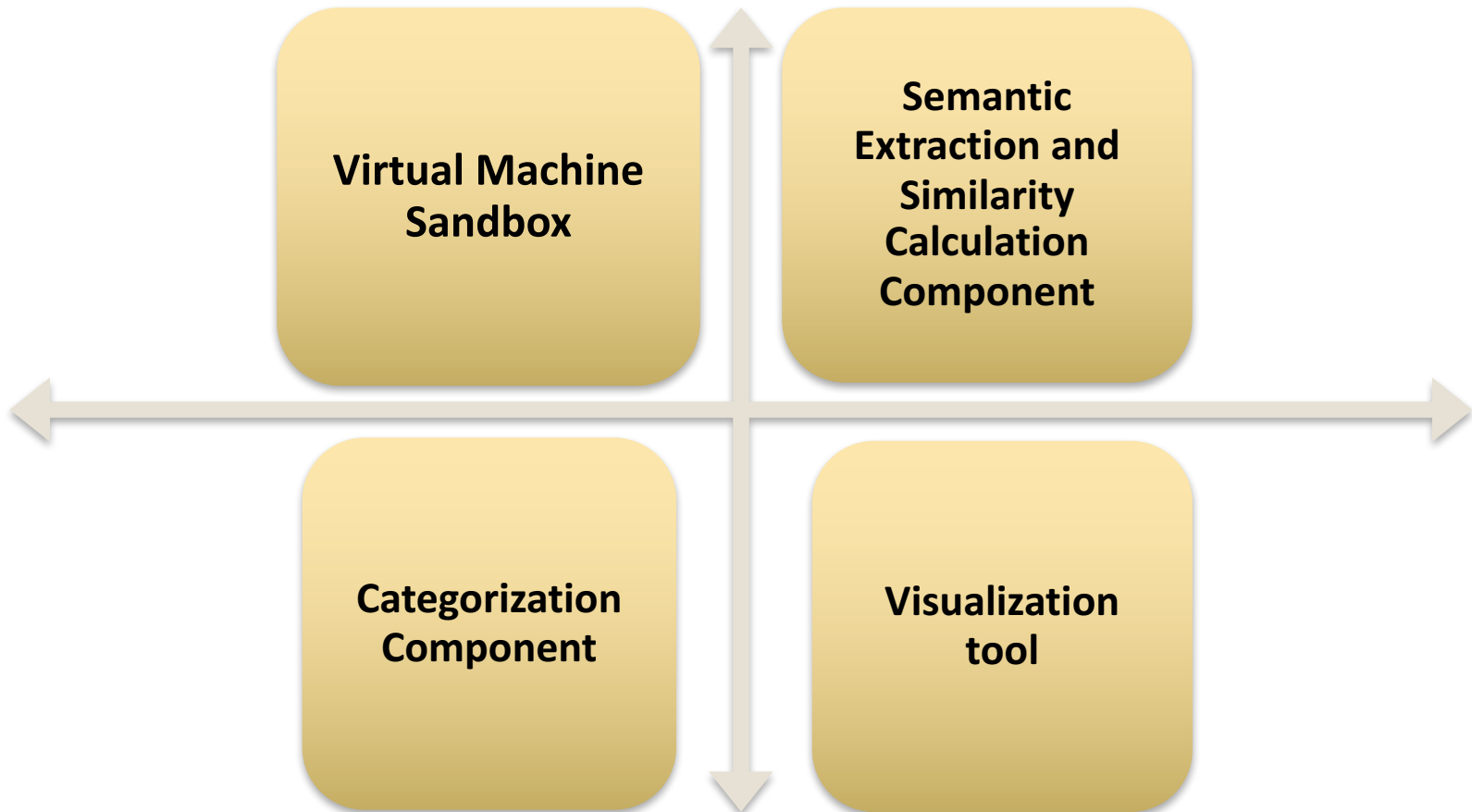
Presenter:

Zicheng Liu

# Background

- Deluge of new malware Variants
- Few effective methods to address this problem
- Need for interpretable visualization

# System Introduction



# Overview


Inspector Panel

Filter Panel

Projection of Similarity Matrix



# Pipeline

- 
- Semantic Sequence Extraction
  - Similarity Matrix Computation
  - Visualization
  - Overall Analysis

# Semantic Sequence Extraction

- Two Intuitions
  - Improbable state transitions
  - Divergent objects locate the meaningful partitions

# Example output

## Example Semantic Subsequence Extraction from Sample of Variant Type "menti-gtmr"

...|

---

```
ReadFile C:/WINDOWS/WinSxS/x86_Microsoft.Windows.GdiPlus_6595b64144ccf1df_1.0.2600.5512_x-ww_dfb54e0c/GdiPlus.dll
ReadFile C:/WINDOWS/WinSxS/x86_Microsoft.Windows.GdiPlus_6595b64144ccf1df_1.0.2600.5512_x-ww_dfb54e0c/GdiPlus.dll
ReadFile C:/WINDOWS/WinSxS/x86_Microsoft.Windows.GdiPlus_6595b64144ccf1df_1.0.2600.5512_x-ww_dfb54e0c/GdiPlus.dll
ReadFile C:/WINDOWS/WinSxS/x86_Microsoft.Windows.GdiPlus_6595b64144ccf1df_1.0.2600.5512_x-ww_dfb54e0c/GdiPlus.dll
ReadFile C:/WINDOWS/WinSxS/x86_Microsoft.Windows.GdiPlus_6595b64144ccf1df_1.0.2600.5512_x-ww_dfb54e0c/GdiPlus.dll
ReadFile C:/WINDOWS/WinSxS/x86_Microsoft.Windows.GdiPlus_6595b64144ccf1df_1.0.2600.5512_x-ww_dfb54e0c/GdiPlus.dll
```

---

```
RegQueryKey HKLM/SOFTWARE/Microsoft/Windows-NT/CurrentVersion/Fonts
```

---

```
RegCreateKey HKU/S-1-5-21-436374069-813497703-1177238915-1004/Software/Microsoft/GDIPlus
```

---

```
RegQueryValue HKU/S-1-5-21-436374069-813497703-1177238915-1004/Software/Microsoft/GDIPlus/FontCachePath
```

---

```
QueryOpen C:/RUNME/ShFolder.DLL
```

```
QueryOpen C:/WINDOWS/system32/shfolder.dll
```

```
CreateFile C:/WINDOWS/system32/shfolder.dll
```

```
CreateFileMap C:/WINDOWS/system32/shfolder.dll
```

```
CreateFileMap C:/WINDOWS/system32/shfolder.dll
```

```
Load Image C:/WINDOWS/system32/shfolder.dll
```

---

...

# Semantic Sequence Extraction(cont'd)


- Three Steps
  - Derive a Markov Chain
  - Compute parameter similarity
  - Insert partitions and extract subsequences



# Derive a Markov Chain

- Sort Malware behavior log
- Define nodes
- Calculate Transition probabilities

# Pipeline

- 
- Semantic Sequence Extraction
  - **Similarity Matrix Computation**
  - Visualization
  - Overall Analysis

# Compute Parameter Similarity

- A score
- Compare the parameter strings of two adjacent System calls
- More different, more lower


# Insert Partitions and Extract Subsequences

- Average the system call transition's parameter similarity score and its transition probability
- Check if this value is below a threshold 0.3
  - If is below, interpret as a partition

# Similarity Matrix Computation

- Boolean Sample Vectors based on the occurrence of variable-length sequences
- Compute a Jaccard Index pairwise based on the vectors

# Pipeline

- 
- Semantic Sequence Extraction
  - Similarity Matrix Computation
  - **Visualization**
  - Overall Analysis

# Visualization

- Sequence Visualization
- Similarity Map
- Filter Panels

# Three Views Are Linked

Mouse Over  
on a System  
Call Sequence

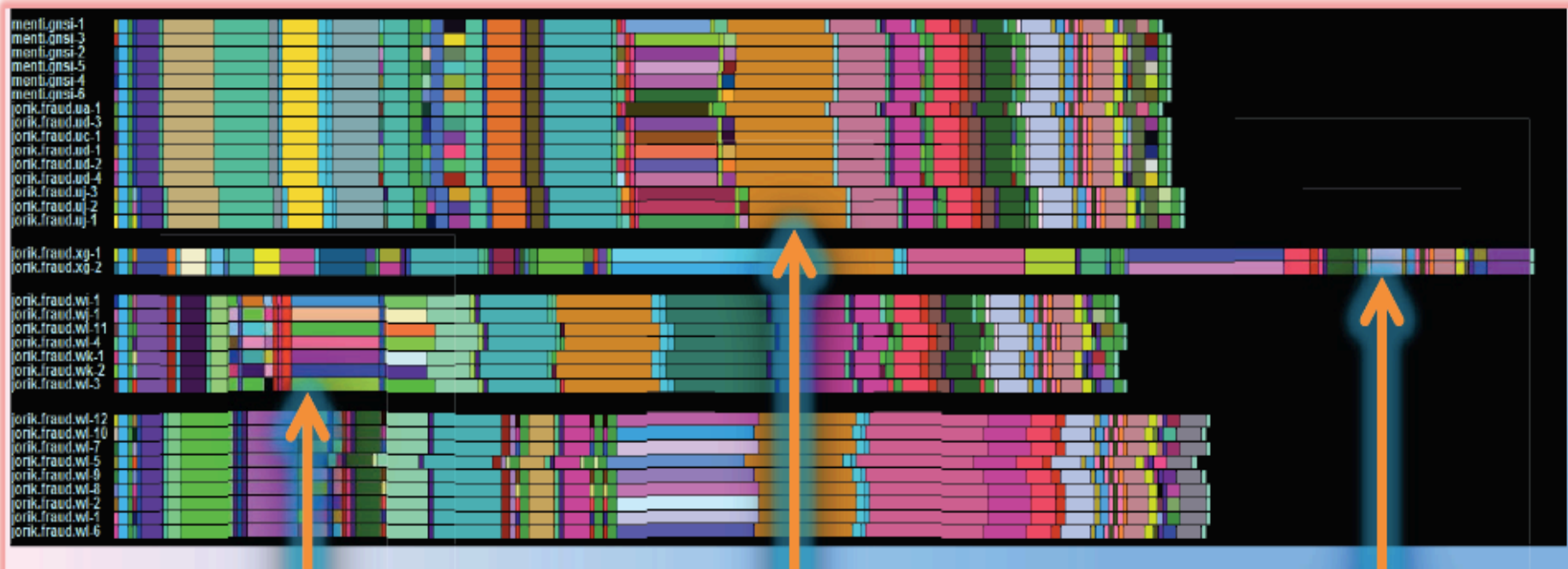
A Sequence  
detail text  
box

Samples that  
executed the  
sequence  
highlight





# Sequence Visualization



Rainbow colored but equal length bands depict variation in sample behavior.

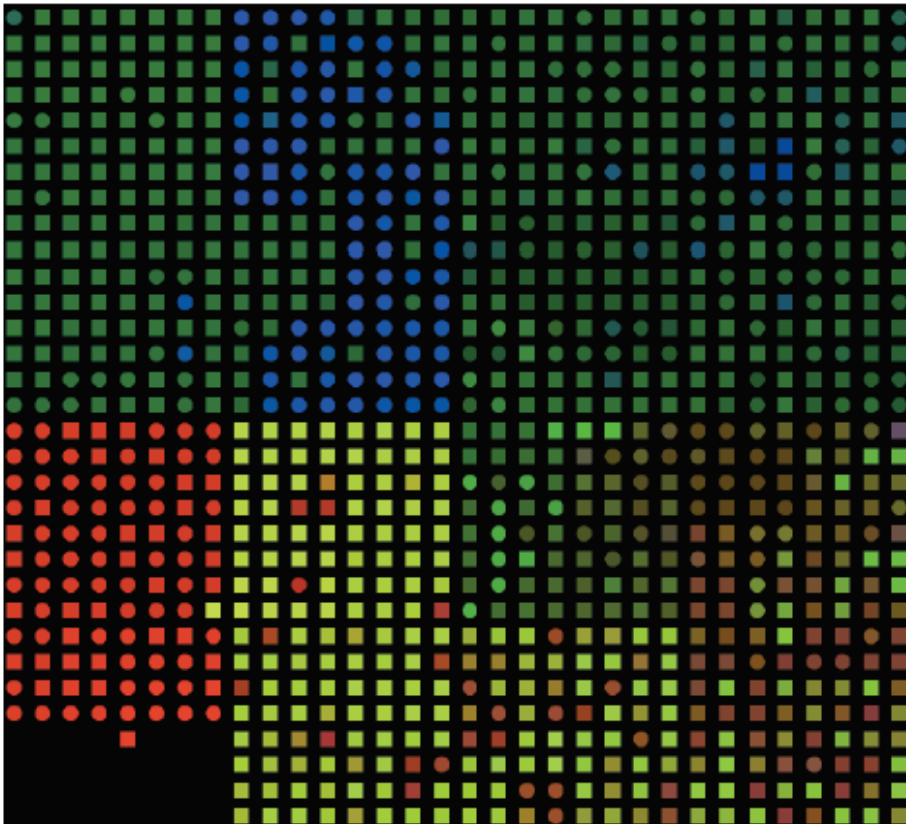
A cluster of malware samples

jorik.fraud samples appear differently

# Sequence Visualization(Cont'd)

- It reveals similarities and differences between malware samples
- Assign each unique sequence a unique color

# Sample Similarity Map



Colored shapes represent clusters

- Circles -- known
- Squares -- unknown

# Sample Similarity Map(Cont'd)

- Principal Component Analysis
- Sort by the first principal component
- Color nodes
  - Green, Red or Blue with respective computed positions

# Filter Panels

- Behavioral Traits are shown on the left
- Support users to insight into the similarities and differences between regions

# Filter Panels(Cont'd)

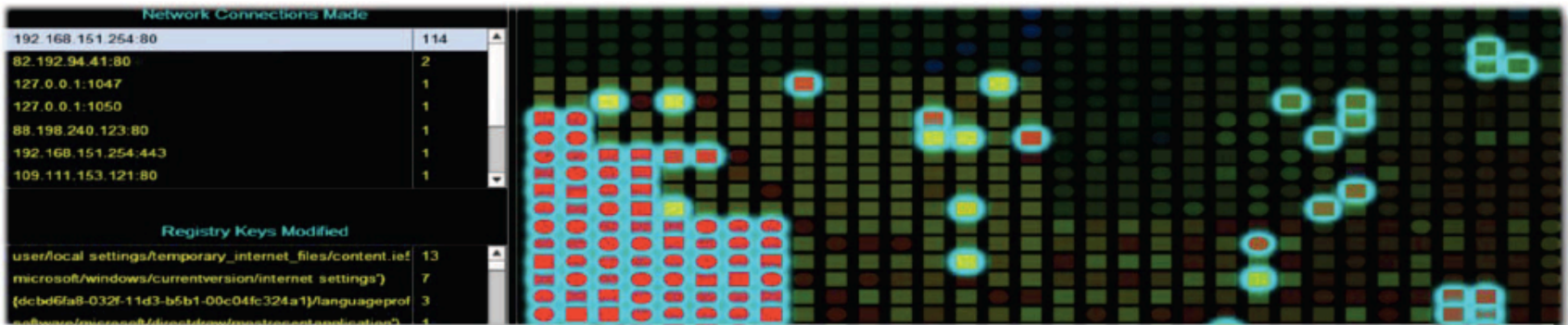


Figure 12. The user brushes an IP address and port on the left, and the samples that connect out to that IP address and port highlight on the right.

# Accomplishment

- Helpful in relating novel samples to known malware samples
- Provide user with visual insight into a focal system call sequence
- Visually discover and investigate cluster structure in malware

Thank You