The Role of Semantics in Systems Integration & Cyber Security

George Karabatis, Professor
Department of Information Systems
Director, Entrepreneurship & Innovation Minor
georgek@umbc.edu

October 27, 2023

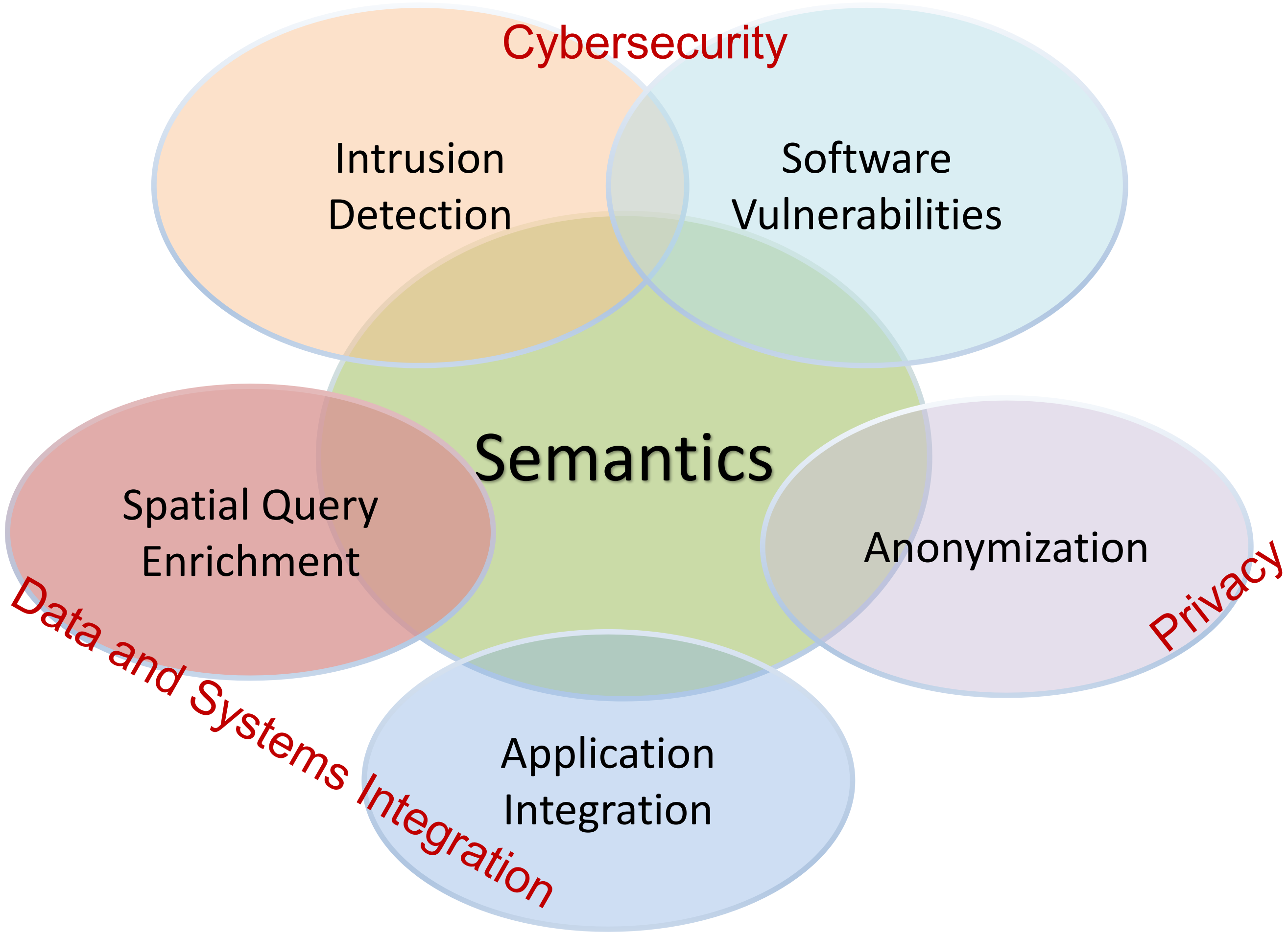
Outline

- Semantics
- Systems and Information Integration
- Cybersecurity
- Anonymization
- Semantics in GIS
- Possible collaborations

Research Experience

- Industrial
  - Bell Communications Research (Bellcore)
    - Spin-off from AT&T Bell Labs, became Telcordia, then acquired by Ericsson (mobile operator)
    - Applied research on database servers for telecommunication applications
- Academic
  - UMBC, since 2002
  - Currently on a sabbatical at TUC

Research Topics



Semantics in Systems Integration

Systems Integration: A Real Problem

- Companies are constantly getting bought, sold, merged
- Their data is not just disparate, but all over (spreadsheets, files on premises, cloud, etc.)
- Market changes require companies to respond:
  - Either with flexibility
  - Or with bankruptcy
- Integration of systems is a major necessity of modern enterprises

Systems Integration: A Real Problem

- Integration projects are hideously
  - time consuming
  - expensive
  - with a dismal 84% failure rate!
- Larger projects have higher failure rates compared to smaller ones
- In general, 50% of IT budget in enterprises goes to integration projects
- **It is not a simple problem when we ignore semantics**

Semantics

- Q: What does ***semantics*** mean?
- A: Simply, the meaning of… words, sentences, text, etc.

- If we type ‘credit’ on Google we get 7.34 billion links (in 0.51secs) **Too generic includes all possible hits – not usable**
- If we provide more context (more semantics)
  - Credit within a university environment: 1.4M links - **still useless**
  - Add more context, e.g., specific course: 116 links - **much better!**

Why Semantics?

Semantics

A human can fully comprehend the intended message by recognizing the actual meaning of the word (within a sentence)

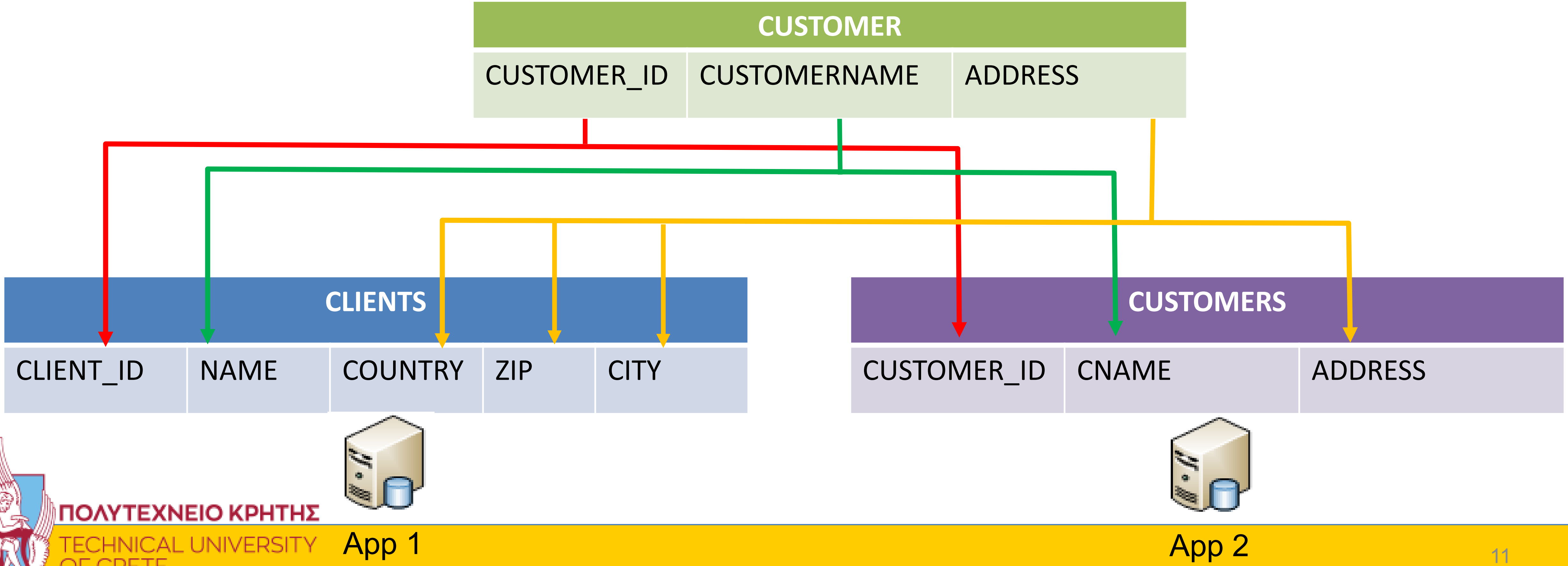
A machine can ‘comprehend’ the intended purpose of the data by recognizing the actual meaning of the data (in a computing environment)

Current Approaches

- Products in general resort to ad-hoc adaptors (not scalable)
- Integration projects rely on programmers to decipher semantics (costly, inefficient, patch)
- **Fundamental issue: Not easy to automatically find the mapping from one object to another**

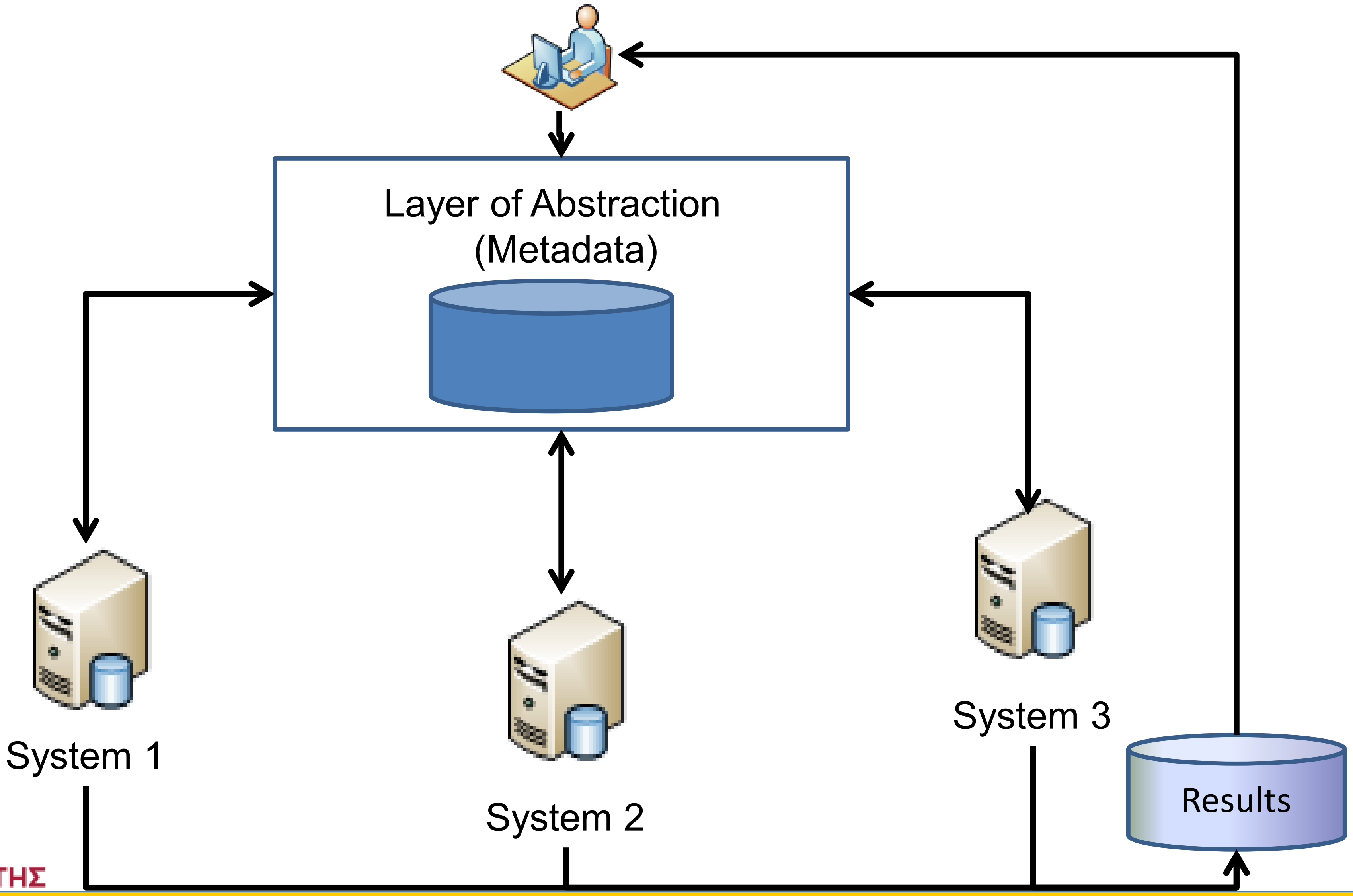
A Simple Integration Problem

Scenario: After Company 1 buys Company 2, create a list with all customer names in both App1 and App2.

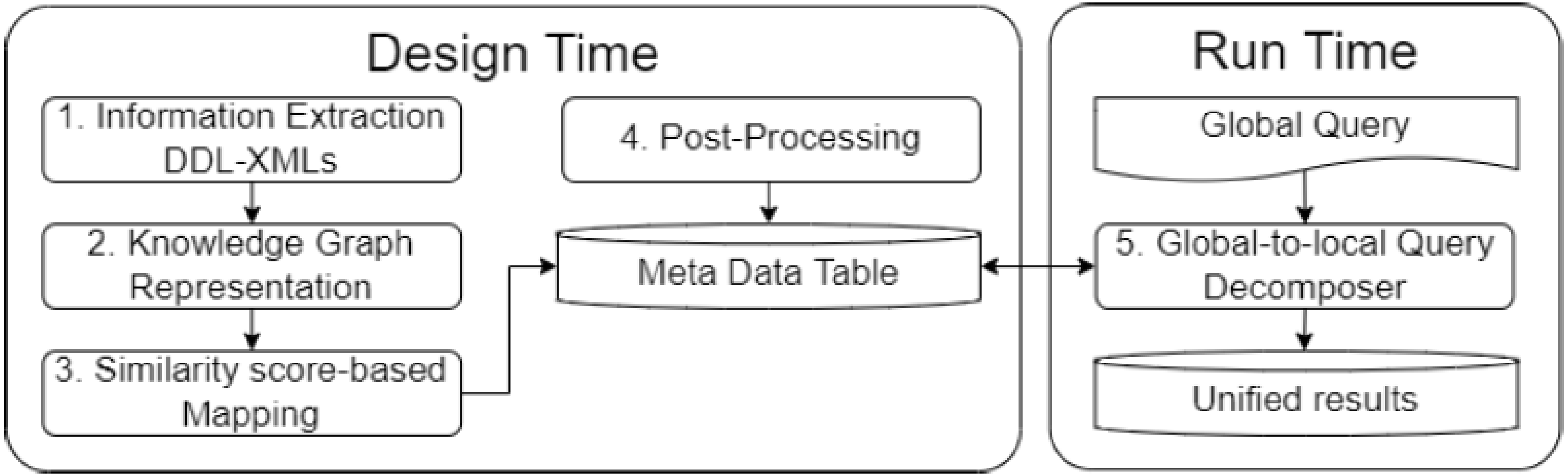


ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Systems Integration



IntePlato



With Leonard Traeger and Andreas Behrend

Automated object mappings between numerous heterogeneous databases

Similarities in IntePlato

#	Similarity function	Attribute	Table
a	Fuzzy search score	✓	✓
b	Synonyms retrieval & intersection	✓	✓
c1	Datatype similarity score	✓	
c2	Constraint similarity score	✓	
c3	Attribute score	$a \oplus b \oplus c1 \oplus c2$	
c4	Attribute with parent table score	$d1$	
c5	Total attribute score	$c3 \oplus c4$	
d1	Table score		$a \oplus b$
d2	Table with attribute children score		$\sum \max(c3)$
d3	Total table score		$d1 \oplus d2$

Similarity functions in IntePlato

IntePlato Mapper

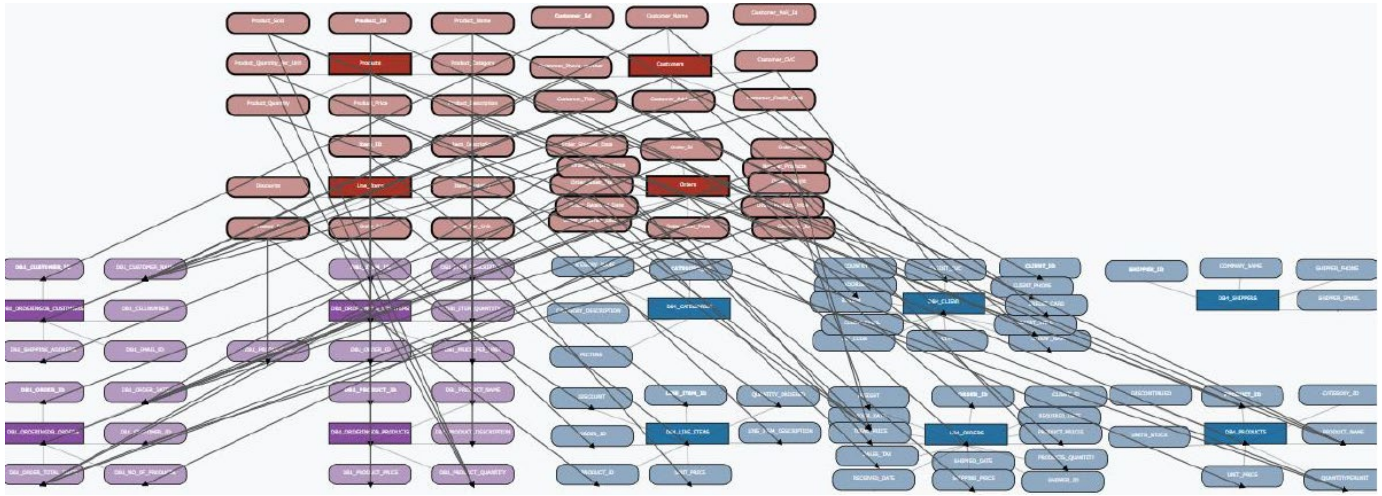
Algorithm 1 ⊕ Generating Similarity Clusters

```
Require: st1, st2 --similarity two-tuples
    --with object structure (conceptID, similarity score)
let union = st1.concat(st2); --UNION
let distinctIDs = new Array;
let setOfClusters = new Array;
for all concept in union do
    if conceptID not in distinctIDs then
        distinctIDs.push(conceptID); --DISTINCT
    end if
end for
for all dConceptID in distinctIDs do
    let cluster = new Object;
    for all concept in union
    where conceptID = dConceptID do
        cluster.addScore(concept); --SUM
    end for
    setOfClusters.push(cluster);
end for
return setOfClusters with SUM of similarity scores
```

Output of potential mappings for CUSTOMER_MAIL_ID (global to local)

id	local_name	refined_local_name	schema	synonyms	fuzzy search	datatype	constraint	tot score attribute	tot score table	tot score	set
CONCEPT_46	DB1_EMAIL_ID	EMAIL ID	SCHEMA1	5/20	0.3874	1	1	2.6374	3	5.6374	
CONCEPT_44	DB1_CUSTOMER_NAME	CUSTOMER NAME	SCHEMA1	2/20	0.5001	1	1	2.6001	3	5.6001	
CONCEPT_47	DB1_SHIPPING_ADDRESS	SHIPPING ADDRESS	SCHEMA1	0	0	1	1	2	3	5	
CONCEPT_83	CLIENT_EMAIL_ID	CLIENT EMAIL ID	SCHEMA2	7/20	0.3631	1	1	2.7131	1.5	4.2131	
CONCEPT_45	DB1_CELLNUMBER	CELLNUMBER	SCHEMA1	0	0	0	1	1	3	4	
CONCEPT_43	DB1_CUSTOMER_ID	CUSTOMER ID	SCHEMA1	7/20	0.5606	0	0	0.9106	3	3.9106	
CONCEPT_75	CLIENT_NAME	CLIENT NAME	SCHEMA2	2/20	0	1	1	2.1	1.5	3.6	

IntePlato Mapper



Algorithm 2 Highest local concept mapper

```
Require: ath {ambiguity tolerance (0..1) hyperparameter}
let lc = new Array; {local clusters}
for all globalConcept in ConceptList do
    for all localSchema do
        if globalConcept is table then
            lc = totalTableScore(globalConcept);
        end if
        if globalConcept is attribute then
            lc = totalAttributeScore(globalConcept);
        end if
        if lc[1] / lc[0] < ath then
            globalConcept.map(lc[0]);
        else
            globalConcept.map(NULL);
        end if
    end for
end for
return Metadata Table;
```

- Input: Clusters of concepts
- Hyperparameters: Ambiguity tolerance (0..1), use of synonyms
- Output: Mapping table (metadata)

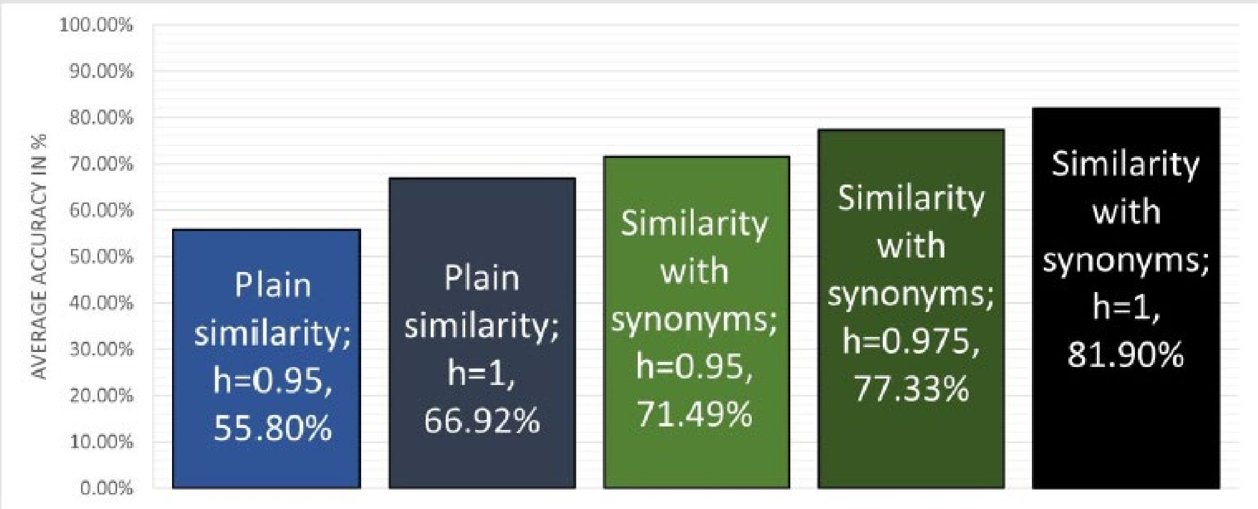
Evaluation of Mappings

Dataset: Three different database schemas were used

Schema 1: Four tables, 21 attributes

Schema 2: Six tables, 47 attributes

Global: Four tables, 37 attributes



Relevant and Future Work

Signature / Vectorization	Apply numerical embedding strategy to all entity profiles (tf-idf, word2vec, etc.)
Scoping	*Sort and constrain entity profiles (filter) to reduce the search pair space*
Blocking	Identify pairs likely to match into buckets (approximate)
Filtering	Discard pairs if they do not match

Relevant and Future Work

U. Brunner and K. Stockinger, “Entity matching with transformer architectures - a step forward in data integration,” Mar. 2020

R. Cappuzzo, P. Papotti, and S. Thirumuruganathan, “Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks,” in Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD ’20

Signature / Vectorization

D. Paulsen, Y. Govind, and A. Doan, “Sparkly: A Simple yet Surprisingly Strong TF/IDF Blocker for Entity Matching,” Proceedings of the VLDB Endowment, vol. 16, no. 6, pp. 1507–1519, Feb. 2023.

Scoping

F. Azzalini, S. Jin, M. Renzi, and L. Tanca, “Blocking Techniques for Entity Linkage: A Semantics-Based Approach,” Data Science and Engineering, vol. 6, no. 1, pp. 20–38, Mar. 2021.

G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, “A Survey of Blocking and Filtering Techniques for Entity Resolution,” Aug. 2020, arXiv:1905.06167

Blocking

S. Thirumuruganathan, H. Li, N. Tang, M. Ouzzani, Y. Govind, D. Paulsen, G. Fung, and A. Doan, “Deep learning for blocking in entity matching: a design space exploration,” Proceedings of the VLDB Endowment, vol. 14, no. 11, pp. 2459–2472, Jul. 2021

S. Lerm, A. Saeedi, and E. Rahm, “Extended Affinity Propagation Clustering for Multi-source Entity Resolution,” BTW 2021, 2021.

Filtering

C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, C. Lofi, A. Bonifati, and A. Katsifodimos, “Valentine: Evaluating Matching Techniques for Dataset Discovery,” in 2021 IEEE 37th International Conference on Data Engineering (ICDE), Apr. 2021, pp. 468–479

Semantics in Cybersecurity

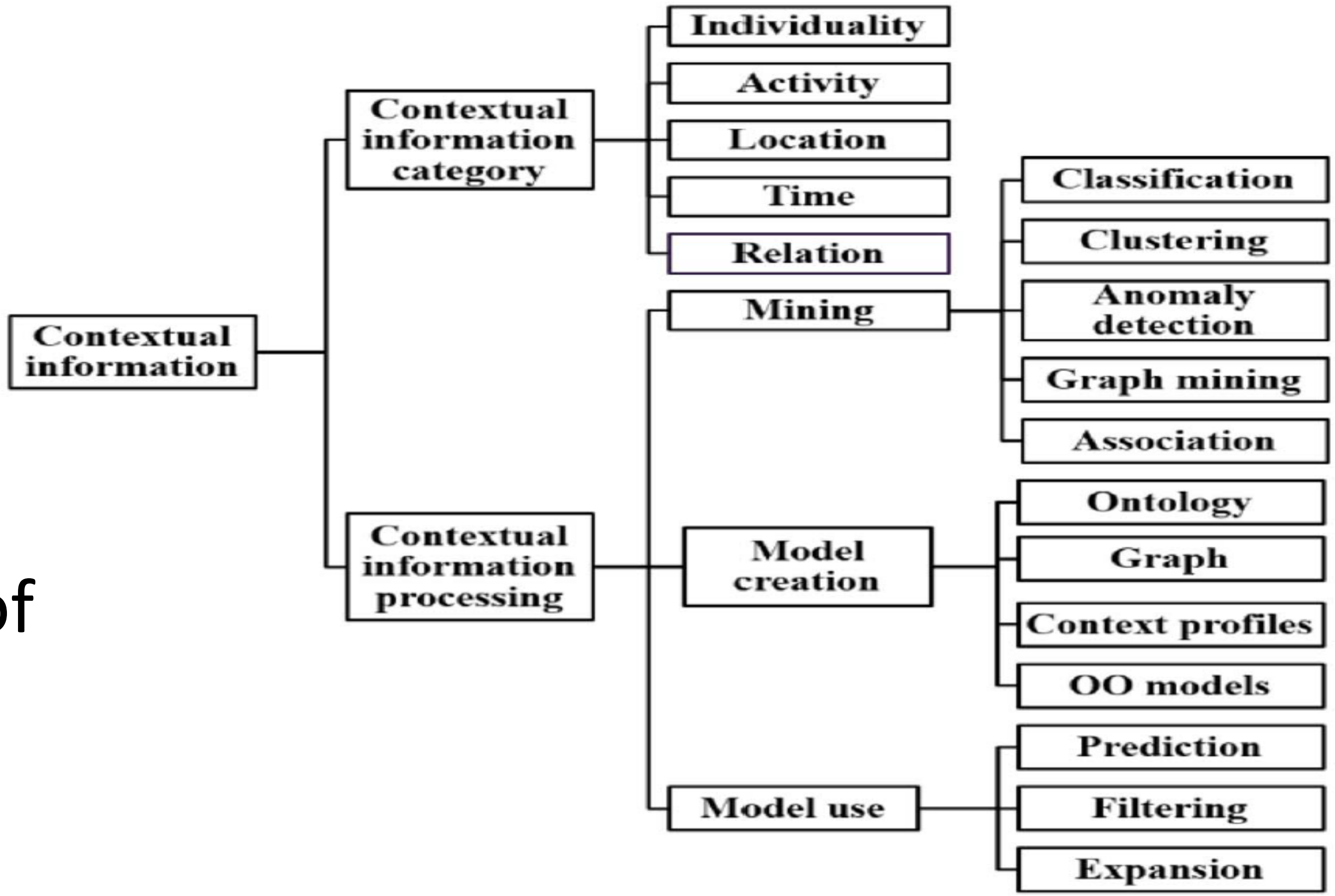
Cyber-attack Detection & Prevention

- Many Intrusion Detection Systems (IDS) help in identifying cyber-attacks but
- IDS trust is low
  – generate too many alerts
  – a lot of false positives
- Use semantic information about attacks
- Provide a more accurate detection of cyber-attacks

With Ahmed Aleroud

Related Work space

Context information is considered to be part of semantics



ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

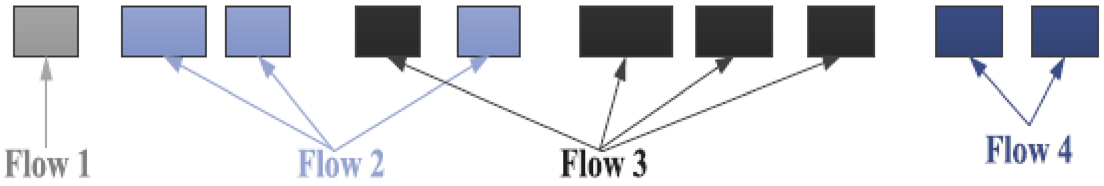
Cyber-attack Detection & Prevention

- Create a system that operates in two phases: Design and Run-time
- Incoming connection passes through the system at run-time
- If the incoming connection is deemed suspicious (confidently similar with set of potential attacks)
  - Mark incoming connection as threat
  - Disallow it from entering the organization network (or send to honeypot)
- else
  - Mark incoming connection as benign
  - Allow incoming connection to proceed

Cyber-attack Detection & Prevention

Design/calibration phase:

- Use an IDS to obtain alerts (along with alert description)

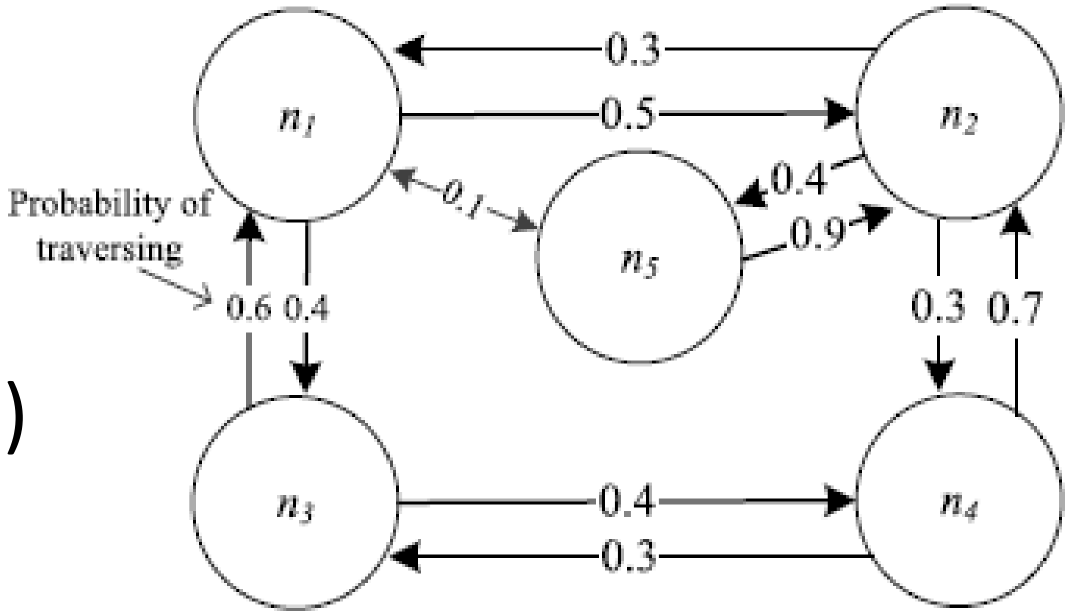


- Generate flows and correlate IDS alerts with flows
- Extract features: Time, loc, pckts, octs, prot, flags, alert description (non-stop words count as features)
- Mining ***semantic relationships based on the description of alerts*** reveals new knowledge that cannot be discovered by traffic features of the flows

Cyber-attack Detection & Prevention

Design/calibration phase:

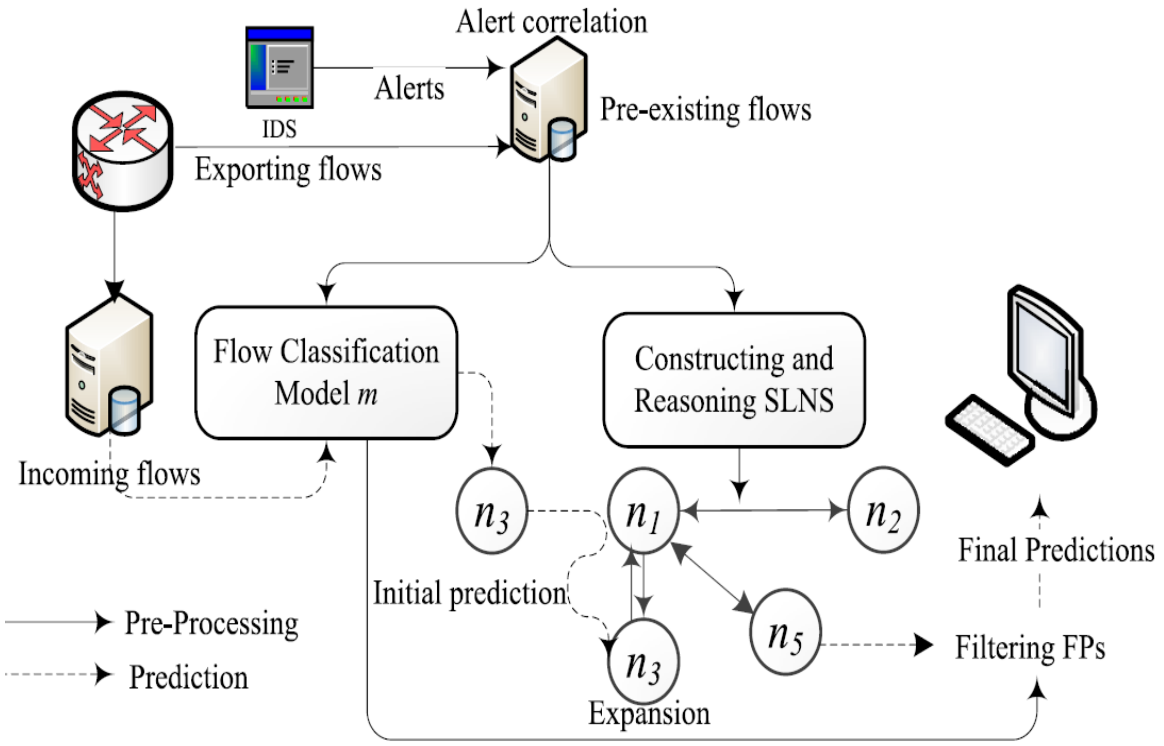
- Create a semantic link network based on similarity (Pearson, Anderberg, etc.) across features
- Edges: each edge identifies connectivity and (possibly multiple) relationships between nodes
- Relevance score *rs* between two nodes: $\mathrm{rs}_{(n_i \rightarrow n_j)} = \sum_{t_l} \prod_{1 \leq i \leq |t_l|} \mathrm{SIM}(n_{l_i}, n_{l_{i+1}})$



Cyber-attack Detection & Prevention

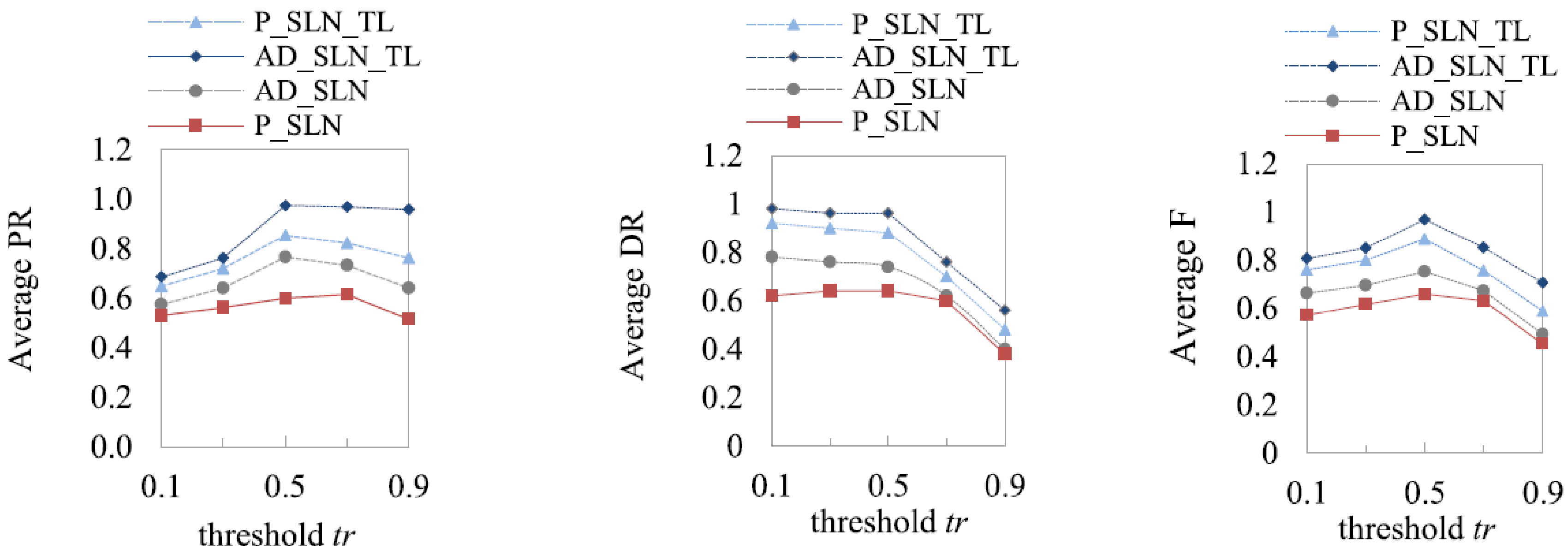
Run-time: Incoming flows are analyzed and marked either as benign or suspicious

- First, classify each flow based on a decision tree classifier
- Initial prediction is passed to SLN
- Expand to include additional semantically related predictions
- Filter out FP using profiles (classifier based on benign activities)



ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Experimental Results



(a) (b) (c)

Dataset: flow data from U. Twente containing several types of SSH and HTTP connection attempts
Approx: 570K suspicious flows, 104K benign flows
Experiments using both Pearson and Anderberg similarity formulas
TL : Time and location features

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Zero-day (unknown) attacks

- A zero-day attack or threat tries to exploit software vulnerabilities that are still unknown (it is an attack exploiting a vulnerability for which no patch exists)
- Question: Can we improve the detection rate of zero day attacks using semantics in our system?

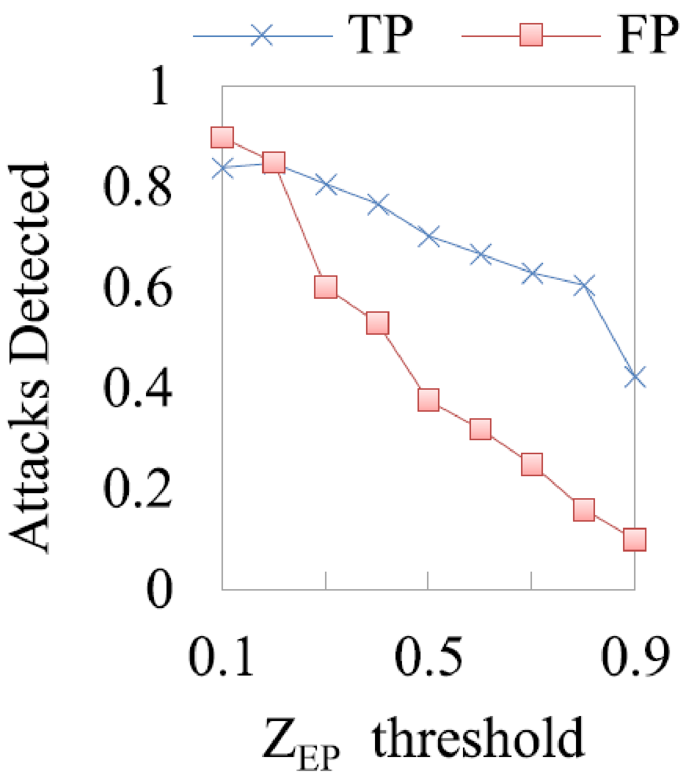
Challenges

Think, design, and innovate techniques that:

- Provide solutions to the problem
  - Identify 0-day attacks with a respectable success rate
  - Measured by TP & FP metrics
- Are practical for computer systems
  - Network data are coming with Gbps speeds
  - Algorithms that take too long to complete are not acceptable
  - Must run at (near) real-time

Detection of 0-day (unknown) attacks

- Remove 5 types of attacks from dataset to simulate 0-day environment
- Train classifiers on resulting dataset
- Measure TP and FP rates of incoming 0-day flows
- Caveat: Works only on 0-days similar to known attacks



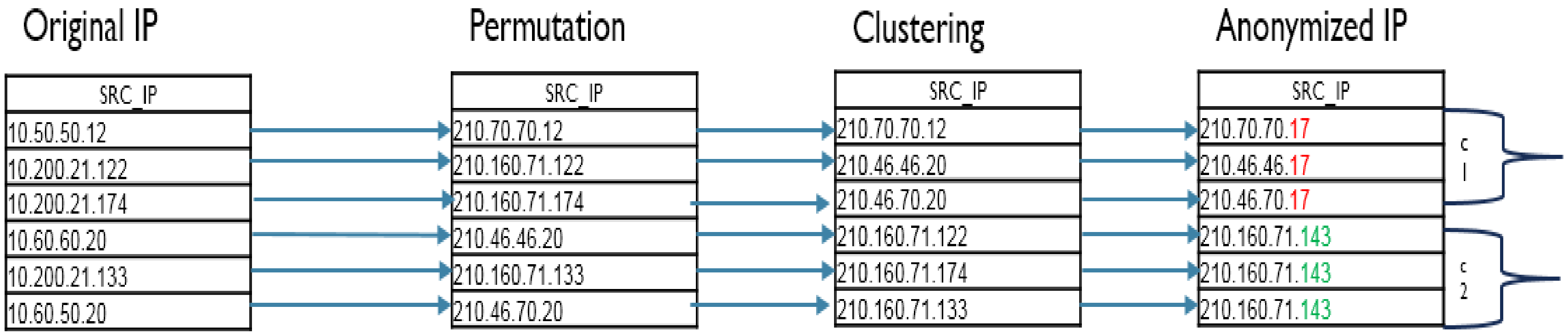
ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Semantics in Anonymization of datasets

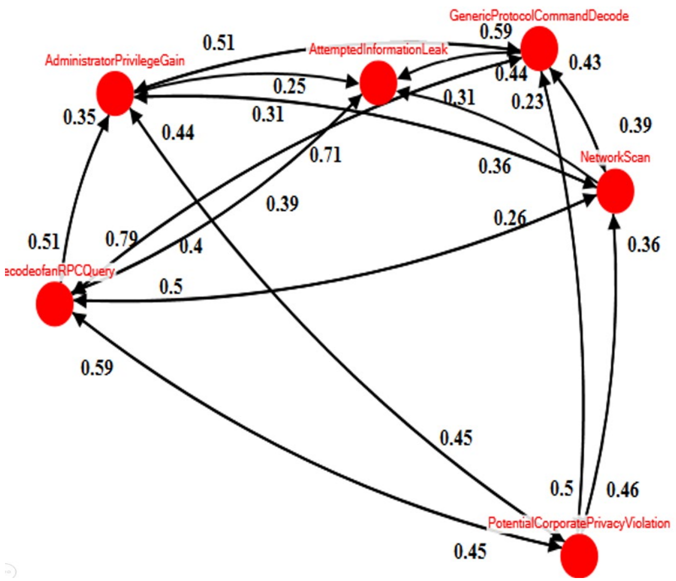
Data Anonymization

- Data contain private/sensitive information
- Most orgs cannot analyze data in-house
- Data analysis by 3rd party
- Privacy implications, law violations, etc.
- Facebook – Cambridge Analytica

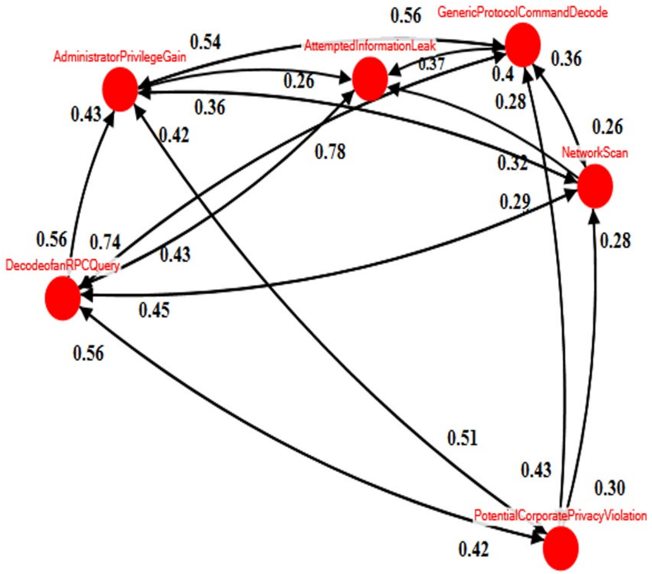
With Zhiyuan Chen, Ahmed Aleroud, Antonios Xenakis



Data Anonymization



SLN Before Anonymization



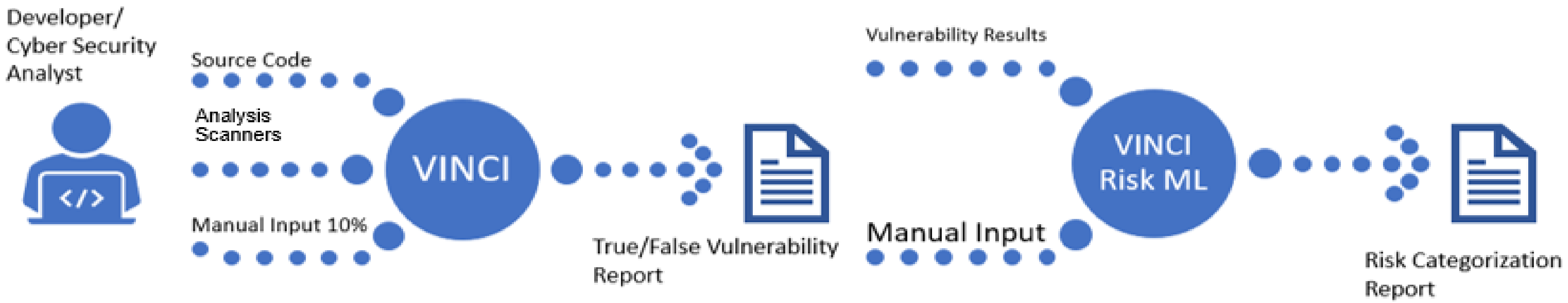
SLN After Anonymization

- Can we analyze anonymized data without losing a lot of information?
- Compare analyses on original vs anonymized datasets

Software vulnerabilities

- Software applications contain vulnerabilities
- Code is scanned with multiple tools to discover vulnerabilities
- Vulnerabilities range in the thousands for an average application
- Scanners produce a lot of FPs
- VINCI removes majority of FPs using an ensemble classifier

With Foteini Argiropoulos



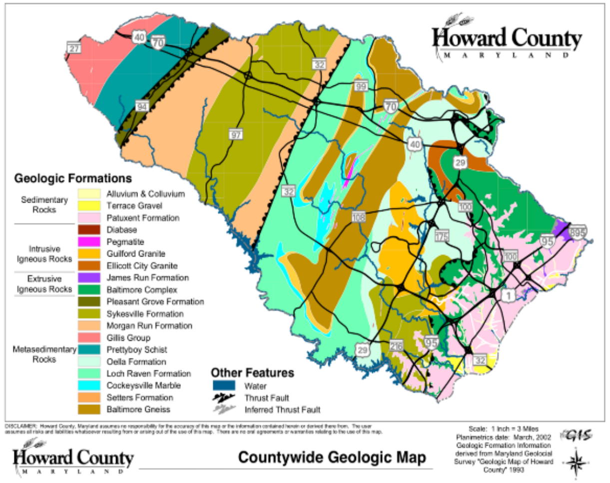
Semantics for Geographical Information (GIS) Systems

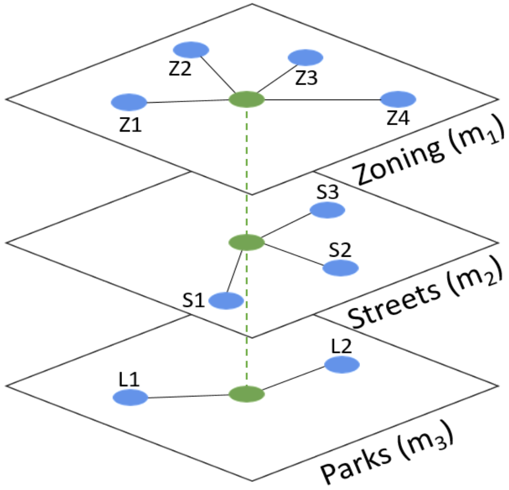
information systems UMBC

Augmenting Spatial Query Results

Project: Spatial Query Semantic Enrichment

With Manesh Pillai





Techniques: Semantic Networks, Multiplex Networks, Context

Thanks to...

- Zhiyuan Chen
- Ahmed Aleroud
- Foteini Argiropoulos
- Sabrina Moumtaz
- Leonard Traeger
- Sai Pallaprolu
- Manesh Pillai
- Antonios Xenakis
- Andreas Behrend
- Several others...

Research outcome

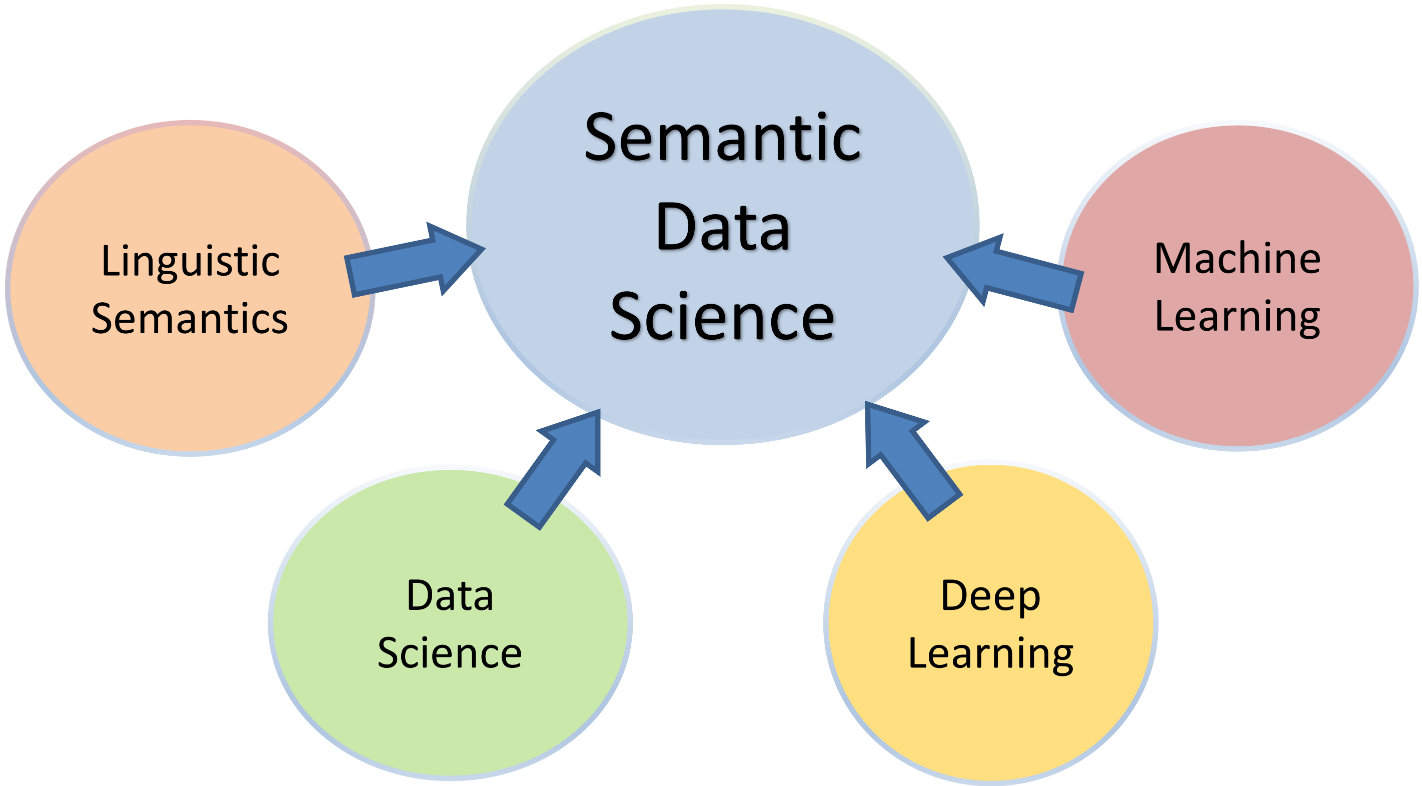
Applied research

- Beyond publications
- Commercializable research (startups, etc.)
- Entrepreneurship
  - Intrusion detection: Cyves, LLC (with A. Aleroud)
  - Software security: ML4Cyber, LLC (with F. Argiropoulos)
  - Dataset anonymization: Anonitech, LLC (with Z. Chen, A. Aleroud)
- Sponsors:

Possible Collaborations

- Faculty & Research level
  - Research publications
  - Proposals for funding
- Institution level
  - Investigate feasibility for student exchanges (study abroad program)
  - Investigate possibility of Memorandum of Understanding (MoU)

Research Collaboration



Goal: Instill more 'common sense' to Automated Systems in:

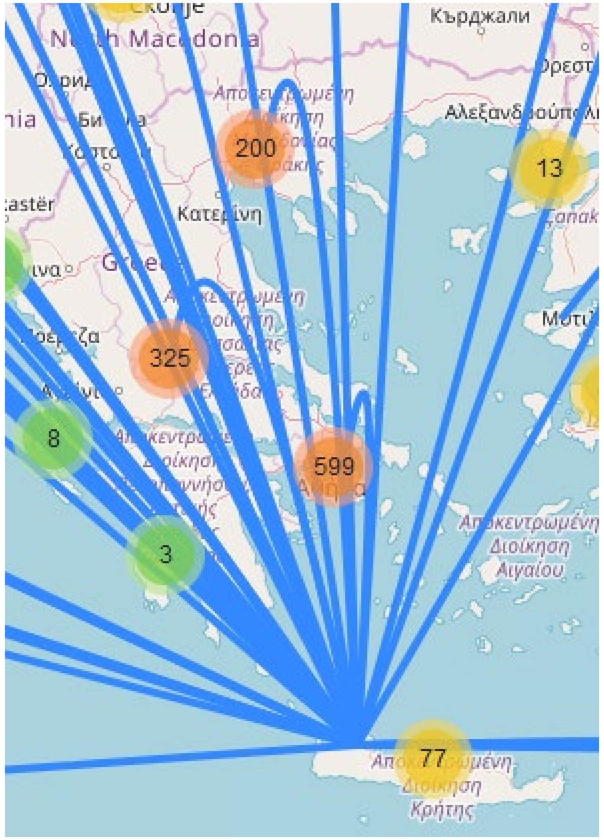
Systems Integration, Cybersecurity, Environmental Sciences, Software Engineering, etc.

EU Horizon

Example: Cluster 5 - Climate, Energy and Mobility

Upcoming deadlines in 2024

- Orchestration of heterogeneous actors in mixed traffic within the CCAM ecosystem (CCAM Partnership) HORIZON-CL5-2024-D6-01-03
- AI for advanced and collective perception and decision making for CCAM applications (CCAM Partnership) HORIZON-CL5-2024-D6-01-04
- Optimising multimodal network and traffic management, harnessing data from infrastructures, mobility of passengers and freight transport HORIZON-CL5-2024-D6-01-06



TUC Horizon collaborations

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Collaboration between Institutions

- Investigate feasibility for student exchanges
  - Undergraduate level (study abroad program)
  - Graduate level (specific courses)
- Investigate possibility of Memorandum of Understanding (MoU) across institutions

Thank You!

Email: georgek@umbc.edu
gkarabatis@tuc.gr