CS259D: Data Mining for Cyber Security

Insider threats: banking/finance sector

- US Secret Service National Threat Assessment Center (NTAC), Insider threat study (2004)
  - 83% of incidents executed from within organization, during normal business hours
  - Financial loss in almost all insider incidents
  - Loss > $500K in 30% of cases

Behavioral Biometrics

- Applications
  - User authentication
  - Intrusion detection
- Advantages
  - Very low impact on usability
- Most useful in multimodal systems
  - Complement to more robust methods
  - Highly sensitive to means of implementation
    - Keyboard hardware

Implementation factors

- Required equipment
  - None
  - Multiple cameras
  - EEG sensors
- Enrollment time
  - Training time for system to recognize the user
- Persistence
  - Time it takes for features to change
- Obtrusiveness
- Error rates
  - False rejection rate (FRR)
  - False acceptance rate (FAR)
  - Equal error rate (ERR): error rate when FRR=FAR

Behavioral biometrics: categories

- Authorship
  - Text or drawing made by user
    - Vocabulary, punctuation, brush strokes
- HCI-based biometrics
  - Input interaction: keystrokes, mouse, haptics
  - Software interaction: strategy, knowledge, skill
- Indirect HCI-based biometrics
  - Low-level system activities
    - System call traces, audit logs, program execution traces, registry access, storage activity, call-stack data analysis
- Kinetics: Motor-skills based biometrics
  - Rely on proper functioning of brain, skeleton, joints, nervous system
- Purely behavioral biometrics
  - Walking style, typing style, gripping style

Authentication

- Knowledge: password, PIN
  - Something you *know*
- Objects: ID card, credit card, access token
  - Something you *have*
- Biometrics
  - Physiological: fingerprints, retina pattern
    - Something you *are*
    - Issues: implementation cost, acceptability, stability
  - Behavioral: signature, gait, keystroke dynamics
    - Something you *do*

Re-authentication

- Continuous authentication of a user for the duration of the user's login session
- Current user same as user who authenticated
- Stolen credentials
- Approaches
  - Indirect: profiling OS and applications
    - System call invocations, call-stack data operation, program trace analysis
  - Direct: profiling valid user
    - Command line input data, keystroke dynamics, mouse activity, GUI events

Deployment scenarios

- Open setting
  - Public library, internet café
  - Unsupervised learning
  - Only data from valid user available for profiling
- Closed setting
  - Corporate office, government building
  - Possible to collect data from all users
  - Supervised learning

Multimodal data analysis

- Combine sources at
  - Data level
    - Can miss on characteristics
  - Feature level
    - Classifier using all features put together
  - Classifier level
    - Separate classifier per data source
    - Voting scheme for final decision

User re-authentication

- Goal: detect and flag anomalies in behavior of current user
- Process
  - Collect clean data from each user
  - Build profile of normal behavior for each user
  - Use profile to compare behavior of current user with that of a valid user
  - Any significant behavioral difference flagged

Data sources

- Keystrokes
- Mouse movements
- GUI events
- Data point: Event ID (assigned by OS), X & Y screen coordinates, event time, application

```
160 1099 1013  6.064E-1 C:\WINDOWS\Explorer.EXE
514 1384   16  6.067E-1 C:\WINDOWS\System32\BROWSEUI.dll
  0 1226  186  6.067E-1 0
  0  518  986  6.067E-1 0
160  526 1012  6.067E-1 C:\WINDOWS\Explorer.EXE
  0  525 1005  6.069E-1 0
513  212  233  6.071E-1 C:\WINDOWS\System32\mshtml.dll
```

ID X Y Time Application

Mouse events

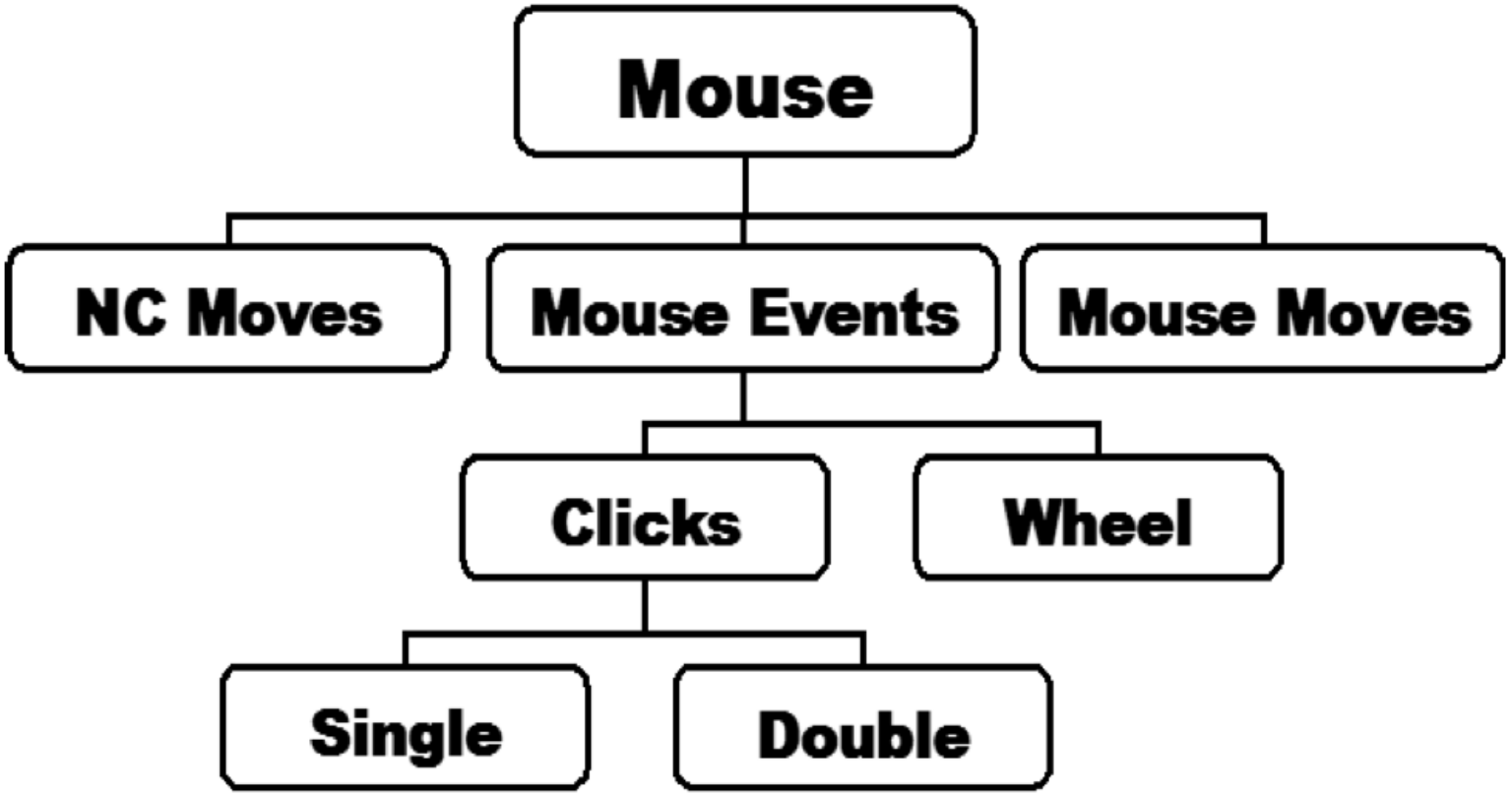
ID	Message Name
513	WM_LBUTTONDOWN
514	WM_LBUTTONUP
515	WM_LBUTTONDBLCLK
516	WM_RBUTTONDOWN
517	WM_RBUTTONUP
518	WM_RBUTTONDBLCLK
519	WM_MBUTTONDOWN
520	WM_MBUTTONUP
521	WM_MBUTTONDBLCLK
522	WM_MOUSEWHEEL

ID	Message Name
161	WM_NCLBUTTONDOWN
162	WM_NCLBUTTONUP
163	WM_NCLBUTTONDBLCLK
164	WM_NCRBUTTONDOWN
165	WM_NCRBUTTONUP
166	WM_NCRBUTTONDBLCLK
167	WM_NCMBUTTONDOWN
168	WM_NCMBUTTONUP
169	WM_NCMBUTTONDBLCLK
160	WM_NCMOUSEMOVE

Mouse data collection

- Client area: area of application window below the menu and toolbars
- Rate of client area mouse movements
  - several hundred movements per second
  - Too large, did not collect all movements
  - Recorded every 100ms if & only if cursor position changed
- Single & double clicks
  - Most infrequent events
  - Higher granularity not helpful

Mouse events hierarchy



Mouse event features

- Over a window of W data points
  - # of events in each mouse category & subcategory
  - Other features extracted from two subsequent events or two events separated by K data points (events of same category or subcategory)
    - K a parameter for each user and type of event
  - Features: Mean, stdv, skewness (3rd moment) of
    - Distance
    - Speed = distance(P,Q)/time(P,Q)
    - Angle of orientation
    - X coordinate
    - Y coordinate
    - N-graph duration (N between 1-8): elapsed time between first & Nth data point
  - Total of 200 features

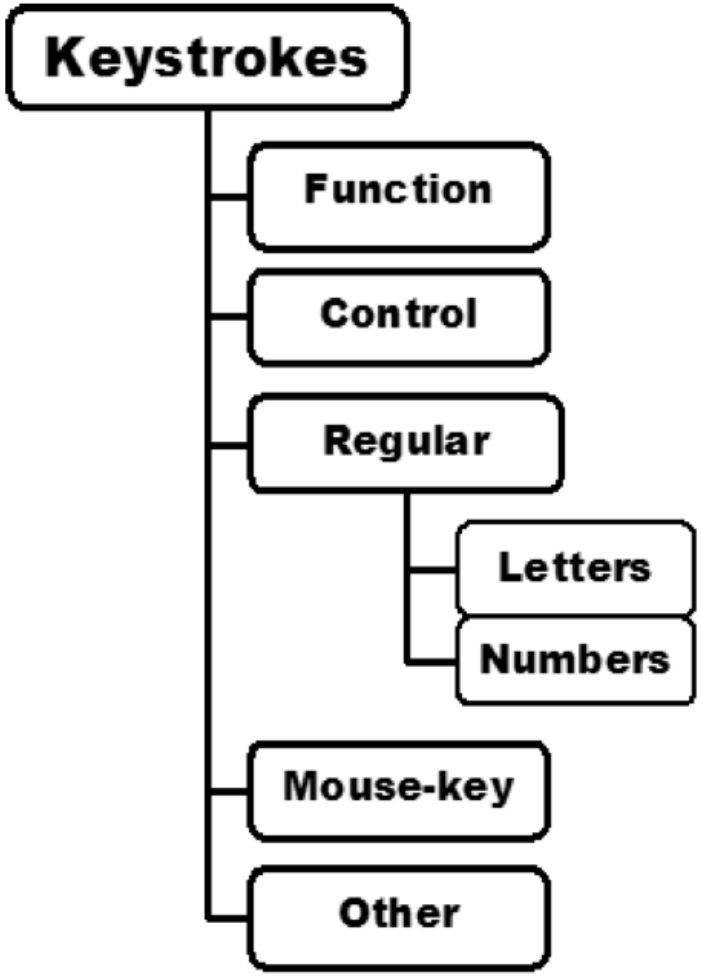
Mouse movements features

- NC moves
  - Similar to event features
  - Distance, speed, angle, etc. computed
    - between two subsequent NC moves
    - two NC moves separated by K data points
  - Total of 40 features
- Client area moves
  - Similar to NC mouse movement features
  - Total of 40 features

Keystroke data

- Two types in Windows OS
  - WM_KEYDOWN
  - WM_KEYUP
  - Event ID uniquely identifying the key

Keystrokes features hierarchy



Keystrokes features hierarchy

- Function keys
  - Example: F1-F12
- Control keys
  - Example: Control, Alt, Delete
- Regular keys
  - Alphabet letters, numbers
- Mouse keys
  - Example: Page Up/Down, Tab, arrow keys
- Other keys
  - Example: punctuation keys, Pause/Break, PrtSc/SysRq

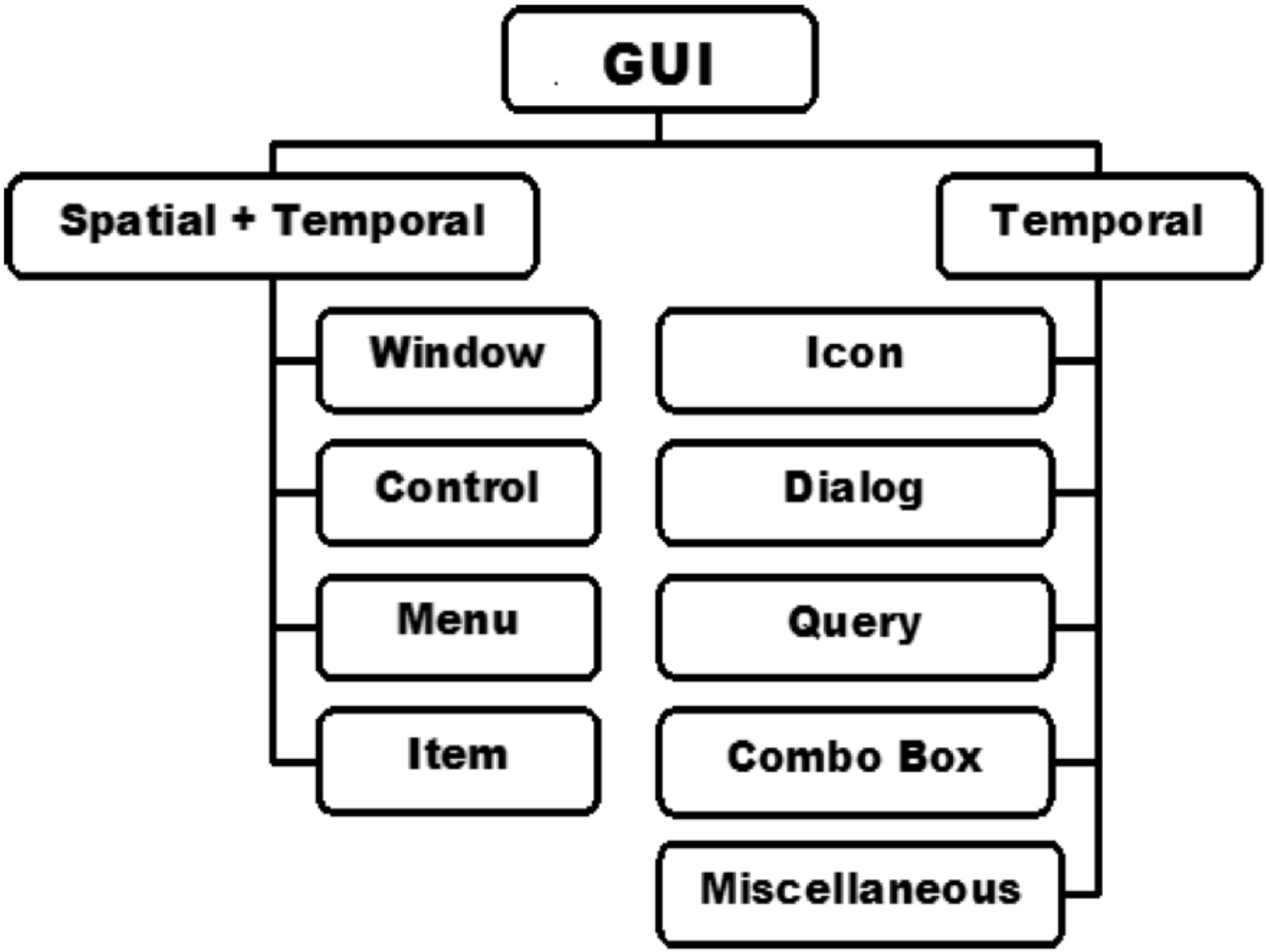
Keystroke features

- Over a window of W data points
  - # of events in each category & subcategory
  - # of occurrences of each letter & each numeral
    - 26 alphabet features, 10 numeric features
  - Mean, stdv, skewness
    - N-graph duration between consecutive keystrokes
      - N between 1-8
  - Total of 236 features

GUI data

- Data induced by 138 GUI events
- Grouped into a hierarchy based on function

GUI events hierarchy



GUI events hierarchy

- Temporal and spatial
  - Window
    - Scroll bar, minimize, maximize, restore, move, etc.
  - Control
    - Application & process control, open/close, etc.
  - Menu
    - Open, select, navigate, close, etc.
  - Item
    - List, button, etc.
- Temporal
  - Icon
  - Dialog
  - Query
  - Combo box
    - Open/close, select, move, resize, etc.
  - Miscellaneous
    - Power up/down, language change, background color change, etc.

GUI spatial features

- Over a window of W data points
  - # of events in each spatial category
  - Subsequent spatial events or spatial events separated by K data points
    - Mean, stdv, skewness
      - Distance
      - Speed
      - Angle of orientation
      - X coordinate
      - Y coordinate
      - N-graph duration (N between 1-8)
  - Total of 200 features

GUI temporal features

- Over a window of W data points
  - # of events in each temporal category
  - Subsequent spatial events or spatial events separated by K data points
    - Mean, stdv, skewness
      - N-graph duration (N between 1-8)
  - Total of 240 features

Summary of feature space

- Window size W: for each user
  - Tested: 100, 300, 500, 1000
  - Chosen: 500
- Frequency K: for each user & subcategory
  - Tested: 1, 5, 10, 15, 20
  - Chosen: 8 for frequent events, 1 otherwise
    - Frequent: induced > 10 times per second
- Candidate feature space: 956 dimensions
- Hierarchical groupings improve performance
  - Removing features from lower levels of mouse hierarchy lowered classifier performance

Classification

- C4.5 decision tree algorithm
  - Feature subset selection, and classification

Data collection

- 61 volunteers
  - Task 1:
    - Reading assignment
    - 20 questions about assignment
  - Task 2:
    - Set of web pages
    - Another set of questions
  - Average 4 hours of data collection
    - Average 92K data points, 4.5MB per user

Implementation schemes

- Two schemes
  1. Classify one feature vector instance at a time
     - If the instance matched normal profile, serve request
     - Else, alert sys admin, ask re-auth, or close session
     - May cause high false alarm rates
  2. Smoothing
     - Look at a window of n (n between 1-11) feature vector instances
     - If m (m between 1,n) of those matched profile, serve request
- Overlapping windows
  - W-s old points, s new points
  - s = 50, equivalent to 5 second intervals
  - Reduces time-to-alarm
- False bell rate
- Evaluation
  - 10-fold cross validation

Pair-wise discrimination

Scheme	FP Rate	FB Rate	FN Rate	Error Rate	Bell Count
Basic	5.12±5.05%	1.32±0.93%	6.76±2.03%	5.48±2.63%	2.30±1.95
Smoothing	4.52±4.58%	1.13±0.65%	5.45±2.05%	4.64±2.43%	1.86±1.18

Anomaly detection

Scheme	FP Rate	FB Rate	FN Rate	Error Rate	Bell Count
Basic	45.70±17.30%	6.28±1.70%	0.54±0.27%	1.21±0.60%	10.85±5.76
Smoothing	23.37±15.97%	1.76±0.94%	1.50±1.40%	1.77±1.01%	2.66±1.82

References

- An examination of user behavior for re-authentication (M. Pusara's PhD thesis, 2007)