Hochschule Zittau/Görlitz
UNIVERSITY OF APPLIED SCIENCES

EAD
ENTERPRISE APPLICATION DEVELOPMENT

Privacy Preserving Data Analysis for Cross-Company Cooperation

Jörg Lässig <jlaessig@hszg.de>

Introduction

Privacy Preserving Data Mining

- discover patterns in large data sets $\Rightarrow$ new data
- science, medicine, business, etc.
- call for privacy policy
- 2 main approaches [GLR10]:
  - anonymization (data for personal identification)
  - secured distributed data mining

Introduction to CoPPDA

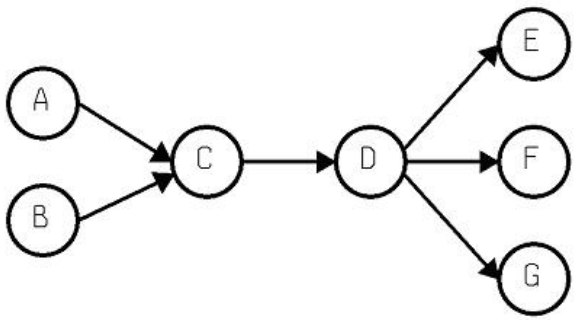
General information

- **Co**rporate **P**rivacy **P**reserving **D**ata **A**nalysis
- **companies** mainly use two approaches for raising their efficiency:
  - modern information and communication systems
  - inter organizational networks
- combinational approach of these two strategies is not very common
- cross company data mining is not possible, because of strategic and juristic restraints (e.g. sharing of in-house data, personal information)
- **CoPPDA** as privacy preserving data mining tool.

Innovation

Scenario

- currently no software implementation
- method described in Schlitter N [SN08]:
  - supplyer A, B and C as producer
  - distribution center D
  - E, F and G as retailers and wholesalers



  - every product gets a RFID chip
  - quality control (class value): A (accurate), D (defective)

Innovation

Scenario

ID	Day	Machine	Parameter
A1	Wednesday	1	29,8
A2	Wednesday	2	31,1
A3	Thursday	1	30,5
A4	Tuesday	2	29,9
A5	Monday	1	30,0
A6	Monday	2	30,3
A7	Friday	2	30,7
A8	Thursday	1	30,8
A9	Monday	2	29,8
A10	Wednesday	1	30,7
A11	Tuesday	1	30,7
A12	Friday	2	30,2
A13	Wednesday	2	31,2

ID	Supplier 's ID	Day	Shift	Temperature
C1	A1,B21	Wednesday	Day	25
C2	A2,B22	Wednesday	Day	20
C3	A3,B23	Thursday	Night	22
C4	A4,B24	Tuesday	Late	23
C5	A5,B24	Monday	Day	21
C6	A6,B26	Monday	Night	25
C7	A7,B27	Friday	Day	21
C8	A8,B28	Thursday	Night	23
C9	A9,B29	Monday	Late	24
C10	A10,B30	Wednesday	Night	23
C11	A11,B31	Tuesday	Day	24
C12	A12,B32	Friday	Night	24
C13	A13,B33	Wednesday	Night	22

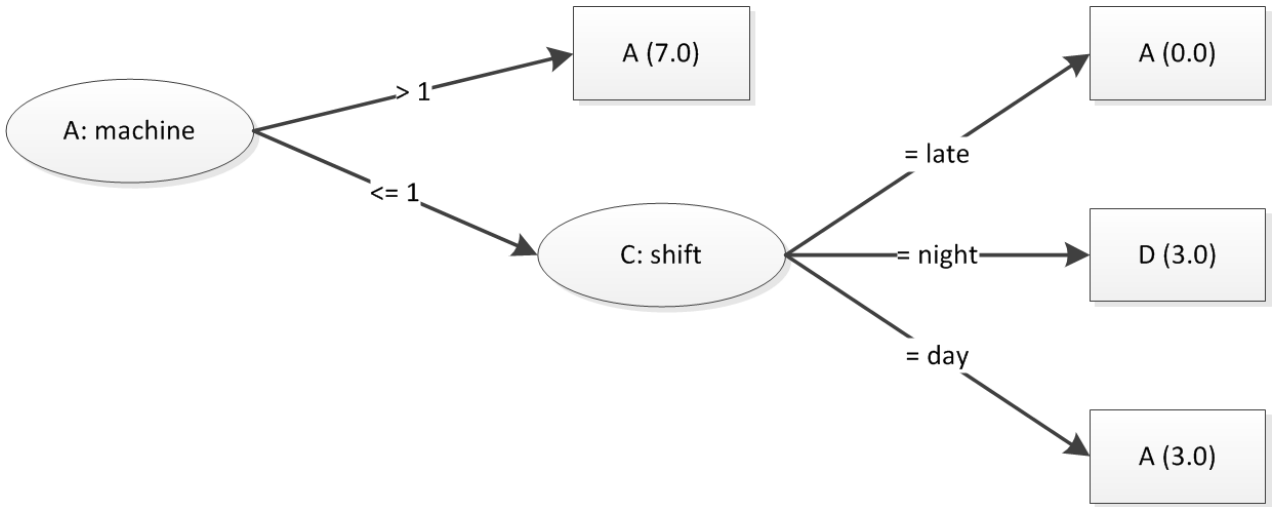
Innovation

Scenario

Unternehmen A			Unternehmen B			Unternehmen C			
Day	Machine	Parameter	Day	Transport	Term	Day	Shift	Temperature	Klasse
Wed.	1	29,8	Wed.	Air	1	Thurs.	Day	25	A
Wed.	2	31,1	Thurs.	Train	4	Mon.	Day	20	A
Thurs.	1	30,5	Thurs.	Truck	4	Mon.	Night	22	D
Tue.	2	29,9	Wed.	Train	4	Sun.	Late	23	A
Mon.	1	30,0	Tues.	Truck	4	Sat.	Day	21	A
Mon.	2	30,3	Tues.	Truck	4	Sat.	Night	25	A
Fri.	2	30,7	Friday.	Train	5	Wed.	Day	21	A
Thurs.	1	30,8	Thurs.	Truck	4	Mon.	Night	23	D
Mon.	2	29,8	Tues.	Air	1	Wed.	Late	24	A
Wed.	1	30,7	Thurs.	Truck	4	Tues.	Night	23	D
Tues.	1	30,7	Wed.	Truck	4	Sun.	Day	24	A
Friday	2	30,2	Friday	Train	5	Wed.	Night	24	A
Wed.	2	31,2	Thurs.	Truck	4	Mon.	Night	22	A

Innovation

Scenario



- products from company A, not machine 1 $\rightarrow$ accurate
- products form company A, machine 1; company C, day and late shift $\rightarrow$ accurate
- products from company A, machine 1; company C, night shift $\rightarrow$ defective

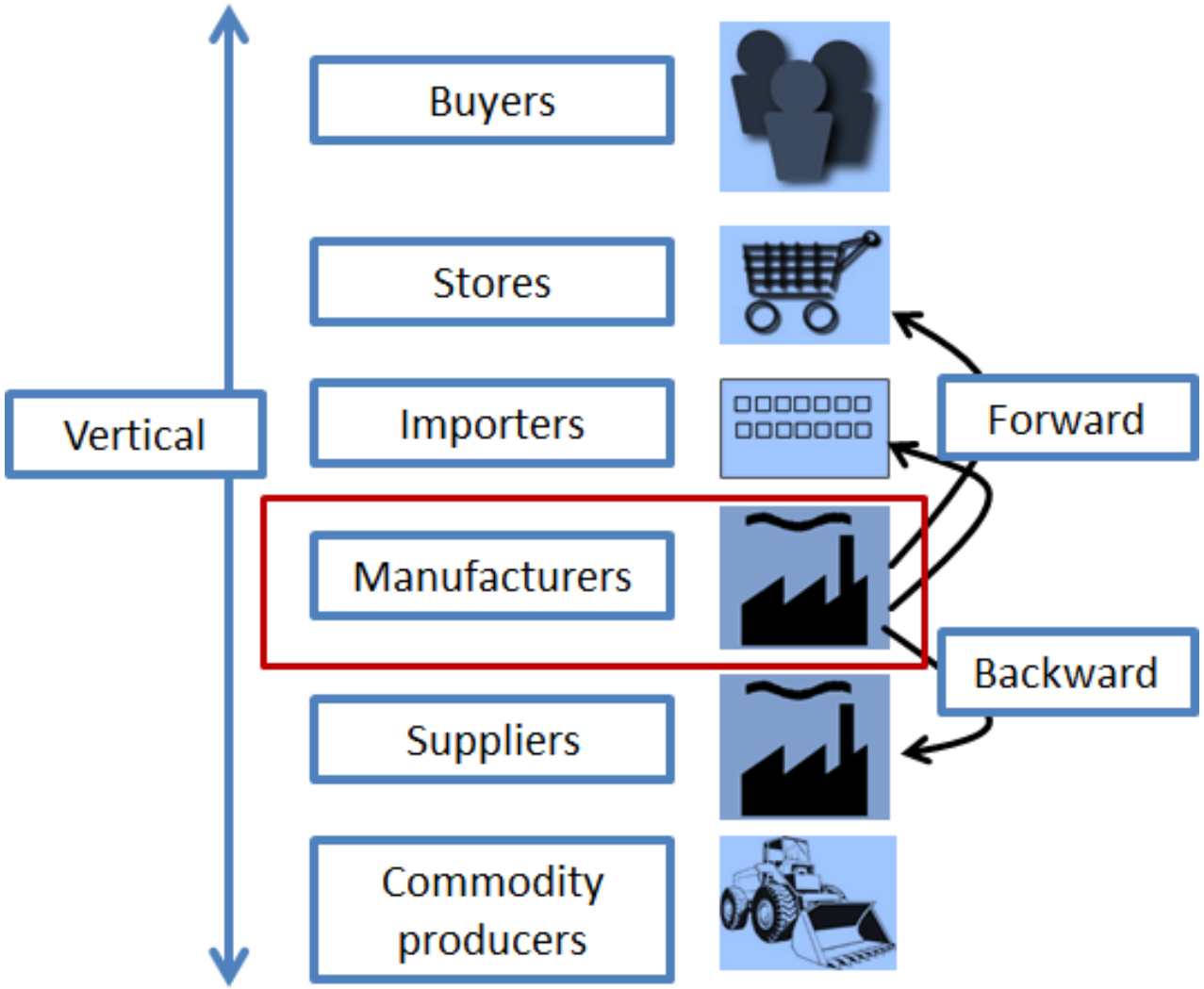
Supply chain: vertical vs horizontal

Vertical cooperation

- companies of different stages of the value chain are working together
- cooperating companies stay legally and economically independent
- can be limited to a part of business of a company
- types of cooperation:
  - forward cooperation: working together with companies closer to the final customer
  - backward cooperation: working together with companies in the direction of procurement

Supply chain: vertical vs horizontal

Vertical cooperation



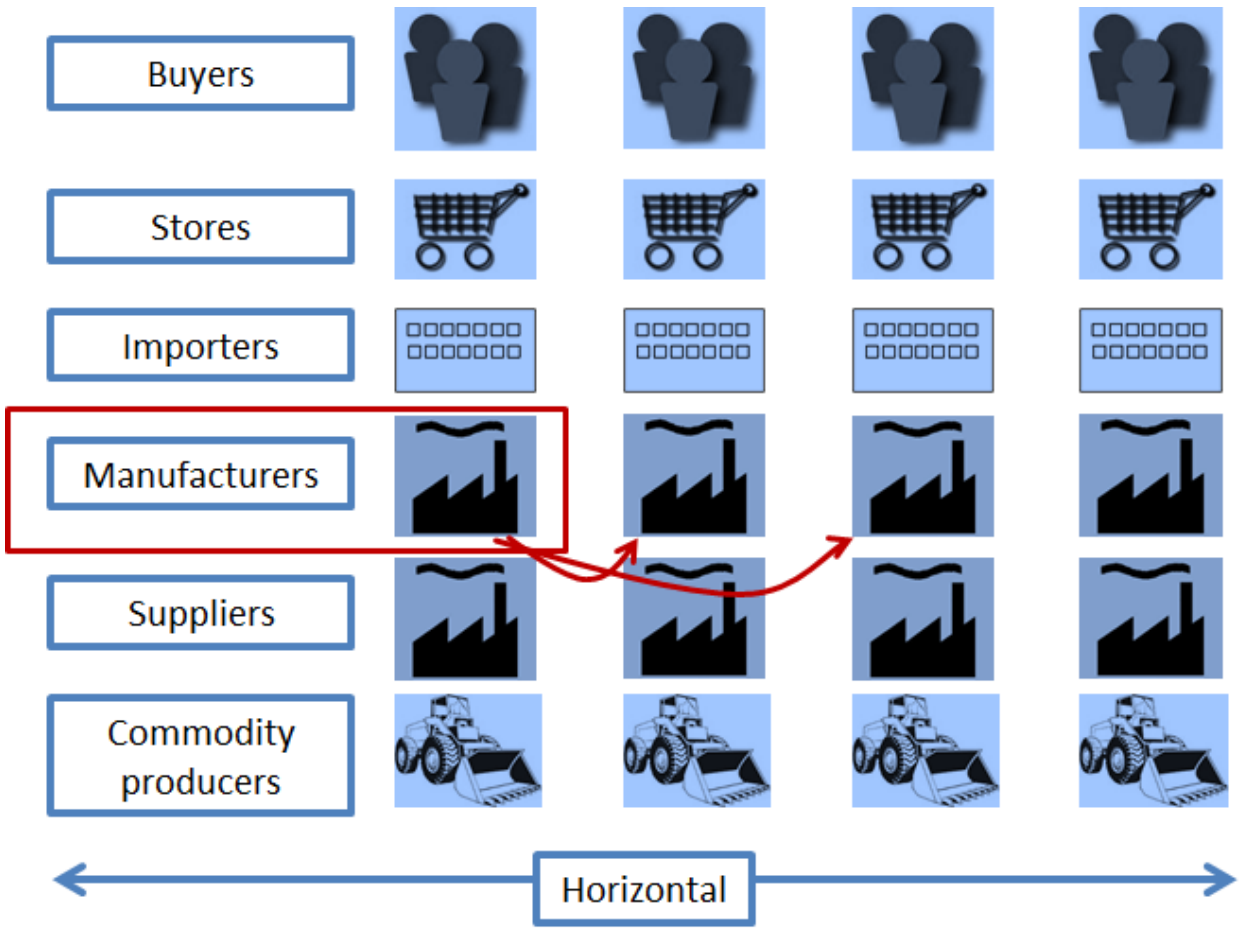
Supply chain: vertical vs horizontal

Horizontal cooperation

- two companies of the same industry and in the same stage of production work together
- companies belong to the same supply chain stage and normally produce or trade the same products
- add their strength to gain benefits
- affects the processes and structure design of distribution networks
- cooperation creates a change of existing hubs
- requires inter-firm coordination

Supply chain: vertical vs horizontal

Horizontal cooperation



Problems

- not fulfill requirements regarding privacy and data security
- exchange of sensitive and confidential data during analysis
- production data of A and B for C visible
- refusal of cooperative analysis $\rightarrow$ no optimization potential
- CoPPDA $\Rightarrow$ no disclosure or exchange of sensitive data
- Secure Multi-Party Computation [DZ02]

CoPPDA

Work plan and milestones

- implementation of ppdm algorithms
  - **ID3** (horizontal, vertical)
  - **Backpropagation Networks** (vertical)
  - **k-means** (horizontal, vertical)
- development of new ppdm methods
- integration in **rapid miner**
- test scenarios

Paillier Cryptosystem

public keys g, n plaintext $m < n$
private key λ ciphertext $c < n^2$

Additive Homomorphic Properties

$$\forall m_1, m_2 \in \mathbb{Z}_n \text{ and } k \in \mathbb{N}$$

$$\mathsf{D}(\mathsf{E}(m_1)\mathsf{E}(m_2) \bmod n^2) = m_1 + m_2 \bmod n$$

$$\left.\begin{array}{r} \mathsf{D}(\mathsf{E}(m_1)^{m_2} \bmod n^2) \\ \mathsf{D}(\mathsf{E}(m_2)^{m_1} \bmod n^2) \end{array}\right\} = m_1 m_2 \bmod n$$

$$\mathsf{D}(\mathsf{E}(m)^{k} \bmod n^2) = km \bmod n$$

$$\mathsf{D}(\mathsf{E}(m_1)g^{m_2} \bmod n^2) = m_1 + m_2 \bmod n$$

Iterative Dichotomiser 3

Information gain computation

- Gini Index

$$Gain(S, A) = Gini(S) - \sum_{i=1}^{n} \frac{|S_{A_i}|}{|S|} Gini(S_{A_i})$$

- Entropy

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_{A_i}|}{|S|} Entropy(S_{A_i})$$

- ID3 algorithm to create decision tree

Iterative Dichotomiser 3

- in [SM09]:„Privacy preserving ID3 using Gini Index over horizontally partitioned data.“
- theoretical approach, using the *Additive Homomorphic Encryption* by Pascal Paillier

Information gain computation (main protocol)

$$Gain(S, A) = Gini(S) - \sum_{i=1}^{n} \frac{|S_{A_i}|}{|S|} Gini(S_{A_i})$$

$$Gini(S_{A_i}) = 1 - \sum_{j=1}^{m} P_{A_i C_j}^2$$

$$Gini(S_{A_i}) = 1 - \sum_{j=1}^{m} \frac{|S_{A_i C_j}|^2}{|S_{A_i}|^2}$$

Iterative Dichotomiser 3

after a whole bunch of derivation (main protocol)

The goal is to calculate the equation:

$$\frac{\left(\sum_{i=1}^{k}|S_{A_iC_1}|\right)^2+\cdots+\left(\sum_{i=1}^{k}|S_{A_iC_m}|\right)^2}{\left(\sum_{i=1}^{k}|S_{A_iC_1}|\right)+\cdots+\left(\sum_{i=1}^{k}|S_{A_iC_m}|\right)}$$

Sub-protocols for privacy preserving ID3

- Multi-party Addition
- Multi-party Multiplication
- Secure multi-party square division

Future work
CoPPDA

- finish pp ID3 (netty - asynchronous event-driven network application framework)
- implement other pp algorithms
- cost and efficiency
- integration in rapid miner
- testing

Thank you.

Jörg Lässig <jlaessig@hszg.de>

References

Henrik Grosskreutz, Benedikt Lemmen und Stefan Rüping. „Privacy-Preserving Data-Mining“. German. In: *Informatik-Spektrum* 33.4 (2010), S. 380–383. ISSN: 0170-6012. DOI: 10.1007/s00287-010-0447-1. URL: http://dx.doi.org/10.1007/s00287-010-0447-1.

Schilz ST Schlitter N. „Strategischer IKT-Einsatz schafft Wettbewerbsvorteile durch unternehmensübergreifendes Data Mining“. In: *Tagungsband ZFPro'08*. M&S-Verlags-OHG, 2008.

Wenliang Du und Zhijun Zhan. „Building decision tree classifier on private data“. In: *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*. Australian Computer Society, Inc. 2002, S. 1–8.

Saeed Samet und Ali Miri. „Privacy preserving ID3 using Gini Index over horizontally partitioned data.“ In: *AICCSA*. IEEE, 12. Mai 2009, S. 645–651. URL: http://dblp.uni-trier.de/db/conf/aiccsa/aiccsa2008.html#SametM08.