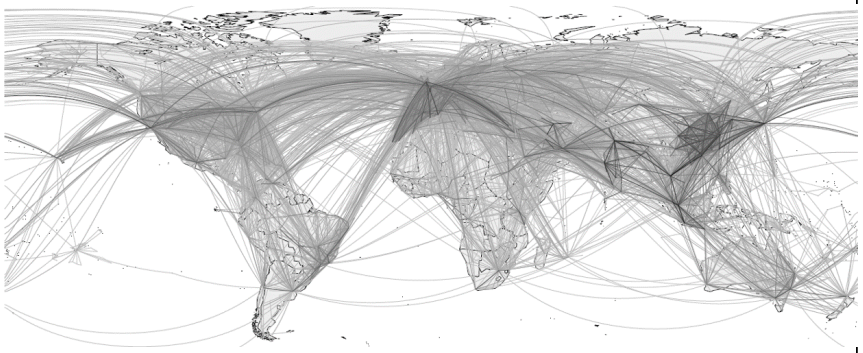
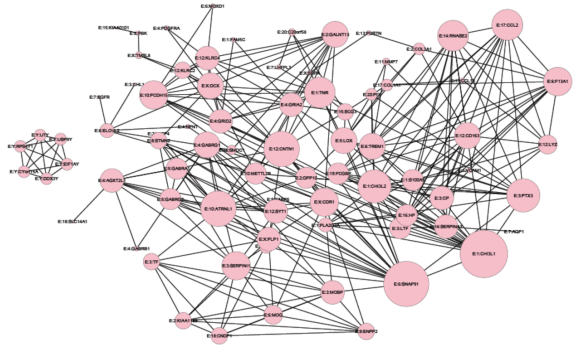
Unsupervised Analysis: Graphical Models

Why Graphical Models?



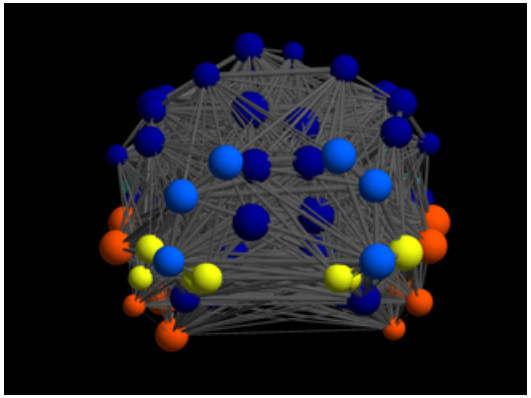
Depicts relationships (edges) between features (nodes).

Why Graphical Models?



Genomics: Relationships between genes.

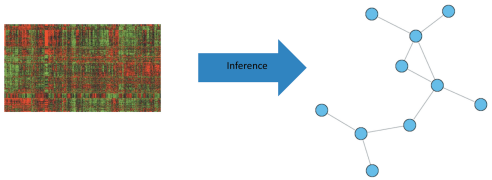
Why Graphical Models?



Neuroimaging: Relationships between brain regions.

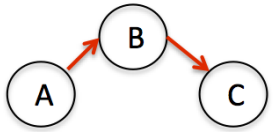
Graphical Models in Unsupervised Learning

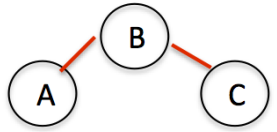
1. Network Data.
   - Examples: Social networks, twitter, citations, surveillance, web links, etc.
2. **Our focus**: Learn network from data.
   - Data matrix: $\boldsymbol{X}_{n\times p}$.
   - Features form p nodes.
   - Goal: Learn edges between nodes (i.e. learn the relationships between features).



Major Types of Graphical Models

Undirected vs. Directed Graphs.





Major Types of Graphical Models

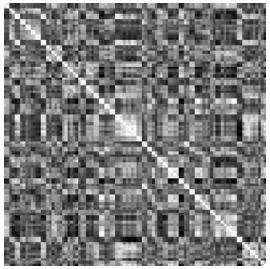
Undirected Graphical Models:

- Most common: Correlation Networks (association networks).
- **Our Focus**: Markov Networks.
- Others: Mutual information networks.

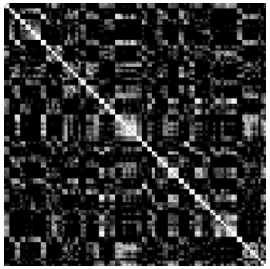
Directed Graphical Models:

- Bayesian Networks (DAG - Directed Acyclic Graphs).

Correlation Networks



Correlation matrix.



Thresholded correlation matrix.

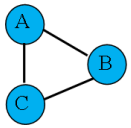
- Correlation matrix - $\mathbf{C} = \mathrm{Cor}(\boldsymbol{X})$: $\mathbf{C}_{ij} = \mathrm{Cor}(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- Measures linear associations between features.

Markov Networks

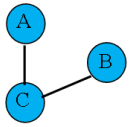
Markov Network

An *undirected graphical model* that characterizes conditional dependence (direct) relationships.

- *Edge*: Two nodes are **conditionally dependent**.
- *No edge*: Two nodes are **conditionally independent**.
- Conditions on all other nodes.



$$A \perp B \mid C$$

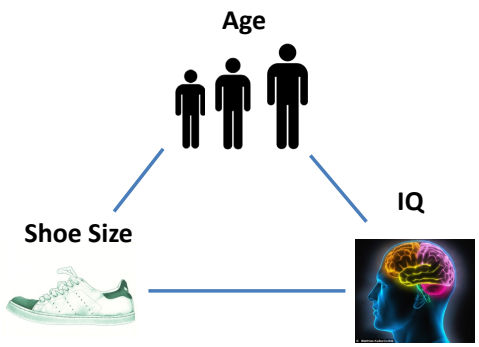


Markov Networks - Conditional Dependence

Regression Interpretation:

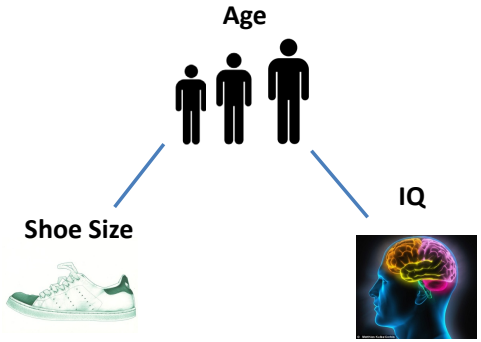
- Imagine trying to predict the observations in Node A (response) by the observations of all other nodes (predictors).
- Node B predictive of Node A (with all other nodes in model).
  - A is conditionally dependent on B.
  - Edge.
- Because of other nodes in model, Node B does not add any predictive value for Node A.
  - A is conditionally independent of B.
  - No Edge.

Markov Networks - Conditional Dependence



Correlation.

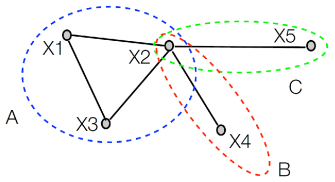
Markov Networks - Conditional Dependence



Conditional Dependence (Partial Correlation).

Markov Networks

- Local Markov Property: Conditional dependencies defined by node-neighborhoods, or the set of nodes connected to a given node via an edge.
- Global Markov Property: Pairwise conditional dependencies and neighborhoods jointly define the global dependence structure (formally defined by separators).
- Hammersley-Clifford Theorem: Density on graph factorizes according to sufficient statistics on cliques. (Probabilistic model!)



Gaussian Graphical Models

Gaussian Graphical Models

GGM:

- Multivariate normal: $\boldsymbol{X} \sim N(0, \boldsymbol{\Theta}^{-1})$
- $\boldsymbol{\Theta}$ the inverse covariance matrix.
- If data standardized, $\boldsymbol{\Theta}$ the partial correlation matrix.
- Zeros in $\boldsymbol{\Theta} \Longrightarrow$ conditional independence!
  - Edges correspond to non-zeros in $\boldsymbol{\Theta}$.
- Inference Goal: Estimate a sparse $\boldsymbol{\Theta}$.

Algorithms:

1. Graphical Lasso - penalized maximum likelihood estimation.
2. Neighborhood Selection - penalized conditional maximum likelihood estimation.

Graphical Lasso

Estimate sparse $\boldsymbol{\Theta}$ via Penalized Maximum Likelihood Estimation (MLE).

Graphical Lasso (Glasso)

$$\underset{\boldsymbol{\Theta}}{\text{maximize}} \quad \log|\boldsymbol{\Theta}| - \text{tr}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1$$

- Blue: Log-likelihood.
- Red: Penalty that encourages zeros in $\boldsymbol{\Theta}$.

[Yuan and Lin (2007), Banerjee et al. (2007), d'Aspremont et al., 2006; Friedman et al., 2008; and many others]

R: `glasso` package.

Neighborhood Selection

- Estimate sparse $\mathbf{\Theta}$ via penalized conditional MLE - by estimating zeros in one row / column of $\mathbf{\Theta}$ at a time.
- For each node $\boldsymbol{x}_j$, find it's node-neighbors (L_1-penalized regression or Lasso):

$$\underset{\beta^j}{\text{minimize}} \quad ||\boldsymbol{x}_j - \boldsymbol{X}_{\neq j}\beta^j||_2^2 + \lambda||\beta^j||_1$$

- Symmetry - β_i^j not always same as β_j^i.
  - Min or max rule.
- Meinshausen and Buhlmann (2006).

Network Sparsity

λ Controls Sparsity.

$$\underset{\mathbf{\Theta}}{\text{maximize}} \quad \log|\mathbf{\Theta}| - \text{tr}(\boldsymbol{X}^T\boldsymbol{X}\mathbf{\Theta}) - \lambda||\mathbf{\Theta}||_1$$

- $\lambda = 0$ gives a dense network (no sparsity).
- As λ increases, network becomes more sparse.

- Modulates trade-off between model fit and network sparsity.

How to Choose λ?

- Cross-Validation - tends to yield overly dense networks.
- Extended BIC - adjusted BIC for high-dimensional settings.

Stability Selection.

- Idea: Choose λ that gives the most **stable network**.

Procedure:

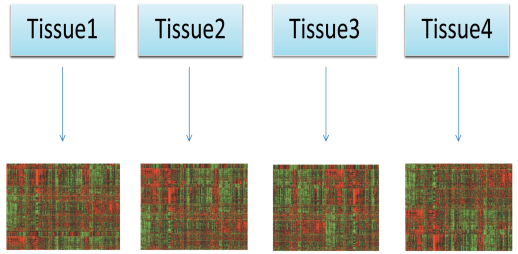
1. Repeatedly re-sample (bootstrapping or sub-sampling) observations.
2. Choose λ that results in the smallest network variability across re-samples.
3. *Stability Score*: For each edge, the proportion of re-samples in which edge was selected.

[Meinshausen & Buhlmann, 2011]
R: `huge` package.

Multiple Gaussian Graphical Models

Multiple Gaussian Graphs

- Multiple data sets available.
- Share common structure.
- Gene networks describing different tissues.



Joint Estimation of Multiple Graphical Models

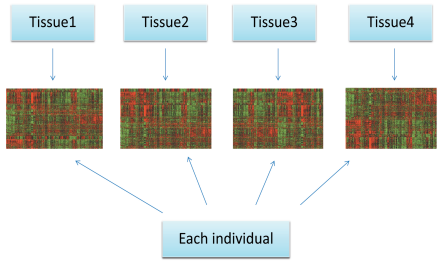
- Independence assumption for multiple data sets; Share common structure.
- Gene networks describing different sources.

$$\text{argmin}\, \mathbf{\Omega}^{(k)} \sum_{k=1}^{K} [\text{trace}(\hat{\mathbf{\Sigma}}^{(k)} \mathbf{\Omega}^{(k)}) - \log\det(\mathbf{\Omega}^{(k)})] + P(\mathbf{\Omega})$$

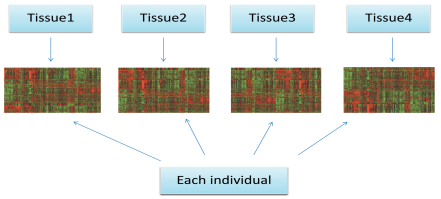
- Encourage common structure through joint regularization (Varoquaux et al, 2007; Guo et al., 2011; Chiquet et al., 2011; Danaher et al., 2012; etc)
  - Guo et al.: Encourage both group sparsity and within group sparsity $P(\mathbf{\Omega}) = \lambda \sum_{i \neq j} \sum_{\ell=1}^{k} \sqrt{|\omega_{i,j}^{(\ell)}|}$.
  - Danaher et al.: Graphical group lasso penalty $\lambda_1 \sum_{\ell=1}^{k} \sum_{i \neq j} |\omega_{i,j}^{(\ell)}| + \lambda_2 \sum_{\ell=1}^{k} \sqrt{\sum_{i \neq j} \omega_{i,j}^{(\ell)2}}$
    Fused graphical lasso penalty $\lambda_1 \sum_{\ell=1}^{k} \sum_{i \neq j} |\omega_{i,j}^{(\ell)}| + \lambda_2 \sum_{\ell < \ell'} \sum_{i,j} |\omega_{i,j}^{(\ell)} - \omega_{i,j}^{(\ell')}|$

Joint Estimation of Dependent Graphical Models

- Multiple data available.
- Data sets from multiple tissues for a group of mice.
- Tissues are the categories.
- Gene expression for four tissues: fat, brain, liver, and muscle.
- Individual is the system.
- Independent assumption is invalid: Related to existing time varying graphs (Zhou et al.; Kolar et al., 2010), but different.



Dependent Graphical Models: Problem Formulation



Xie, Liu, Valdar (2014) considered two layers of graphs:

1. Category specific layer: tissue-specific graphs
2. Systemic layer: whole-body systemic graph

- Let $\mathbf{y}_{k,i} = (y_{k,i1}, ..., y_{k,ip})^T$ be the i-th observed data vector for the k-th category.
- For given i-th mouse, $\mathbf{y}_{k,i}$ are not independent among different tissues k.

Dependent Graphical Models: Problem Formulation

- We model

$$\mathbf{y}_{k,i} = \mathbf{x}_{k,i} + \mathbf{z}_i, \quad i = 1, ..., n \quad k = 1, ..., K$$

  where $\mathbf{z}_i$ is the shared systemic random effect, and $\mathbf{x}_{k,i}$ is the random effect for k-th category.
- $\mathbf{x}_{k,i} \overset{\text{i.i.d.}}{\sim} \mathbf{N}(0, \mathbf{\Sigma}_k)$, $\mathbf{z}_i \overset{\text{i.i.d.}}{\sim} \mathbf{N}(0, \mathbf{\Sigma}_0)$ and $\mathbf{x}_{k,i} \perp\!\!\!\perp \mathbf{z}_i$.
- Only $\mathbf{y}_{k,i}$'s are available.
- Although $\mathbf{x}_k$ and $\mathbf{z}$ are latent variables, $\mathbf{\Omega}_k = \mathbf{\Sigma}_k^{-1}; k = 1, \ldots, K$ are identifiable with $K \geq 2$.
- Terminologies
  systemic network: $\mathbf{\Omega}_0 = \mathbf{\Sigma}_0^{-1}$
  category-specific network: $\mathbf{\Omega}_k = \mathbf{\Sigma}_k^{-1}$
  aggregate network: $\mathbf{\Omega}_{Y_k} = (\mathbf{\Omega}_k^{-1} + \mathbf{\Omega}_0^{-1})^{-1}$

Sparsity Notion for Dependent Graphical Models

- Sparse systemic network $\mathbf{\Omega}_0$: whole-body systemic graph characterizes the body wide dependence structure among genes.
- Sparse category-specific network $\mathbf{\Omega}_k$: tissue-specific graphs characterize the dependence structure among genes within each tissue, after removing the common systemic variation.
- Aggregate network $\mathbf{\Omega}_{Y_k} = (\mathbf{\Omega}_k^{-1} + \mathbf{\Omega}_0^{-1})^{-1}$ may not be sparse, given sparse $\mathbf{\Omega}_0$, $\mathbf{\Omega}_k; k = 1, \ldots, K$.
- Goal: Estimate sparse $\mathbf{\Omega}_0$, $\mathbf{\Omega}_k; k = 1, \ldots, K$.

Application to Mouse Gene Expression Data

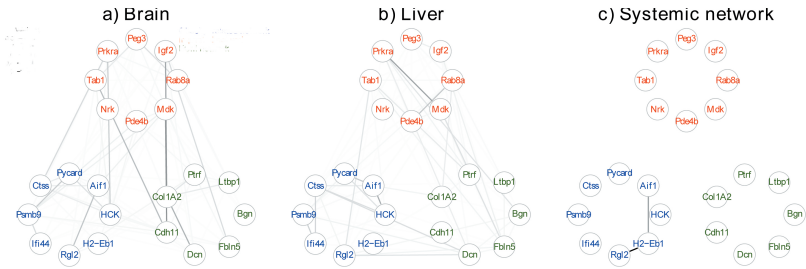
- 301 mice from F_2 cross, varying genes for fat composition (Dobrin et al. 2009).
- Acts like randomized allocation of fat-inducing treatment.
- Gene expression for fat, brain, liver, and muscle.
- For each tissue, over 20,000 gene expression values.
- Three groups of gene networks: the obesity-related network, the imprinting related-network, and the extracellular matrix (ECM)-related network.

Application to Mouse Gene Expression Data

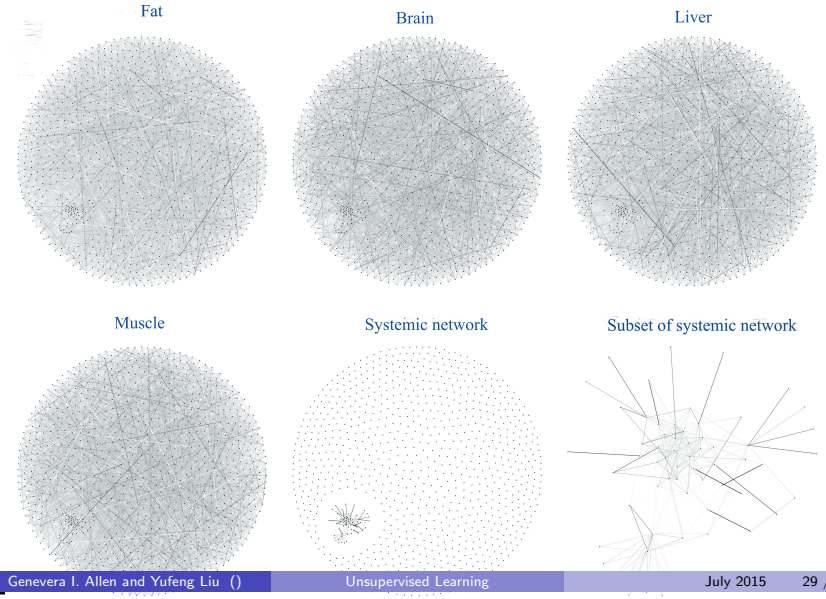
Systemic network should has links that:

1. affect the whole individual
2. reflect the treatment allocation.

Systemic network only has edges for obesity-related network.



Application to Mouse Gene Expression Data: P = 1000



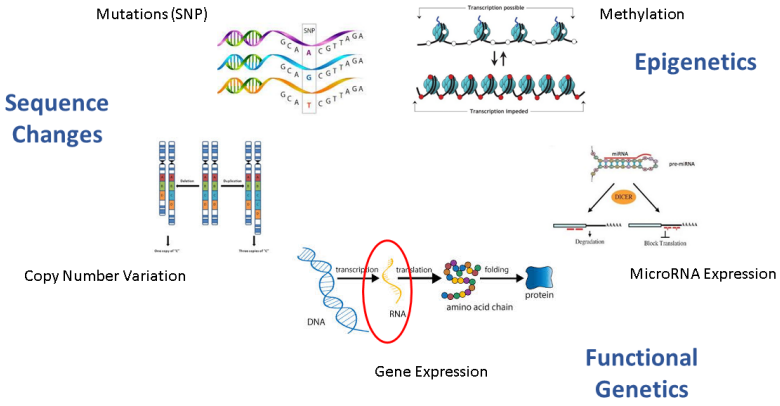
Application to Mouse Gene Expression Data: P = 1000

- Systemic network is sparse (249 edges among 62 genes)
- These 62 genes include some obesity-related genes
- Analysis of gene ontology (GO) enrichment on the systemic network: the network is significantly enriched for genes associated with immune and metabolic processes, consistent with recent studies linking obesity to strong negative impacts on immune response to infection
- Tissue-specific networks are much denser

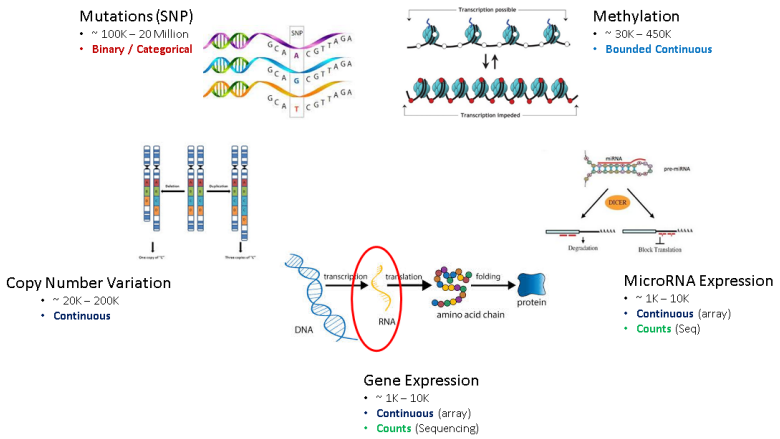
Figure: *Panels a), b), c) and d) display the category-specific networks estimated for adipose, hypothalamus, liver and muscle tissues respectively. Panel e) shows the structure of the estimated systemic network, describing common*

(Mixed) Graphical Models via Exponential Families

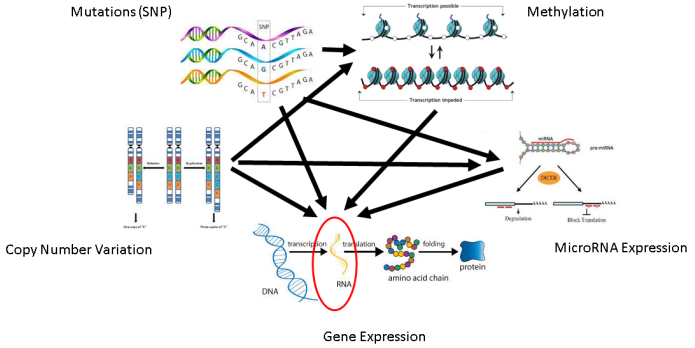
Motivation - Integrated Networks



Motivation - Integrated Networks



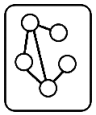
Motivation - Integrated Networks

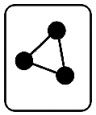


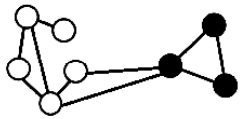
Networks for Different Data Types

Existing Markov Network Types:

1. Gaussian Graphical Models.
2. Ising Models (Binary-Valued).
   - Estimated via L_1 penalized logistic regression - neighborhood selection.
3. Gaussian-Ising Models.







What about other data types?

- Counts? Skewed Continuous? Etc.

Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.

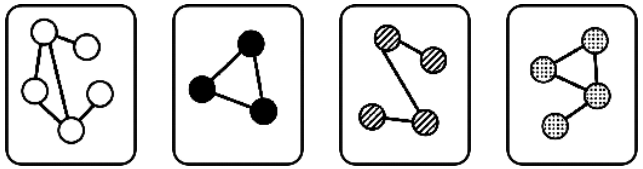
- Assumption: Conditional distributions are Exponential Families.
  - Ex: Gaussian, Bernoulli, Poisson, Exponential, Negative Binomial, etc.

A lot of math . . .

- Theorem: Joint network distribution exists and has a closed form!
  - Dependencies parameterized by products of sufficient statistics.
  - Strong statistical guarantees for network inference.
  - Fast, parallelizable algorithm to learn network structure.

Networks for Different Data Types

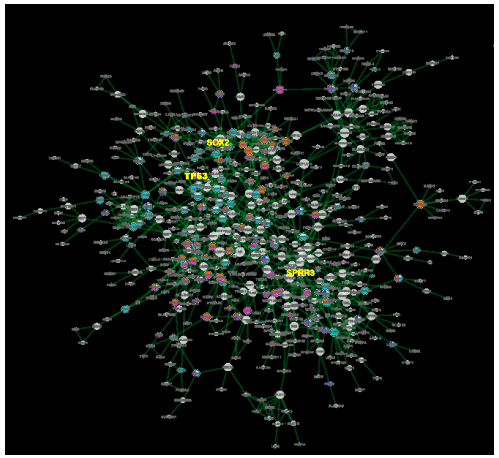
Our Framework: Graphical Models via Exponential Families.



In R: XMRF package.

Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.



Lung Cancer Gene Expression Network (via RNA-Seq).

Integrated Network Models

Block-Directed Graphical Models via Exponential Families.

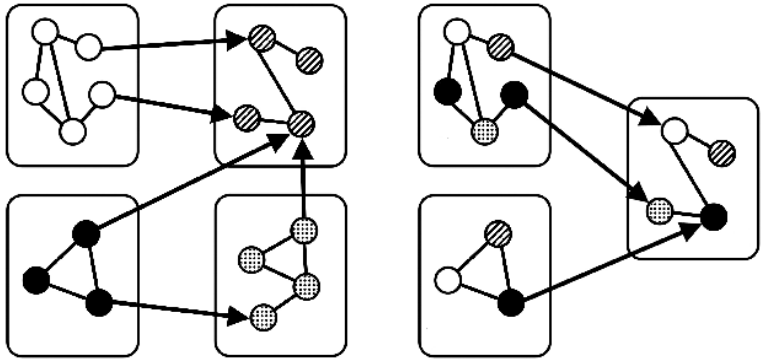
- Assumptions:
  - Conditional distributions are Exponential Families.
  - Variables belong to known groups and the directionality of dependencies between groups is known.

A lot of math . . .

- Theorem: Joint integrated network distribution exists and has a closed form!
  - Dependencies parameterized by products of sufficient statistics from different distributions.
  - Strong statistical guarantees for network inference.
  - Fast, parallelizable algorithm to learn network structure.

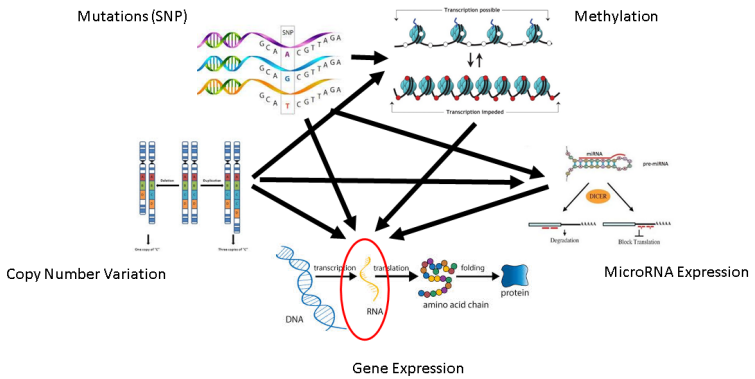
Integrated Network Models

Block-Directed Graphical Models via Exponential Families.



Integrated Network Models

Block-Directed Graphical Models via Exponential Families.



Gene Expression

Integrated Network Models

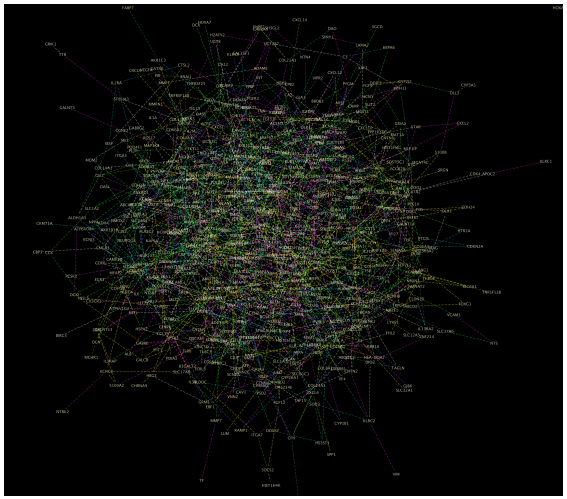
Block-Directed Graphical Models via Exponential Families.

Implication

First multivariate distribution for mixed data types (that directly parameterizes a rich set of dependencies).

Integrated Network Models

Block-Directed Graphical Models via Exponential Families.



Click here for Interactive Network

Integrated Network Models

Block-Directed Graphical Models via Exponential Families.

Breast Cancer Integrated Mutation-Gene Expression Network.

Blue nodes: RNA-sequencing
Yellow nodes: genomic mutations