Event Log Encoding: assessing the state of the art

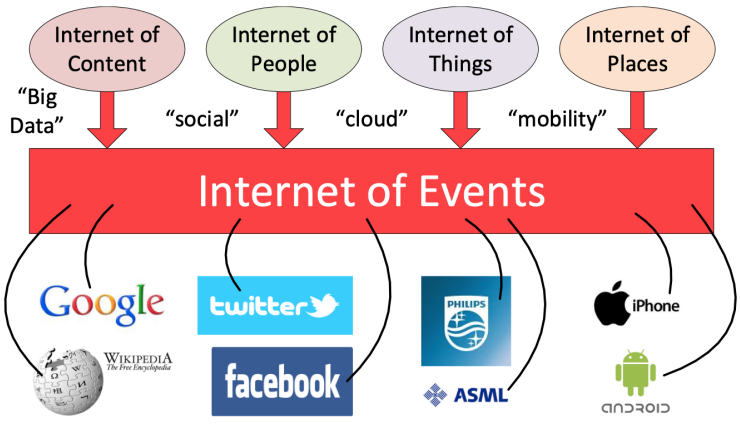
Sylvio Barbon Jr., **Paolo Ceravolo**
Ernesto Damiani, Gabriel Marques Tavares

Londrina State University (UEL)
Università degli Studi di Milano (UNIMI)
Khalifa University (KUST)

Introduction

Internet of Events

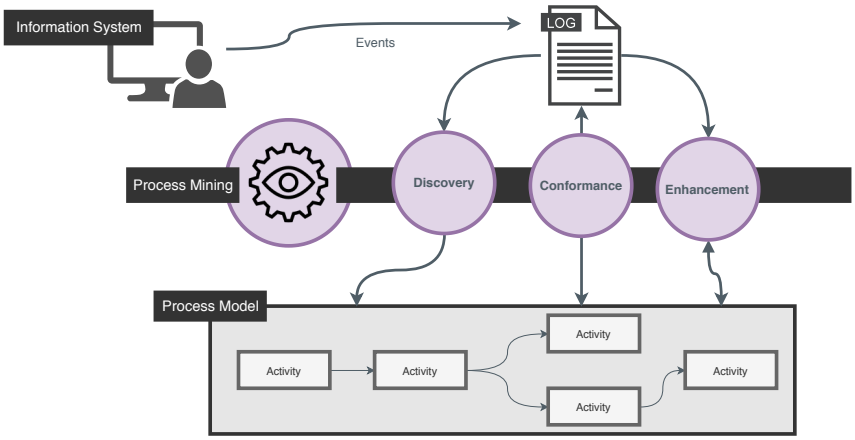
"All event generated by digital technologies. People, contents, sensors, places" (Van Der Aalst, W. ,2014).



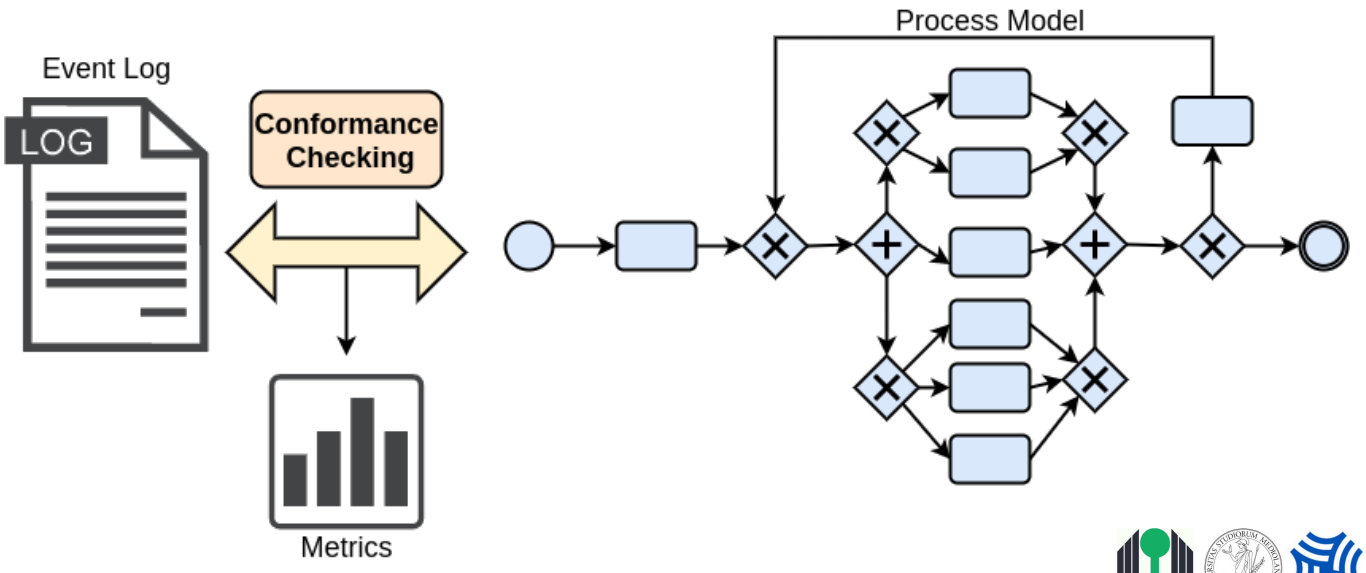
Introduction

Process Mining

“The idea of process mining is to discover, monitor and improve real processes by extracting knowledge from event logs readily available in today’s systems” (Van Der Aalst, W. ,2011).



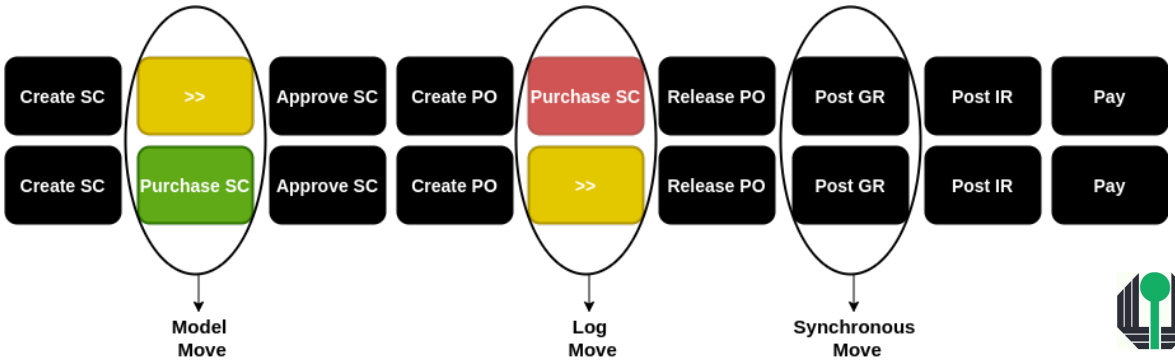
Traditional Feature Encoding - Conformance Checking



Traditional Feature Encoding

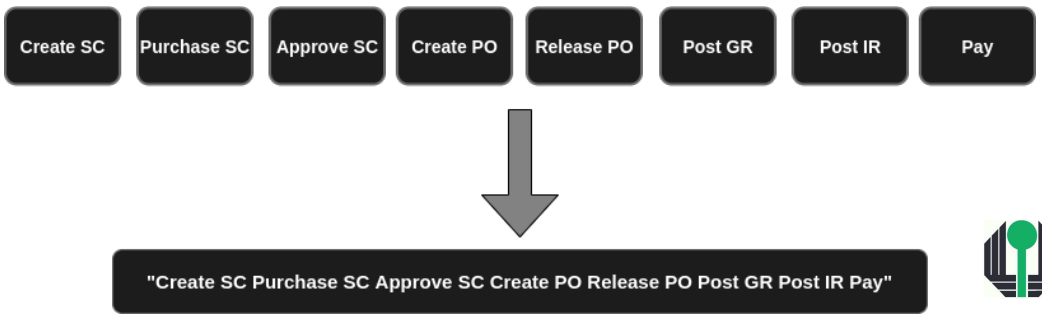
Token-replay: matches a trace to a process model and produces a fitness value along with counting tokens. Features: *trace_is_fit*, *trace_fitness*, *consumed_tokens*, *remaining_tokens*, *produced_tokens*.

Alignment: relates a trace to valid execution sequences in the model computing how synchronous they are. Features: *cost*, *visited_states*, *queued_states*, *traversed_arcs*, *fitness*.



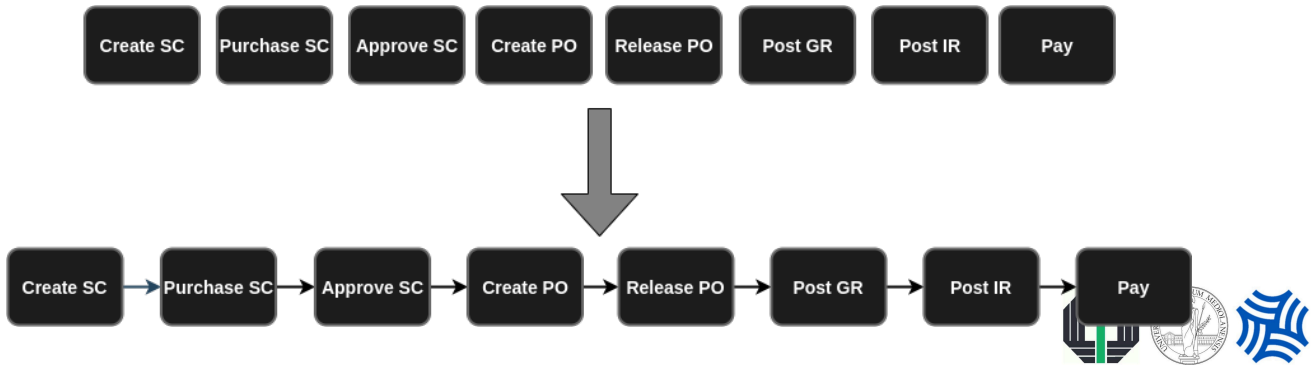
Word embeddings

- Process data contains several layers
- Encoding techniques can provide common grounds for analysis
- Activities describe the action being performed (i.e. words)
- Grounded natural language processing
- Word embeddings capture context given a neighborhood
- Traces and activities are represented as sentences and words



Graph embeddings

- Modeling entities links and long-term relations
- Representation format matches PM necessities
- Can capture additional attributes attached to nodes
- Encode context given the neighbors and neighborhoods



Encoding Methods

encoding	family	feature	type	range
trace replay	PM-based	trace is fit	Boolean	{True, False}
		fitness	Numeric	[0, 1]
		consumed tokens	Integer	[0, ∞[
		remaining tokens	Integer	[0, ∞[
		produced tokens	Integer	[0, ∞[
alignment	PM-based	cost	Integer	[0, ∞[
		visited states	Integer	[0, ∞[
		queued states	Integer	[0, ∞[
		traversed arcs	Integer	[0, ∞[
		fitness	Numeric	[0, 1]
word2vec	Text-based	n-dimensions*	Numeric	]-∞, ∞[
fasttext	Text-based	n-dimensions*	Numeric	]-∞, ∞[
count2vec	Text-based	n-dimensions**	Integer	[0, ∞[
one-hot	Text-based	n-dimensions**	Integer	{0, 1}
tfidf	Text-based	n-dimensions**	Numeric	[0, 1]
hash2vec	Text-based	n-dimensions*	Numeric	[-1, 1]
node2vec	Graph-based	n-dimensions*	Numeric	]-∞, ∞[
edge2vec	Graph-based	n-dimensions*	Numeric	]-∞, ∞[

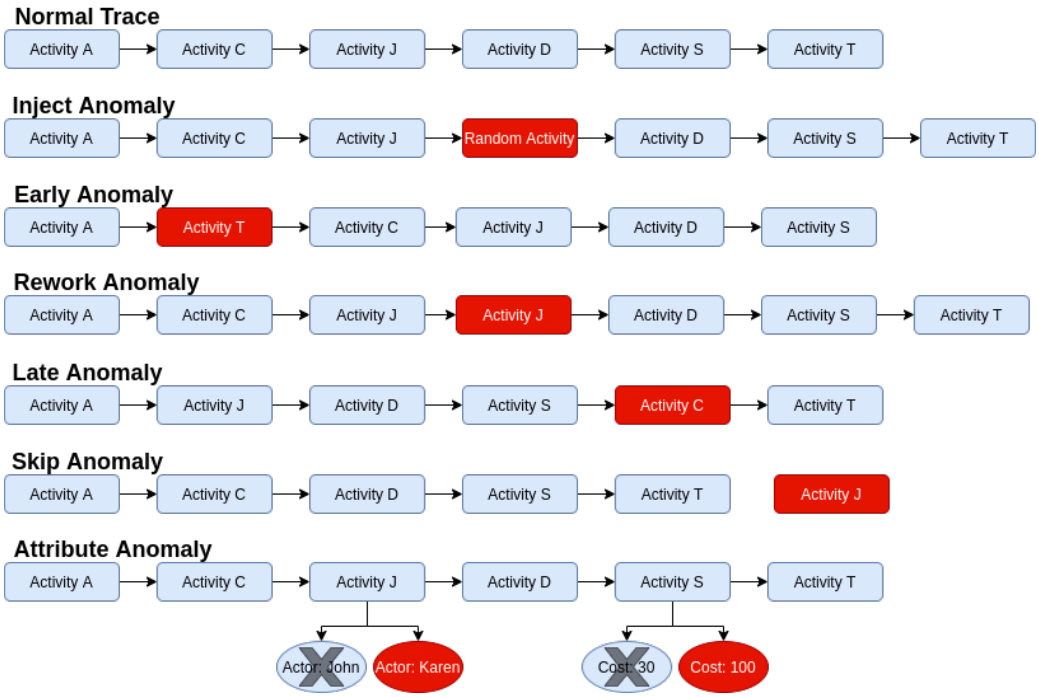
* encoding vector size is determined by a parameter
** encoding vector size is determined by the vocabulary size

Experiments - Event Logs

- 20 event logs with 1k cases were generated for each scenario (total of 100 logs)
- Increasing level of complexity

log name	#gateways	#events	trace size	#activities
scenario 1	8	10k-11k	9-13	22
scenario 2	12	26k	26-30	41
scenario 3	22	43k-44k	42-50	64
scenario 4	30	11k-13k	3-30	83
scenario 5	34	18k-19k	4-37	103

Experiments - Anomaly Types



Experimental Setup

- Goals
  - Assess encoding quality and representativeness
  - Compare encodings in a binary classification task
- Evaluation
  - Maximum Fisher's Discriminant Ratio (F1)
  - Volume of Overlapping Region (F2)
  - The Average Number of Principal Component Analysis (PCA) dimensions compared to the original dimensions (T4)
  - Classification accuracy
  - Time spent in the classification task
- Encoding techniques using standard hyperparameters

Results - Accuracy performance

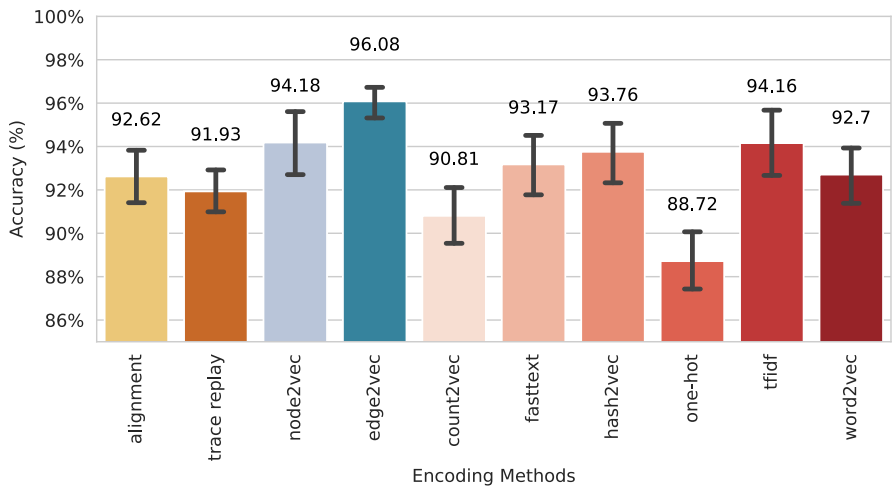


Figure: Average accuracy obtained using all encoding methods across binary problems related to anomaly detection (early, insert, late, rework and skip) affected by four different levels of compromised samples (5%, 10%, 15% and 20%).

Results - Classification time consumption

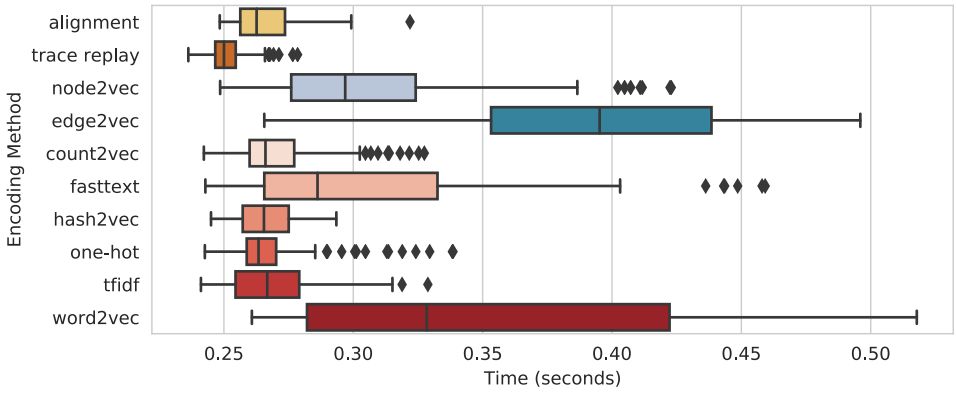


Figure: Average time required in the classification task for each encoding across all scenarios.

Results - Encoding representativeness (F1)

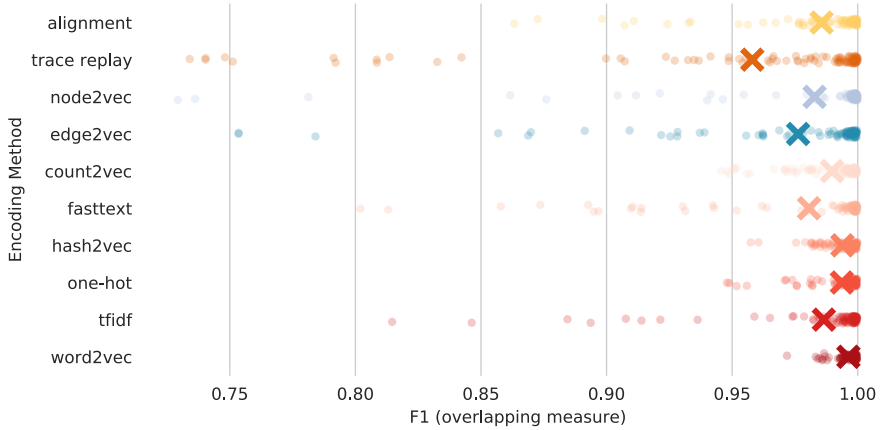


Figure: Maximum Fisher's Discriminant Ratio (F1) values, for the studied scenarios. F1 measures the overlap in classes of the best-disjunct feature.

Results - Encoding representativeness (F2)

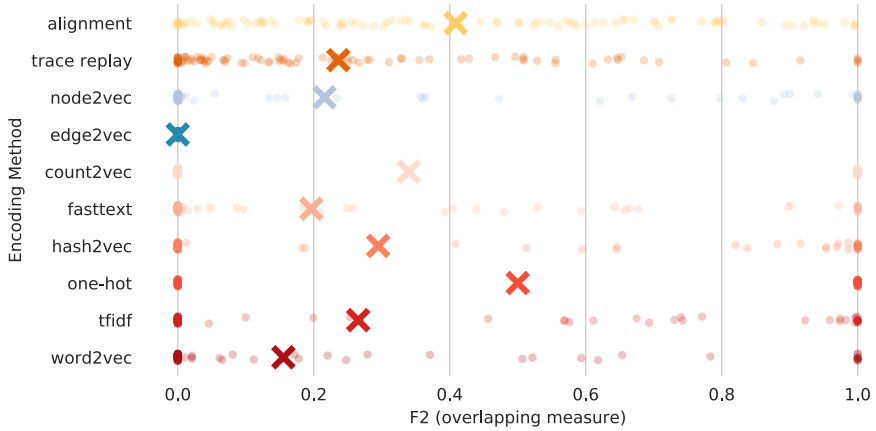


Figure: Volume of Overlapping Region (F2) of the feature values distributions within the problem classes. Low F2 values implies low overlapping.

Results - Encoding representativeness (T4)

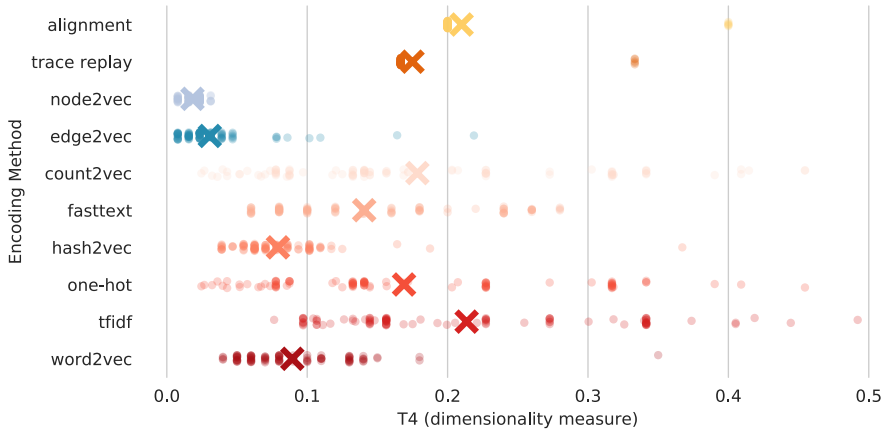


Figure: Ratio of the PCA dimension to the original dimension (T4) of all encoding methods. High T4 means more original features are relevant.

Results - Encoding ranking

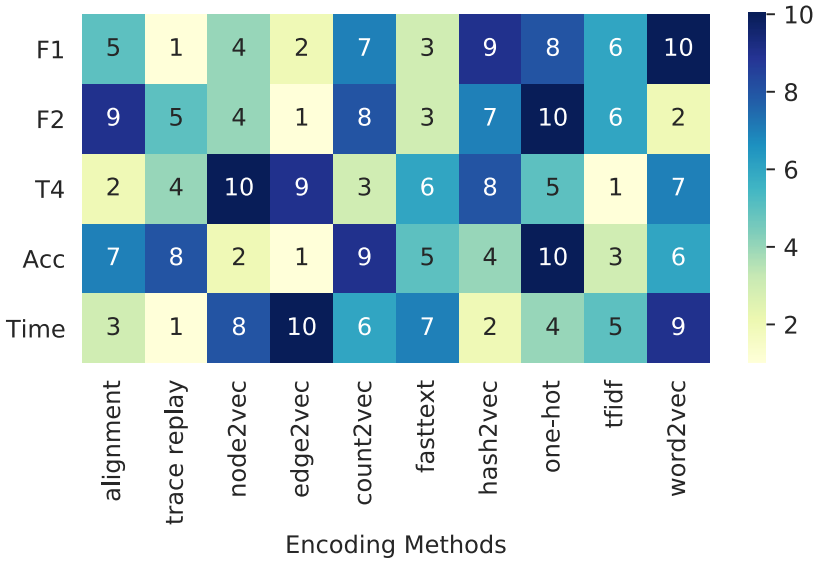


Figure: Ranking of each metric across all encoding methods. The rank ranges from 1 to 10, where the best-ranked position is 1 and the worst-ranked is 10.

Conclusion

- Process Mining requires good quality encoding methods
- Assessment of ten encodings: traditional PM, word embeddings and graph embeddings
- Classification task for anomaly detection
- A good encoding method can improve a wide range of algorithms without the need of tuning
- Future Work:
  - Increase scenarios to explore more complex behavior
  - Evaluate encoding resource consumption
  - Expand encoding families and methods

Conclusion

Open Challenges

- Encoding multiple perspectives (e.g., time and resource)
- Multi-class problems
- Dealing with incomplete traces
- Improving quality assessment
- Organic process encoder
- Variant analysis
- Anomaly detection