Improved Generalization Bounds for Large-scale Structured Prediction

Ben London
Dept. of Computer Science
University of Maryland
College Park, MD 20742
blondon@cs.umd.edu

Bert Huang
Dept. of Computer Science
University of Maryland
College Park, MD 20742
bert@cs.umd.edu

Lise Getoor
Dept. of Computer Science
University of Maryland
College Park, MD 20742
getoor@cs.umd.edu

1 Introduction

Collective inference has been shown empirically to successfully exploit the natural dependencies in relational and network data [5, 11, 12, 13]. Though many collective techniques are capable of induction, and have been shown to be *asymptotically consistent* [16], little to no theory exists concerning the generalization of such methods. Collective inference can sometimes be viewed as large-scale structured prediction, for which there do exist generalization guarantees. The tightest existing uniform convergence rates for structured prediction [2, 10, 14] scale as $\mathrm{O}(\sqrt{\log(mn)/m})$, where m is the number of examples, and n the size of each structure. In practice, however, n may be very large, and m may be few or fixed. For example, in image segmentation, a high-resolution image may have billions of pixels, so the number of labeled images may be limited; in network analysis, the training set may consist of a single labeled connected component. In such cases, existing bounds do not guarantee generalization; yet, this theory contradicts the overwhelming empirical evidence found in the literature.

In this paper, we propose a theory of how collective models generalize in such situations. Our premise is that, if the data exhibits weak dependence within each structured instance, and if the predictor exhibits certain complexity and stability properties, then the empirical risk estimate should concentrate around its mean as n (and/or m) grows. Our analysis leverages recent results in the concentration of dependent random variables. Under certain weak dependence conditions, the effect of dependence amounts to a constant multiplier, thus enabling concentration. We also identify two properties of the hypothesis class—Rademacher complexity, and a new property which we refer to as *collective stability*—as sufficient conditions for generalization. When satisfied, we observe $\mathrm{O}(1/\sqrt{mn})$ uniform convergence to the true risk, which is significantly faster decay than previous bounds when n is very large.

2 Preliminaries

Let $\mathcal{G}$ denote an arbitrary family of undirected graphs, and $\mathcal{G}_n$ the set of all such graphs of order n. Throughout this paper, we define a graph $G \in \mathcal{G}_n$ by a set of nodes $V \triangleq [n]$ (where $[n]$ denotes the set $\{1, \ldots, n\}$) and a set of edges E. Unless otherwise specified, assume that all graphs are undirected. For a node $i \in V$, let $\mathcal{N}(i) \triangleq \{j : \{i, j\} \in E\}$ denote the *neighbors* of i; let $\mathcal{N}^d(i)$ denote the neighbors within graph distance d. Let $\mathcal{X}$ denote a countable input space, $\mathcal{Y}$ a countable output space, and $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ their Cartesian product. We use $z \in \mathcal{Z}$ (or $\mathbf{z} \in \mathcal{Z}^n$) to denote an instance (or instances) of $\mathcal{Z}$.

Definition 1. For an undirected graph $G \triangleq (V, E)$, a *random field* is a set of random variables $\mathbf{Z} \triangleq \{Z_i : i \in V\}$ indexed by V. We say that $\mathbf{Z}$ is *Markovian* if every variable Z_i is conditionally independent of all non-adjacent variables $\mathbf{Z} \setminus \mathbf{Z}_{\mathcal{N}(i)}$, given its neighbors $\mathbf{Z}_{\mathcal{N}(i)}$ (sometimes referred to as its *Markov blanket*); thus, the distribution factorizes over the cliques of G.

For some $n \geq 1$, let $G \in \mathcal{G}_n$ be a fixed graph, and let $\mathbf{Z}$ be a random field on G. Each $Z_i \in \mathbf{Z}$ takes values in $\mathcal{Z}$, and can therefore be expressed as two random variables (X_i, Y_i), taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. Denote by $\mathbb{P}_G$ the distribution of a random field on G. We use $\mathbb{E}_G$ to indicate that the expectation is taken w.r.t. $\mathbb{P}_G$. It is important to distinguish between the *true* (unknown) generating distribution and the *model*-induced distribution. An overly simplistic model may not capture certain dependencies in the true distribution, whereas an excessively complex model may assume dependencies that don't exist (which can result in *overfitting*).

In the canonical learning framework for structured prediction, we are given m independent draws from $\mathbb{P}_G(\mathbf{Z})$—i.e., *realizations* of $\mathbf{Z}$. In our case, we will assume that n is much larger than m, or that n grows and $m = \mathrm{O}(1)$. In network analysis, it is not unusual to learn from a single instance. Note that any number of realizations can be represented as a single realization of a global random field of order mn, whose distribution factorizes over the (identical) marginal distributions of m isomorphic, disjoint random fields of order n. Such is the case in many computer vision tasks, in which the training set consists of multiple images of identical dimensions. Because of this equivalence, unless otherwise noted, we will assume that the training data consists of a single realization $\mathbf{z} \in \mathcal{Z}^n$ of a random field of order n.

Using $\mathbf{z}$ (and possibly G), we learn a hypothesis h from a specified class $\mathcal{H} \triangleq \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$, where $\hat{\mathcal{Y}}$ is not necessarily the same as $\mathcal{Y}$. We are particularly interested in hypothesis classes that perform *collective inference*—that is, joint reasoning over all instances in the input. This means that changes to any single input instance may affect the output predictions on others. One specific class of collective models we will consider are what we refer to as *graph-based*—meaning, inference propagates according to a neighborhood topology derived from the input graph. Formally:

Definition 2. For a hypothesis h, a graph G, and an input $\mathbf{X}$, let random variables $\mathbf{H} \triangleq h_G(\mathbf{X})$ correspond to the prediction vector, and let $\mathbb{P}_{G,h}$ denote the joint distribution of $(\mathbf{X}, \mathbf{H})$. We say that a hypothesis class $\mathcal{H}$ is *(d^{th}-order) graph-based* if, for every $h \in \mathcal{H}$ and $i \in [n]$,

$$\mathbb{P}_{G,h}(H_i \,|\, \mathbf{X}, \mathbf{H} \setminus H_i) = \mathbb{P}_{G,h}(H_i \,|\, X_i, \mathbf{X}_{\mathcal{N}^d(i)}, \mathbf{H}_{\mathcal{N}^d(i)}).$$

In other words, the prediction on X_i is a function of $(X_i, \mathbf{X}_{\mathcal{N}^d(i)}, \mathbf{H}_{\mathcal{N}^d(i)})$ and is conditionally independent of all other inputs and predictions given this set. By implication, for any two subsets $\mathbf{X}_1 \subseteq \mathbf{X}$ and $\mathbf{X}_2 \subseteq \mathbf{X}$ that are disconnected in G, their corresponding predictions $\mathbf{H}_1$ and $\mathbf{H}_2$ are mutually independent. Examples of graph-based hypotheses include many popular graphical models, such as (conditional) Markov random fields [8, 13], and iterative algorithms [11].

Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ be a loss function. Define the *empirical* loss L of a hypothesis h as $L(h, \mathbf{Z}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(\mathbf{X})_i)$. For example, using the *0-1 loss*, $\ell_{\mathbb{1}}(y, \hat{y}) \triangleq \mathbb{1}[y \neq \hat{y}]$, L is equivalent to the normalized Hamming distance. The quantity of interest is the *expected* loss $\overline{L}_G(h) \triangleq \mathbb{E}_G[L(h, \mathbf{Z})]$ (also known as the *risk*) over realizations of a random field $\mathbf{Z}$ on G, which corresponds to the error h will incur on future predictions. In the event that $\mathbf{Z}$ represents m realizations of the same underlying random field (as described above), we may only be interested in the expected loss of a single realization. Using the previous computer vision example, the test instance would be a single image. In such cases, when $\mathcal{H}$ is graph-based, we can easily show that the expected loss on a single realization of this component random field is equal to the expected loss over m realizations.

Lemma 1. *Let $\mathbf{Z}$ be a random field on a graph G. Let $\mathbf{Z}'$ be a random field on $G' \triangleq \bigcup_{j=1}^{m} G_j : G_j \simeq G$, representing m realizations of $\mathbf{Z}$. If $\mathcal{H}$ is graph-based, then, for any $m \geq 1$, any $n \geq 1$, any $G \in \mathcal{G}_n$, and any $h \in \mathcal{H}$, we have that $\mathbb{E}_G[L(h, \mathbf{Z})] = \mathbb{E}_{G'}[L(h, \mathbf{Z}')]$.*

2.1 Concentration Inequalities

Before proceeding to our results, we review some supporting definitions and a theorem on the concentration of dependent random variables. For the following, let $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^{n}$ be a random variables with distribution $\mathbb{P}$, taking values in a countable space $\mathcal{Z}$, and let $f : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function. For $i \in [n]$, $\mathbf{z} \in \mathcal{Z}^{i-1}$ and $a, b \in \mathcal{Z}$, denote by $\Pi^{\mathbf{z}}_{i,a,b}(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)})$ the *maximal coupling* of the conditional distributions $\mathbb{P}(Z_{i+1:n} \,|\, \mathbf{Z}_{1:i-1} = \mathbf{z}, Z_i = a)$ and $\mathbb{P}(Z_{i+1:n} \,|\, \mathbf{Z}_{1:i-1} = \mathbf{z}, Z_i = b)$. (For more on maximal couplings, see [4, Chapter 7.4].) Define the upper triangular *coupling matrix* $\mathbf{\Theta} \in \mathbb{R}^{n \times n}$ as

$$\theta_{i,j} \triangleq \sup_{\mathbf{z} \in \mathcal{Z}^{i-1},\ a,b \in \mathcal{Z}} \Pi^{\mathbf{z}}_{i,a,b}[Z^{(1)}_j \neq Z^{(2)}_j]$$

for all $i < j$, $\theta_{i,j} \triangleq 1$ for $i = j$, and zero elsewhere. For a random field $\mathbf{Z}$ on a given graph G, denote by $\mathbf{\Theta}_G$ the coupling matrix of the conditional distributions induced by the topology of G. Finally, recall the standard definition of the matrix infinity norm, $\|\mathbf{\Theta}\|_\infty \triangleq \sup_{i\in[n]} \sum_{j=1}^n |\theta_{i,j}|$. With these definitions in mind, we present the following adaptation of [3, Theorem 1] and [7, Theorem 1.1].

Theorem 1. *If there exists a constant c such that, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$ that differ only at a single coordinate, $|f(\mathbf{z}) - f(\mathbf{z}')| \leq c/n$, then for any $\epsilon > 0$,*

$$\mathbb{P}\left\{f(\mathbf{Z}) - \mathbb{E}[f(\mathbf{Z})] \geq \epsilon\right\} \leq \exp\left(-2n\epsilon^2/(c\,\|\mathbf{\Theta}\|_\infty)^2\right).$$

3 Generalization

In this section, we prove *probably approximately correct* (PAC) generalization bounds for collective inference. We begin with some definitions. For the following, let $\mathcal{F}$ be an arbitrary class of functions from $\mathcal{Z}^n$ to $\mathbb{R}^n$.

Definition 3. Let $\mathbf{Z}$ be a random field on a graph $G \in \mathcal{G}_n$. Let $\boldsymbol{\sigma} \sim \text{Bin}(n, 1/2)$ be a set of *Rademacher variables*. Define the *empirical Rademacher complexity* of $\mathcal{F}$ as

$$\mathfrak{R}(\mathcal{F}, \mathbf{Z}) \triangleq \mathbb{E}_G\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathbf{Z})_i\right| \mid \mathbf{Z}\right].$$

Define the *Rademacher complexity* of $\mathcal{F}$, w.r.t. realizations of $\mathbf{Z}$, as $\overline{\mathfrak{R}}_G(\mathcal{F}) \triangleq \mathbb{E}_G[\mathfrak{R}(\mathcal{F}, \mathbf{Z})]$.

This differs from the traditional definition [1] by the form of $\mathcal{F}$ and the fact that $Z_1, \ldots, Z_n$ are not assumed to be i.i.d. Stability is a property of algorithms which ensures that small changes to the input result in bounded variation in the output. *Collective* stability applies this concept to vector-valued functions.

Definition 4. We say that $\mathcal{F}$ has *uniform collective stability* β if, for any two inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$ that differ only at a single coordinate, $\sup_{f\in\mathcal{F}} \|f(\mathbf{z}) - f(\mathbf{z}')\|_1 \leq \beta$.

To generalize our results to a variety of loss functions, we will use the following properties.

Definition 5. A loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ is *M-bounded* if, for any $y, y' \in \mathcal{Y}$ and $\hat{y}, \hat{y}' \in \hat{\mathcal{Y}}$, $|\ell(y, \hat{y}) - \ell(y', \hat{y}')| \leq M$.

Definition 6. A loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$ is *λ-admissible* if, for any $y \in \mathcal{Y}$, and any $\hat{y}, \hat{y}' \in \hat{\mathcal{Y}}$, $|\ell(y, \hat{y}) - \ell(y, \hat{y}')| \leq \lambda\,|\hat{y} - \hat{y}'|$.

We now state our main result.

Theorem 2. *If $\mathcal{H}$ has uniform collective stability β, and ℓ is M-bounded and λ-admissible, then, for any $n \geq 1$, any $G \in \mathcal{G}_n$, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over realizations of a random field $\mathbf{Z}$ on G, every $h \in \mathcal{H}$ satisfies*

$$\overline{L}_G(h) \leq L(h, \mathbf{Z}) + 2\lambda\mathfrak{R}(\mathcal{H}, \mathbf{Z}) + (M + 3\lambda\beta)\,\|\mathbf{\Theta}_G\|_\infty \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{1}$$

Using Lemma 1, we can directly apply Theorem 2 to the traditional structured prediction setting, in which the training set consists of multiple independent examples.

Corollary 1. *Let $\mathbf{Z}$ be a random field on a graph G. Let $\mathbf{Z}'$ be a random field on $G' \triangleq \bigcup_{j=1}^m G_j : G_j \simeq G$, representing m realizations of $\mathbf{Z}$. Let $\mathcal{H}$ and ℓ be as in Theorem 2, and further let $\mathcal{H}$ be graph-based. Then, for any $m \geq 1$, any $n \geq 1$, any $G \in \mathcal{G}_n$, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over realizations of $\mathbf{Z}'$, every $h \in \mathcal{H}$ satisfies*

$$\overline{L}_G(h) \leq L(h, \mathbf{Z}') + 2\lambda\mathfrak{R}(\mathcal{H}, \mathbf{Z}') + (M + 3\lambda\beta)\,\|\mathbf{\Theta}_G\|_\infty \sqrt{\frac{\ln(2/\delta)}{2mn}}.$$

Note that $\|\mathbf{\Theta}_{G'}\|_\infty = \|\mathbf{\Theta}_G\|_\infty$ because $\mathbf{\Theta}_{G'}$ is block diagonal.

We will prove Theorem 2 via a series of technical lemmas. Due to space restrictions, we will omit the proofs.

Lemma 2. *If $\mathcal{F}$ has uniform collective stability β, and ℓ is M-bounded and λ-admissible, then $\ell \circ \mathcal{F}$ has uniform collective stability $M + \lambda\beta$.*

For any particular $f \in \mathcal{F}$, let $F(\mathbf{Z}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{Z})_i$, and $\overline{F}_G \triangleq \mathbb{E}_G[F(\mathbf{Z})]$, and define the functions $\Phi(\mathcal{F}, \mathbf{Z}) \triangleq \sup_{f\in\mathcal{F}} \overline{F}_G - F(\mathbf{Z})$, and $\overline{\Phi}_G(\mathcal{F}) \triangleq \mathbb{E}_G[\Phi(\mathcal{F}, \mathbf{Z})]$.

Lemma 3. *If $\mathcal{F}$ has uniform collective stability β, then, for any $n \geq 1$, any $G \in \mathcal{G}_n$, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over realizations of a random field $\mathbf{Z}$ on G,*

$$\Phi(\mathcal{F}, \mathbf{Z}) \leq \overline{\Phi}_G(\mathcal{F}) + \beta \left\|\mathbf{\Theta}_G\right\|_\infty \sqrt{\ln(1/\delta)/(2n)}.$$

Lemma 4. *For a random field $\mathbf{Z}$ on a graph $G \in \mathcal{G}_n$, we have that $\overline{\Phi}_G(\mathcal{F}) \leq 2\overline{\mathfrak{R}}_G(\mathcal{F})$.*

Lemma 5. *If $\mathcal{F}$ has uniform collective stability β, then, for any $n \geq 1$, any $G \in \mathcal{G}_n$, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over realizations of a random field $\mathbf{Z}$ on G,*

$$\overline{\mathfrak{R}}_G(\mathcal{F}) \leq \mathfrak{R}(\mathcal{F}, \mathbf{Z}) + \beta \left\|\mathbf{\Theta}_G\right\|_\infty \sqrt{\ln(1/\delta)/(2n)}.$$

We are now ready to prove Theorem 2.

Proof [Theorem 2] We start with the simple observation that

$$\overline{L}_G(h) \leq L(h, \mathbf{Z}) + \sup_{h'\in\mathcal{H}} [\overline{L}_G(h') - L(h', \mathbf{Z})] = L(h, \mathbf{Z}) + \Phi(\mathcal{F}, \mathbf{Z}).$$

where we let $\mathcal{F} \triangleq \ell \circ \mathcal{H}$. By Lemma 2, $\mathcal{F}$ has uniform collective stability $M + \lambda\beta$. We therefore have from Lemma 3 that, with probability at least $1 - \delta/2$,

$$\overline{L}_G(h) \leq L(h, \mathbf{Z}) + \overline{\Phi}_G(\mathcal{F}) + (M + \lambda\beta) \left\|\mathbf{\Theta}_G\right\|_\infty \sqrt{\ln(2/\delta)/(2n)}.$$

To bound $\overline{\Phi}_G(\mathcal{F})$, we apply Lemma 4 and Talagrand's contraction lemma [9] (since ℓ is λ-Lipschitz w.r.t. its second argument); this yields $\overline{\Phi}_G(\mathcal{F}) \leq 2\overline{\mathfrak{R}}_G(\mathcal{F}) = 2\overline{\mathfrak{R}}_G(\ell \circ \mathcal{H}) \leq 2\lambda\overline{\mathfrak{R}}_G(\mathcal{H})$. Using Lemma 5 (with uniform collective stability β, since we are now dealing with $\mathcal{H}$), we have that, with probability at least $1 - \delta/2$,

$$2\lambda\overline{\mathfrak{R}}_G(\mathcal{H}) \leq 2\lambda\mathfrak{R}(\mathcal{H}, \mathbf{Z}) + 2\lambda\beta \left\|\mathbf{\Theta}_G\right\|_\infty \sqrt{\ln(2/\delta)/(2n)}.$$

Via De Morgan's law and the union bound, all of these bounds hold with probability at least $1 - \delta$, so we combine them to complete the proof. ■

Since binary classification is a common prediction task, we can apply Theorem 2 to measurements of the 0-1 loss, which we denote by a superscript $\mathbb{1}$.

Lemma 6. *The 0-1 loss $\ell_{\mathbb{1}}$ is 1-bounded and $(1/2)$-admissible.*

4 Discussion

To achieve the usual $\mathrm{O}(1/\sqrt{n})$ uniform convergence, we require that $\mathfrak{R}(\mathcal{H}, \mathbf{Z}) = \mathrm{O}(1/\sqrt{n})$, $\beta = \mathrm{O}(1)$ and $\left\|\mathbf{\Theta}_G\right\|_\infty = \mathrm{O}(1)$. In this section, we discuss the circumstances under which these conditions are satisfied.

The empirical Rademacher complexity of many popular hypothesis classes has been thoroughly studied, though not for collective models. Intuitively, the complexity of relational MRFs should be similar to that of linear predictors, which is $\mathrm{O}(1/\sqrt{n})$ [6], since the log likelihood is simply a linear combination of feature functions. Weiss and Taskar [15] offer some insight into this analysis.

It remains to be proven whether popular collective models exhibit acceptable collective stability. Clearly, for non-collective models, $\beta = \mathrm{O}(1)$, though these hypotheses are of little interest to the structured prediction setting. For collective models—in particular, graph-based models—it is reasonable to assume that influence, as a function of graph distance, decays at a geometric rate. Thus, assuming a bounded neighborhood size, the effect of changing any single node should converge to a constant.

Finally, our bounds require that the infinity norm of the coupling matrix be bounded independent of n. Chazottes et al. [3] identify "low-temperature" Ising models as an example of processes that satisfy this condition. It is an open question whether the same can be shown for broader classes of random fields. Intuitively, their analysis should hold for any random field that exhibits geometric strong mixing.

References

[1] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, March 2003.

[2] P. Bartlett, Collins. M., B. Taskar, and D. McAllester. Exponentiated gradient algorithms for large-margin structured classification. In *Advances in Neural Information Processing Systems 17*, pages 113–120, 2004.

[3] J. Chazottes, P. Collet, C. Külske, and F. Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137:201–225, 2007.

[4] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[5] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 593–598, 2004.

[6] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800, 2008.

[7] L. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6):2126–2158, 2008.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[9] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, 1991.

[10] D. McAllester. Generalization bounds and consistency for structured labeling. In G. Bakir, T. Hofmann, B. Scholkopf, A. Smola, B. Taskar, and S. Vishwanathan, editors, *Predicting Structured Data*. MIT Press, 2007.

[11] J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the Workshop on Statistical Relational Learning, 17th National Conference on Artificial Intelligence*, pages 42–49, 2000.

[12] P. Sen, G. Mark Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[13] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 485–492, 2002.

[14] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems 16*, 2004.

[15] D. Weiss and B. Taskar. Structured prediction cascades. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 916–923, 2010.

[16] R. Xiang and J. Neville. Relational learning with one network: An asymptotic analysis. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 779–788, 2011.