

The Use of Frequent Episodes in Intrusion Detection

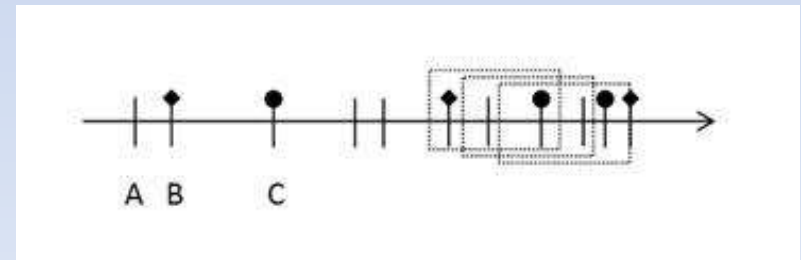
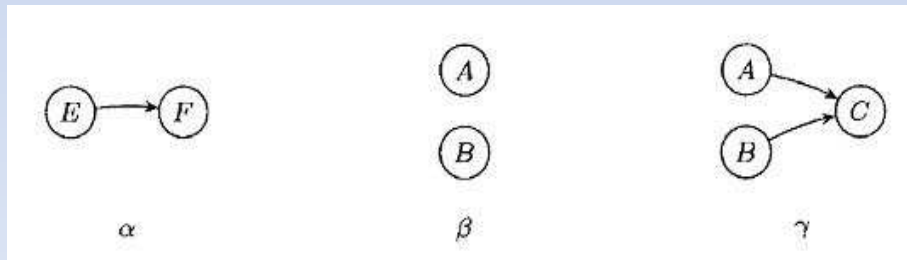
by Liubov Kokorina

Intrusion Detection Systems

- Intrusion detection systems (IDSs) collect and process system and network events in order to detect intrusions.
- Real-time and offline detection
- Misuse: attack signature
- Anomaly: normal profile and anomaly patterns
- Hybrid IDSs
- This work discusses the possibility of building a real-time hybrid IDS based on a data mining approach called frequent episodes.

Frequent Episodes (FE)

- Episode – a **data structure** by Mannila et al., 1995
- A partially ordered set of events that occur relatively close to each other in time.
 - Time window size and frequency threshold
 - Serial and parallel
- **Task:** finding all frequent episodes that belong to a given episode class, from a given sequence of events.
- *Winepi* algorithm: count windows containing the episode
 - Sliding window



Problem Description

- Large amounts of unformatted data
 - Features: behavior distinguishing
 - Patterns: intrusion patterns or normal profile
 - Discovery, matching, pruning
- Misuse-based IDSs are not able to detect unknown attacks.
- Anomaly-based IDSs produce a large amount of false positives and false negatives.
- Combination of the approaches is a big challenge because of the difference in the core ideas.

Justification, Motivation, and Benefits

- Audit data are a sequence of events ordered in time.
- Order relationships between events, in addition to various event characteristics, are supposed to contain evidence of intrusions.
- Episodes are an efficient way of representing partial order relationships between events.
- Frequent episode discovery algorithms are claimed to be fast and precise.
- We can balance between the speed and precision by correspondent thresholds.

Research Questions

- **Is it possible to build an efficient and accurate IDS, based only on FEs?**

Subquestions:

- What information can we get from episode frequencies?
- Are the episode-based attack patterns good enough?
- How can an IDS be constructed only on FEs?

Methods Overview

- Deepen compatibility assessment:
 - Literature study
 - Episode and intrusion analysis
 - Modeling and building an experimental IDS
- Experiment: Build and test the IDS' modules
 - Implement the algorithms (Winepi in C#)
 - Attract real intruders (HoneyNet)
 - Capture traffic data (Wireshark)
 - Pre-processing (Wireshark, C#)
 - Episode Discovery (Winepi implementation)

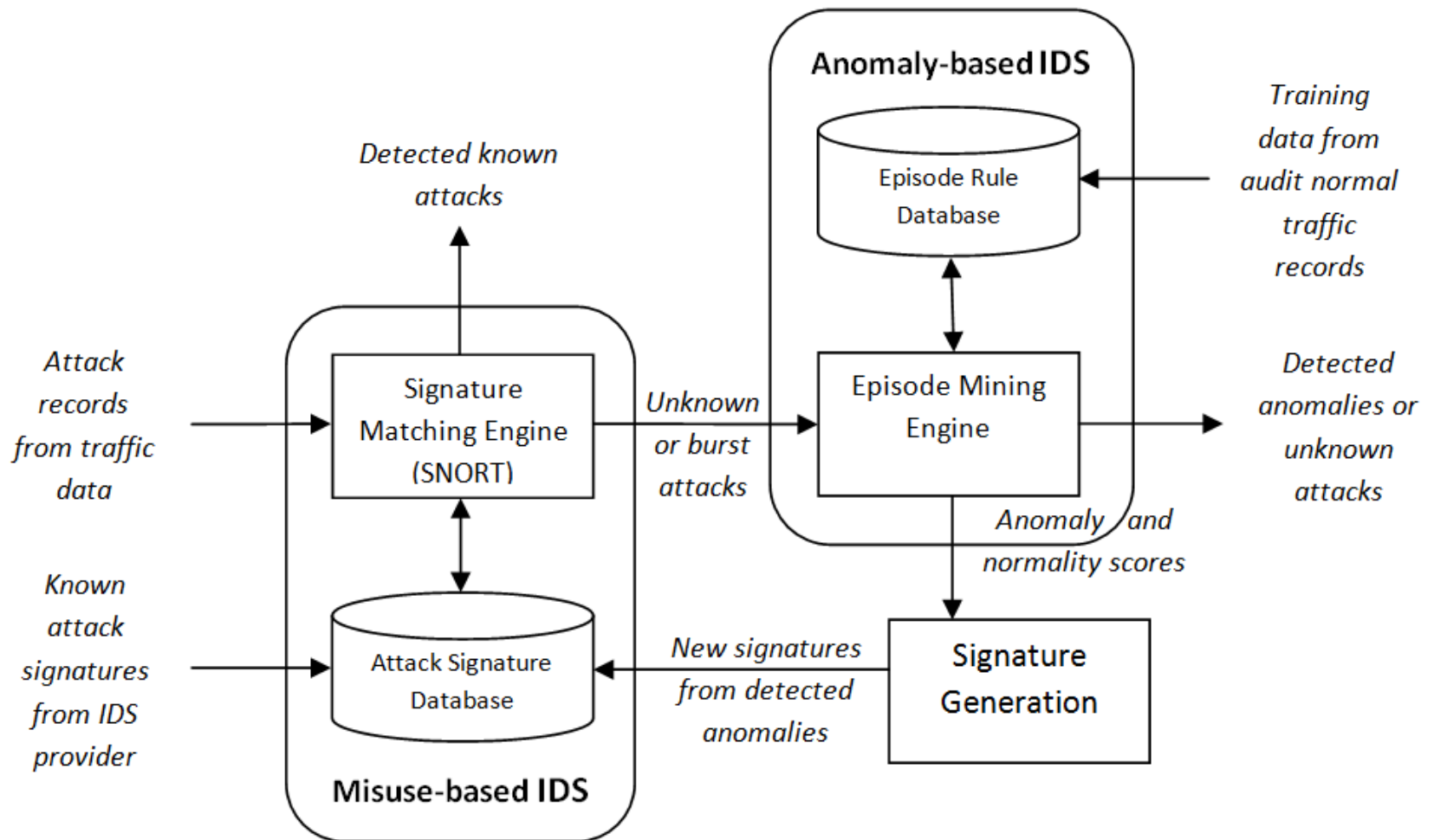
Previous Work

- Feature extraction
- Pattern mining
- Episode modifications:
 - *Domain knowledge*
 - *Time marks*
 - *Fuzzy frequent episodes*
- Alarm investigations
- Hybrid IDS (HIDS) on frequent episode rules

HIDS

- Qin and Hwang 2004, rebuilt in 2007
- Frequent episode rule (FER): Episode *alpha* follows episode *beta* in t seconds with probability p .
- Initial assumptions:
 - “Frequent episodes are mostly resulted from normal users”
 - “A rare episode is likely caused by intruders”
- A new weighted signature generation scheme to capture both attack signatures and normal behavior patterns.

HIDS



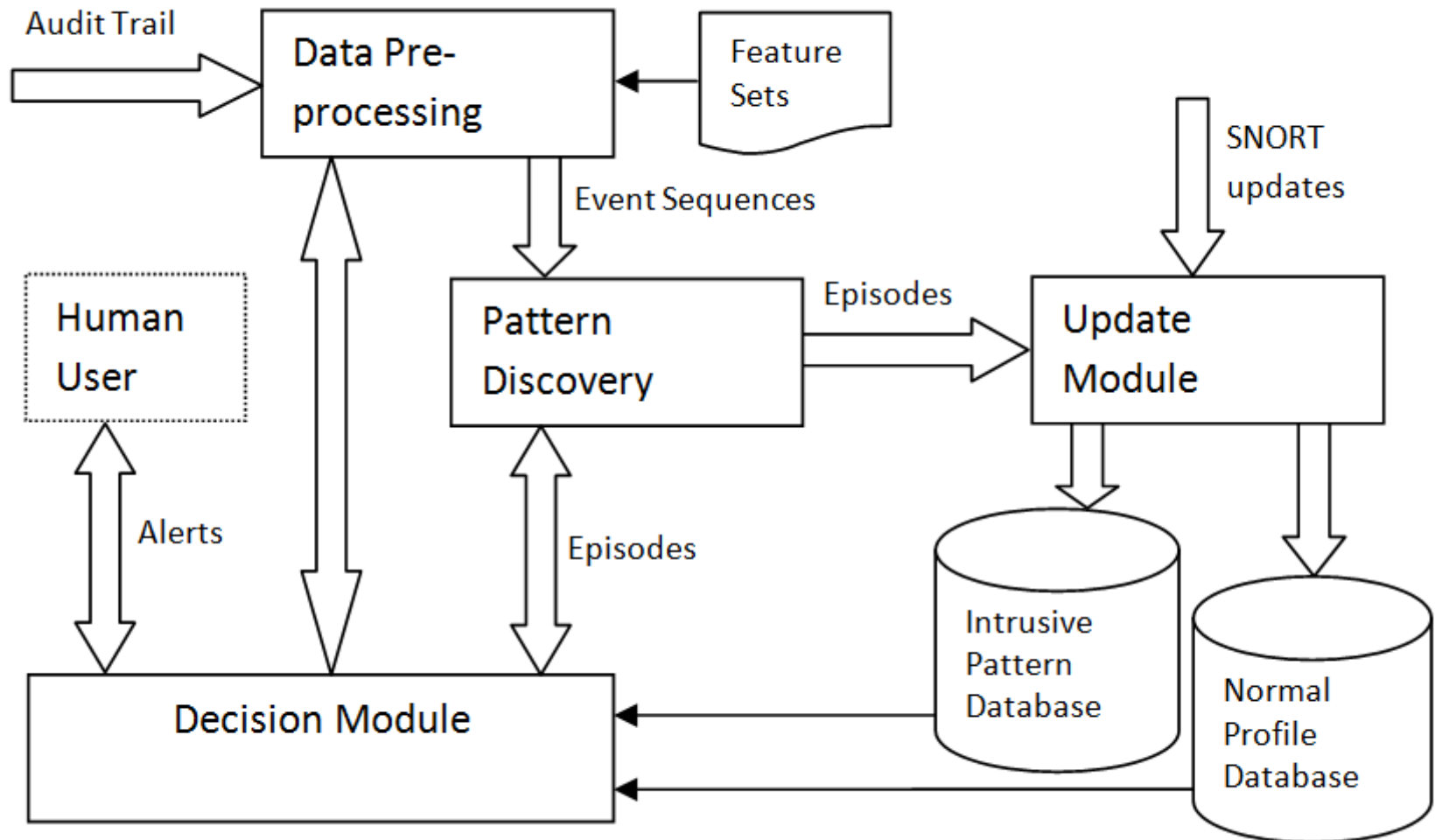
Literature Study Results

Observation	Conclusion
The lack of systematization of episode improvements and episode discovery algorithms.	The episode definition requires generalization to be flexible enough for several application areas, including intrusion detection.
There are few IDS on episodes, and their testing demonstrates outstanding results.	Episode discovery is useful for intrusion detection. But, because of the lack of research and weak discussion, the IDSs are only examples and test objects.
The problem of hidden assumptions, limited view, and anchored decisions.	The use of episodes in IDS needs attention from several researchers to assess the capabilities of episodes.

Our Hybrid IDS

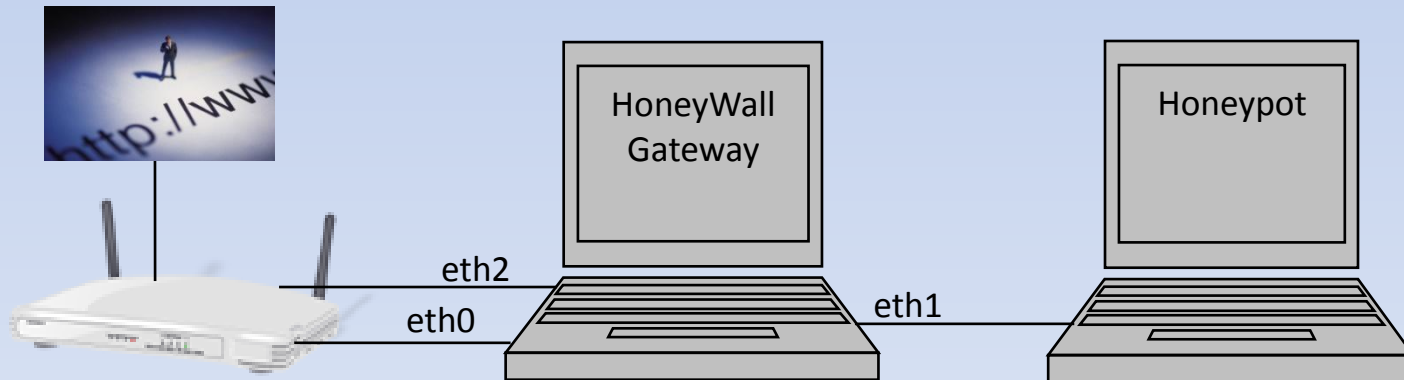
- Data pre-processing
 - continuous discovery in complex sequences
- Pattern discovery
 - attack signatures in form of fixed structures or specific events
 - a modification frequent episode analysis
 - anomalies and normal behavior patterns represented by similar structures
 - a new rare episode analysis technique
 - other useful patterns in form of regularity of activity
 - event distribution analysis on Winepi's properties
- Databases and updates
- Decision module

Our Hybrid IDS



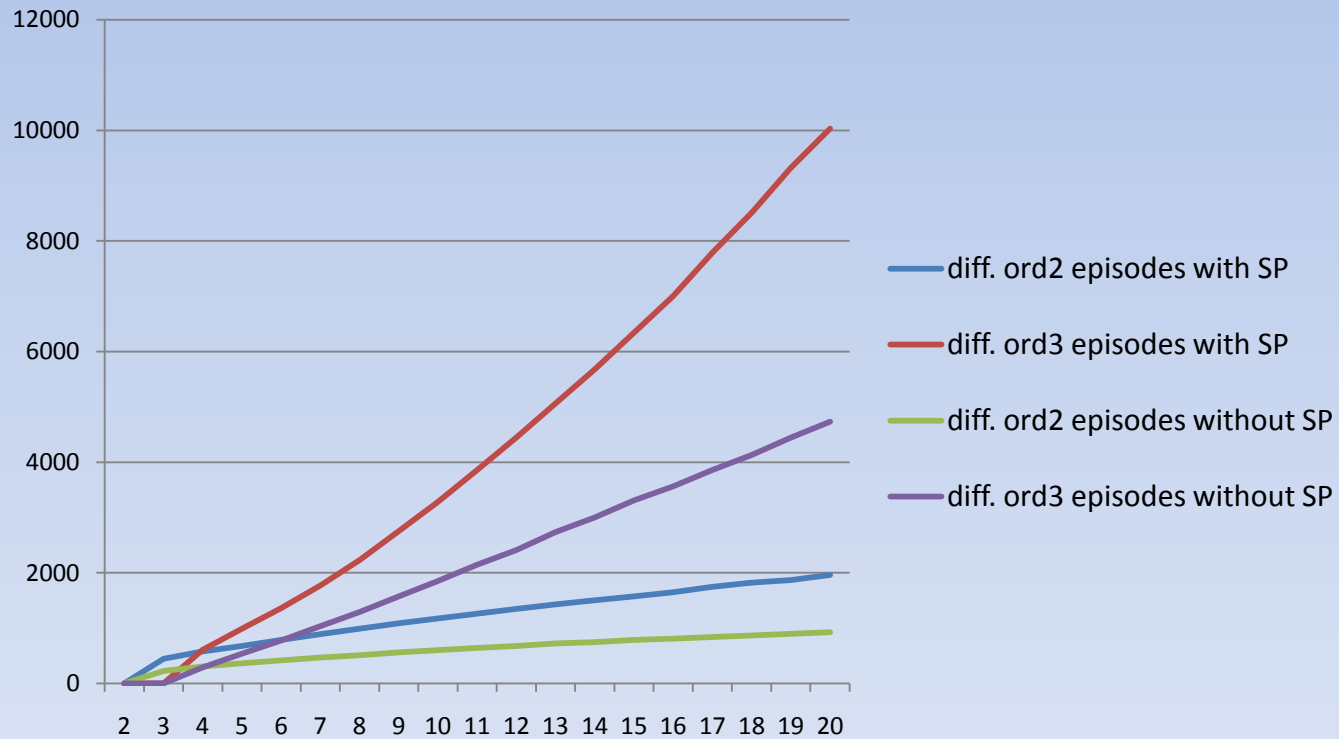
Experiment

- Pre-processing:
 - Ingress
 - Feature set: S-IP, (S-port,) D-port, service
- Threshold Selection



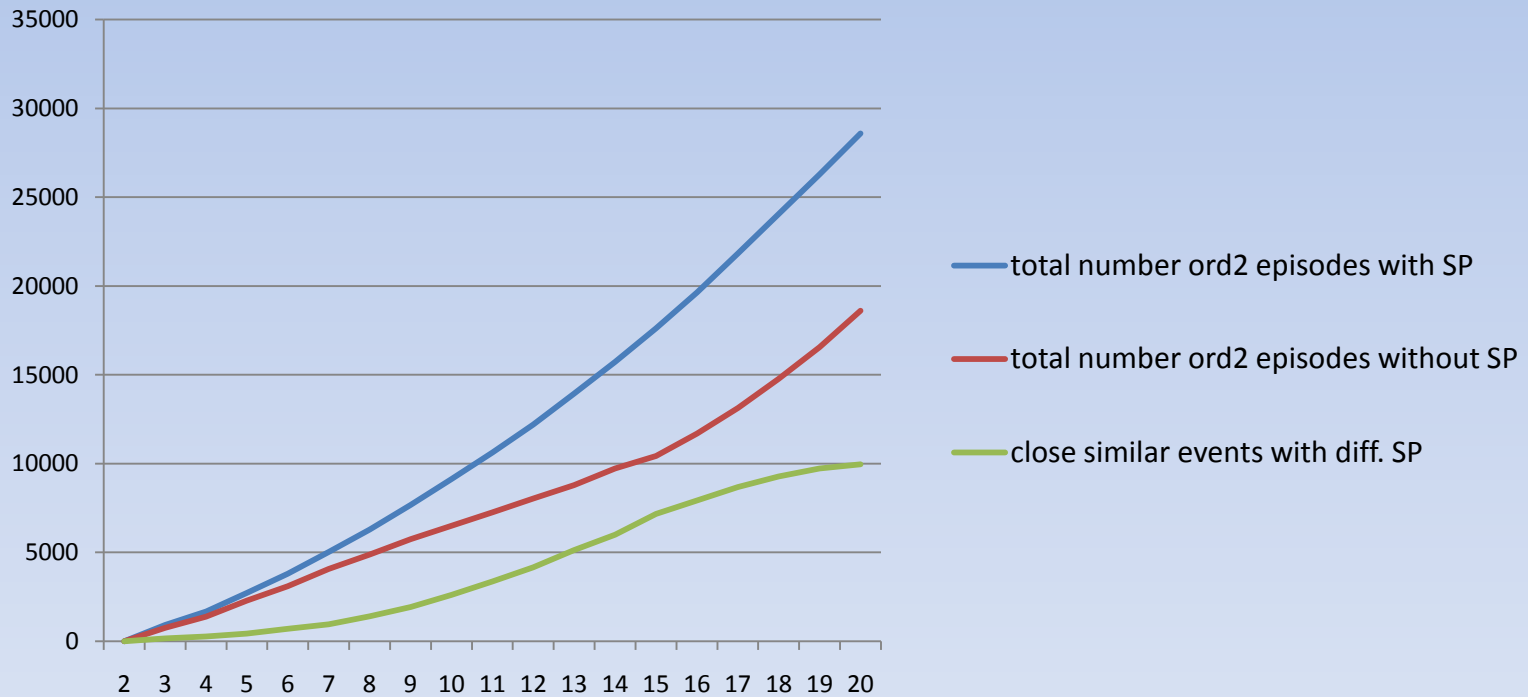
Experiment

Variety of episodes as a function of window size



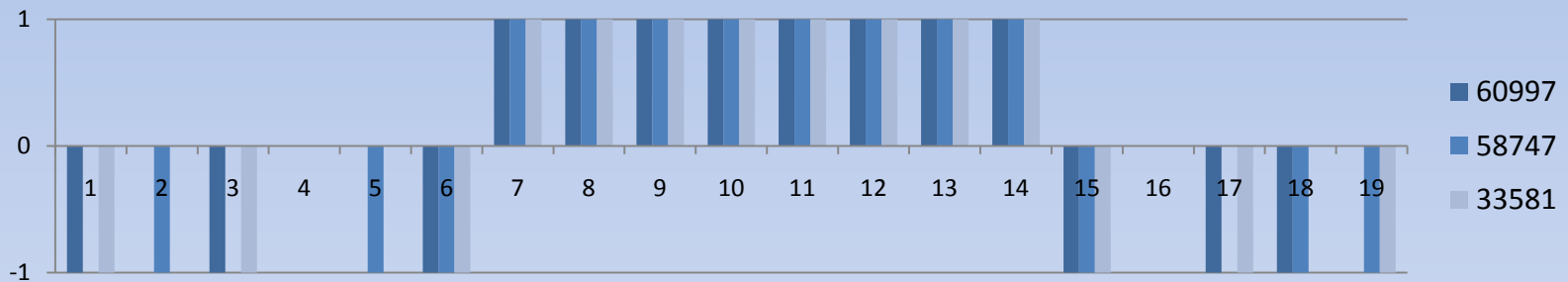
Experiment

Total number of episodes and window size



Traffic Analysis

- FTP Dictionary



FTP packets are 1s, TCP: -1, and other: 0

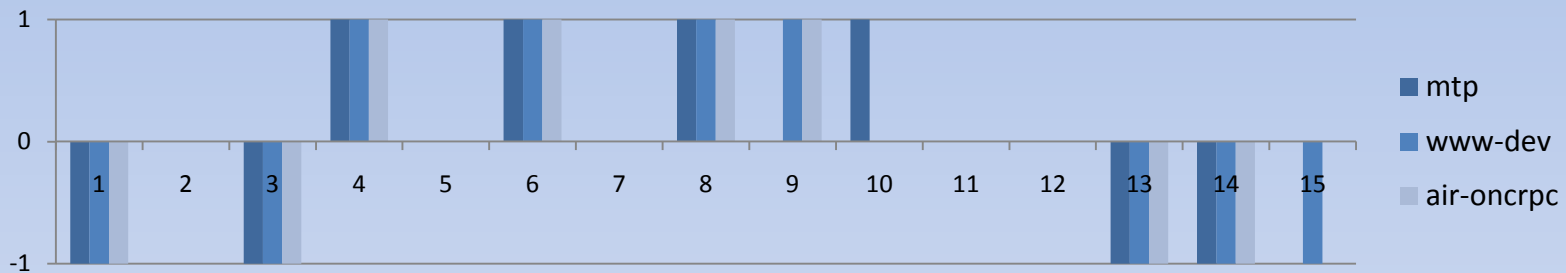
Attack sign is a short (either in packets or in time) FTP session.

Possible signature of episodes : FTP-FTP-FTP and TCP-FTP-TCP.

The episodes FTP-FTP-FTP from the same port to port ftp were the most frequent for any window size higher than 3.

Traffic Analysis

- MS SQL Exploit

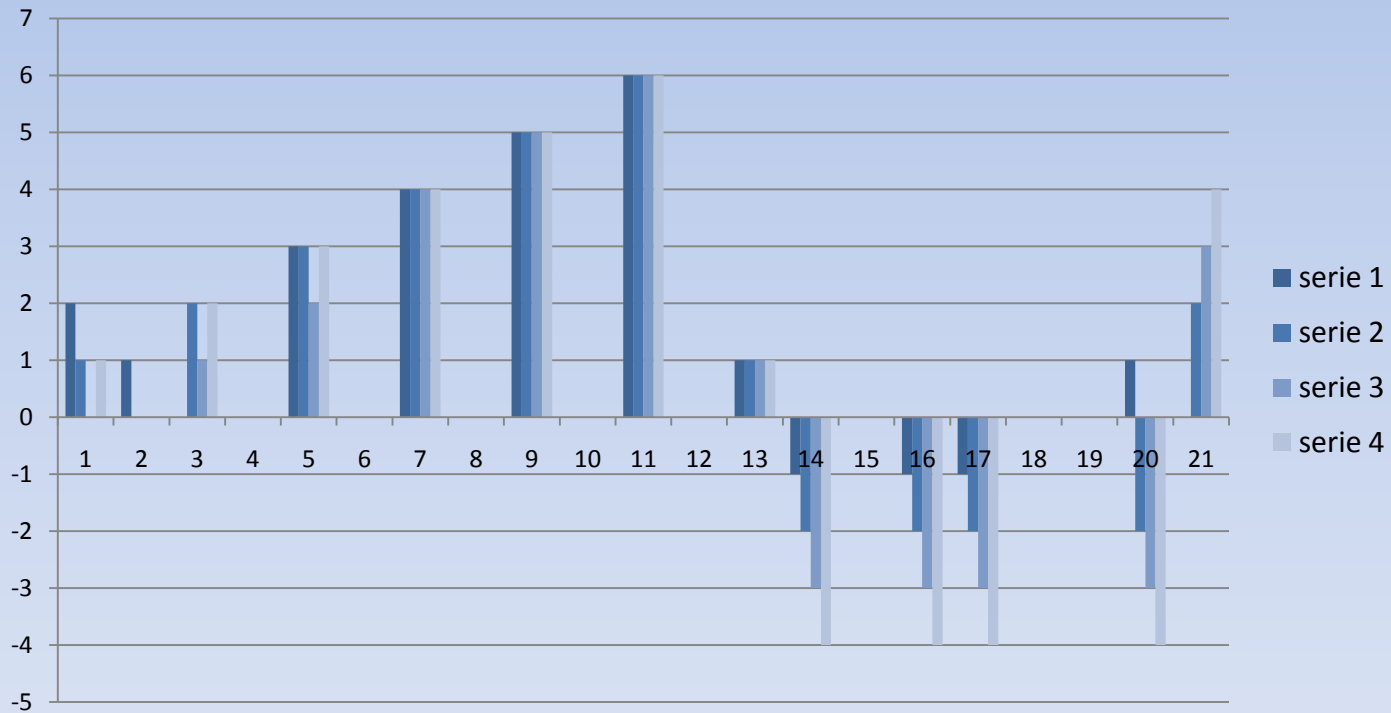


TDS packets are 1s, TCP: -1, and other: 0

Rare episode analysis on a reduced feature set (without source port):
15 packets between the first TDS in one attempt and the first TDS in the next attempt.

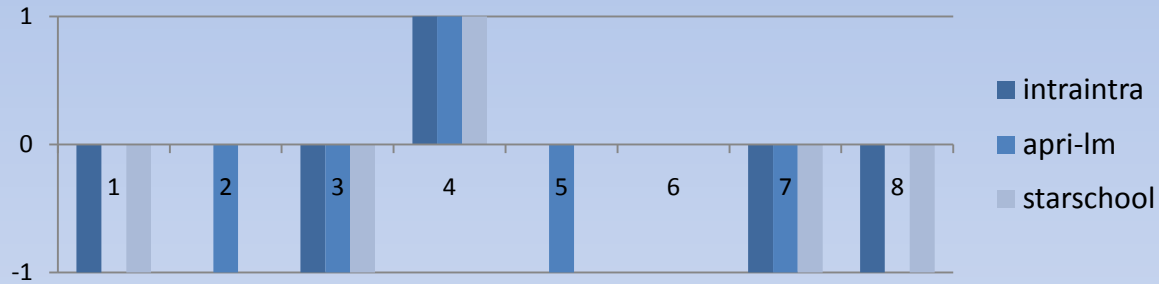
Traffic Analysis

- Portscan



Traffic Analysis

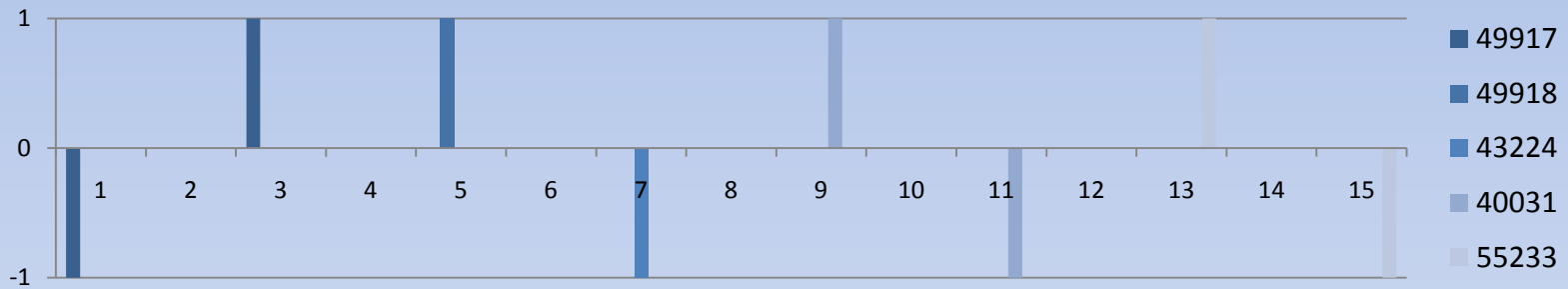
- PHP Probe



HTTP packets are 1s, TCP: -1, and other: 0

Traffic Analysis

- SPIM



Packets to port 1027 are 1s, to port cap: -1, and other: 0

Frequency Analysis

- Portscan

	win=	3	4	5	6	7	8	delta
TCP(s1, d1)-TCP(s1, d2)		5	10	15	20	25	30	5
TCP(x, d1)-TCP(x, d2)		5	10	15	20	25	30	5

The results show that:

1. There are more than $8-3=5$ packets between the port switches, if they were switched.
2. The attacker launched similar sequences of portscans 5 times – meaningless (automated attack).

Frequency Analysis

- PHP Probe

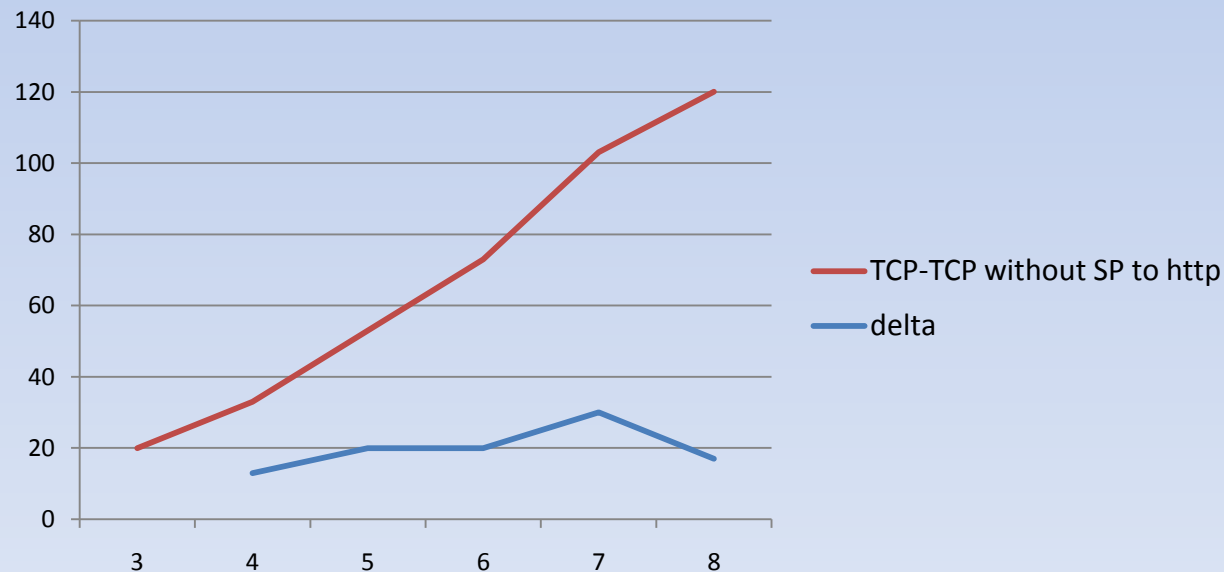
win=	3	4	5	6	7	8	delta
TCP(s1, http)-TCP(s1, http)	4	8	11	14	17	20	3
TCP(x, http)-TCP(x, http)	20	33	53	73	103	120	17
HTTP(s1, http)-HTTP(s2, http)	0	0	0	1	2	3	1
HTTP(x, http)-HTTP(x, http)	0	0	0	8	16	24	8

The results show that

1. Several PHP queries were sent from one port close together.
2. The attacker changed the source port up to 8 times.
3. The distance between the attack series is at least $6-3=3$ packets.

Frequency Analysis

- PHP Probe (cont.)
- The distance between the last TCP of one period and the first TCP of the next one is $7-3=4$ packets



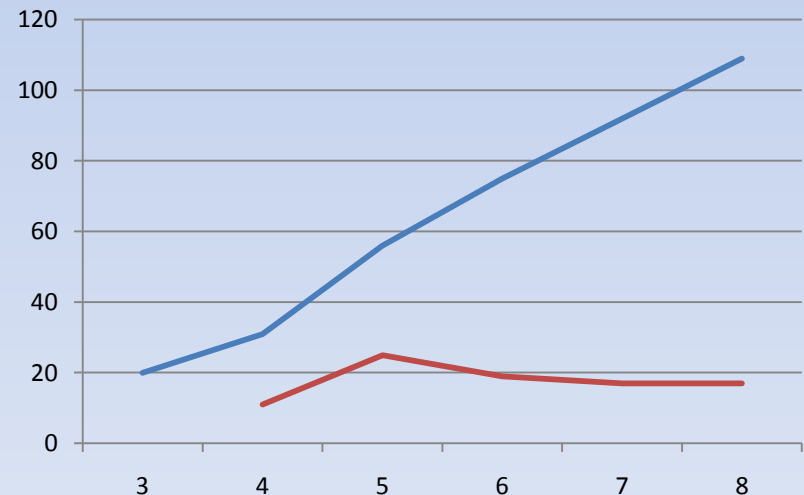
Frequency Analysis

- MS SQL Expolit

win=	3	4	5	6	7	8	delta
TCP(s1,ms-sql-s)-TDS(s2,ms-sql-s)	0	0	0	1	2	3	1
TCP(x, ms-sql-s)-TCP(x, ms-sql-s)	20	31	56	75	92	109	17

The results show that

1. The minimal distance between a TCP of one attack period and a TDS of the next period on another port is at least $6-3=3$ packets.
2. The attack periods are close together: only $5-3=2$ packets in between.



Frequency Analysis

- SPIM (Messenger)

win=	3	4	5	6	7	8	delta
Mes(s1, 1027)-Mes(s2, cap)	1	2	3	4	5	6	1
Mes(x, 1027)-Mes(x, cap)	1	2	3	4	5	6	1
Mes(s1, cap)-Mes(s2, cap)	0	0	1	2	3	4	1
Mes(x, cap)-Mes(x, cap)	0	0	1	2	3	4	1

The results show that

1. The attacker varies only source port, but runs only 2 attempts (i.e. one passage between them)
2. The distance between the attempts: $5-3=2$ packets.
3. Regular? We need at least 3 periods to answer.

Undetected Attack

- Distributed SPIM
- The traffic contains many SPIM series, and we expected high frequencies for the reduced feature set.
- The SPIM attacks were run from many different IP addresses: 202.97.238.xxx.
- It is a distributed attack, which is run by one computer program.
- The attack is relatively slow.

Event Distribution Analysis

- Automated packet generation is a sign of attack
- Detect the attack by regularity assessment
- Frequency for windows from an interval:

$$fr = k * (win) + b$$

Coefficient k depends on the intensity of the attack:

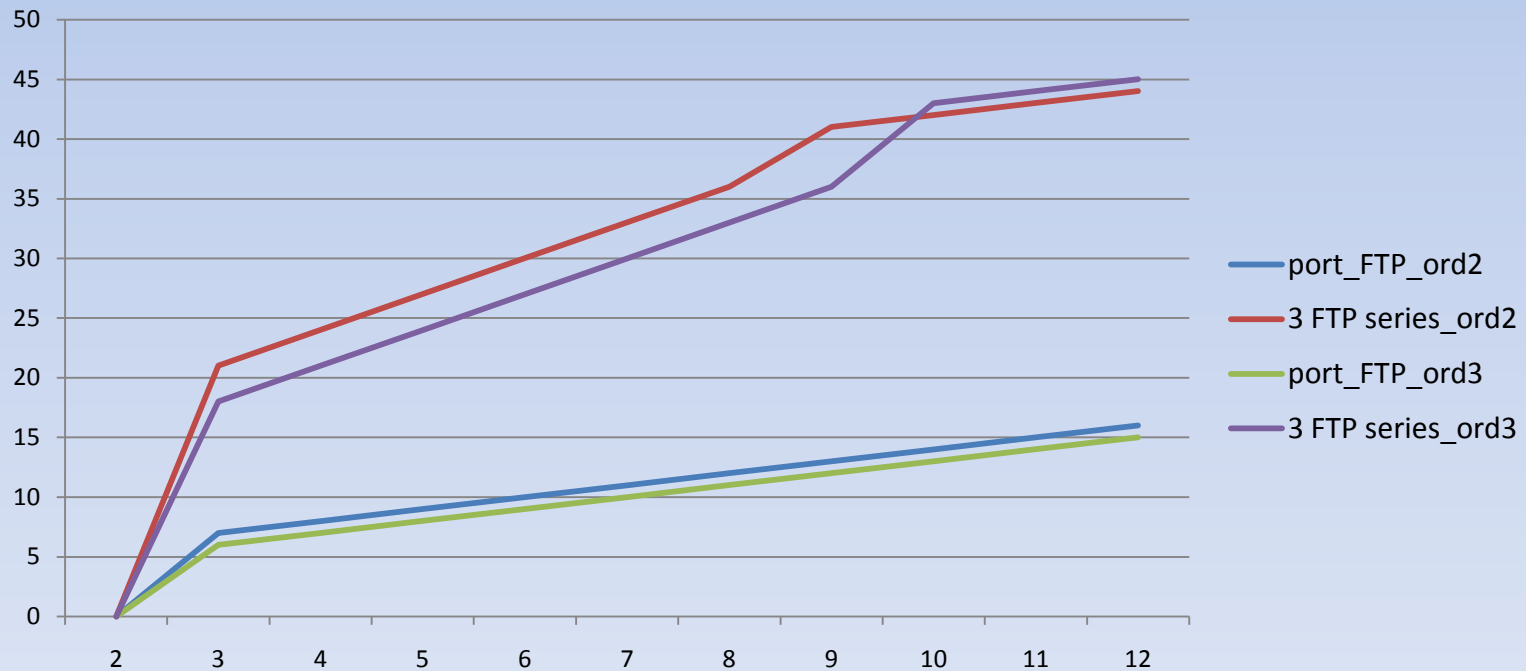
$$k = \text{delta}(fr) / \text{delta}(win)$$

Constant b depends on the attack signature

- **Event distribution analysis technique:**
- Fix the frequency, observe the windows size. For serial attacks:
 $k = k(win)$ and $k \rightarrow 1$ if $win \rightarrow \infty$
- In other words, $k=1$ if the window is so large that all attack series can be covered by one window.

Event Distribution Analysis

- Difference between frequencies with attention to the source port (fixed port results) and without it (any port results): weak order influence.



Event Distribution Analysis

- There are 6 packets between the last FTP of an attack period and the first FTP of the next period. Hence, any window smaller than 9 packets does not cover the two FTP packets from different periods.
- **Episodes (FTP, FTP) from any port:**
- For $\text{win} < 9$, $\text{delta}(\text{fr}) = N$ for $\text{delta}(\text{win}) = 1$, where N is the number of the attack periods
- $\text{fr}(\text{win}=9) - \text{fr}(\text{win}=8) = 2 * N - 1$
- For $\text{win} > 9$, $\text{delta}(\text{fr}) = \text{delta}(\text{win})$



Experiment Summary

- Frequent episode analysis:
 - Pattern discovery
 - Finds similarities in event sequences
- Rare episode analysis:
 - Defines distance between attack series
 - Detects similarities of attack series
- Event distribution analysis:
 - Effective in regularity discovery
 - Helps to detect automated attacks

Conclusions

Research Question	Answer
Are the episode-based attack patterns good enough?	The episode structure is useful in detection of various intrusions, which can be represented by either frequent or rare patterns. The similarity and regularity of most of the episodes related to an attack, allow us to discover and exploit hidden possibilities of episode counting based on sliding window method (Winepi).
What information can we get from FE counting results?	The three pattern discovery techniques that we proposed in the thesis, can be combined to analyze the attacks in depth.
How can an IDS be constructed only on FEs?	The today's episode structure is not flexible enough to be the only structure for IDS building. The episodes' strengths and weaknesses require investigations. Fortifying with other data structures may be necessary.
Is it possible to build an efficient and accurate IDS, based only on FEs?	The effectiveness of episodes in pattern discovery we demonstrated by the experiment. The episode analysis of the results helped to develop and test the new pattern discovery techniques. But there are many unsolved problems and questions, considering the episode applicability to intrusion detection.

Any questions?

