Scalable Latent Tree Model and its Application to Health Analytics

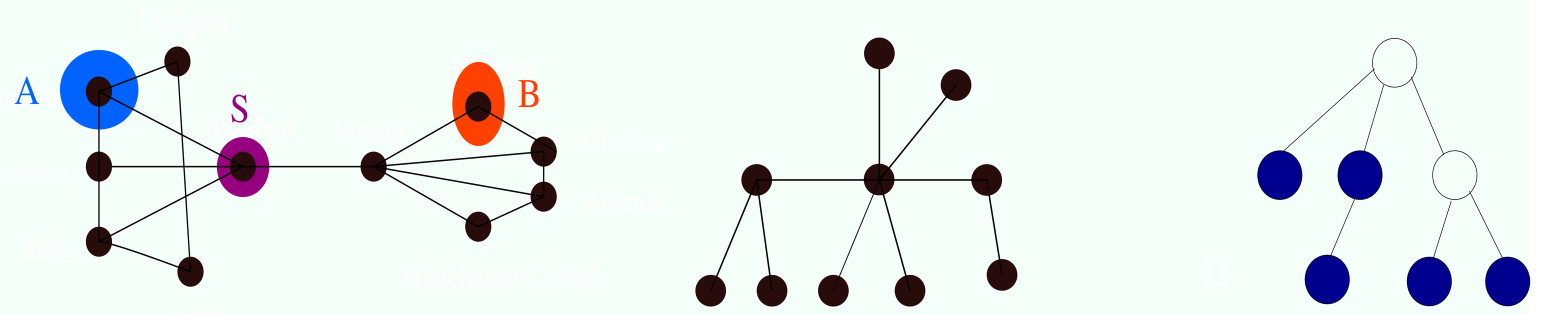
*Furong Huang, Niranjan U.N., Ioakeim Perros,
Robert Chen, Jimeng Sun, Anima Anandkumar

Summary

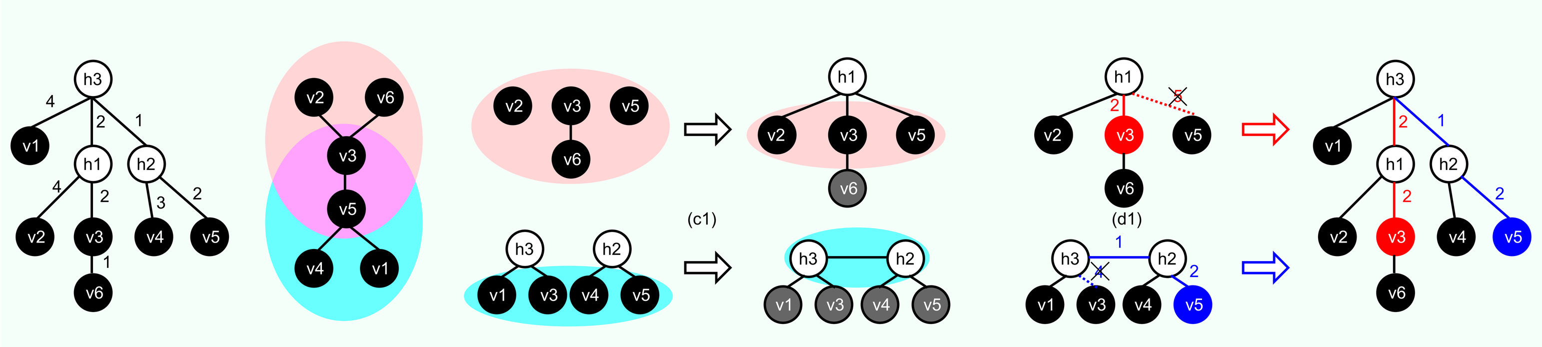
- **Goal:** Efficient learning in latent tree graphical models
- **Methods:** Hierarchical Tensor Decomposition
  - "Divide-and-conquer" strategy
    - Learns over small groups
    - Iteratively merges into global solution
- **Contribution:**
  - Guaranteed local method w. global consistency
  - Bulk asynchronous parallel algorithm
    - log(# variables); linear(dimension)
- **Application:** Human disease hierarchy
  - Generate clinically meaningful disease hierarchies
  - High degree of efficiency and accuracy on electronic health records

Why latent tree?

- Modeling Conditional Independencies through Graphs
  - Learning and inference are NP-hard
- Tractable Models: Tree Models
  - Efficient inference using belief propagation
  - Tree Structure Estimation: MLE $\equiv$ MST, pairwise statistics
- Latent tree graphical model: less restrictive than tree models
  - Number and location of hidden variables unknown
  - Versatile in modeling hierarchical relations



Overview of Approach



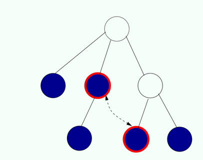
- **(a)** Ground truth latent tree
- **(b)** MST constructed using *information distances*
- Local estimation:
  - **(c1)** Local Recursive Grouping on nbd$[v_3, \text{MST}]$ to get local structure $\mathcal{N}_3$; **(c2)** Local Recursive Grouping on nbd$[v_5, \text{MST}]$ to get local structure $\mathcal{N}_5$
  - Local parameter estimation over triplets : tensor decomposition
- **(d1)(d2)** Merging local sub-trees for global estimation **(e)**

Structure Learning

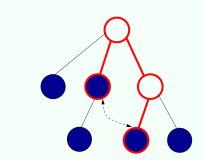
Additive Tree Distance: Information Distances $[d_{i,j}]$ for Tree Models

- Linear multivariate models:

$$d_{ij} := -\log \frac{\prod_{\sigma_i \neq 0} \sigma(\mathbb{E}[y_i y_j^\top])}{\sqrt{\det \mathbb{E}[y_i y_i^\top] \det \mathbb{E}[y_j y_j^\top]}}.$$



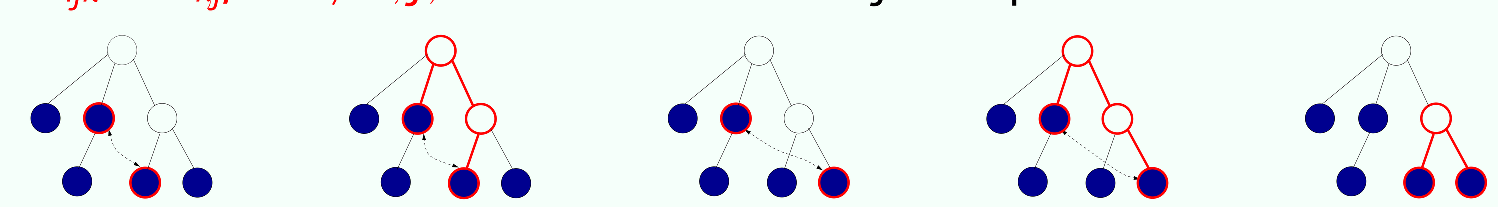
- $[d_{i,j}]$ is an additive tree metric: $d_{k,l} = \sum_{(i,j) \in \text{Path}(k,l;E)} d_{i,j}.$



$\boxed{\text{Learning latent tree using } \widehat{[d_{i,j}]}}$

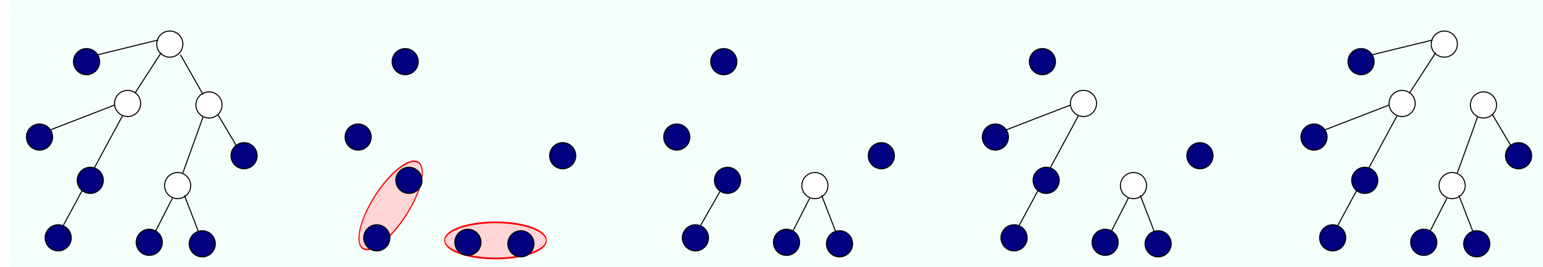
Siblings Test

- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j}\ \forall k, k' \neq i, j, \iff$ i, j leaves with common parent
- $\Phi_{ijk} = d_{i,j}, \forall k \neq i, j, \iff$ i is a leaf and j is its parent.



Recursive Grouping [Choi, Tan, Anandkumar, Willsky 2011]

- Sibling test and remove leaves
- Build tree from bottom up

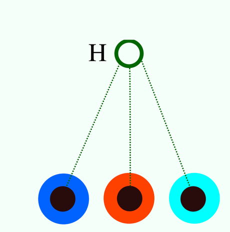


- Consistent structure estimation.
- Serial method, high computational complexity.

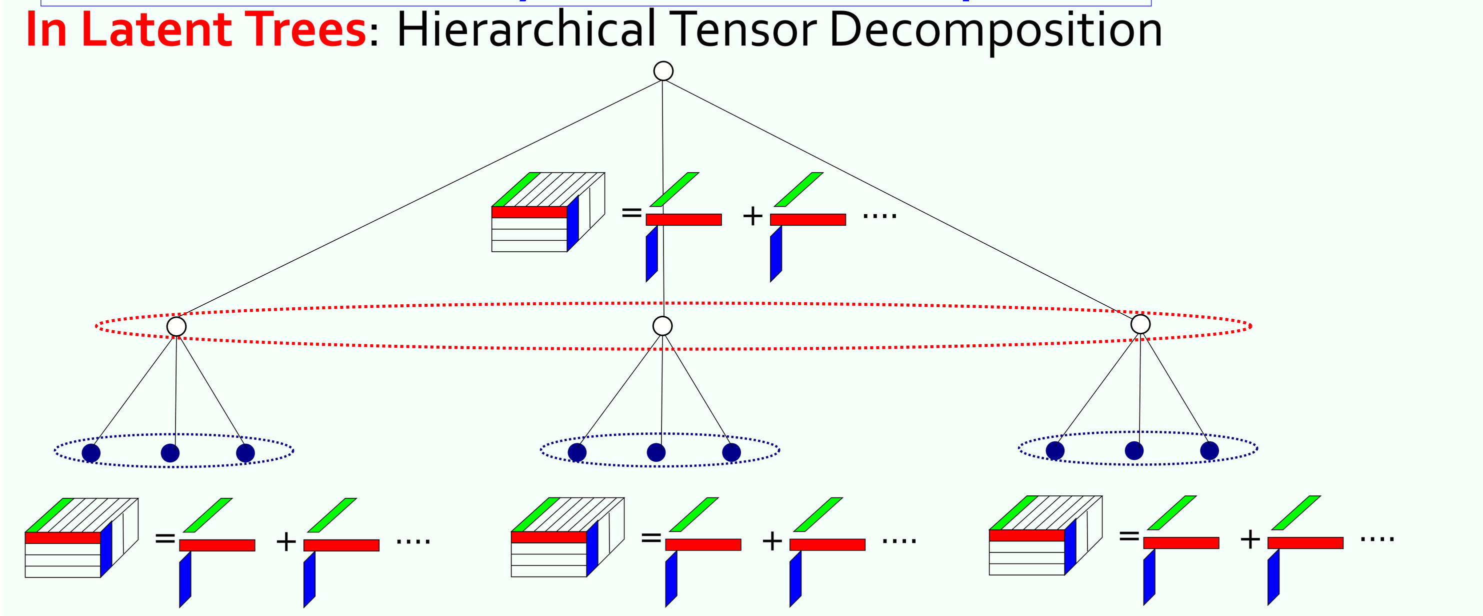
Parameter Learning through Tensor Methods

3-star Linear Multivariate Model

- Transition matrix consists of r components
  - $A_r^u := \mathbb{E}(u|h = e_r)$; $\lambda_r := \mathbb{P}[h = e_r]$.



Guaranteed Recovery via Tensor Decomposition
In Latent Trees: Hierarchical Tensor Decomposition



- Hidden labels permuted across different triplets.
- Solution: Align using common node in triplets

Integrating Structure and Parameter Learning

Divide and conquer

- Find (overlapping) groups of variables
- Learn local subtrees over the groups independently
- Merge subtrees/ tweak parameters $\rightarrow$ global latent tree model

Alignment Correction

- In-group, Across-group, Across-neighborhood

Consistency Guarantees The proposed method consistently recovers the structure with $O(\log p)$ samples and parameters with poly(p) samples.

Computational Complexity under N samples, d dimension, k hidden states, p variables, z non-zero entries per sample, Γ sized groups.

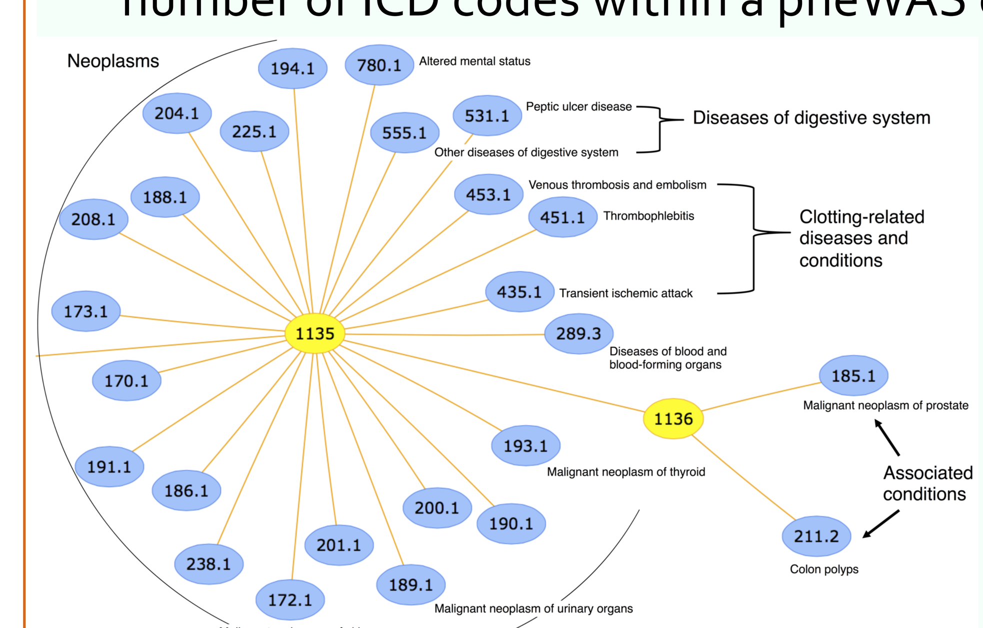
Algorithm Steps	Time per worker	Degree of parallelism
Information Distance Estimation	$O(Nz + d + k^3)$	$O(p^2)$
Structure: Minimum Spanning Tree	$O(\log p)$	$O(p^2)$
Structure: Local Recursive Grouping	$O(\Gamma^3)$	$O(p/\Gamma)$
Parameter: Tensor Decomposition	$O(\Gamma k^3 + \Gamma dk^2)$	$O(p/\Gamma)$
Merging and Alignment Correction	$O(dk^2)$	$O(p/\Gamma)$

Healthcare data analysis

Goal discover a disease hierarchy based on their co-occurring relationships in the patient records.

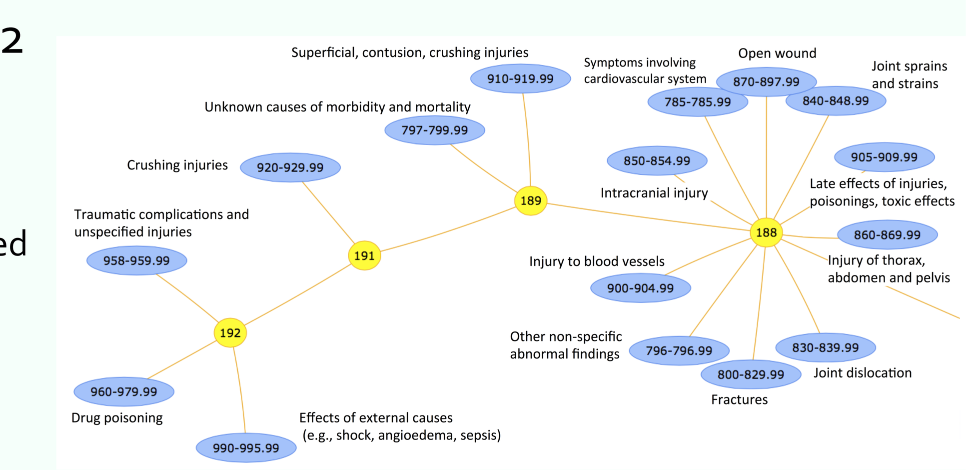
Disease Logs

1. MIMIC2: 30k patients, 314k diagnostic events, 5.6k diseases.
   - Diseases encoded with International Classification of Diseases **(ICD)** codes
   - Patients as samples and groups (varying size) of diseases as variables
   - Existing mapping between **ICD** and higher-level **Phenome-wide Association Study (PheWAS)** codes
   - Node dimension is set to be binary ($d = 2$) or the maximum number of ICD codes within a pheWAS code ($d = 31$)



(a) Case d=2

In figure (a), clotting-related diseases and altered mental status were grouped under the same latent node as several neoplasms. This may reflect the fact that altered mental status and clotting conditions such as thrombophlebitis can occur as complications of neoplastic diseases. The association of malignant neoplasms of prostate and colon polyps, two common cancers in males, is captured under latent node 1136.



(b) Case d =31

Figure (b) shows a portion of the learned tree of four subtrees which all reflect similar diseases relating to trauma.

*Email: furongh@uci.edu

NIPS workshop 2015