KU LEUVEN

Conformance checking using activity and trace embeddings

Jari Peeperkorn, Seppe vanden Broucke, Jochen De Weerdt

KU Leuven, LIRIS research group

jari.peeperkorn@kuleuven.be

Process mining – Conformance checking

- How well does a certain model describe a certain log?

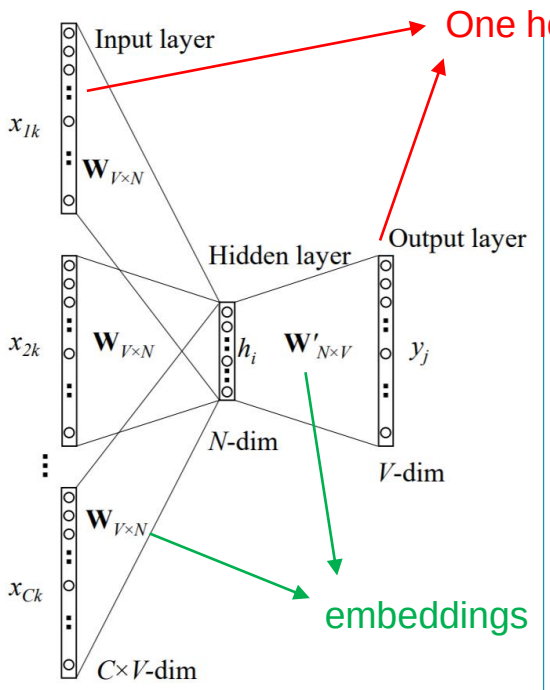
- Novel technique
  - Fully data driven
  - Inspired by Natural Language Processing
  - Meaningful embeddings
  - Three different variants

Vector embeddings

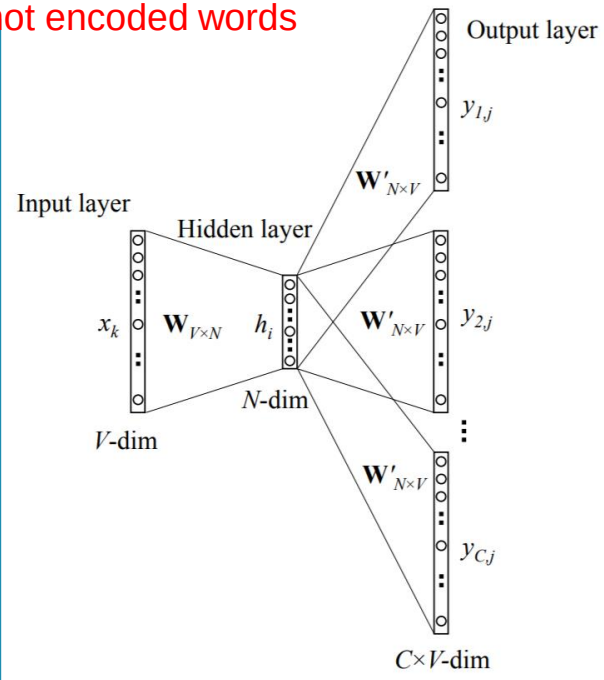
Vector embeddings

- Embeddings: map categorical variables to vectors of real numbers
  - Learned (often self-supervised)
  - Meaningful
  - Low dimensional
  - ≠ One-hot vectors

- Word2vec, Mikolov et al. (2013)
  - Similar words get similar vectors
  - Big amount of text → neural network
  - CBOW – Skip-gram
  - Doc2vec, Le and Mikolov (2014)



CBOW

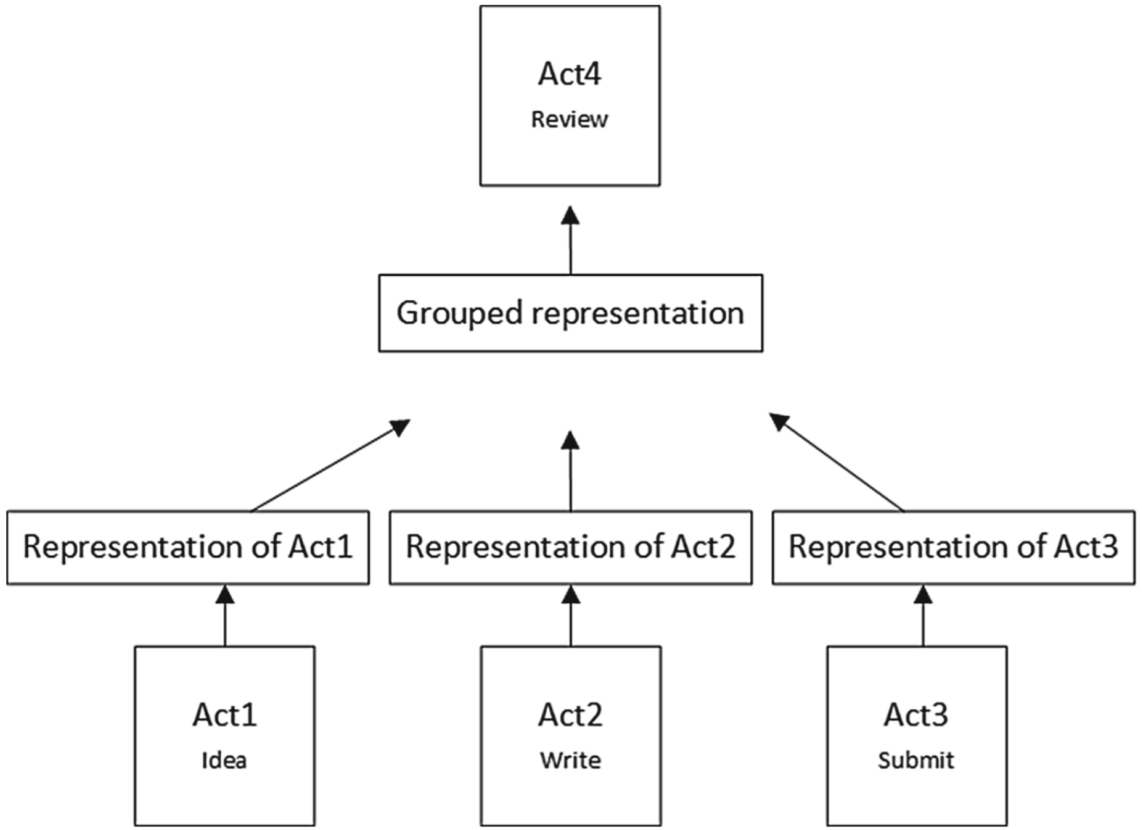


Skip-gram

[5] Rong (2014)

Act2vec

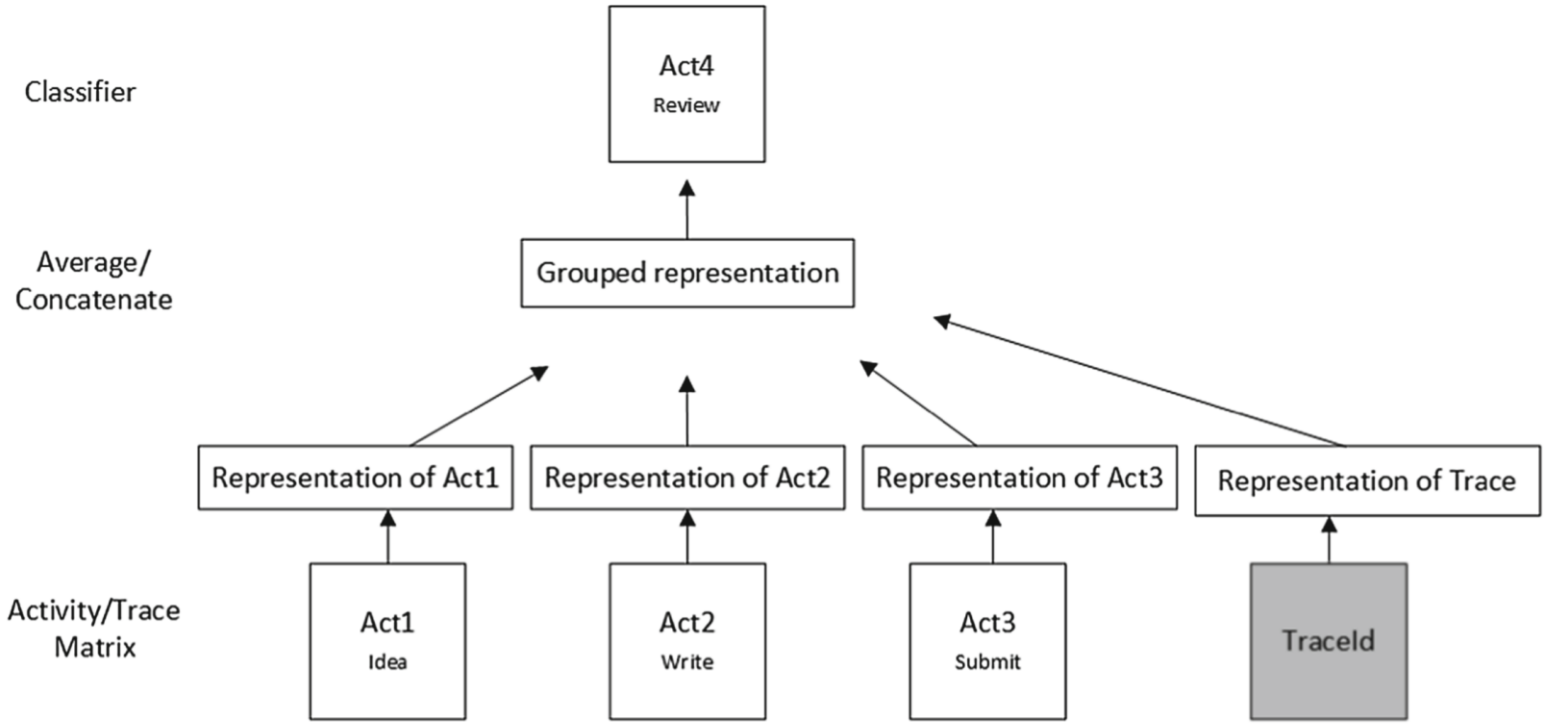
- Like word2vec
  - Words = activities



[1] De Koninck et al. (2018)

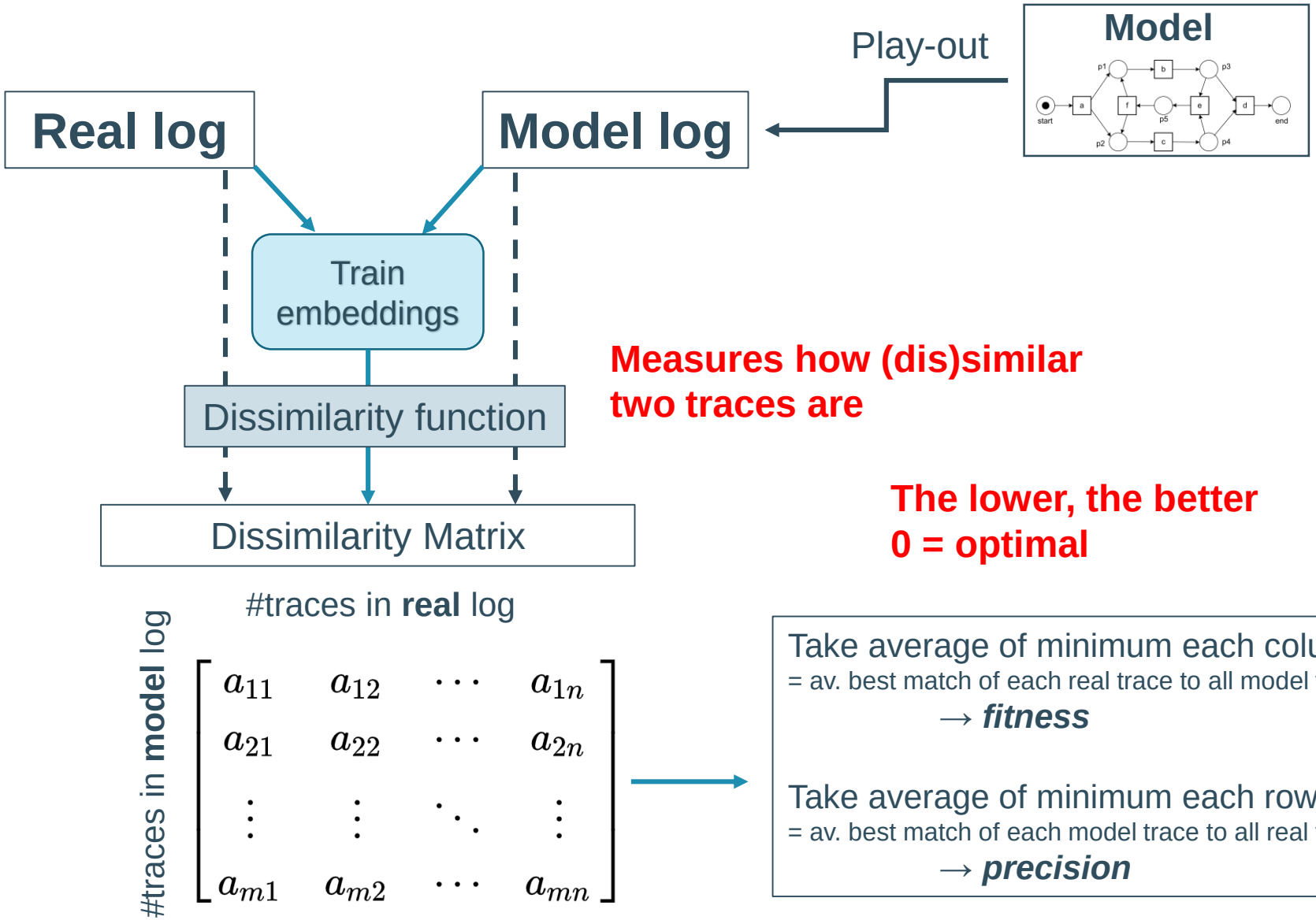
Trace2vec

- Like doc2vec
  - Words = activities
  - Docs = traces



[1] De Koninck et al. (2018)

Conformance checking using activity embeddings


Real log
Model log
Play-out
Model
Train embeddings
Dissimilarity function
Dissimilarity Matrix
Measures how (dis)similar two traces are
The lower, the better
0 = optimal
#traces in real log
#traces in model log
$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$
Take average of minimum each column
= av. best match of each real trace to all model traces
→ fitness
Take average of minimum each row
= av. best match of each model trace to all real traces
→ precision

Conformance checking using activity embeddings

- Activity embeddings trained on both logs together (using act2vec)

- Word Mover's Distance (WMD)
  - Earth Mover Distance (EMD)
  - Minimum transportation cost (total Euclidian distance) to transport every word (activity) from sentence (trace) 1 to sentence (trace) 2
  - Leemans et al. (2019): "Earth Movers' Stochastic Conformance Checking"

- Iterative Constrained Transfers (ICT)
  - More efficient
  - Lower bound WMD

Conformance checking using trace embeddings

Conformance checking using trace embeddings

- Activity **+ trace** embeddings trained both logs together (using trace2vec)
  - Use only distinct traces such that equal traces get only 1 embedding

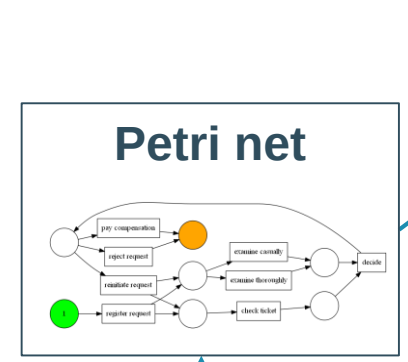
- Use cosine distance between trace embeddings

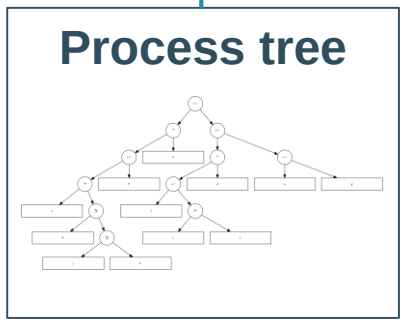
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

- Faster

Experimental Evaluation

Experiment 1: Noise





“Ground truth” log

Add noise

“Noisy” log

Train embeddings

Dissimilarity function

Dissimilarity Matrix

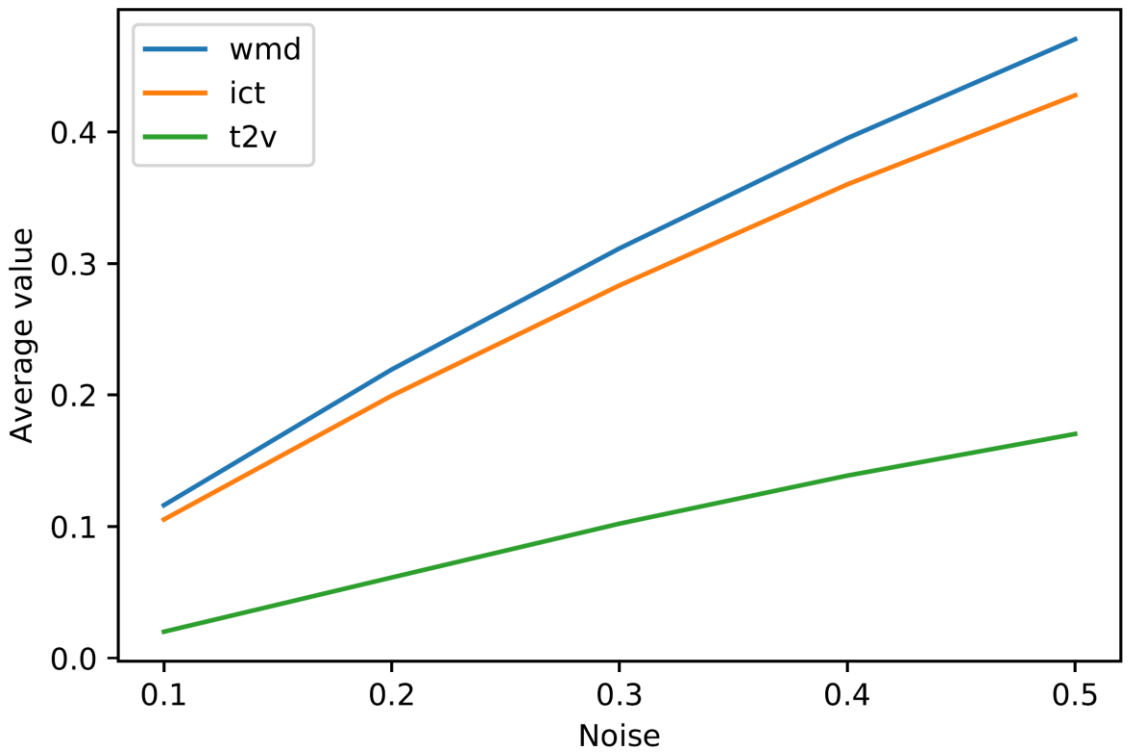
#traces in **“ground truth”** log

#traces in **”noisy”** log

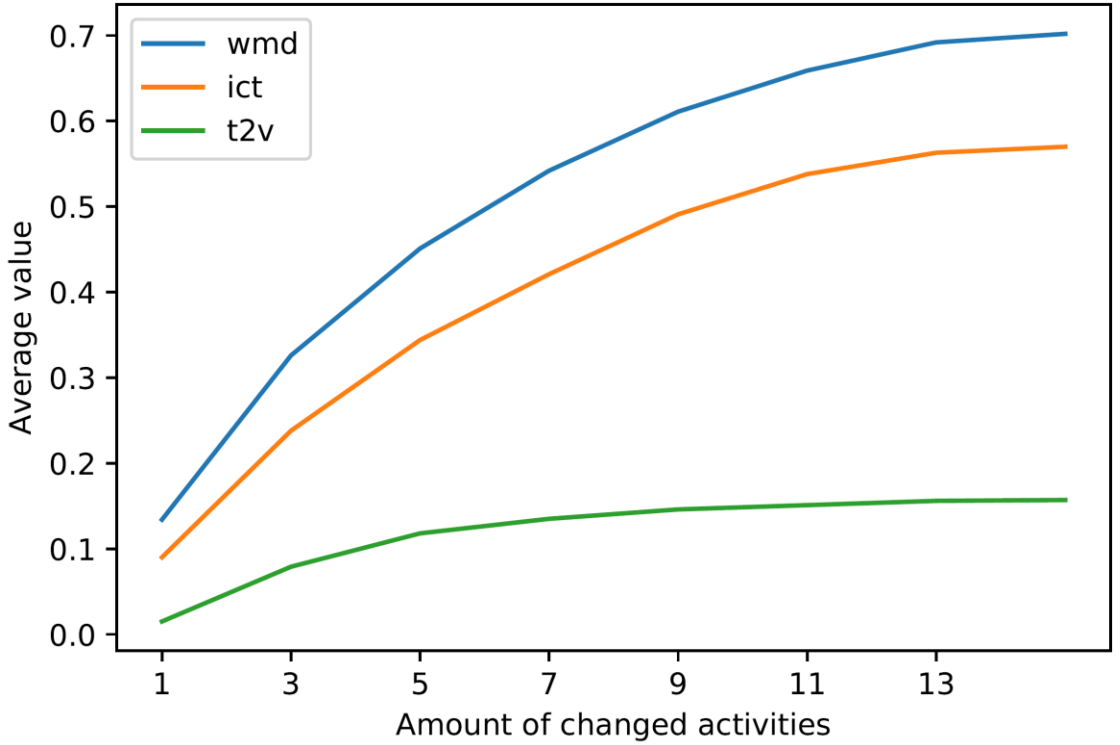
$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Take average of minimum each column
→ *“fitness”*

Take average of minimum each row
→ *“precision”*



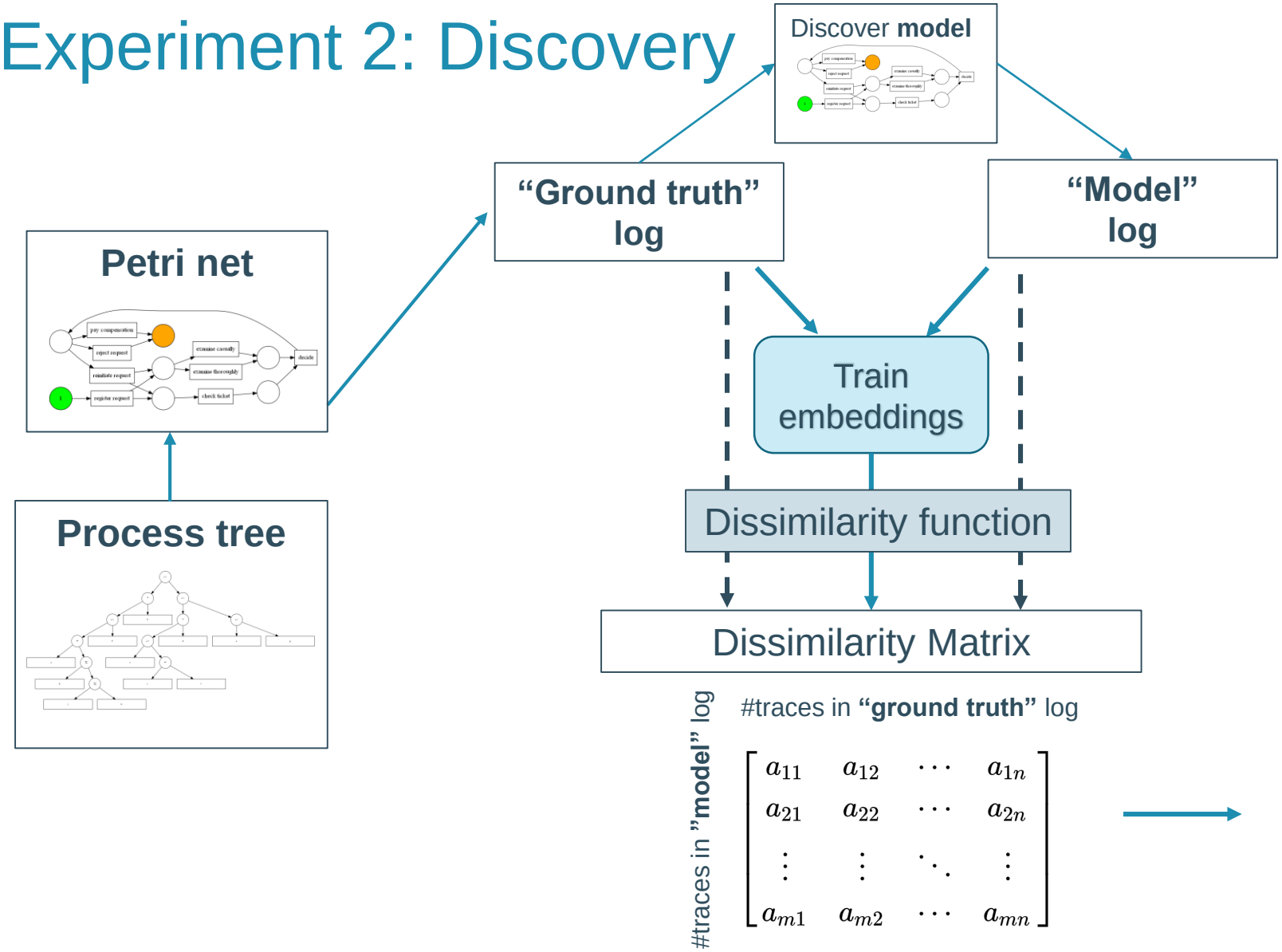
Adding a little noise to different % of traces



Adding different amounts noise to 40% of traces

- Techniques are capable to correctly assess when logs start to differ from each other
- Trace2vec based technique less sensitive to which extend two traces are different

Experiment 2: Discovery



Take average of minimum each column
→ *fitness*

Take average of minimum each row
→ *precision*

- Different “ground truth” logs and different discovery techniques
- Compared to different techniques from literature
- When other techniques agree on perfect fitness/precision
  - Our new techniques usually agree
- When other techniques agrees on which discovered model is best
  - Our new techniques usually agree

Conclusion

- A new conformance checking technique based on representation learning
  - Fully data driven
  - Alternative to classical approaches
  - Neural network embeddings provide an abstraction

- Three variants intrinsically capable of detecting fitness and precision issues:
  - Using act2vec and WMD
  - Using act2vec and ICT
  - Using trace2vec

Future work

- Standardize (value between 0 and 1)
- More explicitly take order into account
  - Ngram embeddings
  - Other distance function
  - …
- Conformance beyond fitness and precision
- Testing on real life log
- Test embeddings in a different context

References

[1] Pieter De Koninck, Seppe vanden Broucke, and Jochen De Weerdt. *act2vec, trace2vec,log2vec, and model2vec: Representation learning for business processes*. In Mathias Weske,Marco Montali, Ingo Weber, and Jan vom Brocke, editors,Business Process Management,pages 305–321, Cham, 2018. Springer International Publishing

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations* in Vector Space.arXiv e-prints, page arXiv:1301.3781, Jan 2013.

[3] Quoc V. Le and Tomas Mikolov. *Distributed Representations of Sentences and Documents*.arXiv e-prints, page arXiv:1405.4053, May 2014.

[4] Leemans, S., Syring, A., Aalst, W.: Earth Movers' Stochastic Conformance Check-ing, pp. 127–143 (07 2019)

[5] Xin Rong, *word2vec Parameter Learning Explained*, page arXiv: 1411.2738, 2014

Table 2. Table showing the run times of the different variations of the algorithm.

Log Size	Dictionary size	wmd	ict	t2v
100	10	1s	1s	1s
	20	2s	2s	1s
	30	3s	2s	1s
500	10	20s	26s	12s
	20	46s	39s	12s
	30	1m15s	45s	12s
1000	10	1m14s	1m43s	42s
	20	2m57s	2m30s	43s
	30	4m50s	2m51s	43s
5000	10	30m38s	37m2s	15m20s
	20	1h10m4s	55m23s	15m24s
	30	1h57m12s	1h3m44s	15m20s