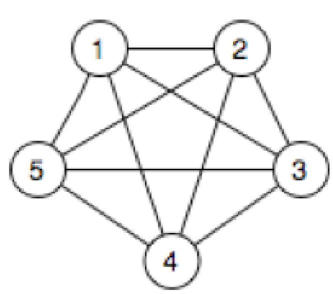
Using Combinatorial Optimization within Max-Product Belief Propagation

Danny Tarlow
Toronto Machine Learning Group
Oct 2, 2006

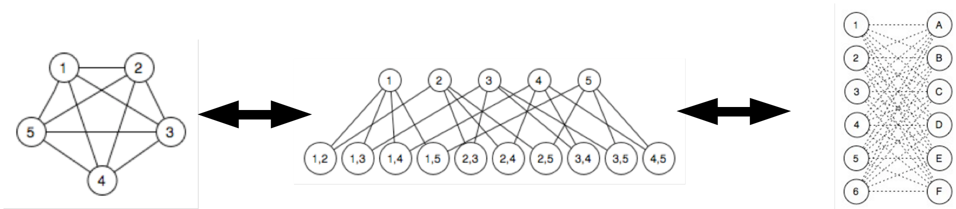
Based on work to be presented at NIPS 06
with John Duchi, Gal Elidan, and Daphne Koller

Motivation

- Markov Random Fields (MRFs) are a general framework for representing probability distributions.
- An important type of query is the *maximum a posteriori* (MAP) query – find the most likely assignment to all variables



Equivalent Representations of Bipartite Matching



MRF ⟷ Cluster Graph ⟷ Bipartite Matching

- Certain problems can be formulated both as a MAP query in an MRF and as a combinatorial optimization problem.
  - MRFs with only *regular* potentials can be formulated as mincut problems.
  - MRFs with only singleton potentials and pairwise mutual-exclusion potentials can be formulated as bipartite matching problems.

Equivalence of MRF and Bipartite Matching

MRF

- MAP problem – find the assignment of values to variables such that the product of their potentials is maximized:

$$\max_{\mathbf{x}\in\mathbf{X}}\left(\prod_i \pi_i(x_i)\right)\cdot\left(\prod_{i,j}\pi_{i,j}(x_i,x_j)\right)$$

$$\max_{\mathbf{x}\in\mathbf{X}}\left(\prod_i \pi_i(x_i)\right) \text{ st } \forall i,j.(i\neq j)\Rightarrow(x_i\neq x_j)$$

$$\max_{\mathbf{x}\in\mathbf{X}}\log\left(\prod_i \pi_i(x_i)\right) \text{ st } \forall i,j.(i\neq j)\Rightarrow(x_i\neq x_j)$$

$$\max_{\mathbf{x}\in\mathbf{X}}\left(\sum_i \log \pi_i(x_i)\right) \text{ st } \forall i,j.(i\neq j)\Rightarrow(x_i\neq x_j)$$

Bipartite Matching

- Maximum weight problem – find the assignment of values to variables such that the sum of the edge weights is maximized

$$\max_{\mathcal{E}}\sum_{e_{ij}\in\mathcal{E}} w(e_{ij}) \text{ st } \forall i,j,k.(e_{ij}\in\mathcal{E}\wedge j\neq k)\Rightarrow e_{ik}\notin\mathcal{E}$$

Set the edge weights in the bipartite matching to be the log of the singleton potentials in the MRF, and both are maximizing the same objective.

Equivalence of MRF and Minimum Graph Cut

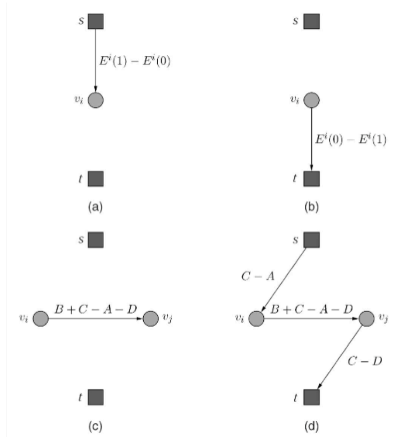
- Similarly, an MRF with only *regular* potentials can be transformed such that MAP inference can be performed by finding a minimum weight graph cut
  - V. Kolmogorov, R. Zabih. “What energy functions can be minimized via graph cuts?” *ECCV 02*.

TABLE 1

$$E^{i,j} = \begin{array}{|c|c|}\hline E^{i,j}(0,0) & E^{i,j}(0,1) \\ \hline E^{i,j}(1,0) & E^{i,j}(1,1) \\ \hline \end{array} = \begin{array}{|c|c|}\hline A & B \\ \hline C & D \\ \hline \end{array}$$

TABLE 2

$$\begin{array}{|c|c|}\hline A & B \\ \hline C & D \\ \hline \end{array} = A + \begin{array}{|c|c|}\hline 0 & 0 \\ \hline C-A & C-A \\ \hline \end{array} + \begin{array}{|c|c|}\hline 0 & D-C \\ \hline 0 & D-C \\ \hline \end{array} + \begin{array}{|c|c|}\hline 0 & B+C-A-D \\ \hline 0 & 0 \\ \hline \end{array}$$



Combinatorial Optimization for MAP Inference

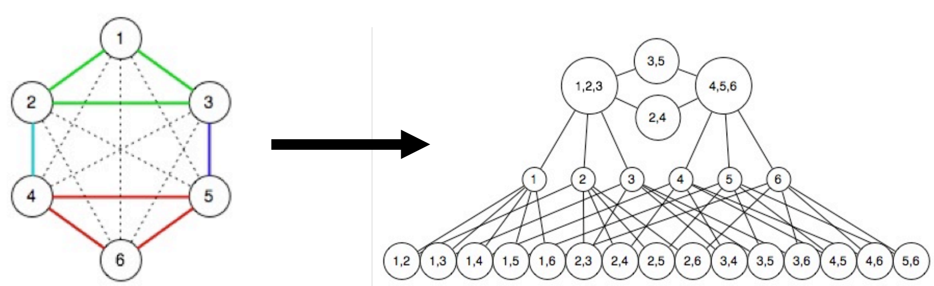
- Moreover, the special purpose formulations allow for more powerful inference algorithms.
  - Mincut based methods for solving regular MRFs outperform traditional inference techniques like Loopy Belief Propagation.
    - R. Szeliski, R. Zabih, et. al. “A comparative study of energy minimization methods for Markov random fields.” ECCV 06.
  - This is also the case with bipartite matching problems.
    - BP doesn't deal well with hard mutual exclusion constraints.

Combinatorial Optimization for MAP Inference

- Why do we care?
  - Combinatorial algorithms are used widely in AI
    - Correspondences (SFM, some tracking problems, object recognition, NLP frame assignment, etc.)
    - Graph cuts (image segmentation, protein-protein interactions, etc.)
  - And for problems that can be formulated as combinatorial optimization problems, the combinatorial formulation often yields the best results

Problem

- Many complex, real-world problems have combinatorial sub-components, but they also have large components that cannot be expressed in a purely combinatorial framework.



Augmented Matching MRF

Augmented Matching Cluster Graph

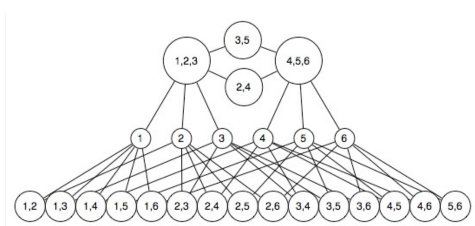
Model to Image Correspondence for Object Recognition

Geremy Heitz, Gal Elidan, Daphne Koller, “Learning Object Shape: From Drawings to Images”. *CVPR 06*.

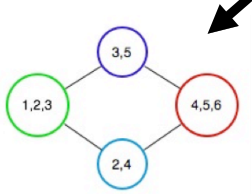
- Two types of constraints
- **Matching**: how well pixel neighborhoods match + mutual exclusion
- **Geometric**: landmarks should be arranged in a shape that looks like a car

How do you use combinatorial algorithms now?

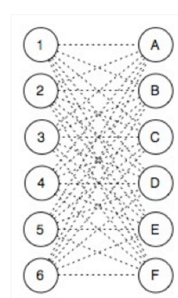
Try Partitioning the Graph?



Original Cluster Graph



+

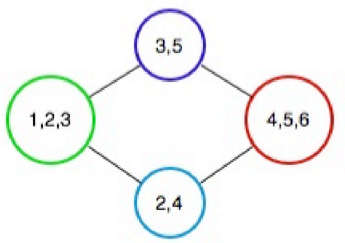


Simpler Cluster Graph

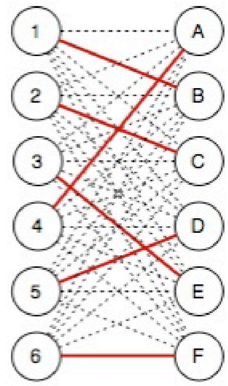
Bipartite Graph

Attempt at Partitioning

- Each component can be solved efficiently alone



Loopy Belief Propagation



Bipartite Matching

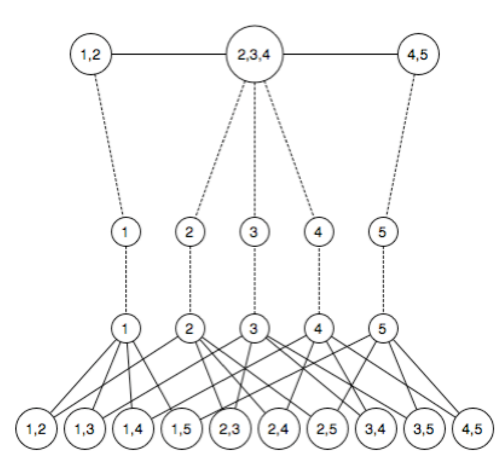
Failure of Partitioning

- We now have two simple subgraphs in which we can do inference efficiently.
- Unfortunately, this doesn't help.
  - Bipartite matching only gives a single assignment.
  - Bipartite matching makes no attempt to quantify uncertainty.
- In order to function within Max-Product BP (MPBP), each subgraph must be able to compute not only the most likely assignment, but also the associated uncertainty.

Limitations of Combinatorial Optimization

- Combinatorial optimization does not work in the presence of non-complying potentials.
- There is some work on truncating non regular potentials in graphs that are *nearly* expressible as min cuts.
- Most often, the solution is to fall back to belief propagation over the entire network.

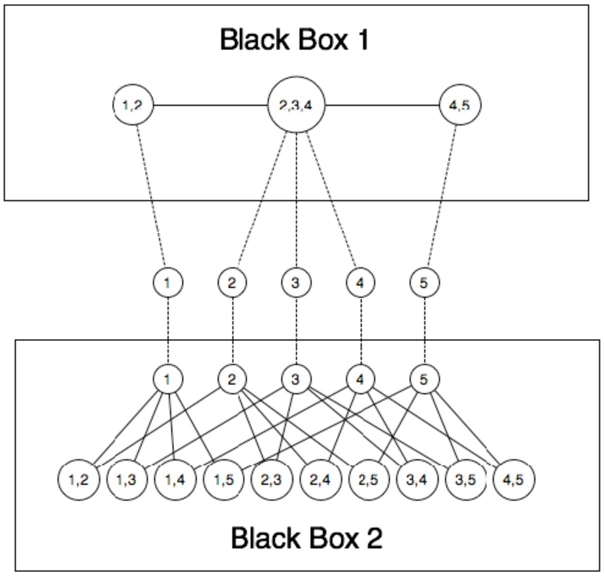
Falling Back to Belief Propagation



BP can handle non-complying potentials without a problem...

...but, must sacrifice the improved performance of combinatorial algorithms

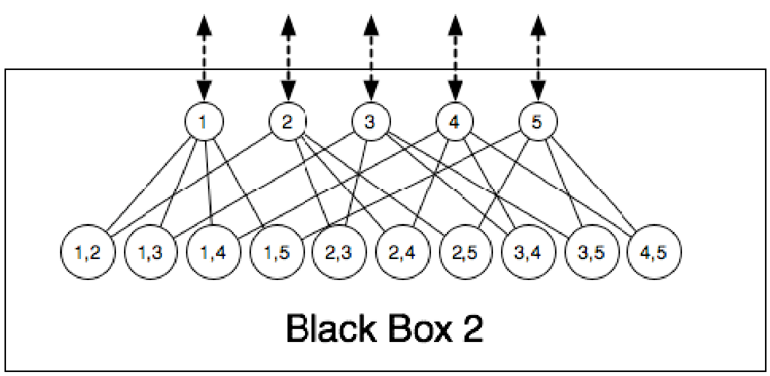
Partitioning: Attempt 2 – BP with Different Scheduling




```
Do until convergence of
interface messages {

    Run BP to convergence
    inside black box i

    Use the resulting
    beliefs to compute
    interface messages.

    Propagate interface
    messages to other
    black boxes

    i <- next black box
}
```

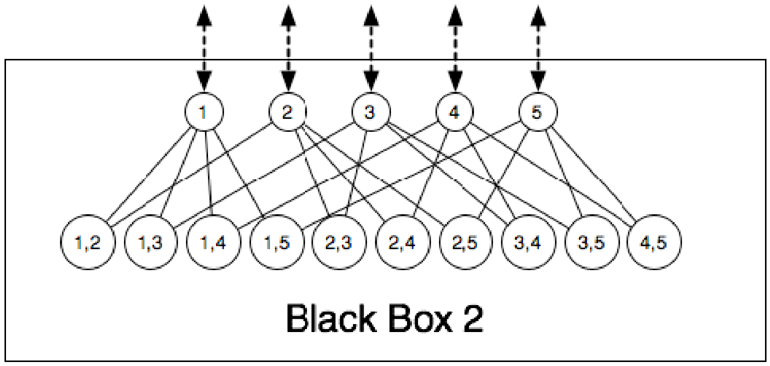
This still is belief propagation!

The Role of the Black Box



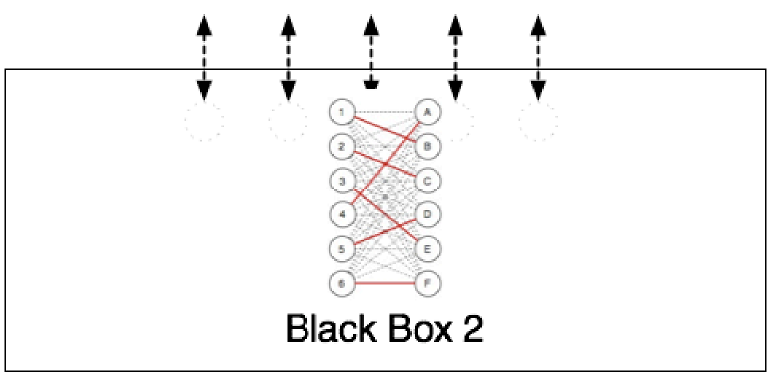
- The communication of the black box with the rest of the network is via the messages that it sends and receives
  - Beyond that, it doesn't matter how the messages are calculated

The Role of the Black Box



- This is a difficult subgraph to do belief propagation in, especially as *n* gets large.
  - Tree width is *n - 1*, very loopy
  - Deterministic mutual-exclusion potentials
- Often doesn't converge or converges to a poor solution.

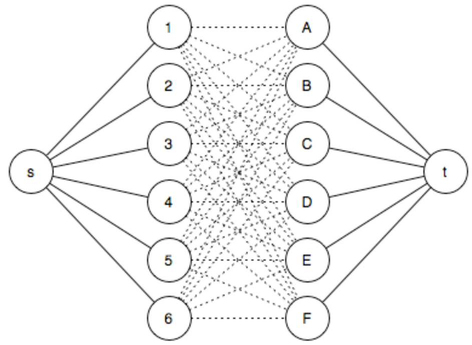
Using a Combinatorial Black Box



- Claim: we can compute exact max-marginals and do it significantly faster than BP by using dynamic graph algorithms for combinatorial optimization.
- The result is exactly MPBP, but faster and more accurate.

Review: Maximum Weight Bipartite Matching

- Problem: given a bipartite graph with weighted edges, maximize the sum of edge weights such that the maximum degree of any node is 1
- Find the maximum weight path in the residual graph. Augment. Repeat until there are no more paths from s to t.
- Include edge (i, j) if it is used in the final residual graph.
- This is guaranteed to be the optimal matching, with a weight of w*.

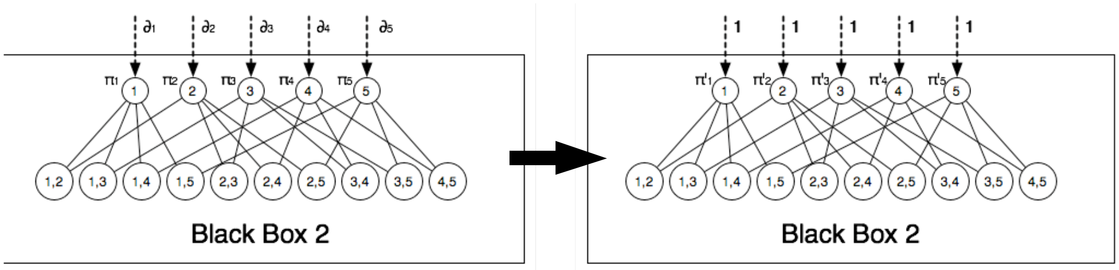


Max-Marginals as All-pairs Shortest Paths

- We need max-marginals
  - For all i, j, find the best score if that X_i is forced to take on value j
- This corresponds to forcing edge (i, j) to be used in the residual graph.
  - If i is matched to j already, then no change from w*.
  - If i is not matched to j, then the resulting weight will be less than or equal to w*.
- The difference is the cost of the shortest path from j to i in the residual graph.
- Negate all edges, then Floyd-Warshall all-pairs shortest paths to compute all max-marginals in $O(n^3)$ time.

Receiving Messages

- All clusters in an MRF's cluster graph must know how to receive messages.
- We need to modify our matching graph to reflect the messages we have received from other parts of the graph.



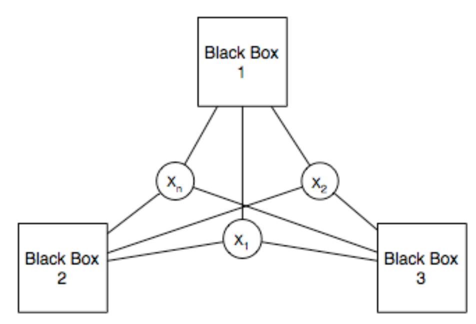
- Just multiply in the incoming messages and set weights in the matching problem to be $\pi'_i = \partial_i \times \pi_i$

Minimum Cut Problems

Minimum cut problems can be formulated analogously.

- Representing MRF MAP query as a min-cut problem
  - V. Kolmogorov, R. Zabih. "What energy functions can be minimized via graph cuts?" *ECCV 02*.
- Computing max-marginals
  - P. Kohli, P. Torr, "Measuring Uncertainty in Graph Cut Solutions – Efficiently Computing Min-marginal Energies Using Dynamic Graph Cuts." *ECCV 06.*
- Receiving messages
  - Same rule as for matchings (premultiply messages then convert modified MRF back to a min-cut)

More General Formulation



- We can perform MPBP in this network as long as each black box can accept *max-marginals* and compute *max-marginals* over the scope of each interface cluster.

Almost Done

- Just need a fancy acronym...

COMPOSE:

Combinatorial Optimization for Max-Product on Subnetworks

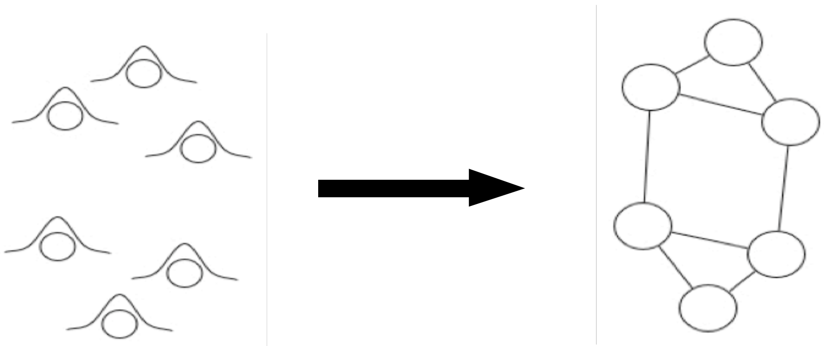
Experiments and Results

- Synthetic Data
  - Simulate an image correspondence problem augmented with higher order geometric potentials.
- Real Data
  - Electron microscope tomography: find correspondences between successive images of cells and markers for 3D reconstruction of cell structures.
    - Images are very noisy
    - Camera and cell components are both moving in unknown ways

Synthetic Experiment Construction

- Randomly generate a set of “template” points on a 2D plane.
- Sample one “image” point from Gaussian centered at each template point.
  - Covariance is $\sigma \mathbf{I}$, σ is increased to make problems more difficult
- Goal is to find a 1-to-1 correspondence between template points and image points.
- Two types of potentials
  - Singleton potentials uniformly generated on [0,1], but the true point is always given a value of .7
  - Pairwise geometric potentials, preferring that pairwise distances in template are preserved
    - Tried tree and line structures, both gave similar results.

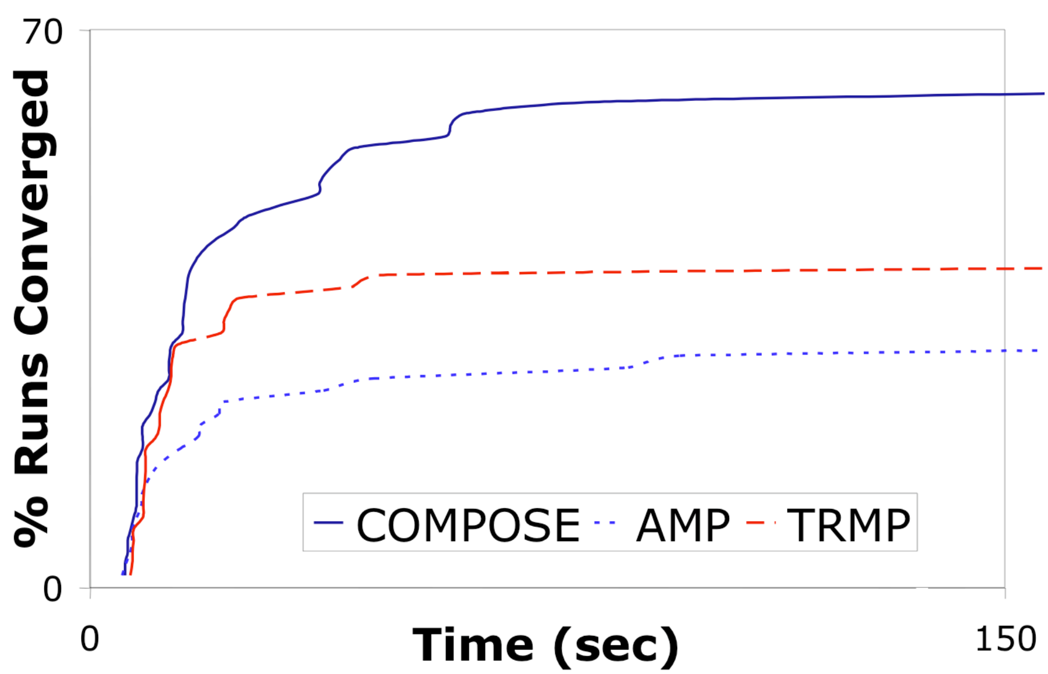
Synthetic Experiment Construction



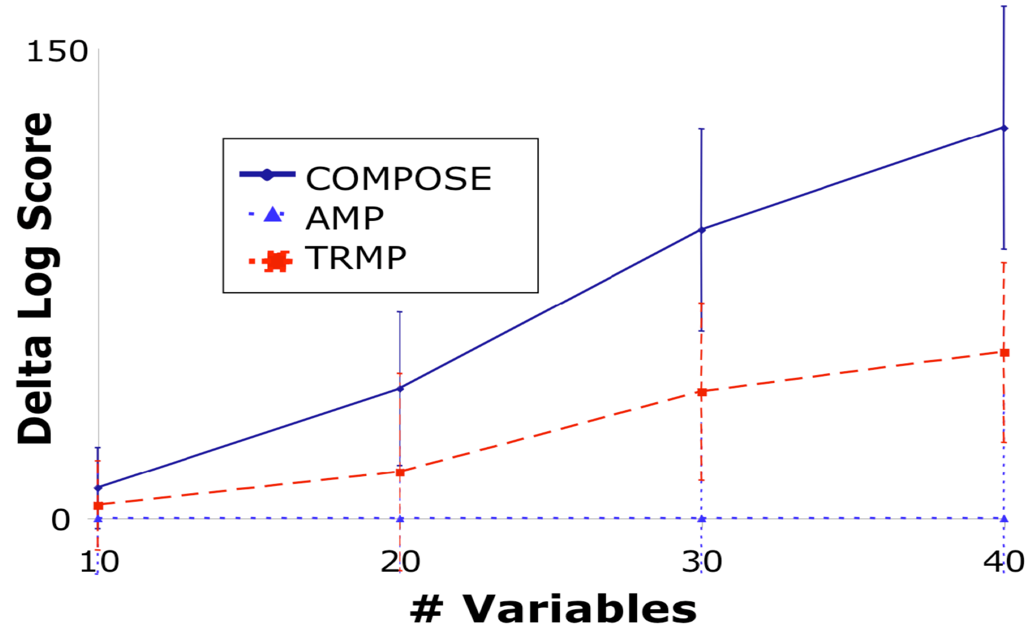
Template points randomly generated then sampled from

Random local potentials, plus geometric potentials to prefer similar relative geometry

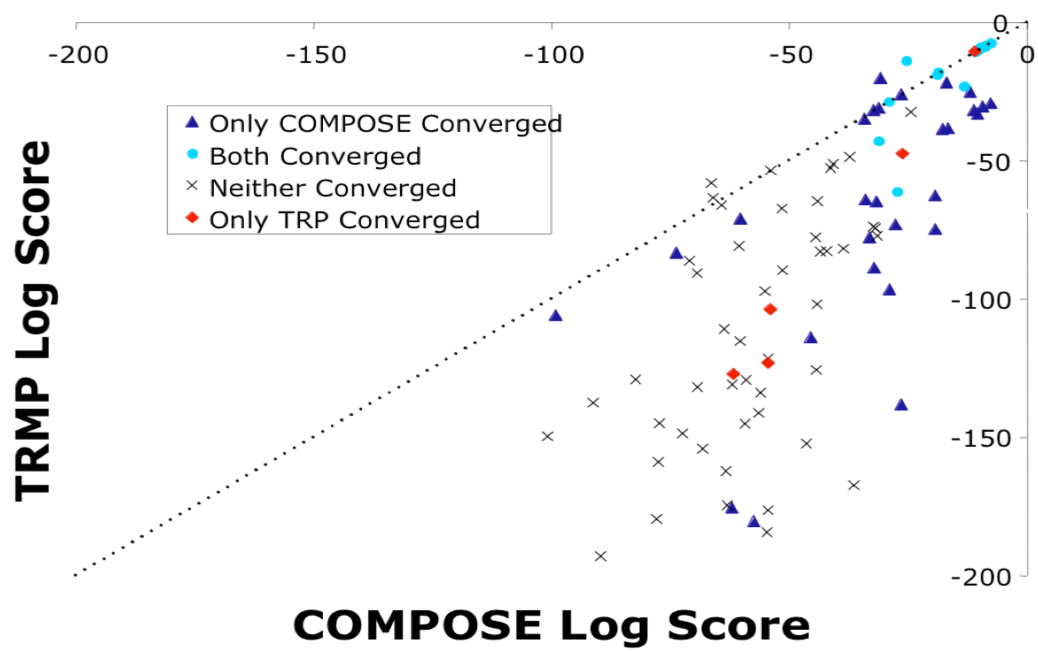
Covergence vs. Time (30 Variables)



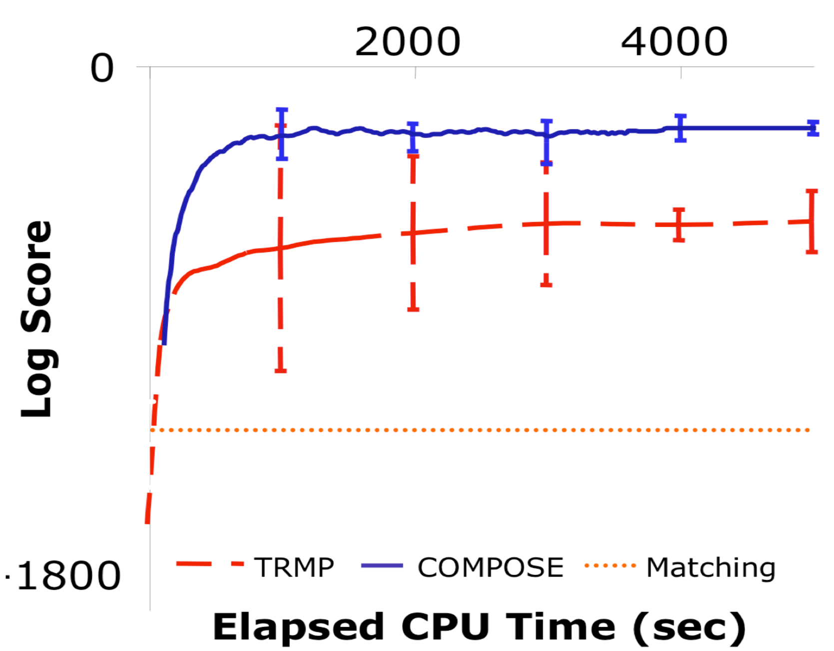
Score vs. Problem Size



Direct Comparison of COMPOSE vs. TRMP

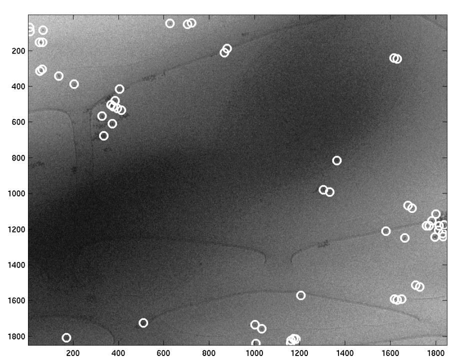


Score vs. Time (100 variables)

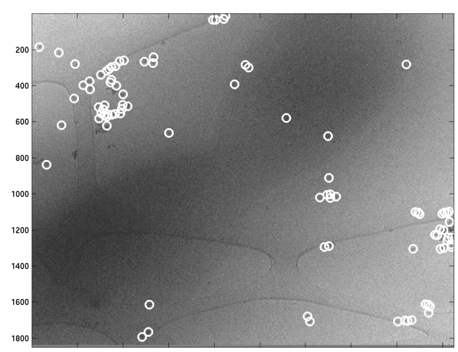


*COMPOSE and TRMP did not converge on any of these problems

Real Data Experiments



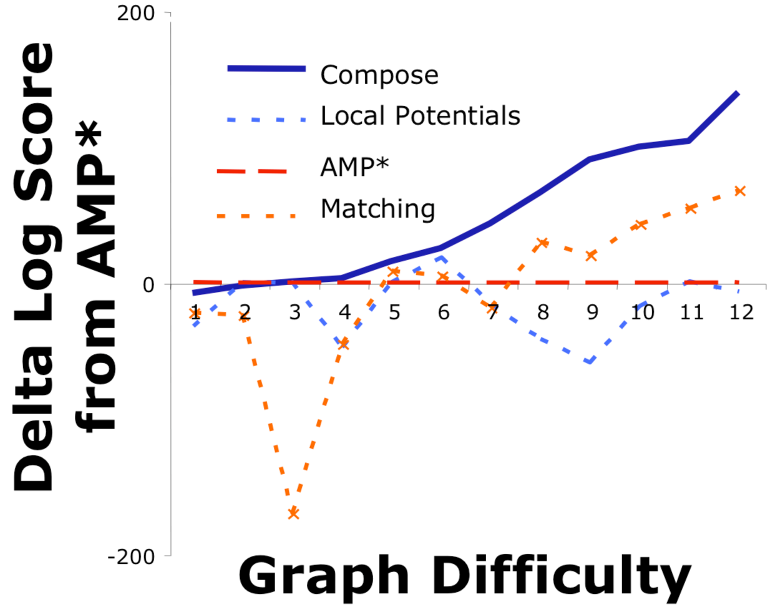
60 Markers



100 Candidates

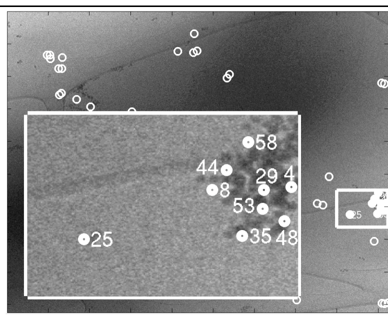
- Biologists preprocessed images to find points of interest
- Problem: find the new location of each left image markers in the right image
- Local potentials – pixel neighborhood + location
- Pairwise geometric potentials – minimum spanning tree

Real Data Scores on 12 Problems

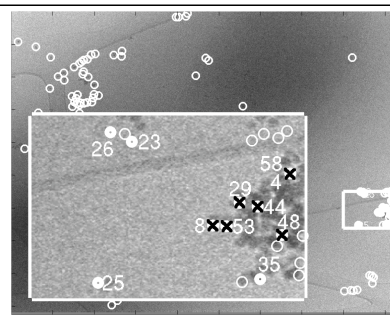


University of Toronto
Department of Computer Science
Machine Learning Group

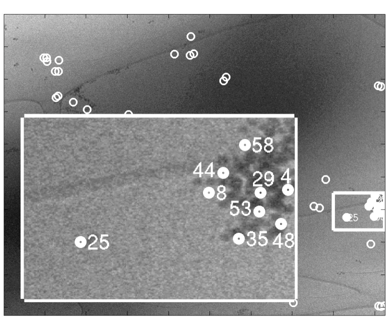
Real Data Results



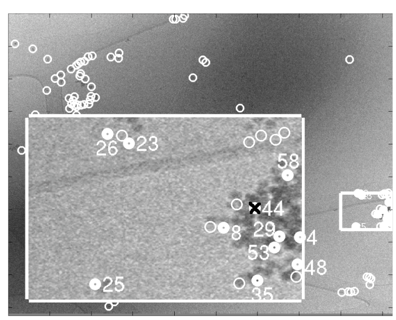
Initial Markers



AMP* Assignment



Initial Markers



COMPOSE Assignment

University of Toronto
Department of Computer Science
Machine Learning Group

Discussion

- All of these are problems where standard BP algorithms perform poorly
  - Small changes in local regions can have strong effects on distant parts of the network
  - Algorithms like TRMP try to address this by more intelligent message scheduling, but messages are still inherently local.
- COMPOSE slices along a different axis
  - Uses subnetworks that are global in nature but do not have all information about any subset of variables
  - Essentially gives a way of making global approximation about one network subcomponent.

Related Work

- Truncating non-regular potentials for mincut problems
  - Must be "close to regular"
  - C. Rother, S. Kumar, V. Kolmogorov, A. Blake. "Digital tapestry." *CVPR 05*.
- Quadratic Assignment Problem (QAP)
  - Encompasses augmented matching problem, but no (known) attempts to use combinatorial algorithms within a general inference procedure
- Exact inference using partitioning as a basis for A* heuristic
  - Our first attempt at solving these problems

Future Work

- Heavier theoretical analysis
  - Are there any guarantees about when we can provide a certain level of approximation?
- Are there other places where we can efficiently compute max-marginals?
  - What else can we do with belief propagation, given this more flexible view of what can be expressed and computed efficiently?

Thanks!

Questions or comments?