Tandem Inference: An Out-of-Core Streaming Algorithm for Very Large-Scale Relational Inference

Sriram Srinivasan*, Eriq Augustine*, & Lise Getoor

University of California, Santa Cruz

*Equal Contribution

Introduction

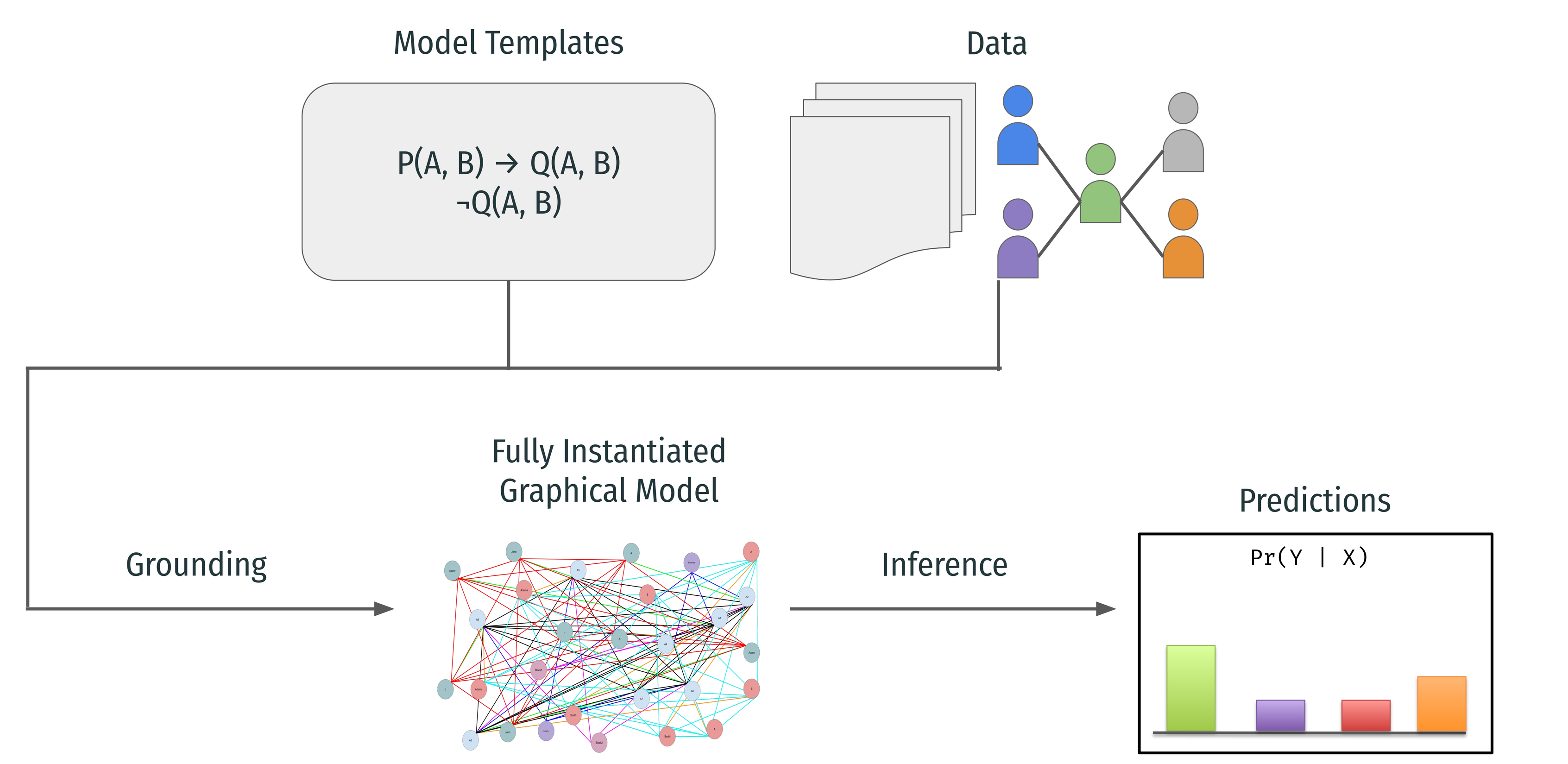
Statistical relational learning (SRL) frameworks allow users to create large, complex graphical models using a compact, rule-based representation. However, these models can quickly become prohibitively large and not fit into memory. In this work we address this issue by introducing a novel technique called **Tandem Inference (TI)**.

Contributions:

1. A general framework, TI, which uses streaming grounding and out-of-core streaming inference to perform memory efficient, large-scale inference in SRL frameworks.
2. Derived a stochastic gradient descent-based inference method (SGD).
3. An efficient streaming grounding architecture and SGD-based out-of-core inference system that runs faster than previous state-of-the-art systems.
4. Performed inference on very-large datasets (whose full models require more than 800 GB of memory) using just 10GB of memory.
5. An empirical evaluation on eight realworld datasets.

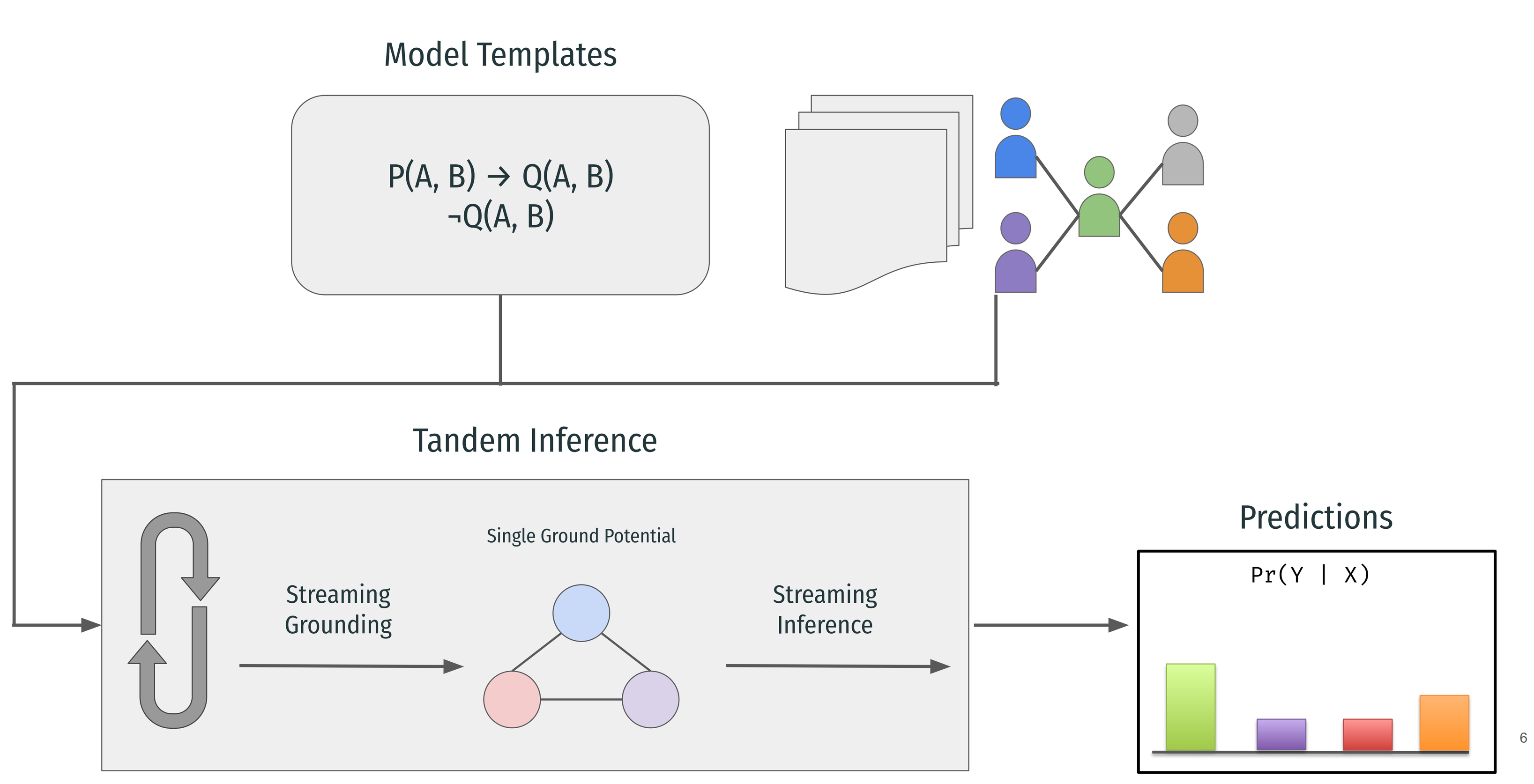
Traditional SRL

Typical SRL systems use this execution pipeline. The full graphical model in instantiated before inference begins. However, the full model may be too large to fit into memory.



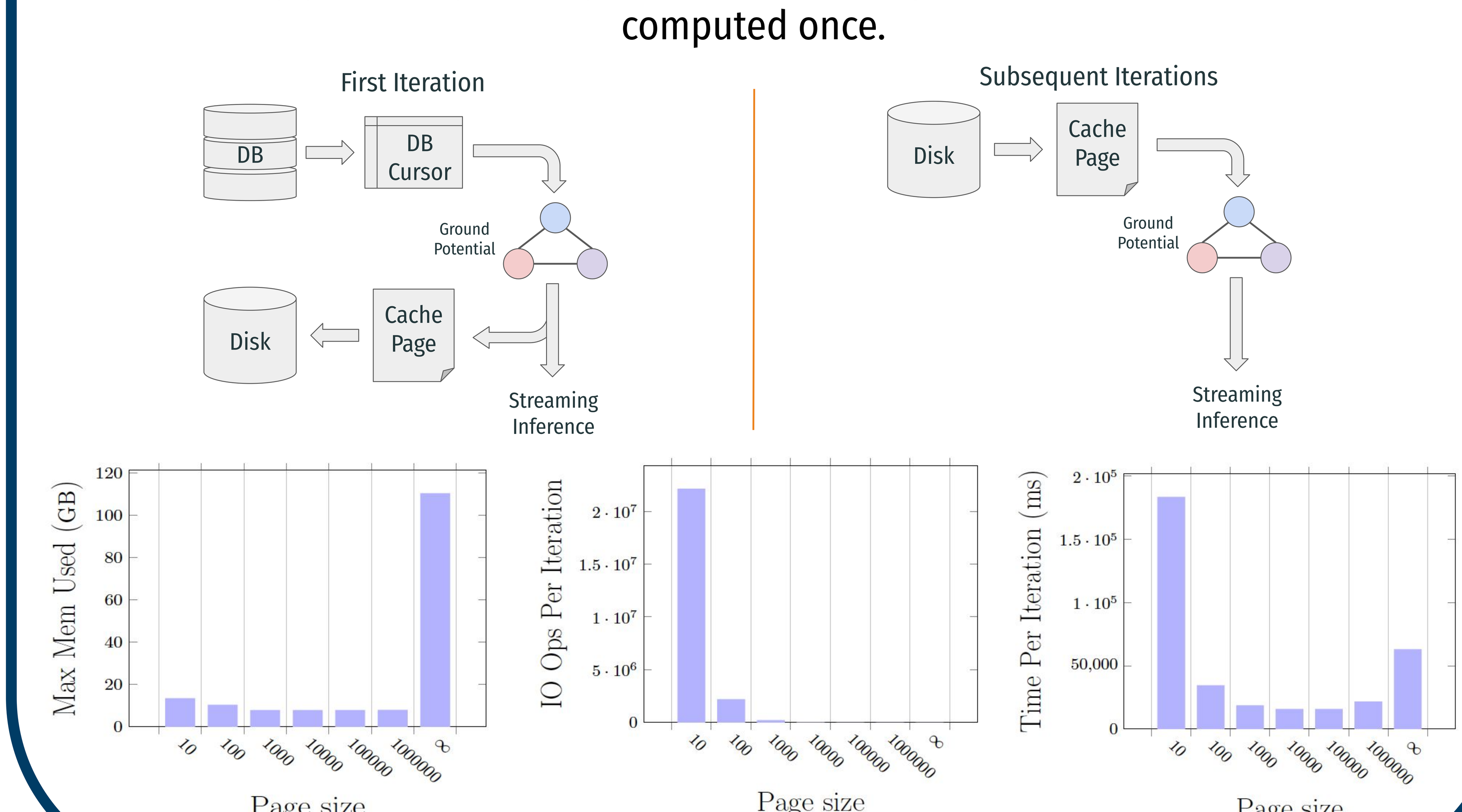
Tandem Inference

TI works by merging the two serial phases of grounding and inference into one iterative process with show iterations. Instead of grounding the full graphical model, only single potentials are ground at a time. Inference then works on single pieces at a time.



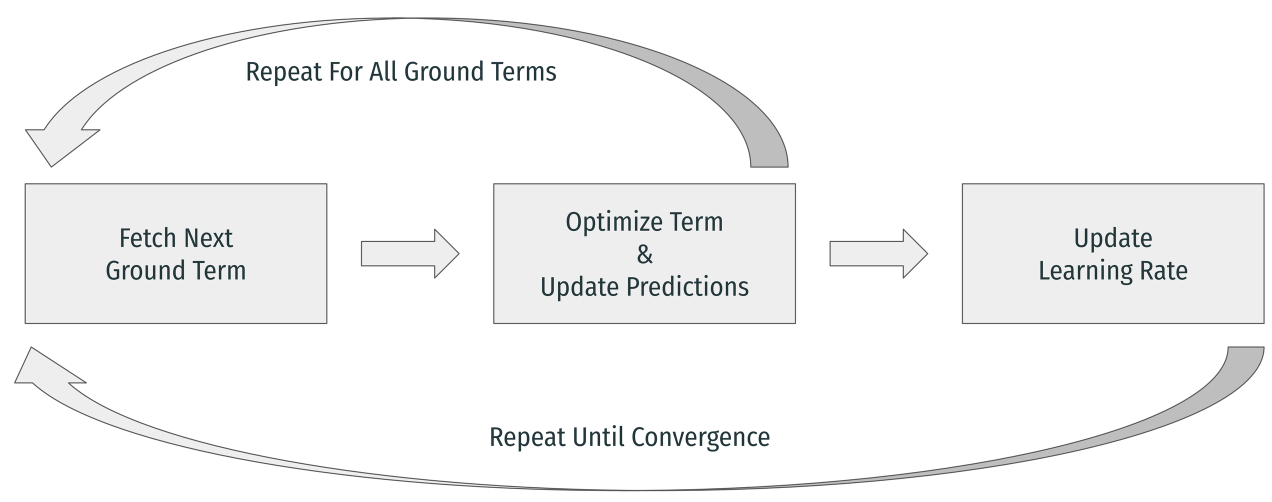
Streaming Grounding

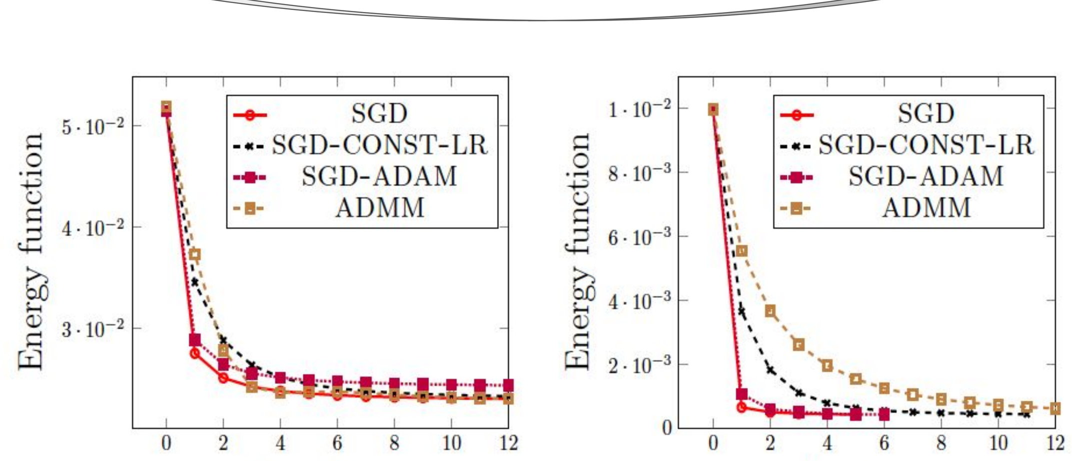
Streaming grounding leverages a disk cache so ground rules only need to be computed once.



Streaming Inference

Streaming inference works by optimizing just a single ground rule at a time until convergence is reached. The simplicity of SGD makes it ideal for TI.



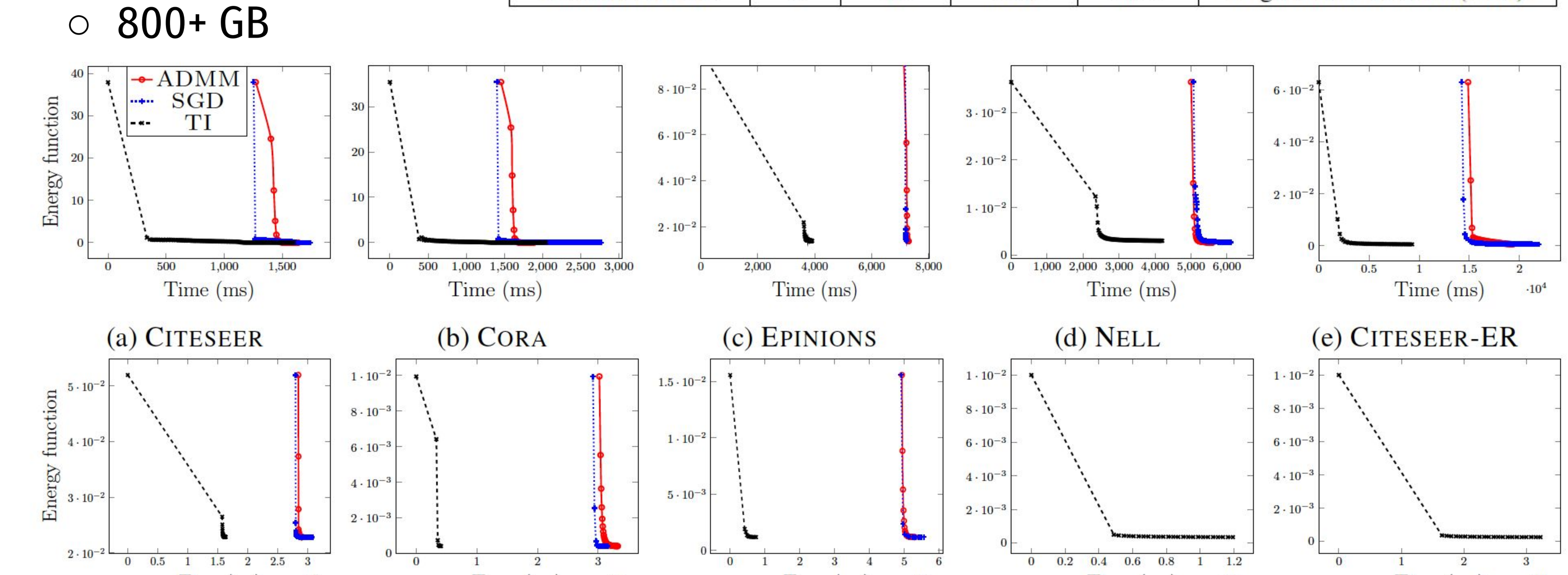


(a) LastFM dataset. (b) Jester dataset.

Empirical Evaluation

- 8 Datasets
  - 6 Realworld
- Largest to-date SRL dataset
  - 1.3 Billion Ground Rules
  - 800+ GB

Dataset	Rules	Ground Rules	Random Variables	Memory (GB)	Source
Citeseer	10	36K	10K	0.10	Bach et al. (2017)
Cora	10	41K	10K	0.11	Bach et al. (2017)
Epinions	20	14K	1K	0.12	Bach et al. (2017)
Nell	26	91K	24K	0.13	Pujara et al. (2013)
Citeseer-ER	9	541K	485K	0.24	Bhattacharya and Getoor (2007)
LastFM	22	1.4M	18K	0.45	Kouki et al. (2015)
Jester	7	1M	50K	0.49	Bach et al. (2017)
Jester-Full	8	110M	3.6M	110	Goldberg et al. (2001)
Friendship-500M	4	500M	4M	400+	Augustine and Getoor (2018)
Friendship-1B	4	1B	7.6M	800+	Augustine and Getoor (2018)



(a) Citeseer (b) Cora (c) Epinions (d) Nell (e) Citeseer-ER

(f) LastFM (g) Jester (h) Jester-Full (i) Friendship-500M (j) Friendship-1B