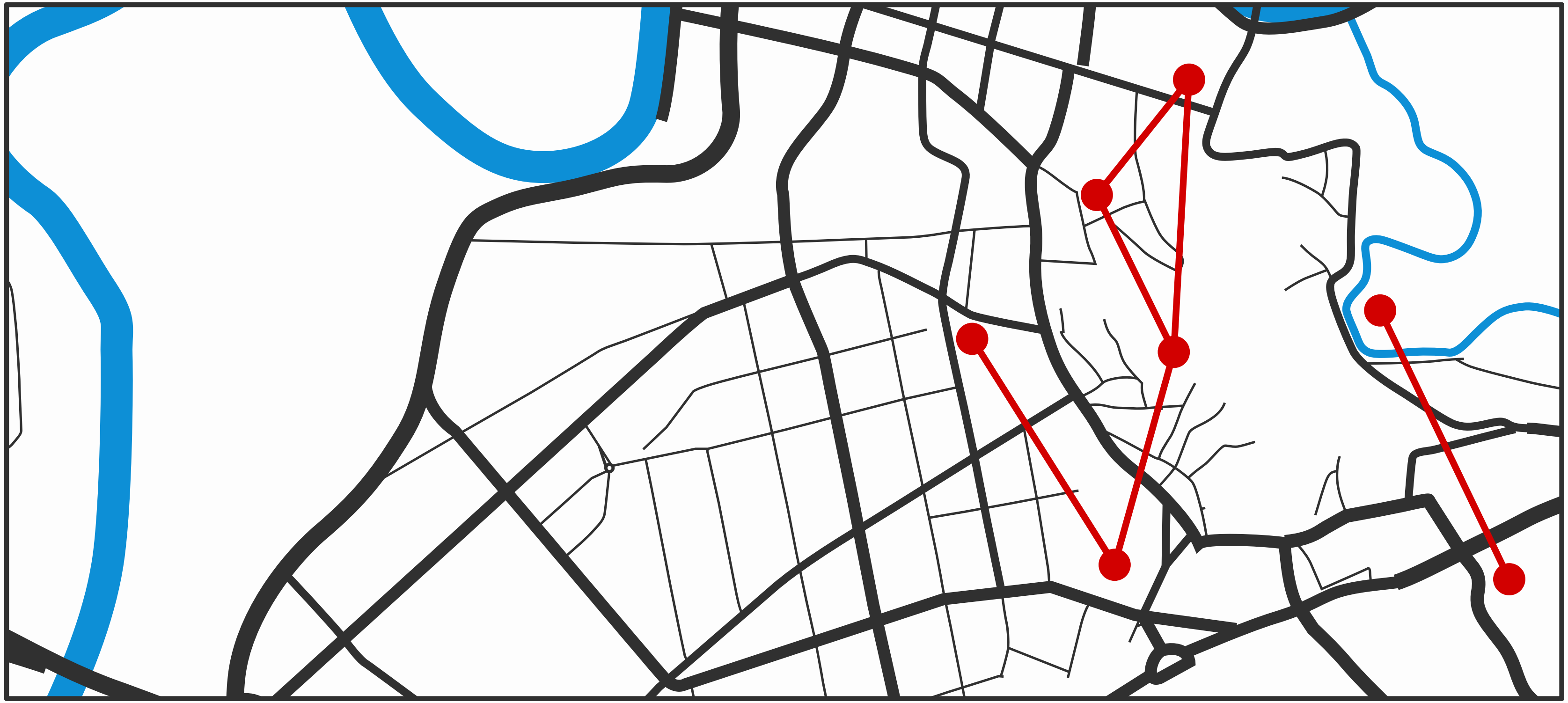
Understanding Domain-Size Generalization in Markov Logic Networks

Florian Chen, Felix Weitkämper, Sagar Malhotra

Motivation

Motivation

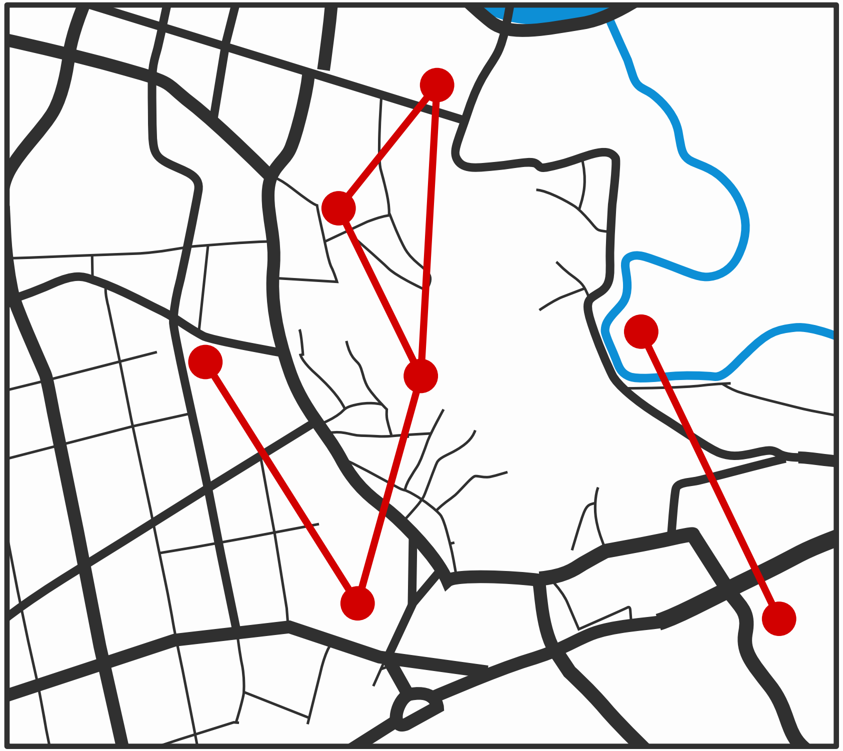


Markov Logic Networks

A Markov Logic Network Φ induces a probability distribution over $\Omega^{(n)}$:

$$P_{\Phi}^{(n)}(\omega) = \frac{1}{Z(n)} \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} a_i N(\phi_i, \omega) \Big)$$

a_i	ϕ_i
a_1	$\texttt{Vaccine}(x) \rightarrow \neg\texttt{Covid}(x)$
a_2	$\texttt{Covid}(x) \wedge \texttt{Contact}(x, y) \rightarrow \texttt{Covid}(y)$



Markov Logic Networks

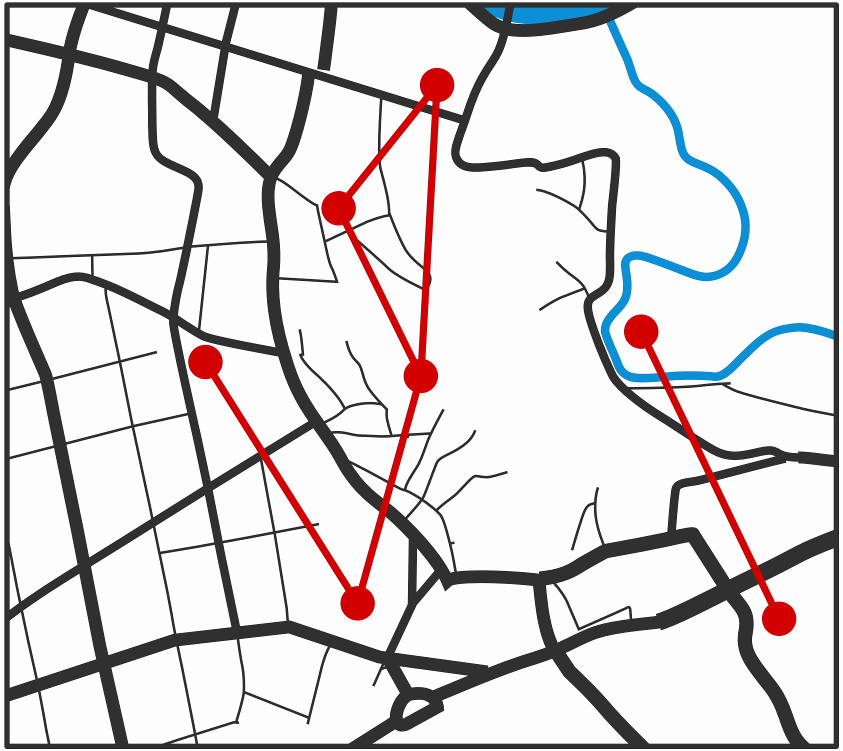
A Markov Logic Network Φ induces a probability distribution over $\Omega^{(n)}$:

$$P_{\Phi}^{(n)}(\omega) = \frac{1}{Z(n)} \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} a_i N(\phi_i, \omega) \Big)$$

Learning is guided by maximum likelihood :

$$\hat{\mathbf{a}} = \operatorname*{argmax}_{\mathbf{a}} P_{\Phi}^{(n)}(\omega)$$

a_i	ϕ_i
a_1	$\texttt{Vaccine}(x) \rightarrow \neg\texttt{Covid}(x)$
a_2	$\texttt{Covid}(x) \wedge \texttt{Contact}(x, y) \rightarrow \texttt{Covid}(y)$



Markov Logic Networks

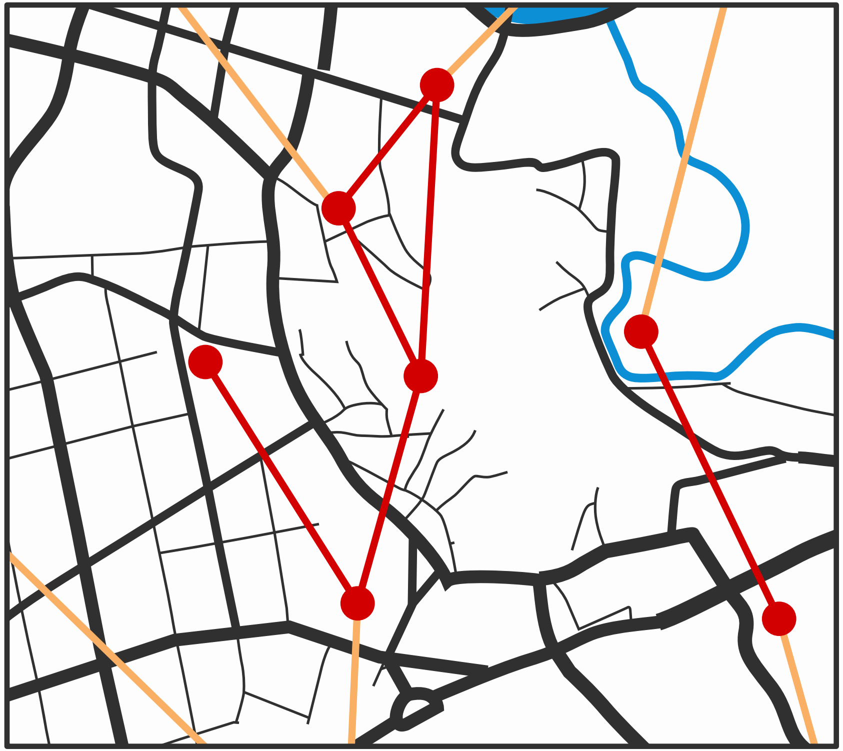
A Markov Logic Network Φ induces a probability distribution over $\Omega^{(n)}$:

$$P_{\Phi}^{(n)}(\omega) = \frac{1}{Z(n)} \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} a_i N(\phi_i, \omega) \Big)$$

Learning is guided by maximum likelihood :

$$\hat{\mathbf{a}} = \operatorname*{argmax}_{\mathbf{a}} P_{\Phi}^{(n)}(\omega)$$

a_i	ϕ_i
a_1	$\texttt{Vaccine}(x) \to \neg\texttt{Covid}(x)$
a_2	$\texttt{Covid}(x) \wedge \texttt{Contact}(x, y) \to \texttt{Covid}(y)$



Markov Logic Networks

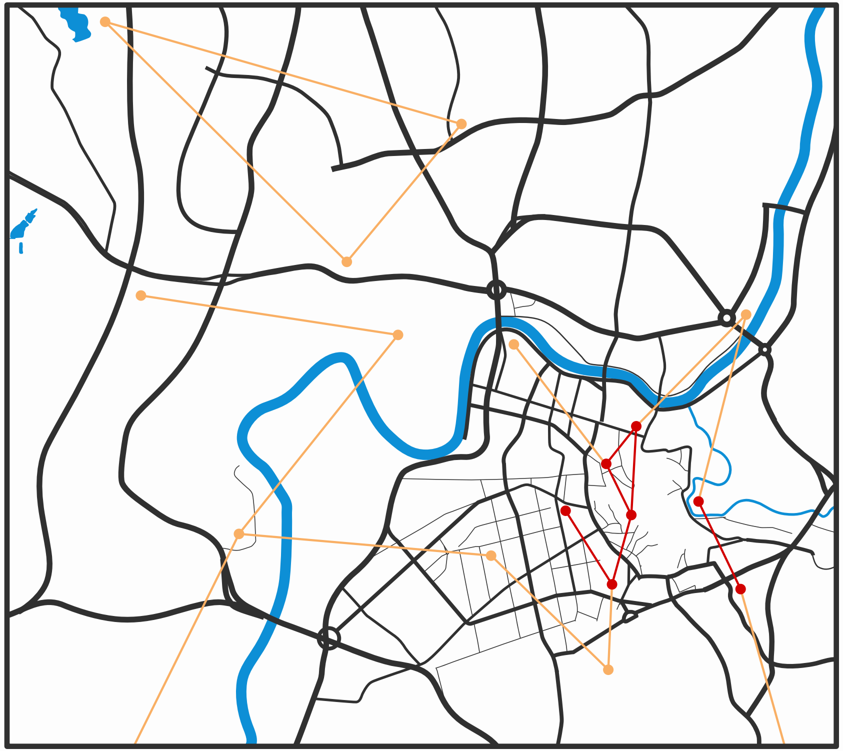
A Markov Logic Network Φ induces a probability distribution over $\Omega^{(n)}$:

$$P_{\Phi}^{(n)}(\omega) = \frac{1}{Z(n)} \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} a_i N(\phi_i, \omega) \Big)$$

Learning is guided by maximum likelihood :

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \, P_{\Phi}^{(n)}(\omega)$$

a_i	ϕ_i
a_1	$\texttt{Vaccine}(x) \rightarrow \neg\texttt{Covid}(x)$
a_2	$\texttt{Covid}(x) \wedge \texttt{Contact}(x, y) \rightarrow \texttt{Covid}(y)$



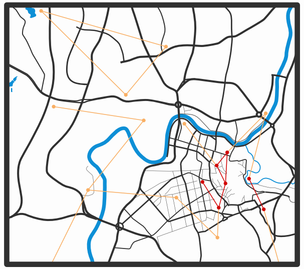
Learning Across Domain Sizes

In most cases, the observed data is substructure of a larger structure.

Our goal is to estimate parameters for the distribution $P_{\Phi}^{(n+m)}$ for some (potentialy large) m:

$$\hat{\mathbf{a}} = \operatorname{argmax}_{\mathbf{a}} P_{\Phi}^{(n+m)} \downarrow [n](\omega)$$

$$P^{(n+m)} \downarrow [n](\omega) = \sum_{\omega' \in \Omega^{(n+m)} : \omega' \downarrow [n] = \omega} P^{(n+m)}(\omega)$$



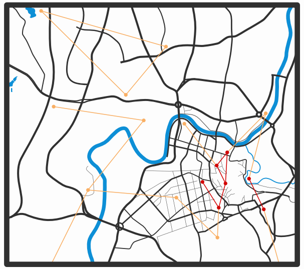
Learning Across Domain Sizes

In most cases, the observed data is substructure of a larger structure.

Our goal is to estimate parameters for the distribution $P_{\Phi}^{(n+m)}$ for some (potentialy large) m:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}}\, P_{\Phi}^{(n+m)} \downarrow [n](\omega)$$

$$P^{(n+m)} \downarrow [n](\omega) = \sum_{\omega' \in \Omega^{(n+m)} : \omega' \downarrow [n] = \omega} P^{(n+m)}(\omega)$$



However, most MLNs are **not projective**[1,2] and hence for most MLNs:

$$\underset{\mathbf{a}}{\operatorname{argmax}}\, P_{\Phi}^{(n)}(\omega) \neq \underset{\mathbf{a}}{\operatorname{argmax}}\, P_{\Phi}^{(n+m)} \downarrow [n](\omega)$$

1. ***Shalizi, C.R., Rinaldo***, A.: Consistency under sampling of exponential random graph models. Ann. Stat. 41 2, 508–535 (2013)
2. ***Jaeger, M., Schulte, O.***: A complete characterization of projectivity for statistical relational models. In: Bessiere, C. (ed.) Proc. IJCAI 2020. pp. 4283–4290. ijcai.org (2020). https://doi.org/10.24963/ijcai.2020/591

Problem Statement

$$\underset{\mathbf{a}}{\operatorname{argmax}} P_{\Phi}^{(n)}(\omega) \neq \underset{\mathbf{a}}{\operatorname{argmax}} P_{\Phi}^{(n+m)} \downarrow [n](\omega)$$

Problems:

- m may be very large
- m may be unknown

➡ This makes it (computationally) prohibitive to make the ML estimate for $P_{\Phi}^{(n+m)} \downarrow [n]$.

Hence, our goal will be to analyze the relation between the distributions $P_{\Phi}^{(n)}$ and $P_{\Phi}^{(n+m)}$ and use this analysis to get better ML estimates for $P_{\Phi}^{(n+m)} \downarrow [n]$.

Main Result

The age old wisdom is true here as well …

Main Result

The age old wisdom is true here as well ...

Regularization* Leads to Better Generalization

i.e., $P_{\Phi}^{(n)}$ approches $P_{\Phi}^{(n+m)} \downarrow [n]$ with regularization.

Weight Decomposition

Reminder - MLN probability distribution:

$$P_{\Phi}^{(n)}(\omega) = \frac{1}{Z(n)} \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} a_i N(\phi_i, \omega) \Big)$$

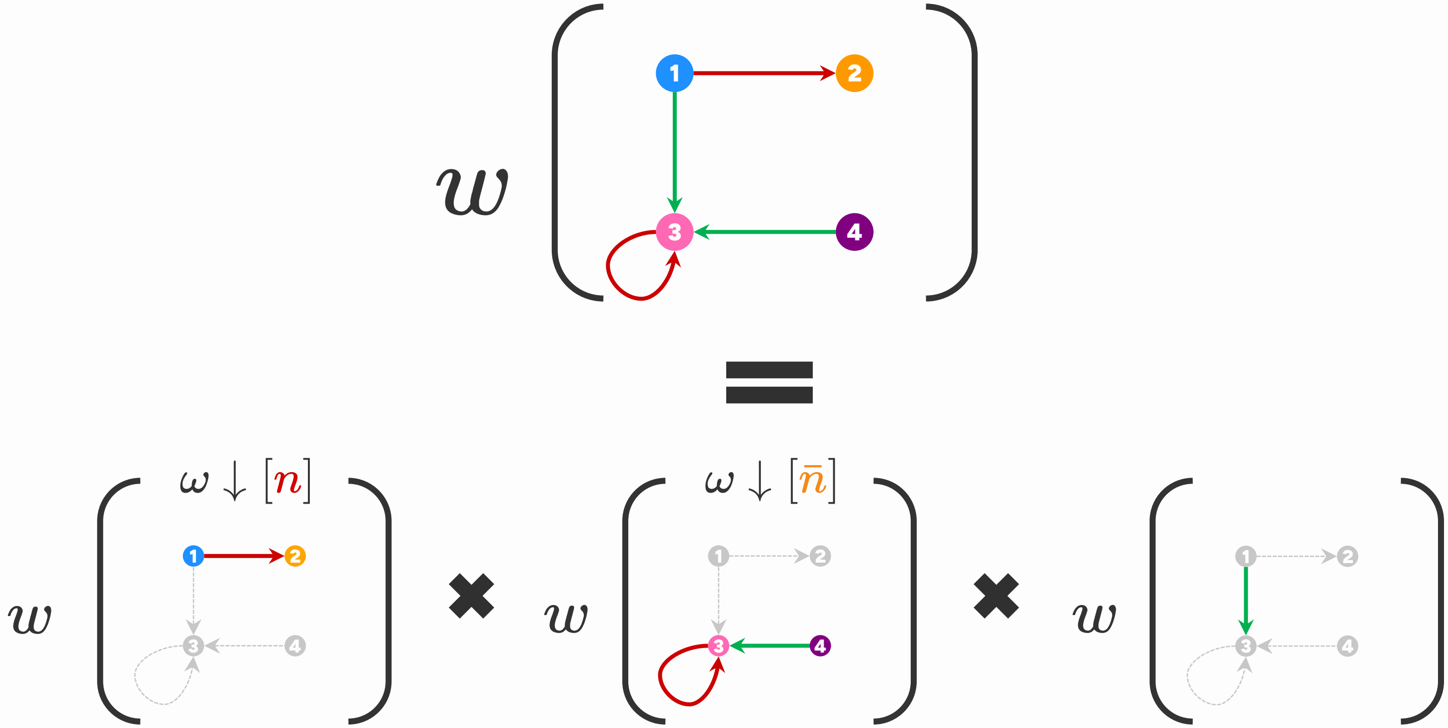
Let us define the **weight** and the ***k*-weights** of a world:

$$w(\omega) = \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} a_i N(\phi_i, \omega) \Big) \qquad w_k(\omega) = \exp\Big(\sum_{(\phi_i, a_i) \in \Phi_k} a_i N(\phi_i, \omega) \Big)$$

We can then decompose the weight of an n+m-world into contributions from the observed substructure, the unobserved structure, and the connections between these two structures:

$$w(\omega) = w(\omega \downarrow [n]) \times w(\omega \downarrow [\bar{n}]) \times \prod_{k \in [d]} \prod_{\mathbf{c} \in \langle n+m \rangle^k \setminus \langle n \rangle^k \cup \langle \bar{n} \rangle^k} w_k(\omega \downarrow \mathbf{c})$$

Weight Decomposition - Example



Bounding the Weight

$$w(\omega) \leq w(\omega \downarrow [n]) \times w(\omega \downarrow [\bar{n}]) \times \prod_{k \in [d]} (w_k^{max})^{\binom{n+m}{k} - \binom{n}{k} - \binom{m}{k}}$$

$$w(\omega) \geq w(\omega \downarrow [n]) \times w(\omega \downarrow [\bar{n}]) \times \prod_{k \in [d]} (w_k^{min})^{\binom{n+m}{k} - \binom{n}{k} - \binom{m}{k}}$$

$$M_{max} = \prod_{k \in [d]} (w_k^{max})^{\binom{n+m}{k} - \binom{n}{k} - \binom{m}{k}}$$

$$M_{min} = \prod_{k \in [d]} (w_k^{min})^{\binom{n+m}{k} - \binom{n}{k} - \binom{m}{k}}$$

For $k = 1$:

$$\binom{n+m}{1} - \binom{n}{1} - \binom{m}{1} = 0$$

$$\frac{M_{min}}{M_{max}} P_{\Phi}^{(n)}(\omega) \leq P_{\Phi}^{(n+m)} \downarrow [n](\omega) \leq \frac{M_{max}}{M_{min}} P_{\Phi}^{(n)}(\omega)$$

Two Notions of Generalization

Using these bounds, we can deduce two natural notions of generalization across domain sizes:

1. Increasing marginal likelihood:

$$-\log P_{\Phi}^{(n+m)} \downarrow [n](\omega) \leq -\log P_{\Phi}^{(n)}(\omega) + \log \Delta$$

2. Decreasing KL divergence:

$$KL(P_{\Phi}^{(n+m)} \downarrow [n] || P_{\Phi}^{(n)}) \leq \log \Delta$$

$$\Delta = \frac{M_{max}}{M_{min}}$$

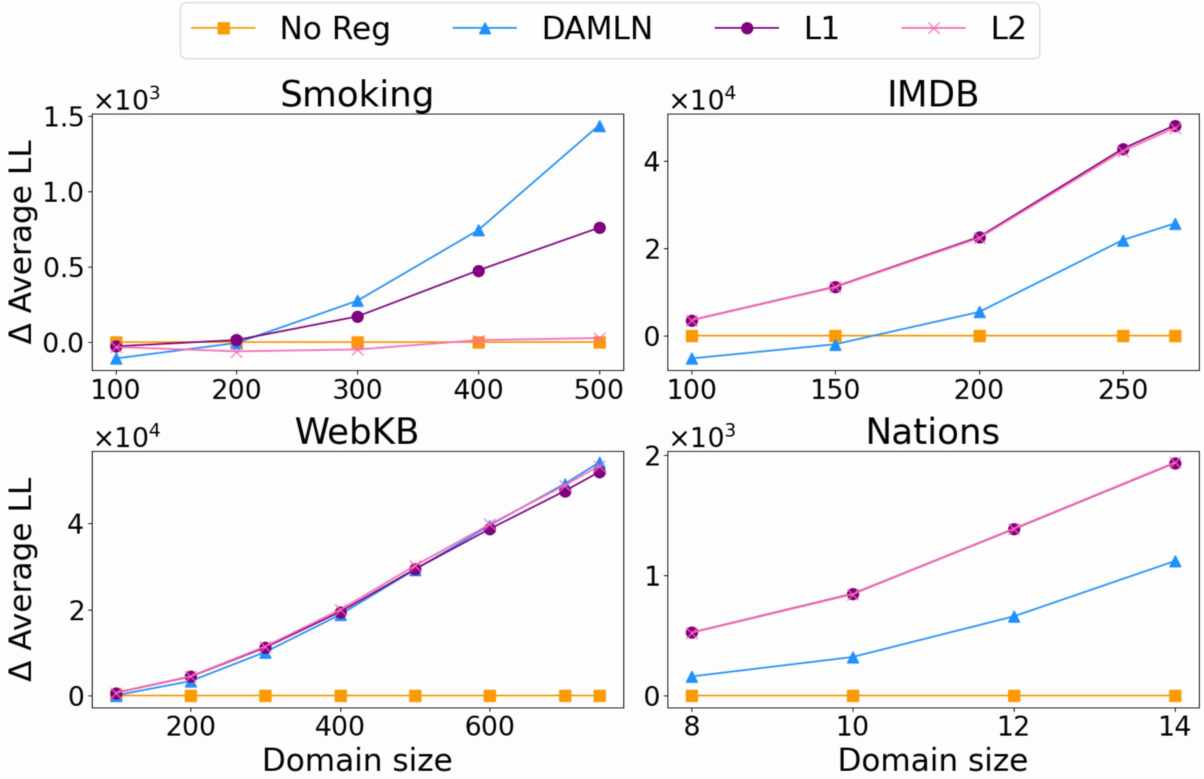
Reducing Parameter Variance

- L1 and L2 regularization
- Domain-Size Aware Markov Logic Networks (DAMLN)[3]
  - Adaptation of regular MLNs
  - Downscale formula weights depending on target domain size

$$P_{\Phi}^{(n)}(\omega) = \frac{1}{Z(n)} \exp\Big(\sum_{(\phi_i, a_i) \in \Phi} \frac{a_i}{s_i(n+m)} N(\phi_i, \omega) \Big)$$

3. **_Mittal, H., Bhardwaj, A., Gogate, V., Singla, P._**: Domain-size aware markov logic networks. In: Chaudhuri, K., Sugiyama, M. (eds.) Proc. AISTATS 2019. Proceedings of Machine Learning Research, vol. 89, pp. 3216–3224. PMLR (2019)

Experimental Results



Conclusion

Relational data does not admit consistency of parameter estimation.

We can bound this inconsistency in terms of the parameter variance.

Decreasing parameter variance allows for better generalization.

Future Work

The bounds are met for the uniform distribution.

However, they can be loose even for some projective distributions.

This indicates that maybe better bounds can be obtained.