

➔ Building Multiple Behavioral Models for Network Intrusion Identification

Wei WANG, Sylvain GOMBAULT et Amine Bsila

Département Réseaux, Sécurité et Multimédia

**École Nationale Supérieure des
Télécommunications de Bretagne, France**



➔ Outline

- **Background: Intrusion Detection System (IDS)**
- **Motivation**
- ***k*NN based Intrusion detection and identification**
- **PCA based intrusion detection and identification**
- **Two methods comparision**
- **Concluding remarks**

Building Multiple Behavioral Models for Network Intrusion Identification

By W. Wang, S. Gombault et A. Bsila

Monam'07, Toulouse, France, November 5-6, 2007

- **Computer Security**

- **Intrusion Detection**

Firewall
Authentication

- **Anomaly intrusion detection model normal behavior**
- **Identify the patterns that deviate from the normal profile.**
- **Can detect new attack**
- **Attract many research group**
- **Pattern recognition problem**

Building Multiple Behavioral Models for Network Intrusion Identification

➔ Motivation

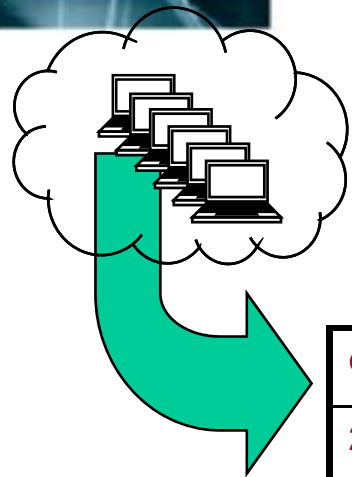
- **Identifying individual intrusions**
 - Most current IDS methods only detect intrusions but cannot identify which type of attack they belong to
 - Intrusion identification is essential for reaction after detection
- **Fast modeling subject behavior and detecting / identifying intrusions**
 - In real environments, a computer system could produce large amounts of audit data in a short time. It is typical for these types of data to be high dimensional.
 - Processing massive data is essential

Building Multiple Behavioral Models for Network Intrusion Identification

By W. Wang, S. Gombault et A. Bsila

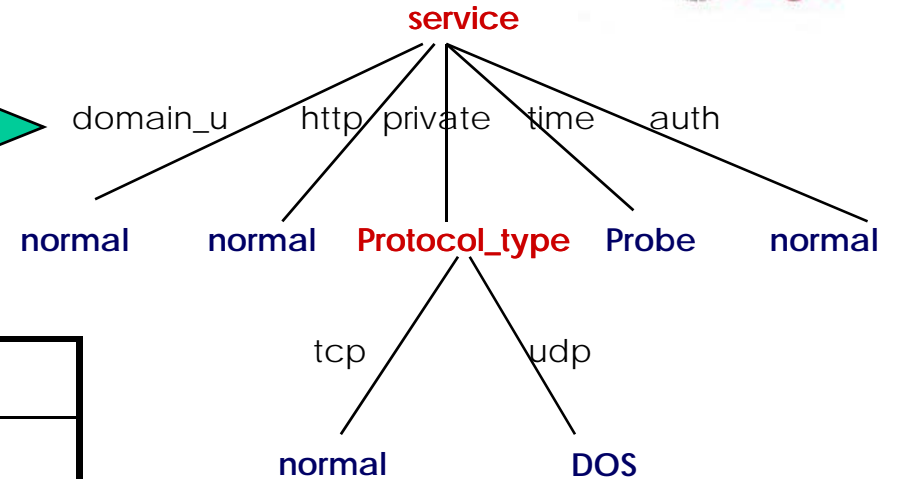
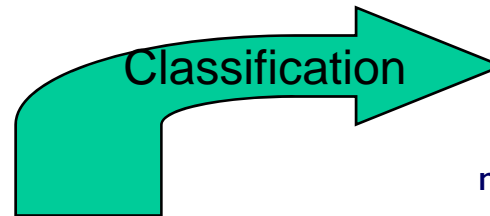
Monam'07, Toulouse, France, November 5-6, 2007

General intrusion detection steps



Transformation
from raw trafic

durée	service	Protocole	Classe
230s	http	tcp	normal
0s	private	udp	DOS



- **Function of the transformation**
 - Choose the important attributes
- **Two steps for intrusion detection / identification**
 - Feature transformation
 - Key steps
 - Extract important and key information from the raw data
 - Classification
 - Classification methods correspond to the features used

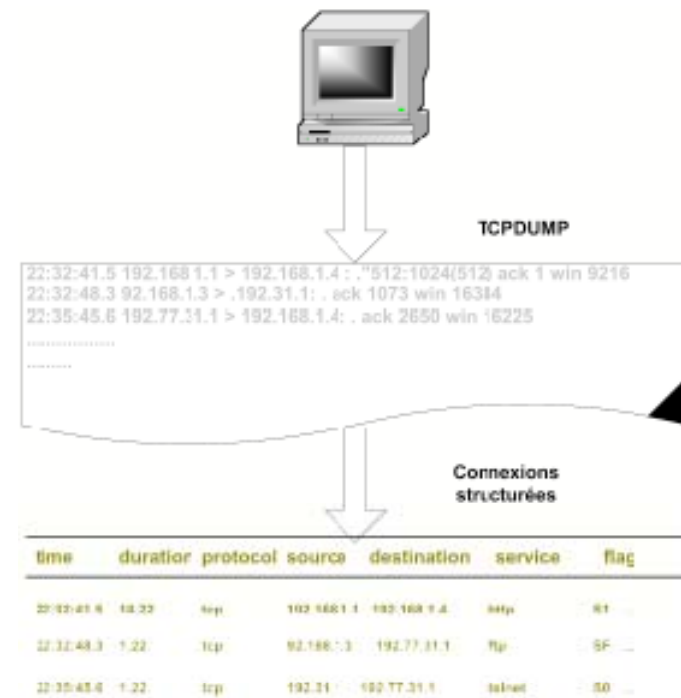
Building Multiple Behavioral Models for Network Intrusion Identification

➔ Function of transformation

- **Data sources:**
 - Raw traffic data
- **The classification function for raw network traffic:**
 - Transformation function T
 - R : ensemble du trafic brut
 - I : ensemble d'items structurés

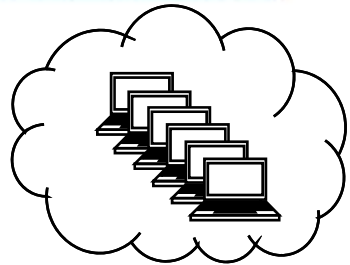
$$T : R \longrightarrow I$$

$$r \longmapsto i(a_1, a_2, \dots, a_n)$$

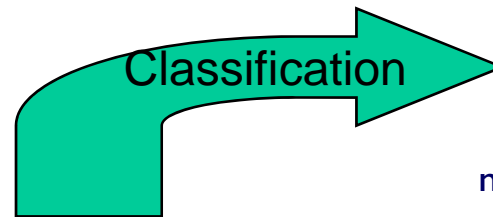


Building Multiple Behavioral Models for Network Intrusion Identification

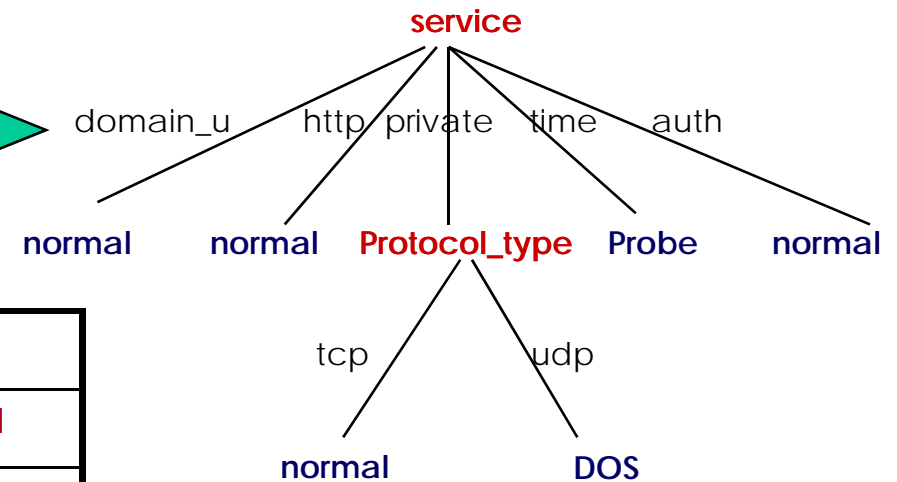
➔ KDD99 data



Transformation of raw traffic by BRO



durée	service	Protocole	Classe
230s	http	tcp	normal
0s	private	udp	DOS



- **DARPA (raw traffic data)**
-> using BRO to construct attributes (designed by W. Lee)
- **Transformation of tcpdump traffic in 41 attributes**

Building Multiple Behavioral Models for Network Intrusion Identification

➔ KDD99 (Continue)

- **Transformation of TCPdump traffic in 41 attributes**
 - Use BRO (policy designed by W. Lee)
- **3 categories of attributes:**
 - Basic features
 - E.g., service, type de protocole (TCP, UDP ou ICMP), ...
 - Content features
 - E.g. ,number of file creation operations, ...
 - Traffic features
 - E.g., number of connections to different hosts

Building Multiple Behavioral Models for Network Intrusion Identification

➔ Learning and test data sets



- **Data description:**
 - 41 attributes + name of the class
 - Text format
- **Data for intrusion *detection* (learning base of kdd99)**
 - Learning data: randomly selected 7000 connections
 - Test data: 4 classes d'attaques + trafic normal
 - Normal data: randomly selected 10,000 normal connections
 - Attack data: all the other attack connections
 - 391,458 DoS attacks, 1,126 R2L attacks, 52 U2R attacks and 4107 Probe attacks.
- **Data for intrusion *identification* (learning base of kdd99)**
 - Learning data:
 - Randomly selected 7,000 normal network connections
 - The former 2,000 back, 10,000 Neptune, 200 Pod, 20,000 Smurf, 800 Teardrop, 40 Guess passwd, 900 Warezclient, 1000 Ipsweep, 900 Portsweep, 1200 Satan, 200 Nmap, 15 Warezmaster, 25 buffer overflow attack
 - Test data
 - All the other network connections of these types of attacks are used for identification.

Building Multiple Behavioral Models for Network Intrusion Identification

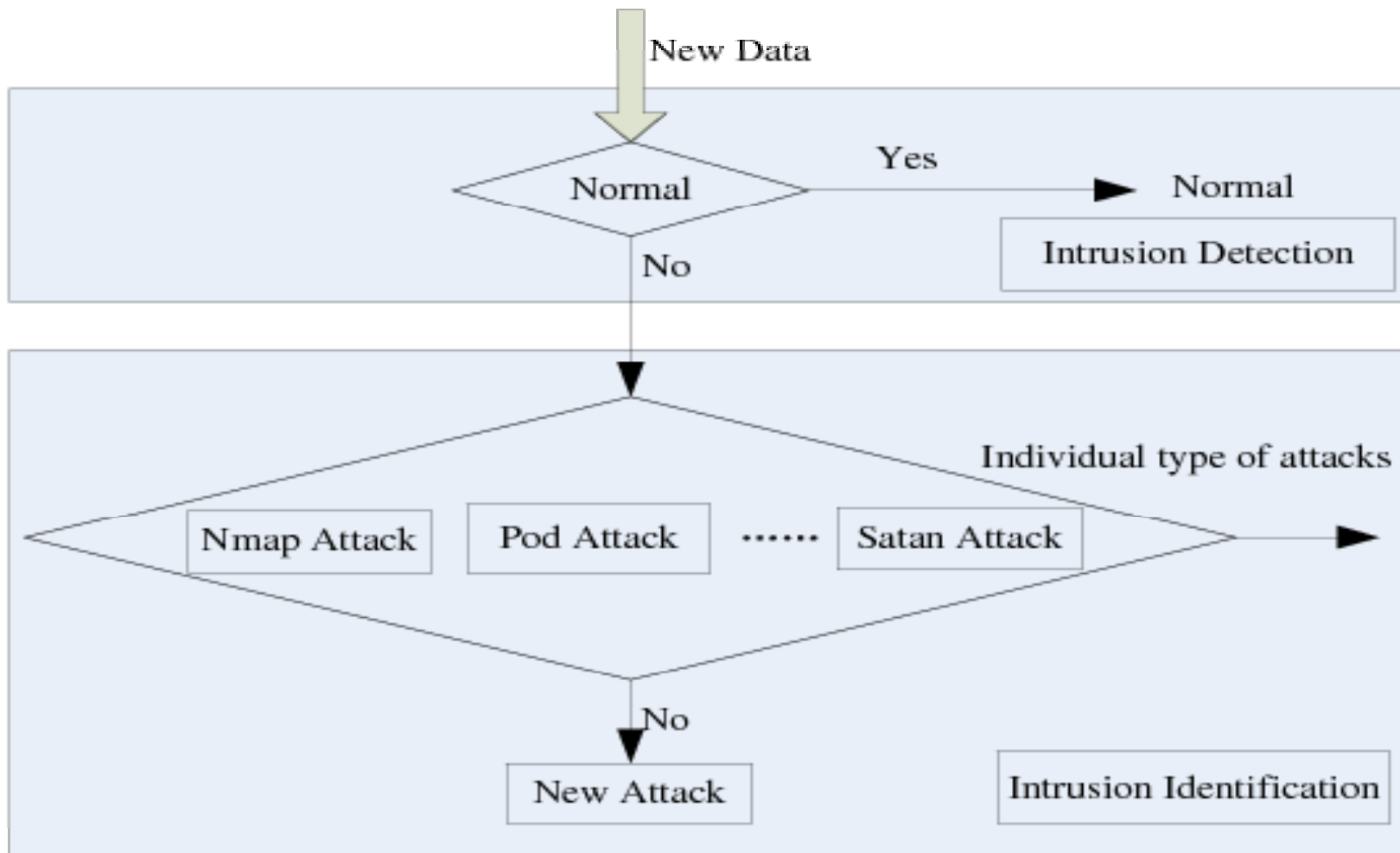
- **Building the normal model based on normal data for *intrusion detection***
- **Building individual attack model based on corresponding attack data for *intrusion identification***

Building Multiple Behavioral Models for Network Intrusion Identification

By W. Wang, S. Gombault et A. Bsila

Monam'07, Toulouse, France, November 5-6, 2007

➔ The general Intrusion detection and Identification Model



Building Multiple Behavioral Models for Network Intrusion Identification

➔ **kNN Based intrusion detection**

- **Building normal behavioral model**

- Calculate the distances between each test vector \mathbf{t} and each vector in the training data set by using Euclidean distance:

$$dis_{eu}(\mathbf{t}, \mathbf{x}_j) = \|\mathbf{t} - \mathbf{x}_j\| = \sqrt{\sum_{i=1}^M (t_i - x_{ij})^2}$$

- Sort the distance and choose the k nearest neighbors.
- Average the k closest distance scores as the *anomaly index*.

- **Detection**

- If the *anomaly index* of a test sequence vector \mathbf{t} is above a threshold ε
 - the test sequence is then classified as abnormal.
 - otherwise it is considered as normal.

Building Multiple Behavioral Models for Network Intrusion Identification

➔ **kNN based intrusion identification**

Define normal and individual attack data sets as D_1, D_2, \dots, D_l ;

Identification:

For each test vector **t do**

Calculate $dis_{eu}(\mathbf{t}, \mathbf{x}_j)$ for \mathbf{X}_j in each training set;

Find k smallest scores of $dis_{eu}(\mathbf{t}, \mathbf{x}_j)$ as k -nearest neighbors;

If more than a half of k nearest neighbors correspond to a specific attack type A_k **then**
t is identified as A_k

Else If the number of smallest distance that corresponds to an attack type A_p is greater than those of others **then**

t is identified as A_p

Else then

t is identified as a new attack

End If

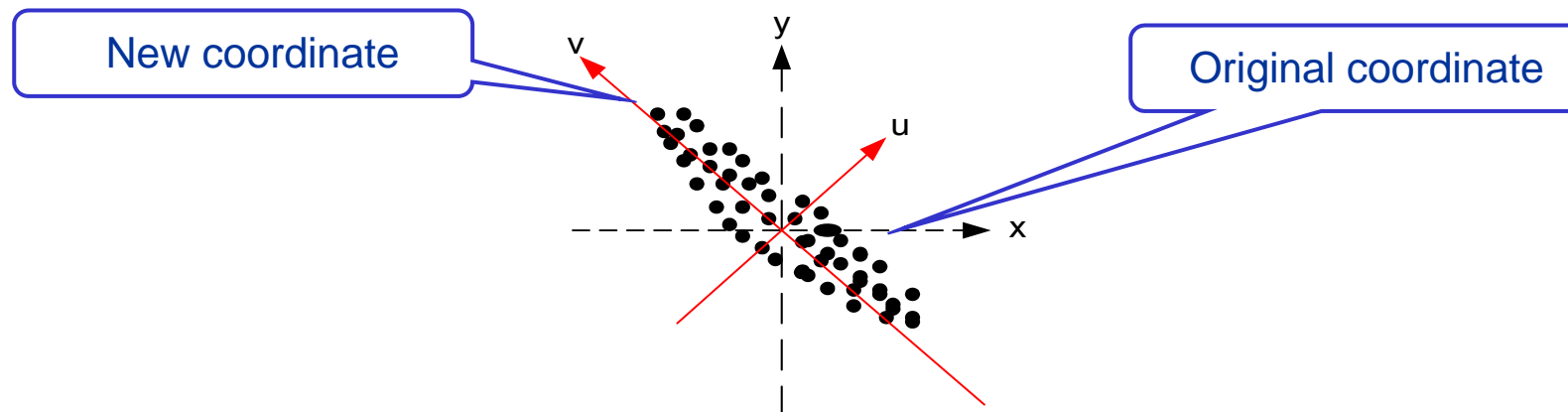
End For

➔ PCA methods for intrusion detection

14 --

➔ Principal Component Analysis

- **Dimension reduction technique for data analysis and compression**
- **New coordinate system to represent the original large data set**
 - The axes are the eigenvectors associated with the several largest eigenvalues
 - without sacrificing valuable information in the data set
- **Have been applied in face recognition, text categorization, etc.**

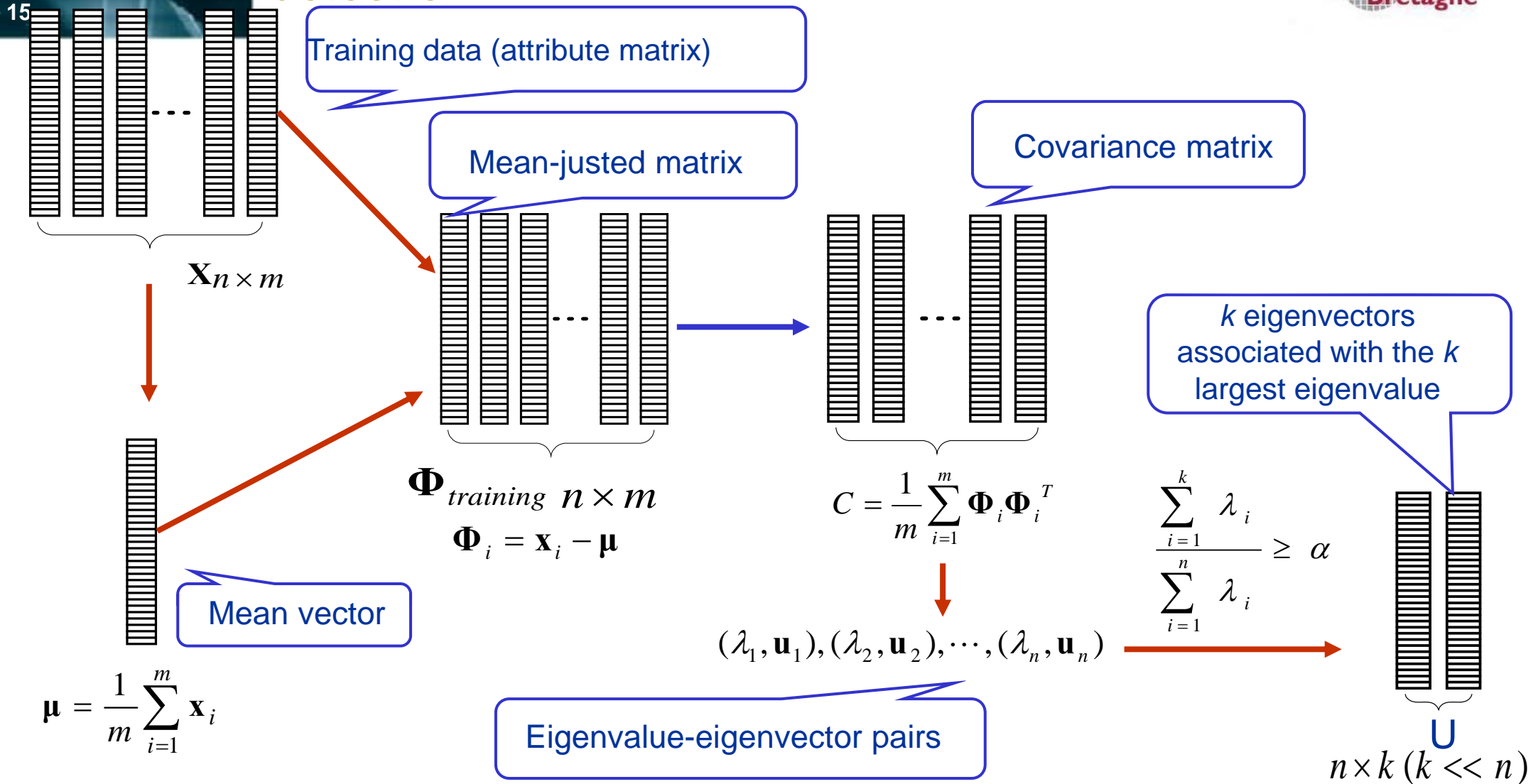


Building Multiple Behavioral Models for Network Intrusion Identification

By W. Wang, S. Gombault et A. Bsila

Monam'07, Toulouse, France, November 5-6, 2007

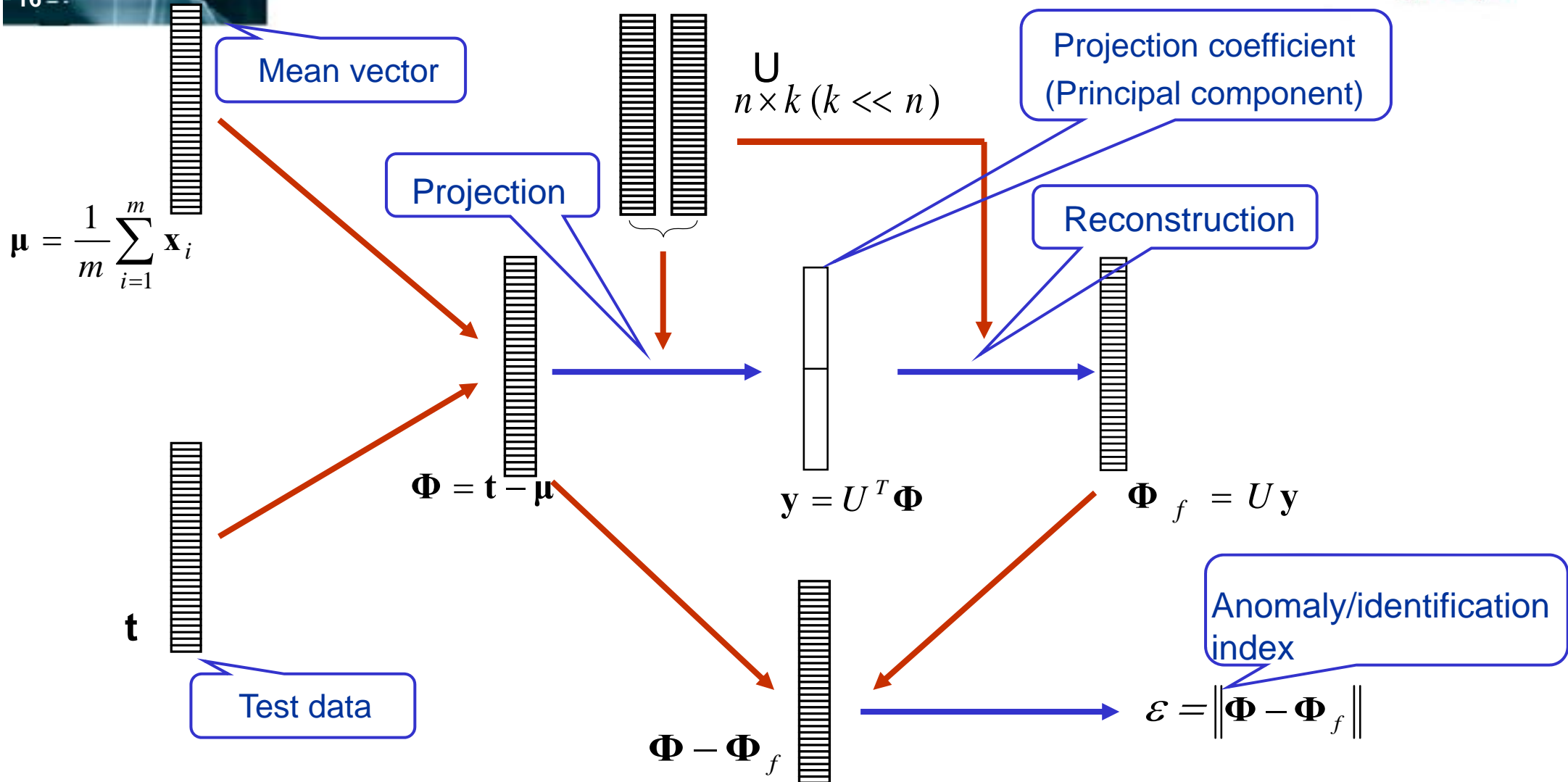
PCA based normal model building for intrusion detection



Building Multiple Behavioral Models for Network Intrusion Identification

➔ Intrusion detection based on PCA model

16 --



Building Multiple Behavioral Models for Network Intrusion Identification

➔ PCA based intrusion detection and intrusion identification



- **Intrusion detection**

- Given a new data vector \mathbf{t} , If its *anomaly index* ε is above a threshold, the test vector is considered as abnormal
- Otherwise, it is classified as normal

- ***Intrusion identification***

- Calculate the Euclidean distance between the test vector and its reconstruction onto each subspace formed by normal data and individual type of attack and set the minimum ε_i as the *identification index*.
- If ε_i is below the predefined threshold θ_i for a certain individual type of attack, the vector is then identified as this type of attack.
- Otherwise it is identified as a new attack.

Building Multiple Behavioral Models for Network Intrusion Identification

⊖ Intrusion detection: results based on PCA and kNN for kdd99 data

Methods	Overall data		DoS		R2L		U2R		Probe	
	DR (%)	FPR (%)	DR (%)	FPR (%)	DR (%)	FPR (%)	DR (%)	FPR (%)	DR (%)	FPR (%)
<i>k</i> NN (k=5)	84.3	2.9	87.1	2.9	37.6	1.6	75	4.1	56.4	18.6
PCA	98.8	0.4	99.2	0.2	94.5	4	88.5	0.6	80.7	4

Building Multiple Behavioral Models for Network Intrusion Identification

By W. Wang, S. Gombault et A. Bsila

Monam'07, Toulouse, France, November 5-6, 2007

⊕ Intrusion identification: results based on PCA and kNN for kdd99 data



Attack type	Attack category	Identification Rate (%)			PCA
		kNN			
		<i>k=5</i>	<i>k=7</i>	<i>k=9</i>	
guess_passwd	R2L	92.3	92.3	92.3	92.3
warezclient	R2L	100	100	100	57.5
warezmaster	R2L	80	80	80	100
back	DoS	98.5	99.5	98	100
neptune	DoS	99.8	99.8	97.7	95.3
pod	DoS	100	96.9	100	95.3
smurf	DoS	100	100	100	80.5
teardrop	DoS	97.7	99.4	97.8	100
buffer overflow	U2R	80	80	80	60
ipsweep	Probe	97.6	99.2	97.6	6.1
nmap	Probe	12.9	12.9	12.9	67.1
portsweep	Probe	100	100	100	0
satan	Probe	88.2	88.2	88.2	91.5

Building Multiple Behavioral Models for Network Intrusion Identification

➔ kNN and PCA methods comparison

- **kNN**

- No need for training
 - Suitable for dynamical environment
- Require large computation in testing stage
 - Need $O(m^2n)$ computation (m – dimensionality of vector; n – number of samples)

- **PCA**

- Need considerable computation for training
- Light weight in testing stage
 - Need $O(mqp)$ computation (p – number of different attack types; q – number principal components)
 - Suitable for detection massive data

Building Multiple Behavioral Models for Network Intrusion Identification

➔ Conclusion



- **Using the 41 attributes can achieve 72% detection rate of Supelec normal data**
- **kNN and PCA achieve good detection and identification results based on kdd99 data**
 - PCA can process massive data sets
 - Identification process needs attack data set (sometimes it is difficult)
- **The 41 attributes may be reduced for light weight detection while remain the detection accuracy**
 - Use some optimization methods for selecting key attributes in future work
- **Early and fast detection of network attacks is important**
 - No need to wait the connection is finished and early detection is our future work

Building Multiple Behavioral Models for Network Intrusion Identification

- 
- ➔ **Thank for your attention!**
 - ➔ **Merci pour votre attention!**
 - ➔ **Questions?**