

User Profiling in Window Title and Process Table

Members : Chien-Chih Lin,
Eun Young Noh, Youngping Yan, and
Dr. Edward Wegman

Outline

- Acknowledgement
- Introduction
- Data
- Methodology
- Implementation
- Conclusion

Acknowledgement

- The authors would like to thank Dr. Ryszard Michalski, Dr. Kenneth Kaufman and Mr. Jarek Pietrzykowski of the Machine Learning and Inference Laboratory for the use of the data set and preprocessing applications in the system.

Introduction

- Now, in the information age, information security is a crucial issue for the business continuity.
- A well-designed information system should develop a security mechanism to protect its intellectual asset.
- Therefore, Computer Intrusion Detection System is needed to reduce the impacts of threats against the computer security and to prevent the misuse of computer system.

Introduction

- User profiling is a system that measures the user interactions with the computers and networks and extracts some useful information that represents the users behavior.
- User profiling is often used to detect the internal misuse by comparing the users behavior to their past patterns.

Introduction

- Other possible measurements are designed for user profiling, such as intervals between keystrokes, mistyping, typing speed, text or command statistics, mouse events, computer usage statistics, window registry access statistics, and email usage.
- The statistical methodology is adapted from the NIDES (Next Generation Detection Expert System) project.

User Profiling in Window Title and Process Table

Data

- Data source is from Window-based computing environment
- Each file in the data set is a session in which a stream of window title and related process table information appears from login to logout for a user.

User Profiling in Window Title and Process Table

Data

Raw Data Set : user1-host5-1_2_03-09_12_09.1s

- (4647) 687.709 explorer pid = 105 Program Manager
:snmp:77:explorer:106:explorer:105:
- (4667) 690.152 outlook pid = 154 Microsoft(<<3142>>) Outlook(<<3142>>) <<7469>>
:explorer:106:outlook:243:outlook:154:
- (4672) 690.162 outlook c 0.020 :outlook:154:
- (4681) 691.093 outlook c 0.030 :outlook:154:
- (4685) 691.394 outlook c 0.040 :outlook:154:
- (4691) 691.995 outlook c 0.050 :outlook:154:
- (4693) 692.596 outlook c 0.060 :outlook:154:
- (4695) 693.497 outlook pid = 154 Inbox - Microsoft Outlook
:explorer:106:outlook:243:outlook:154:
- (4698) 693.497 outlook c 0.070 :outlook:154:
- (4756) 711.563 outlook c 0.080 :outlook:154:
- (4770) 715.198 outlook c 0.100 :outlook:154:
- (4776) 717.602 outlook c 0.110 :outlook:154:
- (4785) 724.211 outlook c 0.120 :outlook:154:
- (4792) 725.413 outlook c 0.130 :outlook:154:
- (4819) 867.517 outlook pid = 154 Microsoft Outlook
:explorer:106:outlook:243:outlook:154:
- (4830) 930.908 outlook pid = 154 Inbox - Microsoft Outlook
:explorer:106:outlook:243:outlook:154:

User Profiling in Window Title and Process Table

Data

- File name:

<username-hostname-month_day_year-hour_minute_second.*s>

<user1-host5- 1_2_03 -09_12_09 .1s>

- Each session has three versions according to the level of detail for the process information.
- The original data set from the first version (1s) has 1,292 sessions for 26 users.
- In this project, six users (User1, 4, 7, 8, 19, 25) who have many sessions were analyzed because of numbers of sessions.
- 1,024 sessions for these selected six users cover 78% of the total sessions.

User Profiling in Window Title and Process Table

Data

- Data was preprocessed in two steps.
 1. Conducted using the processing program by Unix shell script and AWK from the Machine Learning and Inference Laboratory.

32 useful features were extracted.

2. Conducted using the second processing program by Unix shell script and SAS macro.

21 useful features were selected.

User Profiling in Window Title and Process Table

Data

Raw Data Set : user1-host5-1_2_03-09_12_09.1s

- First Processing Data

```
user1,host5,Thu,09,688,explorer,n,N/A,N/A,N/A,N/A,105,,N/A,N/A,2,1.09861,0,0,0,690,6.53814,14,1,2,1,0,0,1,0.693147,2,0
user1,host5,Thu,09,690,outlook,n,N/A,N/A,N/A,N/A,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,690,outlook,c,0,0,0,short,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,691,outlook,c,0,1,0.693147,short,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,691,outlook,c,0,0,0,short,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,692,outlook,c,0,1,0.693147,short,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,693,outlook,c,0,1,0.693147,short,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,693,outlook,o,N/A,N/A,N/A,N/A,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,693,outlook,c,0,1,0.693147,short,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,712,outlook,c,0,18,2.94444,short,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,715,outlook,c,0,4,1.60944,short,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,718,outlook,c,0,2,1.09861,short,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,724,outlook,c,0,7,2.07944,short,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,725,outlook,c,0,1,0.693147,short,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,868,outlook,o,N/A,N/A,N/A,N/A,154,,N/A,N/A,63,4.15888,0,175,5.17048,931,6.83733,16,1,2,1,0,0,2,1.09861,2,0
user1,host5,Thu,09,931,outlook,o,N/A,N/A,N/A,N/A,154,,0,0,1,0.693147,0,63,4.15888,932,6.83841,21,1,3,1,2,1.09861,2,1.09861,3,0
```

- Second Processing Data (Unix Shell Script)

```
user1,host5,Thu,09,688,explorer,n,N/A,N/A,N/A,N/A,105,,N/A,N/A,2,1.09861,0,0,0,690,6.53814,14,1,2,1,0,0,1,0.693147,2,0
user1,host5,Thu,09,690,outlook,n,N/A,N/A,N/A,N/A,154,,0,0,3,1.38629,0,2,1.09861,693,6.54247,16,0.4,5,0.4,5,1.79176,2,1.09861,2,3
user1,host5,Thu,09,693,outlook,o,N/A,N/A,N/A,N/A,154,,0,0,175,5.17048,0,0,0,868,6.76734,21,1,3,1,6,1.94591,2,1.09861,3,0
user1,host5,Thu,09,868,outlook,o,N/A,N/A,N/A,N/A,154,,N/A,N/A,63,4.15888,0,175,5.17048,931,6.83733,16,1,2,1,0,0,2,1.09861,2,0
user1,host5,Thu,09,931,outlook,o,N/A,N/A,N/A,N/A,154,,0,0,1,0.693147,0,63,4.15888,932,6.83841,21,1,3,1,2,1.09861,2,1.09861,3,0
```

User Profiling in Window Title and Process Table

Data

- Second Processing Data (SAS)

userid	f1	f2	f3	f4	f5	f6	f11	f15	f17	f18	f20	f22	f23	f24	f25	f26	f28	f30	date	time
user1	host5	Thu	09	688	explorer	n	105	2	0	0	690	14	1	2	1	0	1	2	03_01_02	09_12_09
user1	host5	Thu	09	690	outlook	n	154	3	0	2	693	16	0.4	5	0.4	5	2	2	03_01_02	09_12_09
user1	host5	Thu	09	693	outlook	o	154	175	0	0	868	21	1	3	1	6	2	3	03_01_02	09_12_09
user1	host5	Thu	09	868	outlook	o	154	63	0	175	931	16	1	2	1	0	2	2	03_01_02	09_12_09
user1	host5	Thu	09	931	outlook	o	154	1	0	63	932	21	1	3	1	2	2	3	03_01_02	09_12_09

User Profiling in Window Title and Process Table

Data

Summary of data structure and Preprocessing

	Users	Sessions	Records	Features	Version
Original data	26	1,292	-	-	Three Version
Select Users and Versions	6	1,014	-	-	First Version
Preprocessing1	6	1,014		32	First Version
Preprocessing2	6	1,014	129,313	21	First Version
	User1	287	31,370		
	User4	134	26,170		
	User7	193	23,289		
	User8	167	17,597		
	User19	134	10,497		
	User25	99	20,390		

Methodology

- Score value
- Individual measure S
- Individual statistic Q

Methodology

-- Score value

- The NIDES (Next Generation Detection Expert System) project has a statistical subsystem that profiles the past behavior of users on a computer system and learns the pattern of each user.
- The meaning of the recent past is that the current statistics are calculated by exponential weights in which the weight reduces as the record is further from the current record.
- The large score value means that the behavior of user is abnormal and the score value close to zero means that the behavior is normal.

Methodology

-- Score value

- A single statistic value that is called as score value is generated for each record.
- Score Value is calculated as a kind of sum of squares of n individual measures.
- If all the individual measures are independent and the correlation matrix is identity matrix, the score value can be simplified as sum of squares.

$$IS = S_1^2 + S_2^2 + \cdots + S_n^2$$

Methodology

-- Individual measure S

- Individual measure S that reflects the degree of the abnormality of a particular type of recent behavior is derived from a corresponding Q statistics. For example, if S is the measure of the degrees of abnormality of CPU time, then the corresponding Q is the actual CPU time used.
- The probability distribution of each component is constructed by the relative frequencies of the corresponding component during the profiled periods.
- The relative frequencies are computed in the pre-defined intervals, the width of which can be determined differently for the users, and the last interval does not have upper bounds.

Methodology

-- Individual measure S

- Let the P_i , ($i = 1$ to n) be the relative frequency for the i -th interval.
- Let $P(i)$ be the order statistics of P_i from small to large relative frequencies.
- Let $CUMP(i)$ be the cumulative relative frequencies of $P(i)$.
- S_i is defined by the following formula.

$$CUMP_i = \Pr \{ |N(0,1)| \geq s_i \} = 1 - \Pr \{ |N(0,1)| \leq s_i \}$$

$$\Rightarrow \Pr \{ |N(0,1)| \leq s_i \} = 1 - CUMP_i$$

$$\Rightarrow \Phi(s_i) - \Phi(-s_i) = 2\Phi(s_i) - 1 = 1 - CUMP_i$$

$$\Rightarrow \Phi(s_i) = 1 - \frac{CUMP_i}{2}$$

$$\Rightarrow s_i = \Phi^{-1} \left(1 - \frac{CUMP_i}{2} \right)$$

User Profiling in Window Title and Process Table

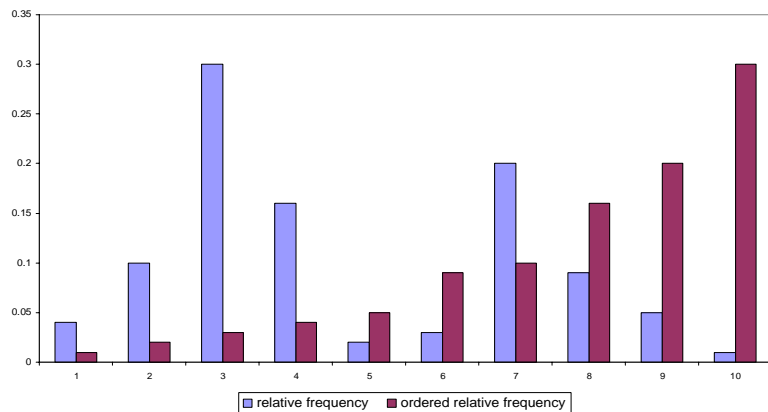
Methodology

-- Individual measure S

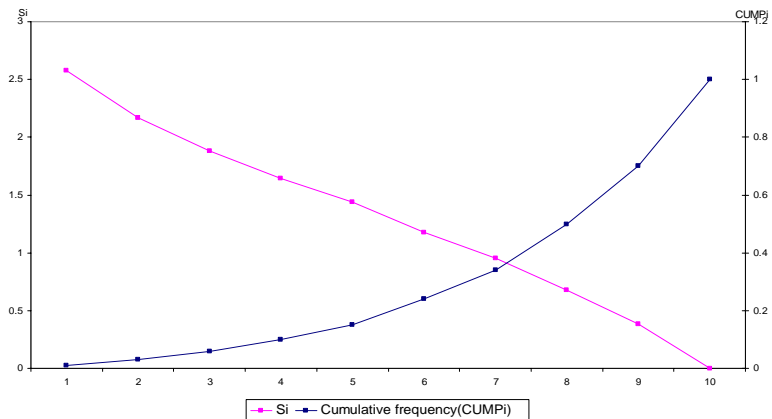
Example : Frequency tables for the data with 10 intervals

I	1	2	3	4	5	6	7	8	9	10
P_i	0.04	0.1	0.3	0.16	0.02	0.03	0.2	0.09	0.05	0.01
$P(i)$	0.01	0.02	0.03	0.04	0.05	0.09	0.1	0.16	0.2	0.3
(i)	10	5	6	1	9	8	2	4	7	3
$CUMP(i)$	0.01	0.03	0.06	0.1	0.15	0.24	0.34	0.5	0.7	1
S_i	2.567	2.170	1.881	1.645	1.440	1.175	0.954	0.675	0.386	0

Relative frequencies and ordered



Relationship between S_i and $CUMP(i)$



Methodology

-- Individual statistic Q

- Q statistics in NIDES is the sum of the record values in the recent past with exponential weights that give more weight to the record closer to the current value.
- In our project, the weighted sum of recent past record values was not applied because of some restrictions of the data and the computational burdens .
- Instead we used record value itself as individual statistic Q.

Implementation

- Candidate features for user profile
- Period of user profile
- Selecting features for user profile
- Test

Implementation

-- Candidate features for user profile

- From the 21 features, 18 features except userid, date, and time are selected as candidate features for the user profile.
- For the ordinary measure, the relative frequencies are computed in 10 equally spaced intervals.
- The width of interval was decided considering the distribution of the corresponding feature and the same interval was applied to all users.
- For the categorical measure, the relative frequencies are computed in each possible categories for each categorical feature.

User Profiling in Window Title and Process Table

Implementation

-- Candidate features for user profile

Candidate features for user profile

Name	Description	Measure
F1	Host machine ID	Category
F2	Day of week	Category
F3	Time of day(hour)	Category
F4	Number of seconds from the start of the session	Ordinary
F5	Window process name or Process name	Category
F6	Window status	Category
F11	Window process ID	Ordinary
F15	Total elapsed time in window	Ordinary
F17	Ratio of Cpu time accrued by process within window to Total elapsed time in window or 0	Ordinary
F18	Delta time between window titles whenever NEW window is opened	Ordinary
F20	Elapsed time since login whenever NEW window is opened	Ordinary
F22	Number of characters in protected words	Ordinary
F23	Number of characters in protected words / total number of characters	Ordinary
F24	Total number of words in window title	Ordinary
F25	Number of protected words / Total number of words in window title	Ordinary
F26	Number of process-level records in a single window unit	Ordinary
F28	Total number of windows opened	Ordinary
F30	Number of protected words in window title	Ordinary

Implementation

-- Period of user profile

- Data during eight months from November 2001 to June 2002 were used in making the past profile for the first period.
- July 2002 data were tested by using this profile.
- By doing so, users past profiles can be constructed from the data with same time periods.
- Furthermore, moving the time period for user profile can reflect the changing patterns of users behavior.

Implementation

- Selecting features for user profile
 - If the distributions of a feature are quite different among the users, corresponding feature can be thought to have good properties to identify the user.
 - While a feature that has very similar distributions among the users has not good properties to identify the user.

Implementation

-- Selecting features for user profile

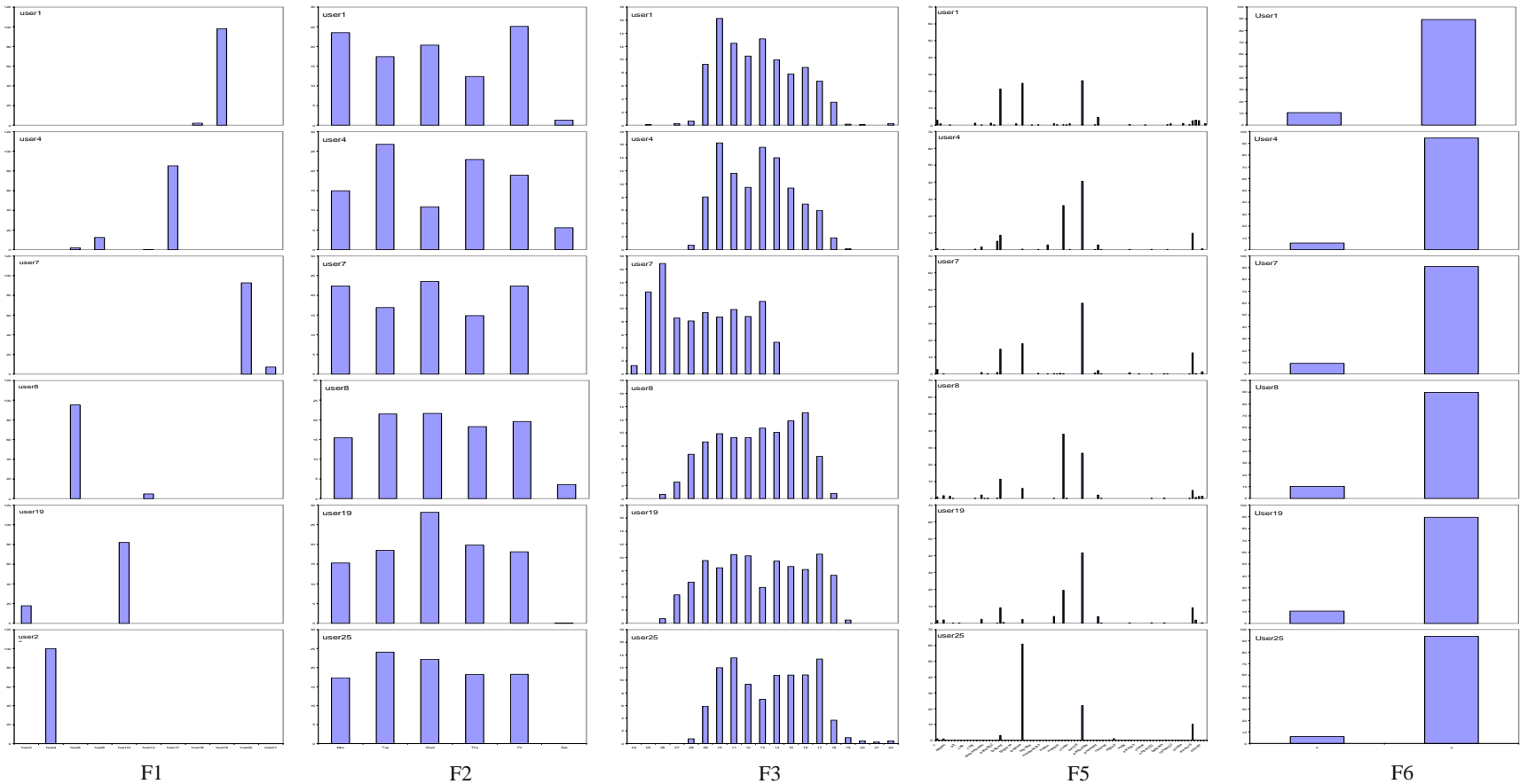
- Histograms of F1(host machine ID), F2(day of week), F3(time of day), and F5(window process name) have different patterns among the users. Therefore, these features can be selected as user profile and anomaly detecting features.
- However, F6(window status), which has only two categories was not selected because it has almost the same frequencies in two categories for all six users.
- F6 was excluded.

User Profiling in Window Title and Process Table

Implementation

-- Selecting features for user profile

Histograms for the features with categorical measure



Implementation

-- Selecting features for user profile

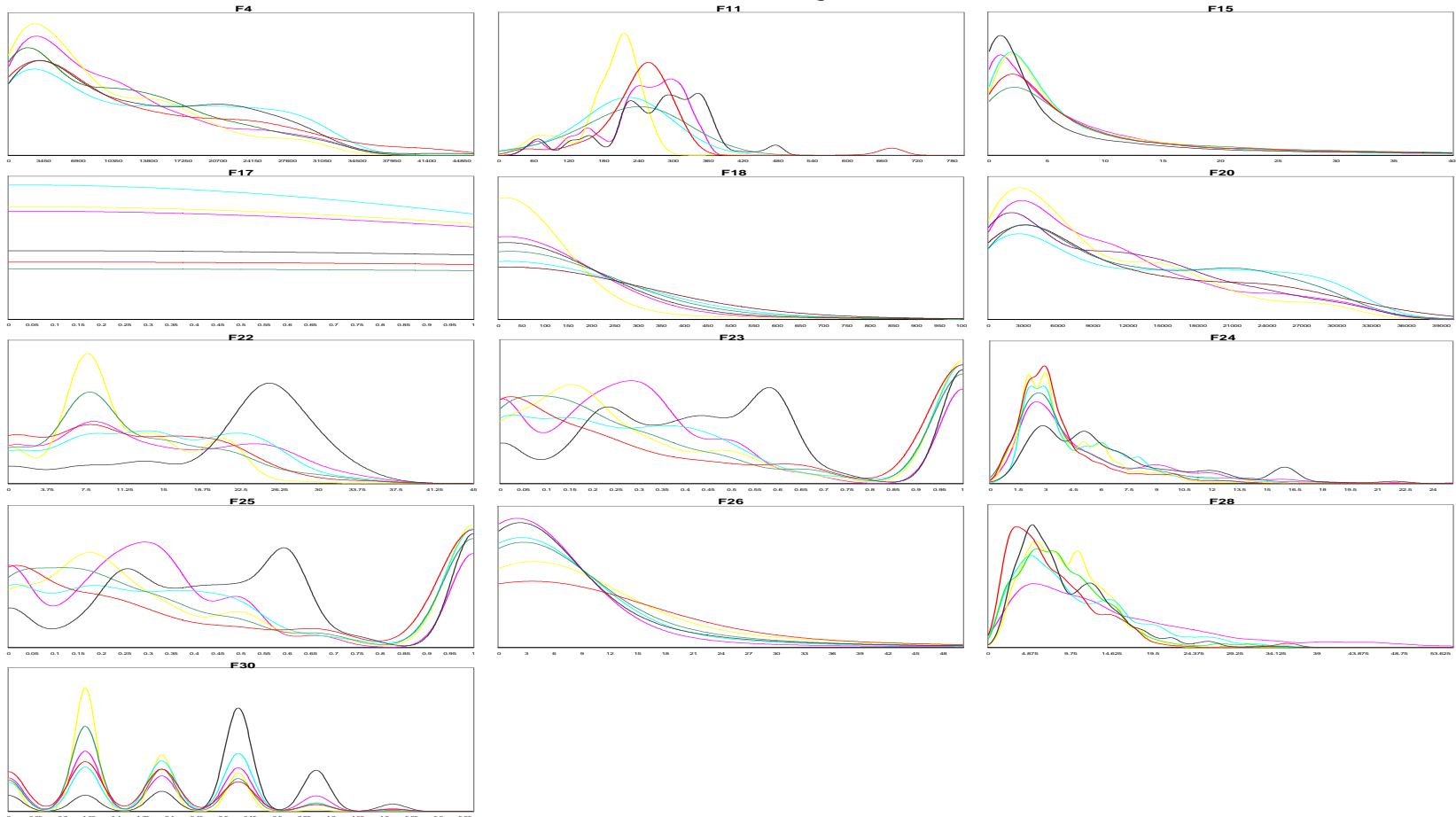
- The densities of F17 for each users are different but each densities are not informative because they are uniformly distributed.
- F4(number of seconds from the start of the session) and F20(elapsed time since login whenever new window is opened) have almost same densities for each user, which means these two features are highly correlated.
- F23(ratio of number of characters in protected words to total number of characters in window title) also has almost same densities with F25(ratio of number of protected words to total number of words in window title).
- So, F17, F20, and F23 were not included in the final feature sets.

User Profiling in Window Title and Process Table

Implementation

-- Selecting features for user profile

Densities for the features with ordinary measure



Implementation

-- Selecting features for user profile

- Finally, 14 features were selected.
- Categorical features : F1, F2, F3, F5
- Ordinary features : F4, F11, F15, F18, F22, F23, F24, F26, F28, F30
- Based on these features, users past profile and the score values were computed by the methodology to detect unauthorized users.

Implementation

-- Test

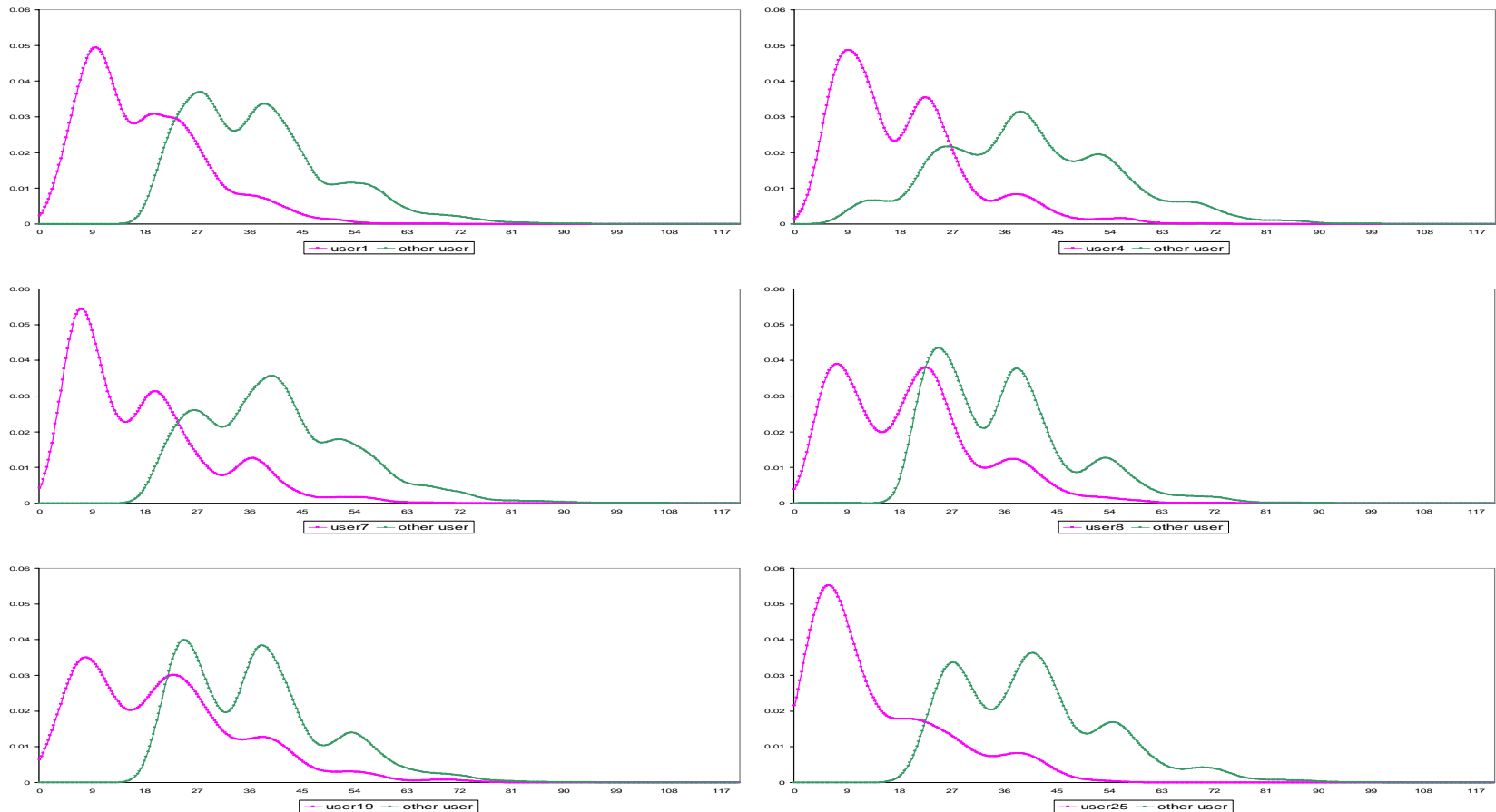
- Based on score value, performance is evaluated by how accurately the unauthorized users or the misuses of the authorized user are detected from authorized usages under the allowed false alarm rate.
- The false alarm is to detect authorized user as unauthorized user.
- The following figures represent the estimated probability densities of score values for each user and other users group.

User Profiling in Window Title and Process Table

Implementation

-- Test

Densities of score values for each 6 user and other users groups



User Profiling in Window Title and Process Table

Implementation

-- Test

False Alarm Rate		1%	5%	10%
User1	Threshold Value	47.4	38.2	31.9
	Detecting Rate	19%	42.6%	60.4%
User4	Threshold Value	54.8	40.2	34.3
	Detecting Rate	16.7%	46.3%	64.8%
User7	Threshold Value	54.5	39.7	35.9
	Detecting Rate	15.4%	48.9%	62.0%
User8	Threshold Value	53.3	41.3	37.3
	Detecting Rate	10.3%	26.3%	41.1%
User19	Threshold Value	58.3	45.0	39.8
	Detecting Rate	6.3%	22.1%	36.2%
User25	Threshold Value	43.8	38.8	31.5
	Detecting Rate	35.0%	53.1%	70.1%

Conclusion

- The detecting performances for User1, User4, User7, User25 are better than User8, and User19.
- The detecting rates under the false alarm rate of 1% are very low, and the detecting rates under 10% false alarm rate are about 60~70% for User1, User4, User7, User25.
- This test is based on the authorized users. Therefore if data from unauthorized users are available, those data will give larger detecting rates.
- Since exponential weighting method was not applied in our system, the score values have irregular component.

User Profiling in Window Title and Process Table

Question

- ????????
- ?????????????
- ?????????????????