KU LEUVEN

Research Foundation Flanders
Opening new horizons

CAiSE '26

JOURNAL-FIRST · CAiSE

Model-driven Stochastic Trace Clustering

Jari Peeperkorn, Johannes De Smedt, Jochen De Weerdt

LIRIS (Research Center for Information Systems Engineering), KU Leuven

Full paper

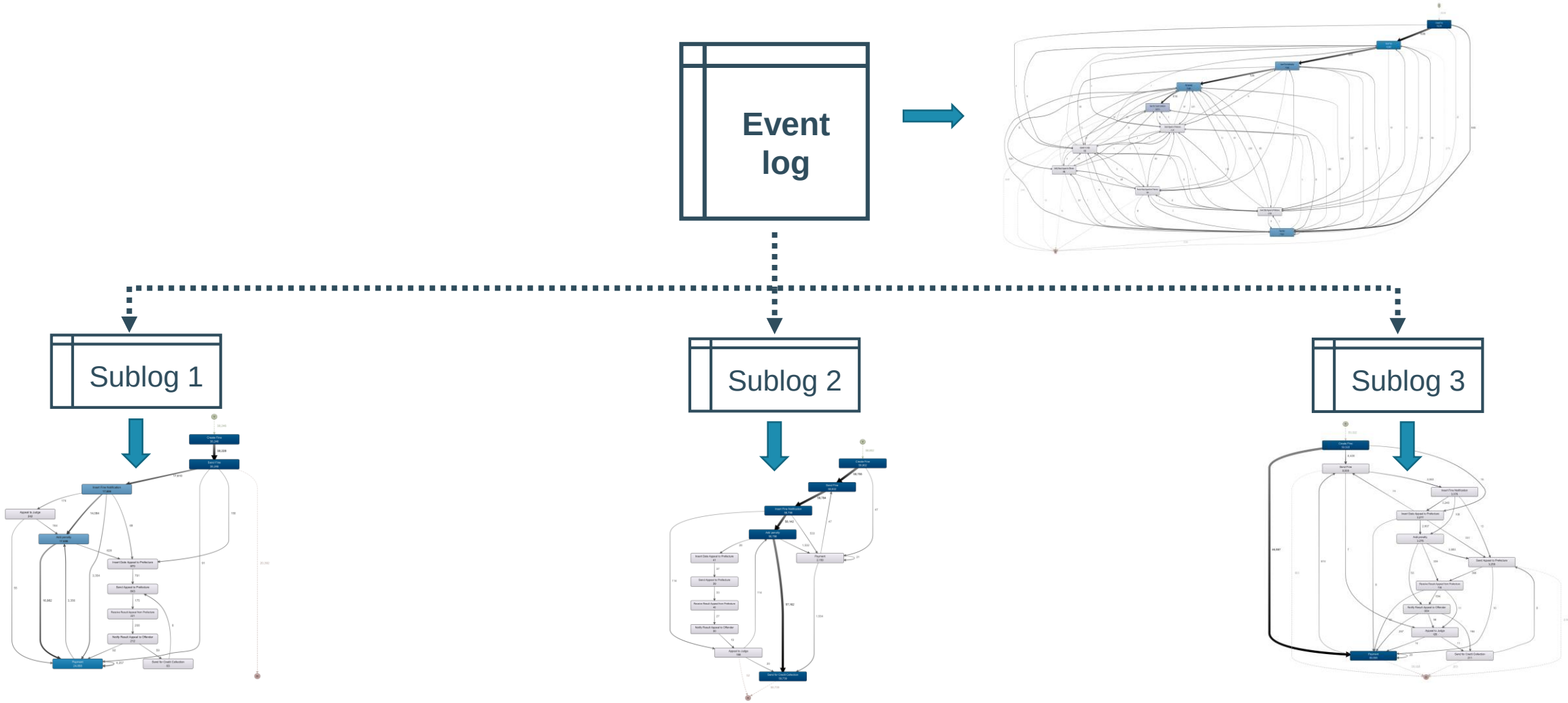
Information Systems, Volume 139, 2026

Research Center for Information Systems Engineering (LIRIS), Faculty of Economics and Business

Context

Full paper

- **Trace clustering goal:** more readable process models per cluster



Source: Disco

Context

Full paper

- **Trace clustering goal:** more readable process models per cluster
- Novel **model-driven** trace clustering using **stochastic conformance**
- Stochastic process mining = account for behavior **likelihood**

How?

Full paper

- The algorithm directly optimizes for the representativeness of a cluster's **Directly-Follows Graph (DFG)**

- Trace assignment guided using (adapted) **Entropic Relevance (ER)** [1]
  - Stochastic conformance checking (information compression costs)
  - Per trace: $-\log_2(p)$ $(+ \text{other terms})$
  - Higher p = lower ER = better fit

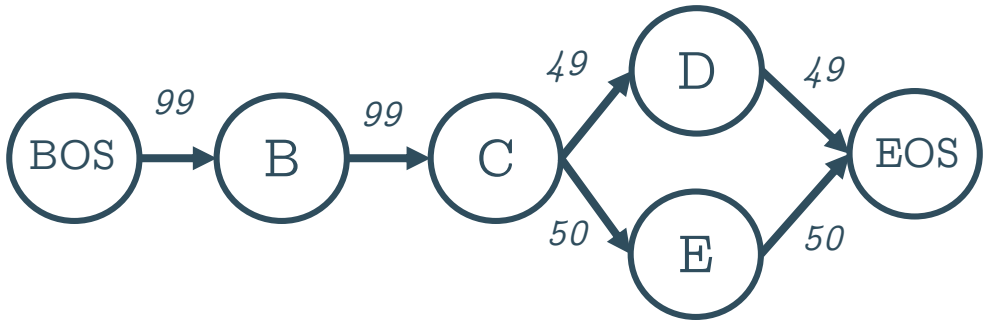
(stochastic) replay probability
= probability that this model (system) would produce this trace

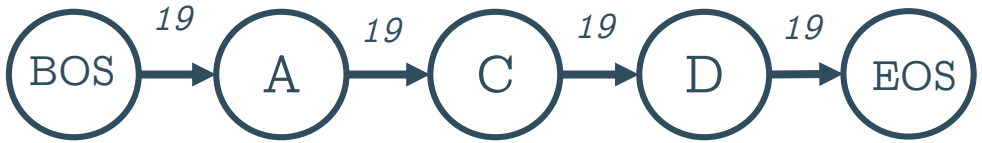
[1] Alkhammash, H., Polyvyanyy, A., Moffat, A., & García-Bañuelos, L. (2022). Entropic relevance: A mechanism for measuring stochastic process models discovered from event data. Information Systems, 107, 101922. https://doi.org/10.1016/j.is.2021.101922

Core idea

Full paper

- For each cluster we keep the current DFG
- Iterate though all control-flow variants in the log



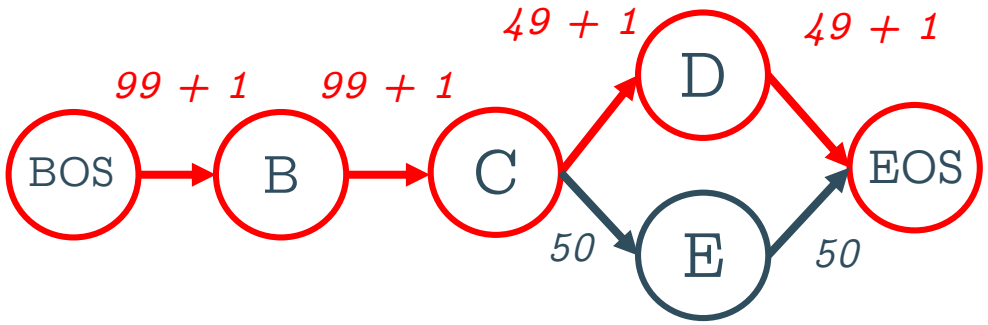


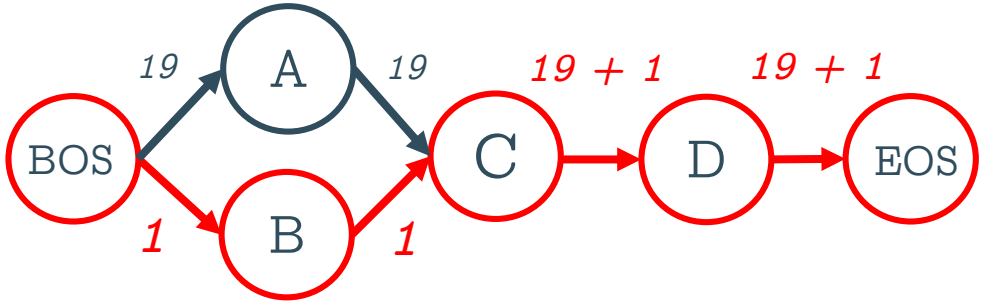
Core idea

Full paper

- The variant is temporarily added to all clusters
- Update DFGs

New variant: <BOS, B, C, D, EOS>
frequency 1



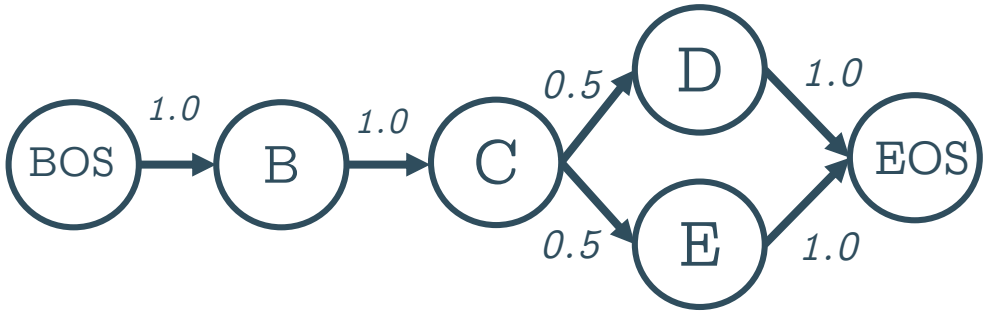


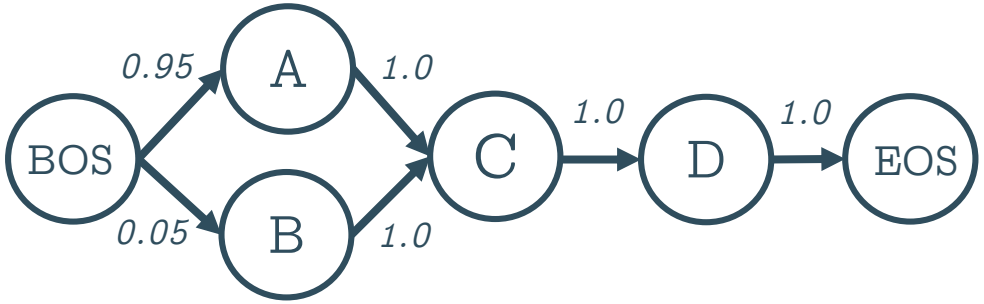
Core idea

Full paper

- Convert DFG to a “stochastic model”
  - Divide each DF (edge) by source activity (node) count

New variant: <BOS, B, C, D, EOS>
frequency 1





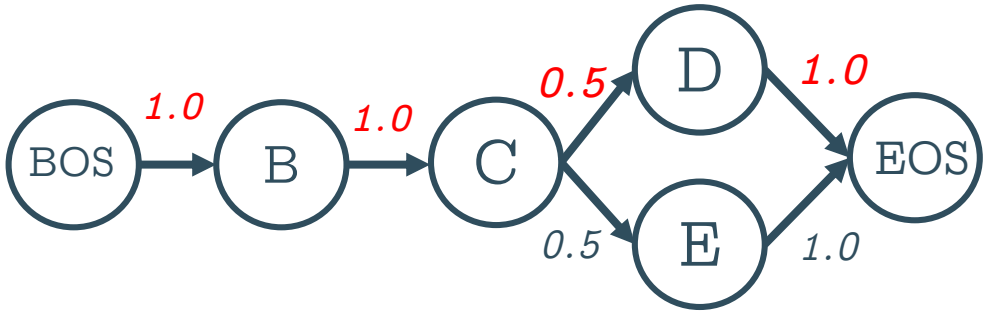
Core idea

Full paper

- Calculate “playout probability” by multiplying DF probabilities

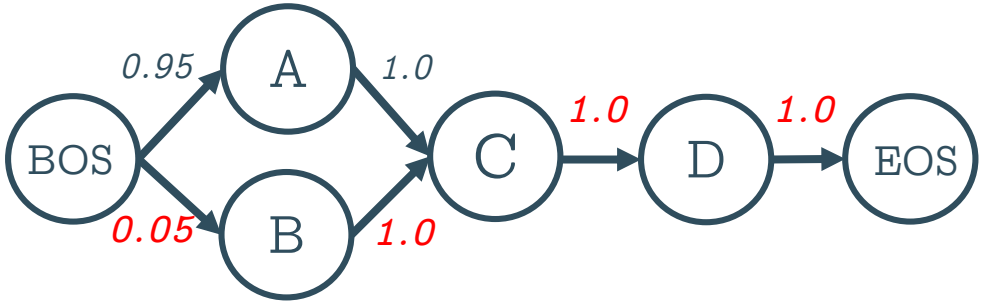
(and calculate ER)

New variant: <BOS, B, C, D, EOS>
frequency 1



$1.0 * 1.0 * 0.5 * 1.0 = 0.5$

$\rightarrow \mathrm{ER} = 1.0$



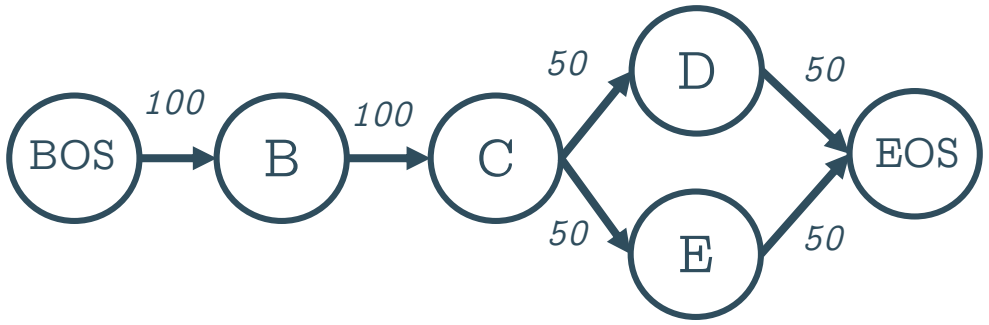
$0.05 * 1.0 * 1.0 * 1.0 = 0.05$

$\rightarrow \mathrm{ER} \approx 4.32$

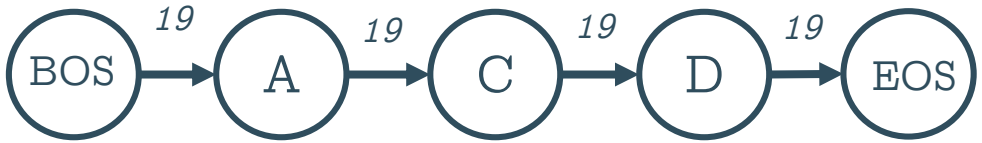
Core idea

Full paper

- Greedy assignment: add trace to optimal cluster, remove from other ones
- Paper contains more details & configurations!



→ This cluster is selected, trace added
→ Updated DFG saved



→ Not selected, updated DFG discarded

Experiment

Full paper

- Tested on 8 event logs
- Benchmarked against clustering algorithms from literature
  - Random, activity frequency-based, Trace2Vec [2], and ActiTraC [3]

- For each cluster we mine
  - Petri Net (inductive miner)
  - DFG

- We assessed these models and their conformance (with the traces)
  - DFG: ER + graph density & graph entropy
  - Petri Net: arc-degree simplicity + replay/alignment based fitness & precision

[2] De Koninck, P., vanden Broucke, S., De Weerdt, J. (2018). act2vec, trace2vec, log2vec, and model2vec: Representation Learning for Business Processes. BPM 2018. Lecture Notes in Computer Science(), vol 11080
[3] J. De Weerdt, S. vanden Broucke, J. Vanthienen and B. Baesens, "Active Trace Clustering for Improved Process Discovery," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 12, 2013

Main results

Full paper

Entropic clustering outperforms

- On specific metrics:
  - ER
  - Graph entropy & density
  - Petri Net simplicity

(statistically significant)

The trade-off

- Worse on purely structural Petri Net precision

(no statistical significance)

Conclusion

Full paper

- First stochastic, conformance-based model-driven trace clustering
- **Key result:** Entropic Clustering produces stochastically uniform clusters and the simplest downstream process models
- **Insight:** Optimal clusters differ depending on whether you optimize for structural vs. stochastic conformance
- **Industry ready!** Scales linearly with event log size and directly optimizes the DFG, the dominant model in commercial process mining tools

Full paper

https://www.sciencedirect.com/science/article/abs/pii/S0306437926000116

Contact me!

https://jaripeeperkorn.github.io/

Code & experiments

https://github.com/jaripeeperkorn/EntropicClustering

CAiSE '26