Graph Reconstruction with Degree-Constrained Subgraphs

Stuart J. Andrews **Tony Jebara**
Computer Science, Columbia University
New York, NY 10027

Given observations about a collection of nodes, the goal of graph reconstruction is to predict a set of edges that connect the nodes in a realistic fashion. A number of similar formulations of the problem have been introduced across research areas, notably social sciences, epidemiology and biology. What is common across domains is the understanding that interesting properties of the system are known to depend on the graph structure.

As a example, consider the cell signaling network previously studied in [4, 2]. Gene expression modulations are a by-product of a living network of activity. In this application, the expression levels of 11 signaling molecules have been measured using a technology called flow cytometry, and the goal is to recovery the causal influences of one molecule on another in this network. An important aspect of this data is the inclusion of interventions, whereby external factors are used to perturb the expressions of individual molecules, as these help to identify the directionality of influence between two molecules.

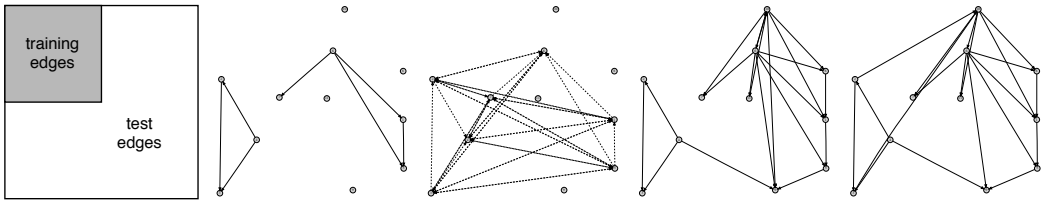
The authors use Bayesian networks to infer the causal relations. In this framework, the expression levels of signaling molecules in a cell are modeled as discrete random variables whose dependency structure is assumed to form a directed acyclic graph. Then, using this well-established probabilistic model, they demonstrate with remarkable success how structure learning is able to explain the causal relationships in an unsupervised fashion. The authors point out two limitations of their approach: 1) the inferred dependency graph is acyclic, and 2) effective inference requires many observations.

We adopt an alternative approach to analyze this cell signaling data based on the structured outputs framework [1]. In this setting $\mathbf{X} \in \mathbb{X}$ is an input feature vector representation of an object and $\mathbf{Y}$ is a multi-variate output that specifies the structure of the object. Note that $\mathbf{Y}$ is restricted to a combinatorial family of structures $\mathbb{Y}$ defined for each application, where $\mathbb{Y} \subsetneq \{0,1\}^N$. Common structures and the applications that use them, listed in parentheses, include: chains (label-sequences), trees (parse trees), matchings (word and sequence alignments), and partitions of graphs (image segmentations).

Due to the natural occurrence of cycles in biological networks, which are not easily modeled with chains, trees, matchings etc., we have chosen to combine *degree-constrained subgraphs* (DCS) with the structured outputs framework. DCS are described mathematically in terms of n^2 structure variables $y_{j,k}$ corresponding to the graph's 0-1 adjacency matrix. A graph is degree constrained if $\sum_j y_{j,k} = \boldsymbol{\delta}_k^{in}, \forall k$, and $\sum_k y_{j,k} = \boldsymbol{\delta}_j^{out}, \forall j$, where the $2n$ constants $\boldsymbol{\delta}_k^{in}, \boldsymbol{\delta}_j^{out}$ are node in-degrees and out-degrees that are assumed to be known. DCS have received considerable attention in operations research due to their attractive computational properties; namely, there is a $\mathcal{O}\left(n^3\right)$ algorithm that identifies the maximum-weight DCS: $\operatorname{argmax}_{\mathbf{Y} \in \text{DCS}} \sum_{j,k} y_{j,k} s_{j,k}$, given a set of weights $s_{j,k}$.

Using DCS in this framework allows us to learn a function that maps intervention data into a causal graphs. First, we parametrize the edge weights $s_{j,k} = \mathbf{w}^T \mathbf{x}_{j,k}$ using a linear parameter $\mathbf{w}$. Then, as in standard learning to predict structured outputs, a risk minimization framework is typically employed, leading to a max-margin problem formulation. For details see [1].

Flow-cytometry data was collected and discretized into 3 levels by [4]. The data is summarized by two 11×5400 matrices. The first contains discrete expression levels 1-3, and the second contains



(a) training edges (b) positive training (c) negative training (d) our completion (e) true completion

Figure 1: The train-test partition of the adjacency matrix is depicted in (a). Most of the entries in the training subset will be 0, which we call negative edges. The positive and negative edges used for training are shown in (b) and (c). The completed graph after learning (d) and the ground truth in (e). False-positives are demarked by open arrows.

T	2700	1350	675	364	*i.i.d.* SVM T=2700
auc (recall)	0.97 (80)	0.96 (79)	0.96 (77)	0.96 (77)	0.94 (73)

Table 1: Results of structured-outputs model averaged over 5 repeated trials. A trial randomly selects two disjoint $11 \times T$ slices of the expression data for training and for testing. The final column is the performance of an *i.i.d.* support vector machine trained on the edge features.

the assumed perfect intervention state. We use pairwise flow-cytometry measurements to describe each possible directed edge $j \rightarrow k$ by a feature vector $\mathbf{x}_{j,k}$. Our representation is: 1) invariant to the ordering and number of flow-cytometry measurements; 2) sensitive to correlations but invariant to their sign; and 3) sensitive to both parent and child interventions, under the assumption that they are not coincident. The second and third conditions are imposed so that we can predict causal influences, and not simply correlations, between signaling molecules. To construct $\mathbf{x}_{j,k}$ we first remove pairs where both molecules have been intervened. Next, assuming the expression levels for j and k are positively correlated on the non-intervened measurements, we construct 3x3 normalized histograms over all remaining measurements. We use a separate histogram for each of the three intervention cases: $j \wedge k$, $\neg j \wedge k$, and $j \wedge \neg k$, for a total of 27 dimensions. If the expressions levels are negatively correlated on the non-intervened measurements, the expression levels of the child node are flipped before constructing the histogram features.

For our first experiment, the goal was to see if we could generalize across different subsets of the expression data. We randomly sampled two disjoint subsets of size T from the original 5400 measurements, for training and testing. Using structured output learning on the training sample, we found the weight vector $\mathbf{w}$ that robustly separated the true network from the remaining DCS. Switching to the testing sample, and computing all new edge features $\mathbf{x}_{j,k}$, we computed the maximum-weight DCS for comparison. The results included in Table 1 show that we can generalize well from sample to sample, and that performance does not degrade for small sample sizes. In fact, the average recall of 77% can likely be attributed wholly to the degree constraints. For comparison, we have included the performance of an *i.i.d.* SVM on subsets of size T=2700. Note that the true network is *not used* during testing, although it is used during training. Also note that the node degrees of the true network are used during both training and testing to define the set of DCS.

For our second experiment, we tested whether we could complete a network that was partially observed, this time using all 5400 samples. We followed the protocol of [5], using a 2:1 bipartite split of the nodes (see Figure 1). Here, we discovered that we could complete the network with an average area under the curve (AUC) of 0.93 and recall of 80%.

This work combines DCS with the structured outputs framework for learning to complete graphs. This model is able to learn cycles, and appears to be robust to small sample sizes. Therefore, this approach is promising in situations where degree information is available, which is not that unusual. Two examples are: 1) the social networking web site LinkedIn (www.linkedin.com), and 2) the structural interaction network (SIN) of [3]. Further comparisons of the structured outputs model and the Bayesian structure learning are needed.

References

[1] G. Bakir, T. Hofmann, B. Schölkopf, and S. V. N. Vishwanathan. *Predicting Structured Data*. MIT Press, Cambridge, Massachusetts, 2007.

[2] D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics, AISTATS*, March 2007.

[3] P.M. Kim, L.J. Lu, Y. Xia, and M.B. Gerstein. Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science*, 314(5807):1938, 2006.

[4] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721):523–529, 2005.

[5] Y. Yamanishi, J. P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 2004.