Learning Latent Groups with Hinge-loss Markov Random Fields

Stephen H. Bach BACH@CS.UMD.EDU
Bert Huang BERT@CS.UMD.EDU
Lise Getoor GETOOR@CS.UMD.EDU
University of Maryland, College Park, Maryland 20742, USA

Abstract

Probabilistic models with latent variables are powerful tools that can help explain related phenomena by mediating dependencies among them. Learning in the presence of latent variables can be difficult though, because of the difficulty of marginalizing them out, or, more commonly, maximizing a lower bound on the marginal likelihood. In this work, we show how to learn hinge-loss Markov random fields (HL-MRFs) that contain latent variables. HL-MRFs are an expressive class of undirected probabilistic graphical models for which inference of most probable explanations is a convex optimization. By incorporating latent variables into HL-MRFs, we can build models that express rich dependencies among those latent variables. We use a hard expectation-maximization algorithm to learn the parameters of such a model, leveraging fast inference for learning. In our experiments, this combination of inference and learning discovers useful groups of users and hashtags in a Twitter data set.

1. Introduction

Hinge-loss Markov random fields (HL-MRFs) are a powerful class of probabilistic graphical models, which combine support for rich dependencies with fast, convex inference of most-probable explanantions (MPEs). They achieve this combination by expessing dependencies among variables with domain [0,1] as hinge-loss potentials, which can generalize logical implication to continuous variables. While recent advances on inference and learning for HL-MRFs allows these models to produce state-of-the-art performance on various problems with fully-observed training data, methods for parameter learning with latent variables are currently less understood. In particular, there is need for latent-variable learning methods that leverage the fast, convex inference in HL-MRFs. In this work, we introduce a hard *expectation-maximization* (EM) strategy for learning HL-MRFs with latent variables. This strategy mixes inference and supervised learning (where all variables are observed), two well-understood tasks in HL-MRFs, allowing learning with latent variables while leveraging the rich expressivity of HL-MRFs.

HL-MRFs are the formalism behind the probabilistic soft logic (PSL) modeling language (Broecheler et al., 2010), and have been used for collective classification, ontology alignment (Broecheler et al., 2010), social trust prediction (Huang et al., 2013), voter opinion modeling (Bach et al., 2012; Broecheler & Getoor, 2010), and graph summarization (Memory et al., 2012). PSL is one of many tools for designing relational probabilistic models, but is perhaps most related to Markov logic networks (Richardson & Domingos, 2006), which use a similar syntax based on first-order logic to define models.

When learning parameters in models with hidden, or latent, variables, the standard approach is to maximize the likelihood of the observed labels, which involves marginalizing over the latent variable states. In many models, directly computing this likelihood is too expensive, so the variational method of expectation maximization (EM) provides an alternative (Dempster et al., 1977). The variational interpretation of EM iteratively updates a proposal distribution, minimizing the Kullback-Leiber (KL) divergence to the empirical distribution, interleaved with estimating the expectation of the latent variables. The variational view allows the possibility of EM with a limited but tractable family of proposal distributions and provides theoretical justification for what is commonly known as "hard EM". In hard EM, the proposal distribution

Presented at the International Conference on Machine Learning (ICML) workshop on *Inferning: Interactions between Inference and Learning*, Atlanta, Georgia, USA, 2013. Copyright 2013 by the author(s).

comes from the family of point distributions: distributions where the probability is one at a single point estimate and zero otherwise. Since HL-MRFs admit fast and efficient MPE inference, they are well-suited for hard EM.

We demonstrate our approach on the task of group detection in social media data, extending previous work that used fixed-parameter HL-MRFs for the same task (Huang et al., 2012). Group detection in social media is an important task since more and more real-world phenomena, such as political organizing and discourse, take place through social media. Group detection has the potential to help us understand language, political events, social interactions, and more.

2. Hinge-loss Markov Random Fields

In this section, we review hinge-loss Markov random fields (HL-MRFs) and probabilistic soft logic (PSL). HL-MRFs are parameterized by constrained hinge-loss energy functions. The energy function is factored into hinge-loss potentials, which are clamped linear functions of the continuous variables, or squares of these functions. These potentials are weighted by a set of parameter weights, which can be learned and can be templated, i.e., many potentials of the same form may share the same weight. Additionally, HL-MRFs can incorporate linear constraints on the variables, which can be useful for modeling, for example, mutual exclusion between variable states. For completeness, a full, formal definition of HL-MRFs is as follows.

Definition 1. *Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be a vector of n variables and $\mathbf{X} = (X_1, \dots, X_{n'})$ a vector of n' variables with joint domain $\mathbf{D} = [0,1]^{n+n'}$. Let $\phi = (\phi_1, \dots, \phi_m)$ be m continuous potentials of the form*

$$\phi_j(\mathbf{Y}, \mathbf{X}) = [\max\{\ell_j(\mathbf{Y}, \mathbf{X}), 0\}]^{p_j}$$

where ℓ_j is a linear function of $\mathbf{Y}$ and $\mathbf{X}$ and $p_j \in \{1,2\}$. Let $C = (C_1, \dots, C_r)$ be linear constraint functions associated with index sets denoting equality constraints $\mathcal{E}$ and inequality constraints $\mathcal{I}$, which define the feasible set

$$\tilde{\mathbf{D}} = \left\{ \mathbf{Y}, \mathbf{X} \in \mathbf{D} \;\middle|\; \begin{array}{l} C_k(\mathbf{Y}, \mathbf{X}) = 0, \forall k \in \mathcal{E} \\ C_k(\mathbf{Y}, \mathbf{X}) \geq 0, \forall k \in \mathcal{I} \end{array} \right\}.$$

For $\mathbf{Y}, \mathbf{X} \in \tilde{\mathbf{D}}$, given a vector of nonnegative free parameters, i.e., weights, $\lambda = (\lambda_1, \dots, \lambda_m)$, a constrained hinge-loss energy function *f_λ is defined as*

$$f_\lambda(\mathbf{Y}, \mathbf{X}) = \sum_{j=1}^{m} \lambda_j \phi_j(\mathbf{Y}, \mathbf{X}) \, .$$

Definition 2. *A hinge-loss Markov random field P over random variables $\mathbf{Y}$ and conditioned on random variables $\mathbf{X}$ is a probability density defined as follows: if $\mathbf{Y}, \mathbf{X} \notin \tilde{\mathbf{D}}$, then $P(\mathbf{Y}|\mathbf{X}) = 0$; if $\mathbf{Y}, \mathbf{X} \in \tilde{\mathbf{D}}$, then*

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\lambda)} \exp\left[-f_\lambda(\mathbf{Y}, \mathbf{X})\right] \, ,$$

where $Z(\lambda) = \int_{\mathbf{Y}} \exp\left[-f_\lambda(\mathbf{Y}, \mathbf{X})\right]$.

Thus, MPE inference is equivalent to finding the minimizer of the convex energy f_λ.

Probabilistic soft logic (Broecheler et al., 2010; Kimmig et al., 2012) provides a natural interface to represent hinge-loss potential templates using logical conjunction and implication. In particular, a logical conjunction of Boolean variables $X \wedge Y$ can be generalized to continuous variables using the hinge function $\max\{X + Y - 1, 0\}$, which is known as the *Lukasiewicz t-norm*. Similarly, logical implication $X \Rightarrow Y$ is relaxed via $1 - \max\{Y - X, 0\}$. PSL allows modelers to design rules that, given data, ground out possible substitutions for logical terms. The groundings of a template define hinge-loss potentials that share the same weight. PSL rules take the form of these soft logical implications, and the linear function of the HL-MRF potential is the ground rule's *distance to satisfaction*, $\max\{Y - X, 0\}$. We defer to Broecheler et al. (2010) and Kimmig et al. (2012) for further details on PSL.

Inference of the *most probable explanation* (MPE) in HL-MRFs is a convex optimization, since the hinge-loss potentials are each convex and the linear constraints preserve convexity. Currently, the fastest known method for HL-MRF inference uses the alternating direction method of multipliers (ADMM), which decomposes the full objective into subproblems each with their own copy of the variables and uses augmented Lagrangian relaxation to enforce consensus between the independently optimized subproblems (Bach et al., 2012). The factorized form of the HL-MRF energy function naturally corresponds to a subproblem partitioning, with each hinge-loss potential and each constraint forming its own subproblem.

A number of methods can be used to learn the weights of an HL-MRF. Currently, two main strategies have been studied: approximate maximum likelihood (Broecheler et al., 2010) and large-margin estimation (Bach et al., 2013). In this work, we focus on approximate maximum likelihood using voted perceptron gradient ascent. The gradient for the likelihood of training data contains the expectation of the log-linear features, which we approximate via the MPE solution. Thus far, these learning methods require all unknown variables to have labeled ground truth during training. In the next section, we describe one method to relax this restriction and learn when only part of the unknown variables have observed labels.

3. Learning with Latent Variables

Probabilistic models can have latent variables for a number of reasons. In some cases, large models have many unknowns to the point that it is impractical to collect ground truth for them all. In other cases, the latent variables represent values that can never be measured, since they are inherently latent and may have no real-world analogue. In both of these scenarios, the standard strategy is to maximize the likelihood of the available labeled data. Let $\mathbf{X}$ be the observed variables, $\mathbf{Y}$ be the target variables (i.e., labels), and $\mathbf{Z}$ be latent variables, which are available for inference but for which we have no ground-truth labels. The standard learning objective for learning parameters λ is

$$\max_{\lambda} P(\mathbf{Y}|\mathbf{X};\lambda).$$

Evaluating this objective function in many graphical models, including HL-MRFs, is intractable in general, so direct optimization is often not an available option.

3.1. Expectation Maximization

Expectation maximization (EM) is a general framework for learning in the presence of latent variables (Dempster et al., 1977). EM maximizes a lower bound on the marginal likelihood $P(\mathbf{Y}|\mathbf{X};\lambda)$. For models where evaluating $P(\mathbf{Y}|\mathbf{X};\lambda)$ is intractable; it is often possible to work with the original distribution $P(\mathbf{Y},\mathbf{Z}|\mathbf{X};\lambda)$. Thus, EM alternates between fitting a distribution $q(\mathbf{Z})$ to the posterior distribution $P(\mathbf{Z}|\mathbf{Y},\mathbf{X};\lambda)$ and maximizing $\mathbb{E}_{q(\mathbf{Z})}\left[P(\mathbf{Y},\mathbf{Z}|\mathbf{X};\lambda)\right]$, the expected complete data likelihood with respect to $q(\mathbf{Z})$.

For general, undirected graphical models, this maximization is intractable. One way to make it more tractable is to use a "hard" variant of EM, in which $q(\mathbf{Z})$ is restricted to the family of delta distributions which place all mass on a single assignment to the variables. With zero density assigned to all points except one, maximizing the expected complete data likelihood is equivalent to supervised learning using the assignment to the latent varaibles with non-zero density. Hard EM is attractive, especially in models such as HL-MRFs, because it alternates between inference and supervised learning, two tasks we have efficient algorithms to solve. While hard EM also maximizes a lower bound on the marginal likelihood, the delta distribution restriction makes it prone to finding local optima. This makes initializing the model wit reasonable parameters important. For example, parameters can be initialized with expert seed knowledge. The hard EM algorithm is summarized in Algorithm 1.

Algorithm 1 Hard Expectation Maximization

```
Input: model $P(\mathbf{Y},\mathbf{Z}|\mathbf{X};\lambda)$, initial parameters $\lambda^0$
$t \leftarrow 1$
while not converged do
    $\mathbf{Z}^t = \arg\max_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{Y},\mathbf{X};\lambda^{t-1})$
    $\lambda^t = \arg\max_{\lambda} P(\mathbf{Y},\mathbf{Z}^t|\mathbf{X};\lambda)$
    $t \leftarrow t+1$
end while
```

4. Evaluation

To evaluate our proposed approach, we build a model over a rich social-media data set collected from South American users in the 48 hours around the 2012 Venezuelan presidential election. The data set is collected from the social medium Twitter and is composed of short, public messages or tweets written by users. Some tweets express opinions, some mention other users, some are retweets (rebroadcast messages), and many contain hashtags (user-specified annotations). Such social-media data sets can contain a wealth of information about public opinions, social structures, and language. We extract some of this information by including in our model the notion of latent user groups which mediate probabilistic dependencies between hashtag usage and social interactions such as retweets and mentions.

Our goal is to learn model parameters that explain a set of users' interactions with a smaller set of top users of interest, e.g., political figures, news organizations, and entertainment accounts, given the users' hashtag usage and their interactions with others outside the set of top users. Since our data set focuses on a presidential election, we assume that their are two latent groups, one associated with each major candidate, and we will interpret the learned parameters as the strengths of associations between each group and particular hashtags or top users.

4.1. Data Set

Our data set is roughly 4,275,000 tweets collected from about 1,350,000 Twitter users via a query that focused on South American users. The tweets were collected from October 6 to October 8, 2012, a 48-hour window around the Venezuelan presidential election on October 7. The two major candidates in the election were

Hugo Chávez, the incumbent, and Henrique Capriles. Chávez won with 55% of the vote.

To learn a model relating hashtag usage and interacations with top users, we first identify 20 users as top users based on being the most retweeted or, in the case of the state-owned television network's account, being of particular interest. We then identify all other users that either retweeted or mentioned at least one of the top users and used at least one hashtag in a tweet that was not a mention or a retweet of a top user. Filtering by these criteria, the set contains 1,678 regular users (i.e., users that are not top users).

Whether each regular user tweeted a hashtag is represented with the PSL predicate UsedHashtag. Tweets that mention or retweet a top user are not counted. For example, if we observe that User 1 tweeted a tweet that contains the hashtag `#hayuncamino` then $\textsc{UsedHashtag}(1, \texttt{\#hayuncamino})$ has an observed truth value of 1.0. The PSL predicate RegularUserLink represents whether a regular user retweeted or mentioned *any* user in the full data set that is not a top user, regardless of whether that mentioned or retweeted user is a regular user. Whether a regular user retweeted or mentioned a top user is represented with the PSL predicate TopUserLink. Finally, the latent group membership of each regular user is represented with the PSL predicate InGroup.

4.2. Latent Group Model

We now describe our model for predicting the interactions of regular users with top users via latent group membership. We describe the HL-MRF in terms of PSL rules and constraints. (See Section 2.) We implement the hard EM algorithm (Algorithm 1) by treating atoms with the UsedHashtag or RegularUserLink predicate as the set of conditioning variables $\mathbf{X}$, atoms with the TopUserLink predicate as the set of target variables $\mathbf{Y}$, and atoms with the InGroup predicate as the set of latent variables $\mathbf{Z}$.

When defining our model, we will make reference to the set H of hashtags used by at least 15 different regular users ($|H| = 33$), the set T of top users ($|T| = 20$), and the set of latent groups $\mathcal{G} = \{g_0, g_1\}$.

We first include rules that relate hashtag usage to group membership. For each hashtag in H and each latent group, we include a rule of the form

$$w_{h,g} : \textsc{UsedHashtag}(U, h) \rightarrow \textsc{InGroup}(U, g) \qquad \forall h \in H, \forall g \in \mathcal{G}$$

so that there is a different rule weight governing how strongly each commonly used hashtag is associated with each latent group.

Next, we leverage one the advantages of our approach to learning with latent variables: the ability to easily include interesting dependencies among latent variables. We use the rule

$$\begin{aligned} w_{\text{social}} : &\textsc{RegularUserLink}(U_1, U_3) \\ &\wedge \textsc{RegularUserLink}(U_2, U_3) \wedge U_1 \neq U_2 \\ &\wedge \textsc{InGroup}(U_1, G) \rightarrow \textsc{InGroup}(U_2, G) \end{aligned}$$

to encode the intuition that regular users who interact with the same people on Twitter are more likely to belong to the same latent group.

Third, we include rules of the form

$$w_{g,t} : \textsc{InGroup}(U, g) \rightarrow \textsc{TopUserLink}(U, t) \qquad \forall g \in \mathcal{G}, \forall t \in T$$

for each latent group and each top user so that there is a parameter governing how strongly each latent group tends to interact with each top user.

Finally, we include a simple rule that acts as a prior penalizing strong assignments of regular users to either group.

$$w_{\text{prior}} : \neg\textsc{InGroup}(U, G)$$

Since the potentials defined by all our rules are squared, this prior indicates that given little or weak evidence, users belong to all groups evenly. I.e., the model will only infer strong group assignments if there is strong evidence.

In addition to our rules, we include two sets of constraints. The first set constrains the InGroup atoms for each regular user to sum to 1.0, making InGroup a mixed-membership assignment. The second set constrains the TopUserLink atoms for each regular user to sum to the number of interactions with top users observed for that regular user. This makes the inference task to predict *which* interactions occurred, since the constraint fixes how many interactions occurred.

All that remains to use the hard EM algorithm is to specify initial parameters λ_0. We initialize w_{prior} to 3.0. We initialize $w_{h,g}$, w_{social}, and $w_{g,t}$ to 2.0 for all hashtags, groups, and top users, *except* two hashtags and two top users which we initially assign as seeds. We initially associate the top user `hayuncamino` (Henrique Capriles's campaign account) and the hashtag for Capriles's campaign slogan `#hayuncamino` with Group 0 by initializing the parameters associating them with Group 0 to 10.0 and those associating them with Group 1 to 0.0. We initially associate

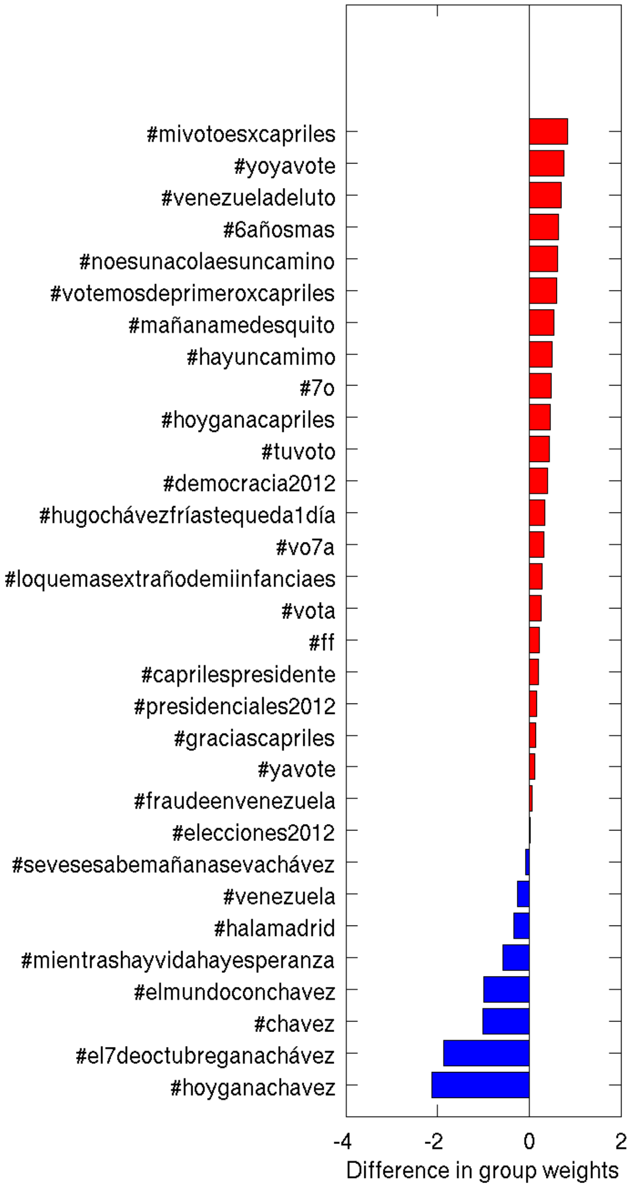

Figure 1. Learned parameters associating latent groups with hashtags. Shown is the value $w_{h,g_0} - w_{h,g_1}$ for each hashtag $h \in H$.

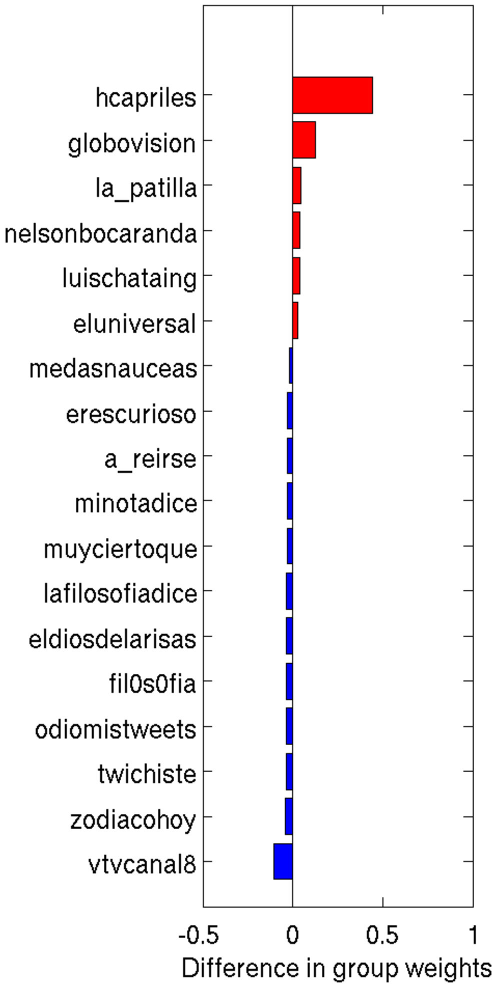

Figure 2. Learned parameters associating latent groups with interactions with top users. Shown is the value $w_{g_0,t} - w_{g_1,t}$ for each target user $t \in T$.

the top user `chavezcandanga` (Hugo Chávez's account) and the hashtag for Chávez's campaign slogan `#elmundoconchávez` with Group 1 in the same way.

4.3. Results

To learn associations between latent groups and hashtags or interactions with top users, we perform ten iterations of hard EM on the HL-MRF defined in the previous subsection over the entire data set. Each maximizer for $P(\mathbf{Z}|\mathbf{Y}, \mathbf{X}; \lambda)$ is easy to find exactly via convex optimization. To find each λ, we perform ten steps of voted-perceptron gradient ascent approximating the expected values of the log-linear features by their values in the MPE state. This approximate maximization works well because as long as it improves the log-likelihood of $\mathbf{Y}$ and $\mathbf{Z}$, the marginal likelihood $P(\mathbf{Y}|\mathbf{X};\lambda)$ will improve.

Figure 1 shows the learned parameters associating hashtag usage with latent groups (excluding the two seeded hashtags). The hashtags are sorted by differences in parameter values from Capriles to Chávez. Our assignment of seeds associated pro-Capriles users with Group 0 and pro-Chávez users with Group 1. The results show a very clean ordering of hashtags based on ideology. Many of the hashtags most strongly associated with the latent Capriles group are explicitly pro-Capriles, e.g., `#mivotoesxcapriles`, `#votemosdeprimeroxcapriles`, and `#hayuncamimo`, an alternative spelling of Capriles's campaign slogan. Others are also clearly anti-Chávez: `#venezueladeluto` ("Venezuela in mourning" after Chávez's reelection) and `#hugochávezfríastequeda1día` (roughly "Hugo Chávez has one day left"). One surprising result is that `#6añosmas` ("six more years") is strongly associated with the latent Capriles group despite superficially appearing to support the incumbent. However, upon inspection of the tweets that use this hashtag, most in our dataset use it ironically, predicting "six more years" of negative outcomes of Chávez's reelection. On the other hand, the hashtags strongly associated with the Chávez group are all explicitly pro-Chávez or just his name. Interestingly, the semantically neutral hashtags promoting voter turnout, such as `#tuvoto`, `#vota`, and `#vo7a`, are inferred to favor the Capriles group. We hypothesize that these may be because the social media campaign for increasing voter turnout was stronger from the Capriles side.

Figure 2 shows the learned parameters associating interactions with top users with latent groups (again excluding the two seeded top users). According to the learned model, users in the latent Capriles group are most likely to interact with `hcapriles` (Capriles's personal account) and independent media outlets and journalists such as `globovision`, `la_patilla`, `nelsonbocaranda`, `luischataing`, and `eluniversal`. On the other side of the spectrum, users in the latent Chávez group are most likely to interact with `vtvcanal8`, the Twitter account of the state-owned television network.

Our results include both obviously correct and surprising, but verifiable, elements. The learned parameters are useful for understanding the language used by and the social associations among people with different political opinions. They are also easy to compute. This experiment uses an HL-MRF with approximately 37,000 variables and 146,000 potentials and constraints, but ten iterations of hard EM takes only a few minutes on a workstation.

5. Conclusion

By learning a model that orders top hashtags and top users according to associations with two latent political groups, we show the power of HL-MRFs for understanding rich, real-world data. We advance the state of the art by learning an HL-MRF that includes latent variables, using a hard EM procedure that benefits from the fast, convex inference available for HL-MRFs. These latent variables are made interpretable by their construction with PSL, demonstrating the utility of HL-MRFs for many fields, including computational social science. Directions for future work include developing richer proposal distribution families compatible with HL-MRFs, modeling latent variables in other tasks, and quantifying the effects of modeling latent variables on prediction tasks.

Acknowledgments

This work was supported by NSF grant CCF0937094 and IARPA via DoI/NBC contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Bach, S., Broecheler, M., Getoor, L., and O'Leary, D. Scaling MPE inference for constrained continuous Markov random fields with consensus optimization. In *Neural Information Processing Systems*, 2012.

Bach, S., Huang, B., London, B., and Getoor, L. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence*, 2013.

Broecheler, M. and Getoor, L. Computing marginal distributions over continuous Markov networks for statistical relational learning. In *Neural Information Processing Systems*, 2010.

Broecheler, M., Mihalkova, L., and Getoor, L. Probabilistic similarity logic. In *Uncertainty in Artificial Intelligence*, 2010.

Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

Huang, B., Bach, S., Norris, E., Pujara, J., and Getoor, L. Social group modeling with probabilistic soft logic. In *NIPS 2012 Workshop - Social Network and Social Media Analysis: Methods, Models, and Applications*, 2012.

Huang, B., Kimmig, A., Getoor, L., and Golbeck, J. A flexible framework for probabilistic models of social trust. In *Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction*, 2013.

Kimmig, A., Bach, S., Broecheler, M., Huang, B., and Getoor, L. A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

Memory, A., Kimmig, A., Bach, S., Raschid, L., and Getoor, L. Graph summarization in annotated data using probabilistic soft logic. In *Proceedings of the International Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*, 2012.

Richardson, M. and Domingos, P. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.