



CS259: Data Mining for Cyber Security



Mouse dynamics vs physiological biometrics

- **Benefit: Transparent collection**
 - No hit on usability
- **Challenge: Intrinsic behavioral variability**
 - Intrinsic human factors
 - Biological or emotional status of the user
 - External environmental variables
 - Software environment, task, interaction mode



Mouse events

- System messages sent to receiving applications
 - Inform current cursor position & mouse button status
- Types
 - Mouse Down
 - Mouse Up
 - Mouse Wheel
 - Mouse Move



Mouse actions

- Single click
 - Mouse down followed by mouse up
- Double click
 - Mouse down, up, down, up
- Common movement
 - General mouse movement with no clicks
- Point and click movement
 - Mouse movement followed by single/double click
- Drag and drop movement
 - Mouse down, movement, mouse up
- Silence
 - No mouse operation

Mouse operation

- **Tuple:**

<action type, app type, screen-area, window-position, timestamp>

Attribute	Description	Encoding
Mouse action type	As in previous slide	0-9
Application type	Internet surfing, word processing, online chatting, gaming	0-3
Screen area	Area of screen, evenly divided to 9 regions	0-8
Window position	Position of window, including client area, close area, maximum area, minimum area, menu, toolbar, title bar	0-6
Timestamp	Time of action	--



Mouse behavior pattern

- Behavior pattern: recurring & fixed segments
 - Micro-habitual patterns
 - Subconscious/habitual factors urging GUI interactions
 - Task-intended patterns
 - Operating habits under certain applications (e.g., using certain function of an application)
 - Example: creating a new document in a word processing app



Hypothesis

- Measurements from behavior patterns more stable than measurements from holistic behavior
 - Better characterize discriminating user features

Mouse behavior pattern mining


- $W = \{w_1, w_2, \dots, w_n\}$ a set of all mouse ops
- Operation-set: a set of mouse operations
 - Example: $\{(1,3,4,0), (2,1,4,0), (3,1,4,0)\}$
- Sequence: ordered list of operation sets by user ID and timestamp
 - $s = \{s_1, s_2, \dots, s_k\}$, each s_j
 - An operation set (subset of W)
 - Called an element of sequence s
 - $s_j = \langle x_1, x_2, \dots, x_m \rangle$, each x_t a mouse operation
- Length of sequence: # mouse operation instances
 - L-sequence: A sequence of length L
- Example:
 - $s = \{\langle (1,3,4,0) \rangle, \langle (1,3,4,0), (2,1,4,0), (3,1,4,0) \rangle, \langle (2,1,2,1) \rangle\}$
 - $\text{length}(s) = 5$

Mouse operation sequence database

User ID	Sequence ID	Sequence
1	1	{<(1,3,4,0)>, <(1,3,4,0), (2,1,4,0), (3,1,4,0)>, ..., <(2,1,2,1)>}
1	2	{(2,1,4,0), (1,1,5,1), ..., (3,1,4,0)}
1	3	{(2,3,4,0), (1,1,3,0), ..., (1,2,2,5)}
...

Mouse patterns

- $\langle uID, sID, s \rangle$ contains sequence q if q a subsequence of s
 - Example: $\{\langle (1,3,4,0) \rangle\}$ is a subsequence of $\{\langle (1,3,4,0), (2,1,4,0), (3,1,4,0) \rangle\}$
- $\text{Support}_{DB}(q) = \# \text{ tuples in DB that contain } q$
- Sequential pattern: $\text{Support}_{DB}(q) \geq \text{min_supp}$
 - min_supp a given threshold
- L-pattern: sequential pattern of length L



Problem: Mouse behavior pattern mining

- Input
 - Mouse operation sequence DB
 - Threshold *min_supp*
- Output
 - Set of all frequent mouse behavior patterns in DB

Pattern sequences: Example

- Set of items = {a, b, c, d, e, f, g}, $min_supp = 2$
- $s = \langle a \text{ (abc) (ac) d (cf)} \rangle$
 - $Length(s) = 9$
 - $\langle a \text{ (bc) d f} \rangle$ a subsequence of s
 - Sequences 1, 3 contain $q = \langle (ab) c \rangle$
 - $Support(q) = 2$

User ID	Sequence ID	Sequence
1	1	$\langle a \text{ (abc) (ac) d (cf)} \rangle$
1	2	$\langle (ad) c \text{ (bc) (ae)} \rangle$
1	3	$\langle (ef) \text{ (ab) (df) c b} \rangle$
1	4	$\langle e g \text{ (af) c b c} \rangle$

Algorithm GSP: Generalized Sequential Patterns (1996)

4th scan, 6 candidates
4 length-4 sequential patterns

<a(bc)a> <(ab)dc> **<efbc>**

3rd scan, 64 candidates
21 length-3 sequential patterns
13 candidates not appear in database at all

<aab> <a(ab)> <aac>

2nd scan, 51 candidates
22 length-2 sequential patterns
9 candidates not appear in database at all

<aa> <ab> <af> <ba> **<bb>** **<ff>** <(ab)> **<(ef)>**

1st scan, 7 candidates
6 length-1 sequential patterns

<a> <c> <d> <e> <f> **<g>**

Candidate cannot pass support threshold

Candidate does not appear in database at all





Algorithm GSP: Drawbacks

- Too many candidates generated
- Candidates may not appear in database at all



PrefixSpan: Prefix-Projected Sequential Patterns Mining (2004)

- Project the database into a set of smaller databases
 - Based on set of patterns mined so far
- Mine locally frequent patterns in each projected database

PrefixSpan: Example

prefix	projected (suffix) database	sequential patterns
$\langle a \rangle$	$\langle (abc)(ac)d(cf) \rangle$, $\langle (-d)c(bc)(ae) \rangle$, $\langle (-b)(df)cb \rangle$, $\langle (-f)cbc \rangle$	$\langle a \rangle$, $\langle aa \rangle$, $\langle ab \rangle$, $\langle a(bc) \rangle$, $\langle a(bc)a \rangle$, $\langle aba \rangle$, $\langle abc \rangle$, $\langle (ab) \rangle$, $\langle (ab)c \rangle$, $\langle (ab)d \rangle$, $\langle (ab)f \rangle$, $\langle (ab)dc \rangle$, $\langle ac \rangle$, $\langle aca \rangle$, $\langle acb \rangle$, $\langle acc \rangle$, $\langle ad \rangle$, $\langle adc \rangle$, $\langle af \rangle$
$\langle b \rangle$	$\langle (-c)(ac)d(cf) \rangle$, $\langle (-c)(ae) \rangle$, $\langle (df)cb \rangle$, $\langle c \rangle$	$\langle b \rangle$, $\langle ba \rangle$, $\langle bc \rangle$, $\langle (bc) \rangle$, $\langle (bc)a \rangle$, $\langle bd \rangle$, $\langle bdc \rangle$, $\langle bf \rangle$
$\langle c \rangle$	$\langle (ac)d(cf) \rangle$, $\langle (bc)(ae) \rangle$, $\langle b \rangle$, $\langle bc \rangle$	$\langle c \rangle$, $\langle ca \rangle$, $\langle cb \rangle$, $\langle cc \rangle$
$\langle d \rangle$	$\langle (cf) \rangle$, $\langle c(bc)(ae) \rangle$, $\langle (-f)cb \rangle$	$\langle d \rangle$, $\langle db \rangle$, $\langle dc \rangle$, $\langle dcb \rangle$
$\langle e \rangle$	$\langle (-f)(ab)(df)cb \rangle$, $\langle (af)cbc \rangle$	$\langle e \rangle$, $\langle ea \rangle$, $\langle eab \rangle$, $\langle eac \rangle$, $\langle each \rangle$, $\langle eb \rangle$, $\langle ebc \rangle$, $\langle ec \rangle$, $\langle ecb \rangle$, $\langle ef \rangle$, $\langle efb \rangle$, $\langle efc \rangle$, $\langle efc \rangle$.
$\langle f \rangle$	$\langle (ab)(df)cb \rangle$, $\langle cbc \rangle$	$\langle f \rangle$, $\langle fb \rangle$, $\langle fbc \rangle$, $\langle fc \rangle$, $\langle fcb \rangle$



Reference-behavior pattern generation and matching

- Pattern generation
 - For each user
 - Mine behavior patterns from each session
 - Collect all patterns as reference behavior pattern
- Pattern matching
 - Given a new operation sequence match against mined patterns

Minimum support

Minimum support	Length-1 pattern	Lenth-2 Pattern	Other Patterns	All Patterns
2%	23.64%	32.15%	25.22%	81.01%
5%	16.08%	22.06%	24.90%	63.04%
8%	12.29%	17.65%	17.34%	47.19%
20%	0.95%	1.26%	0%	2.21%



Feature construction from patterns

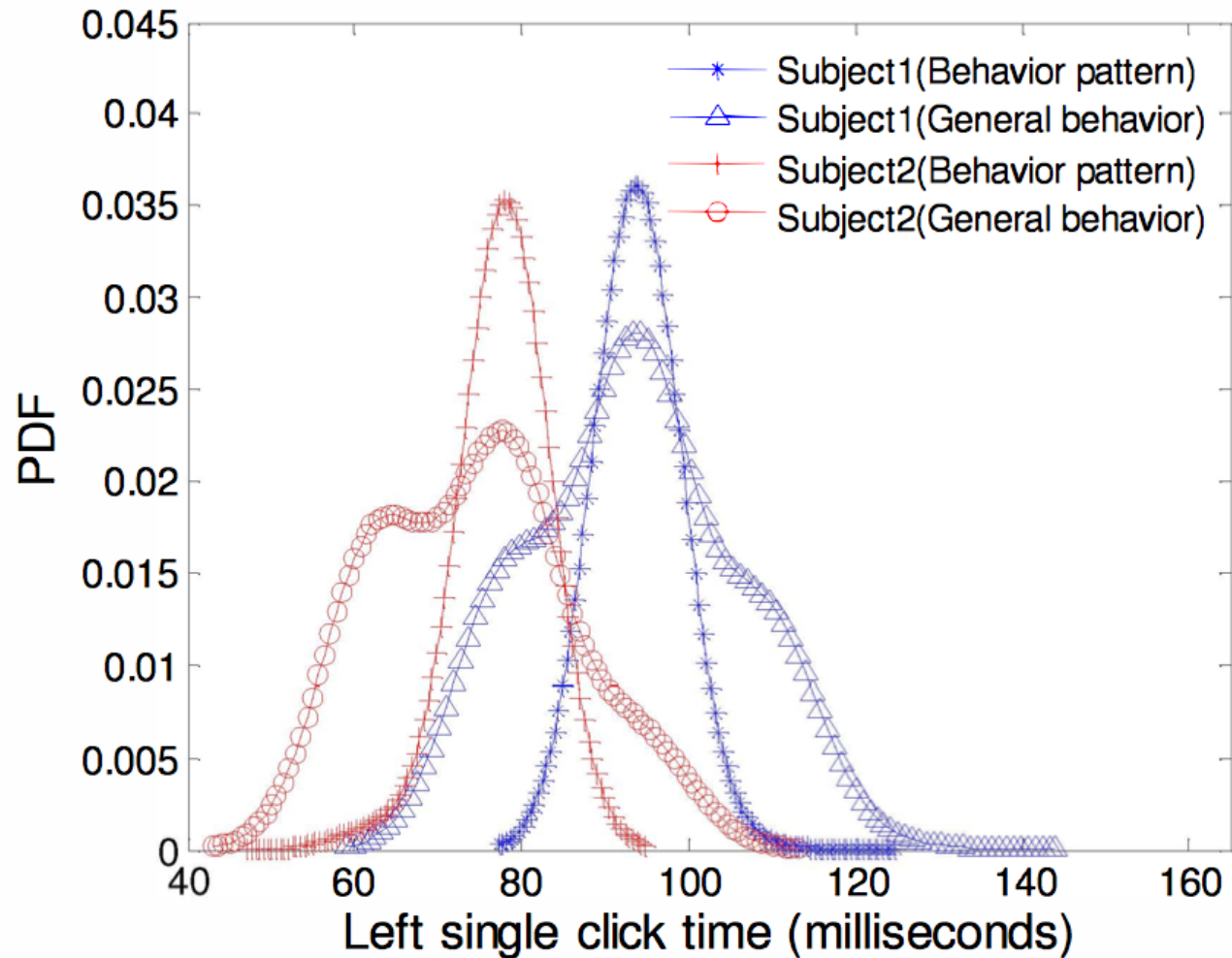
- Click elapsed time
 - Time spent by user to perform a click action
 - Single click: mean, stdv of overall time
 - Double click: mean, stdv of overall & 3 interval times
- Movement speed
 - Average movement speed for different types of mouse movement
 - 24 types: 8 directions, 3 distance ranges
- Movement acceleration
 - Average acceleration for different types of mouse movement
 - Similar to movement speed
- Relative position of extreme speed
 - Example: 0.5 for middle position of movement speed curve



Features

- 20 click-related features
- 24 movement-related features
 - Only from common & point-and-click movements
- 24 acceleration-related features
- 24 extreme-speed-related features
- Total: 92 features

Behavior pattern features: Stability and discrimination power



Detectors

- Nearest neighbor detector
 - Anomaly score: Mahalanobis distance between test & training feature vectors
- Neural network detector
 - Single hidden layer, 1 output node
 - Train with every input feature vector & output=1.0
 - Test-vector fed into network, output \sim 1.0 or -1.0
- One-class SVM



Data collection

- 28 participants
 - ~90,000 mouse actions/user
 - 30 sessions
 - Each 30 minute
 - Internet surfing, word processing, online chatting, programming, online gaming
 - Between 30-60 days per participant
- Data record
 - Event type (e.g., mouse move/click), position, timestamp, application information

Results

<i>Detector</i>	Behavior Pattern		Holistic Behavior	
	<i>FAR</i>	<i>FRR</i>	<i>FAR</i>	<i>FRR</i>
Nearest Neighbor	2.73% (0.0124)	3.67% (0.0089)	8.87% (0.0936)	9.63% (0.0923)
Neural Network	0.89% (0.0057)	2.15% (0.0065)	6.36% (0.0534)	6.95% (0.0653)
One-Class SVM	0.37% (0.0062)	1.12% (0.0067)	5.57% (0.0502)	6.73% (0.0493)

Results

Operation length	FAR	FRR	Authentication time
100	44.65%	34.78%	about 1 minute
500	7.78%	9.45%	about 5 minutes
1000	2.75%	3.39%	about 10 minutes
2000	1.22%	1.69%	about 20 minutes
3000	0.37%	1.12%	about 30 minutes



Administrativa

- HWI is out!



Reference

- Continuous Authentication for Mouse Dynamics: A Pattern-Growth Approach (C. Shen et al., 2012)