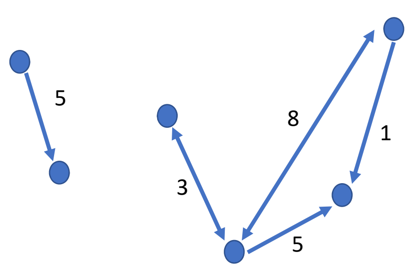
Systemic networks for high-dimensional exposures, mediators, and health outcomes

Jul. 25th 2018

Presenter: Jai Woo Lee

Graphs - Systemic Networks



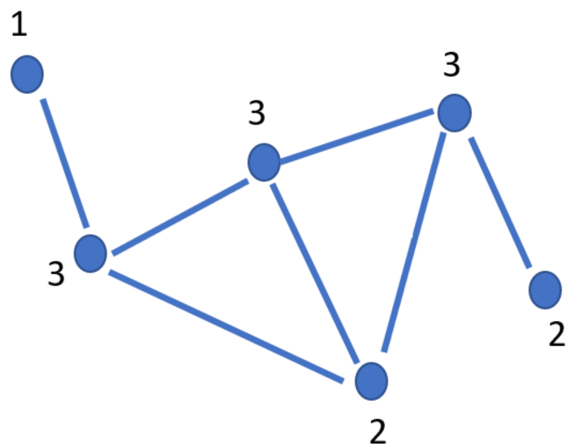
Nodes: elements, metabolites, genes, or sequence positions

Edges: Interactions, regulations, activations or inhibitions

Weight: Strength

Other conditions: Denseness or Sparseness, disconnected components, or walks

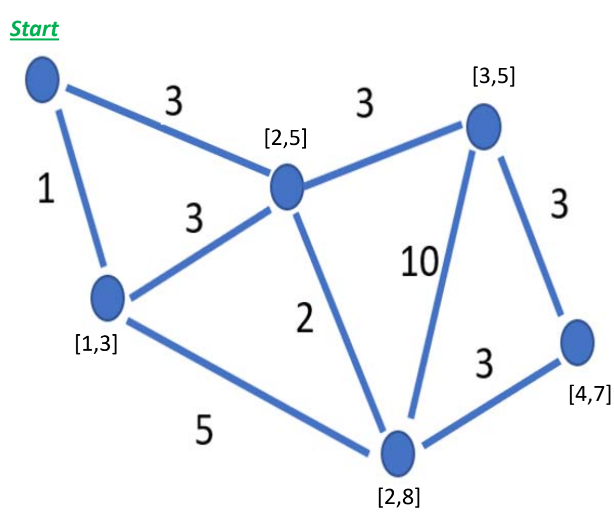
Theoretical Problem – Mathematical Games (With Math & CS Prof. Alan Frieze at CMU)



Rules:
1) Two players take alternative turns
2) Each player can start on any node
3) Each player can move to adjacent nodes of the current node
4) The player who removed final chips in the network wins
5) The upper bound of chips which each player can remove is fixed

Problems:
1) Design the strategy the first player wins.
2) Prove that the strategy is correct

Application Problem – Travelling Salesman Problem (With Business Prof. John Hooker at CMU)



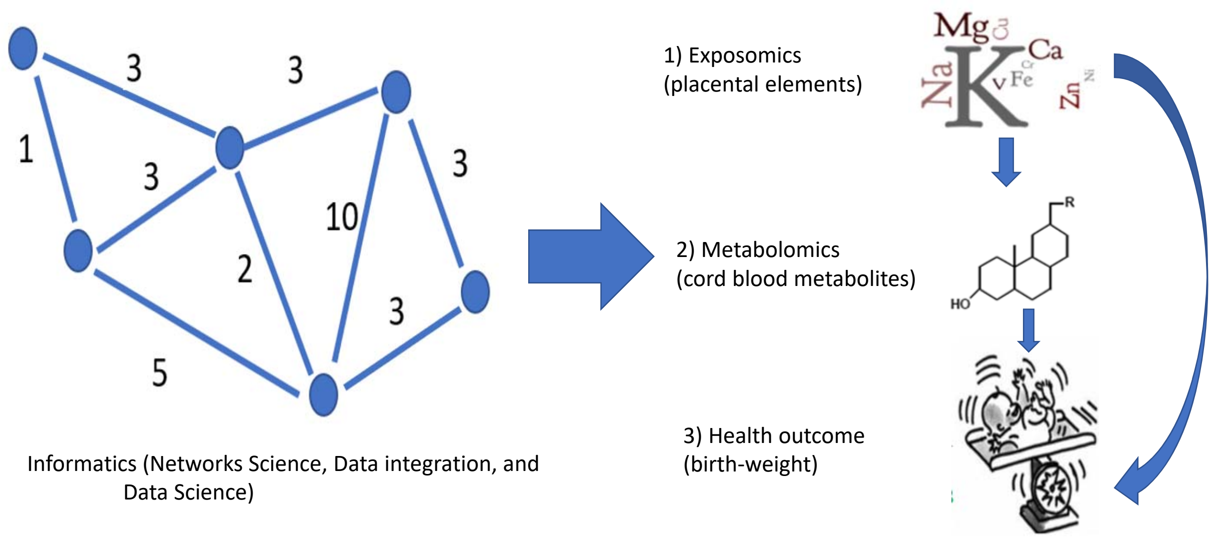
Problem:

Find the shortest trip not violating time limits for each city

Methods:

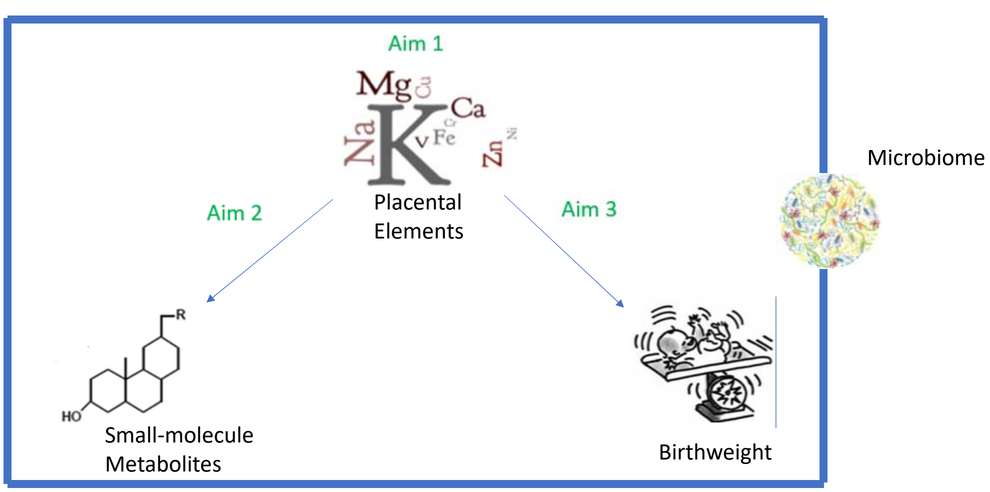
1) Linear Programs with AMPL (A Mathematical Programming Language)
2) Parallel computing for a large problem

My Current Research

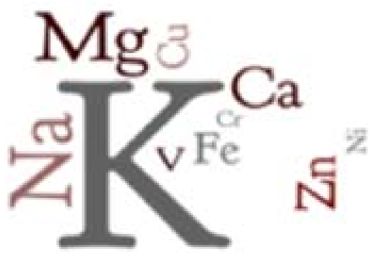


Background

Placental Elements: A vital role in biological systems involved in major processes as oxygen transport, or as catalysts like enzymes, in many metabolic reactions
Metabolomics: High-throughput analysis of metabolites, Functional end-point of physiology



Aim 1



Placental
Elements

Aim 1: Ideas

Partial Correlation

It's raining outside

My bike in the front yard is wet due to rain

My car in the front yard is wet due to rain

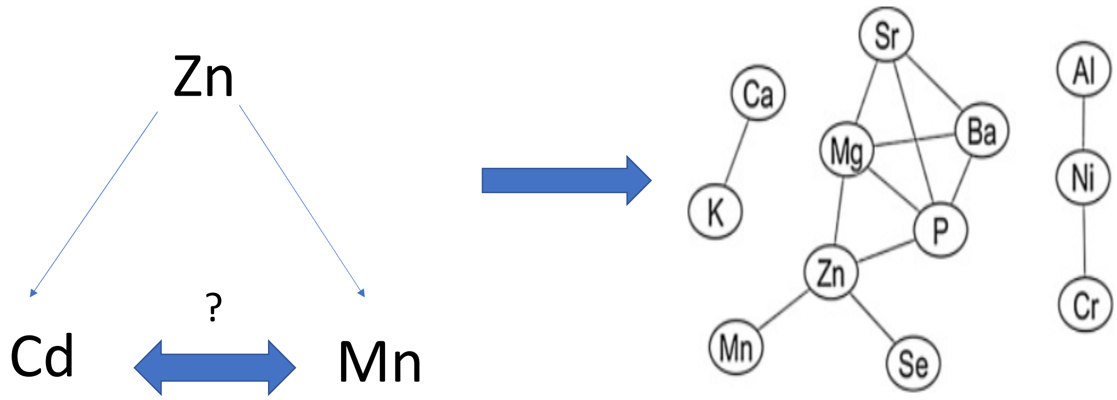
It's raining outside

?

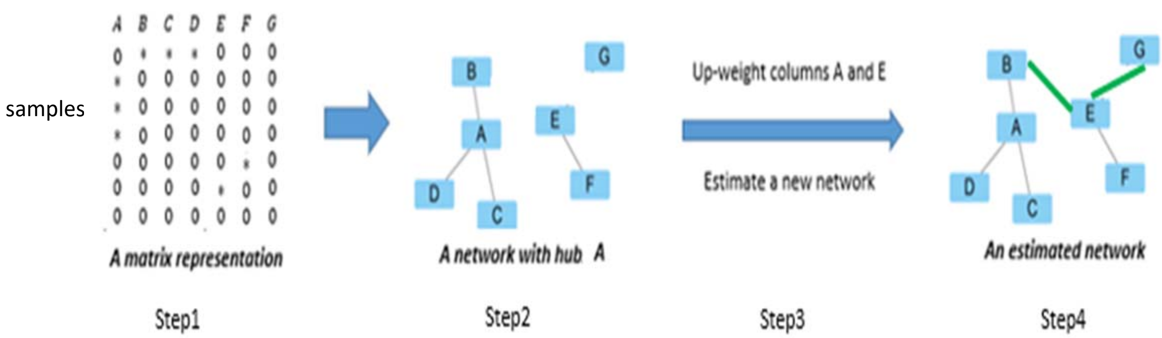
My bike in the front yard is wet due to rain

My car in the front yard is wet due to rain

Aim 1: Problem Description

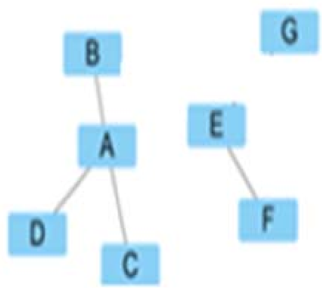


Aim1: Methods (key word: Weight)



Note: In step 3, weight hubs or nodes of biological interest

Aim1: Definition- hub



A network with hub A

Why is “node A” a hub in this network?

The degree (the number of neighbors) of each node should be checked.

deg(A) = 3 <- Max!
deg(B) = 1
deg(C) = 1
deg(D) = 1
deg(E) = 1
deg(F) = 1
deg(G) = 0

Aim1: Simulated Data

Define 40 features with 120 samples with following conditions

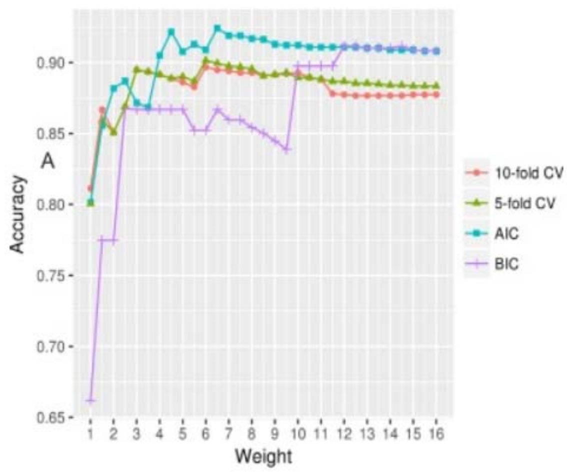
a) Define 3-5 hub nodes with 8-10 edges

b) Define 10-15 nodes with 1-2 edges

c) All the other nodes have no edges (totally disconnected)

d) Using eigenvalues and eigenvectors, generate the simulated data

Aim1 : Simulation result 1

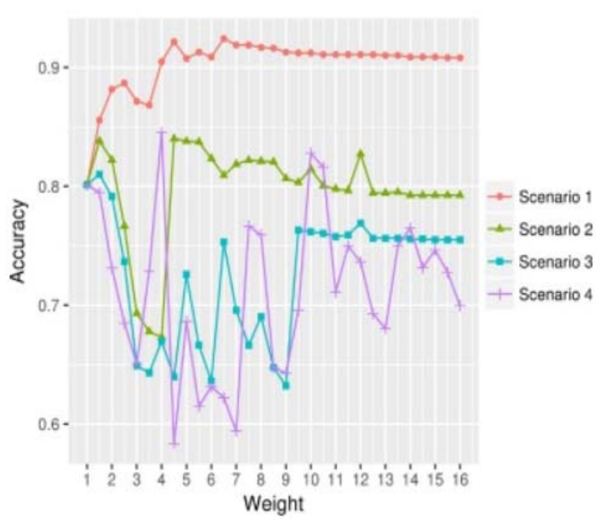


Why/How does AIC outperforms BIC and CV in this problem?

1) High sample size makes BIC penalize harsh

2) Low feature size makes CV less effective (vs. Gene expressions)

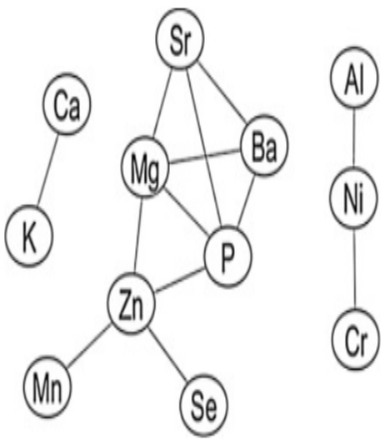
Aim1 : Simulation result 2



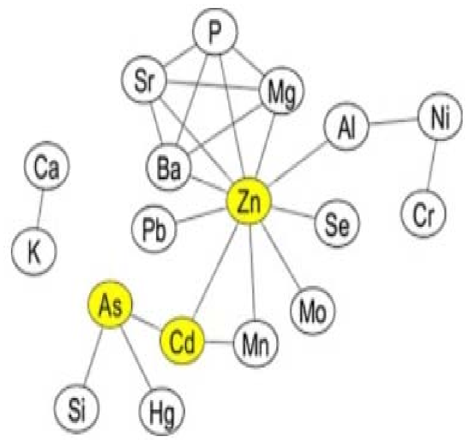
Scenario 1: Weight only hubs
Scenario 2: Weight hubs and non-hubs
Scenario 3: Weight only non-hubs
Scenario 4: Weight randomly chosen hubs

These scenarios were tested after testing other parameters such as sample size, the number of hubs, density of network and so on.

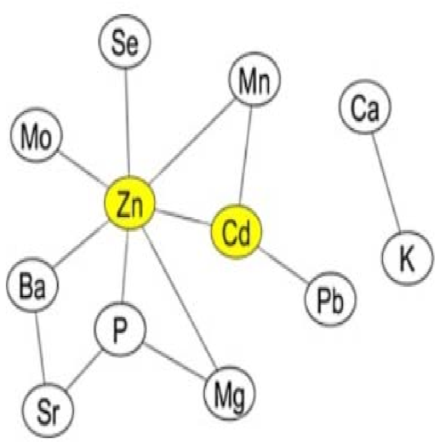
Aim1: Real Data application result 1



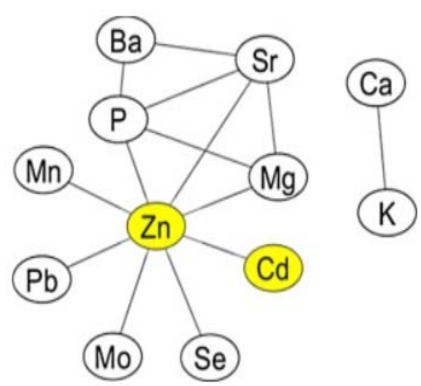
As, Cd, and ZN are 3-weighted



Aim1: Real Data application result 2



3-weighted female babies



3-weighted male babies

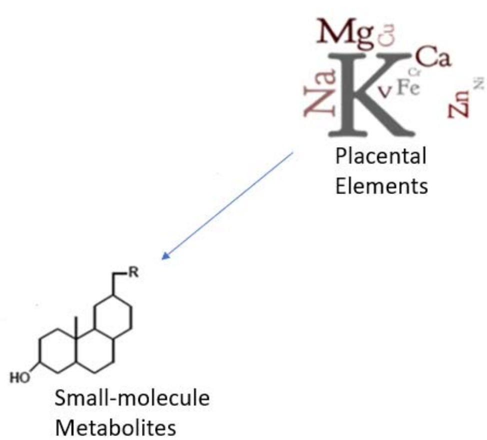
Aim1: Future Direction

1) Can we get a tight bound of weight values for hubs?
 (a more mathematically rigorous bound?)

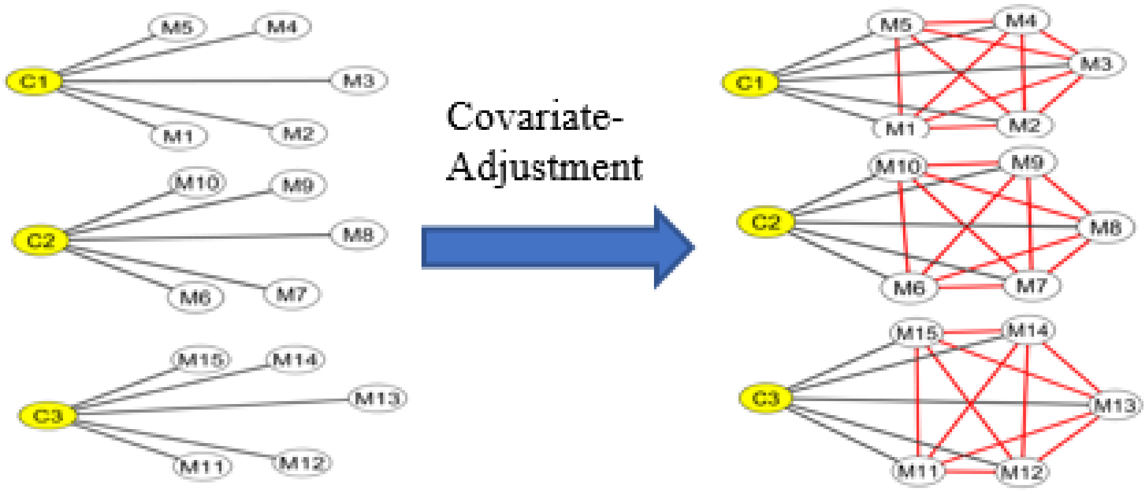
2) This method can be applied to other observed data for babies at transition?

3) This project was accepted and published as a paper,
"Penalized Estimation of sparse concentration matrices based on prior knowledge with applications to placenta elemental data"
Jai Woo Lee, Tracy Punshon, Erika L. Moen, Margaret R. Karagas, and Jiang Gui
in Computational Biology and Chemistry journal.

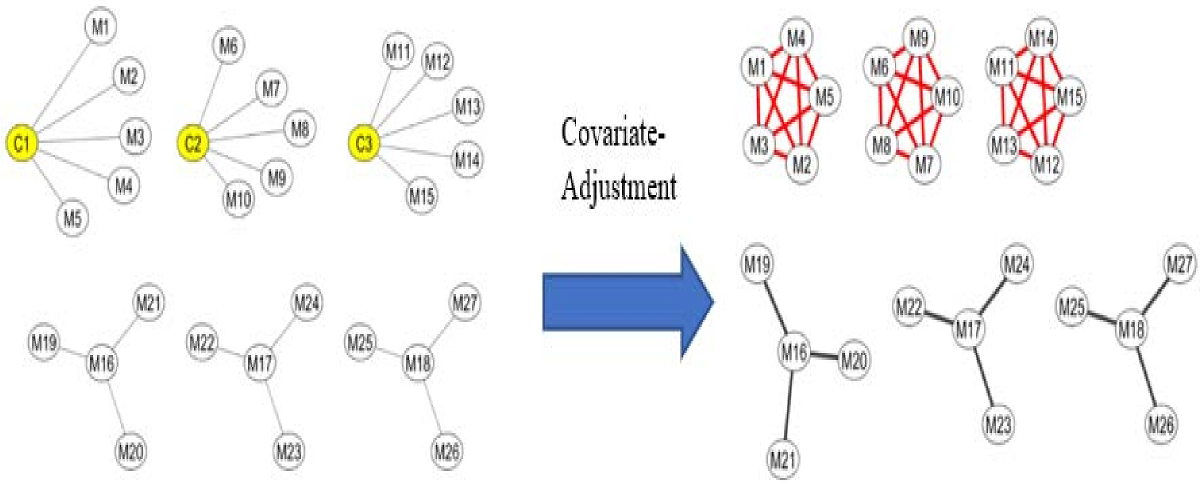
Aim 2



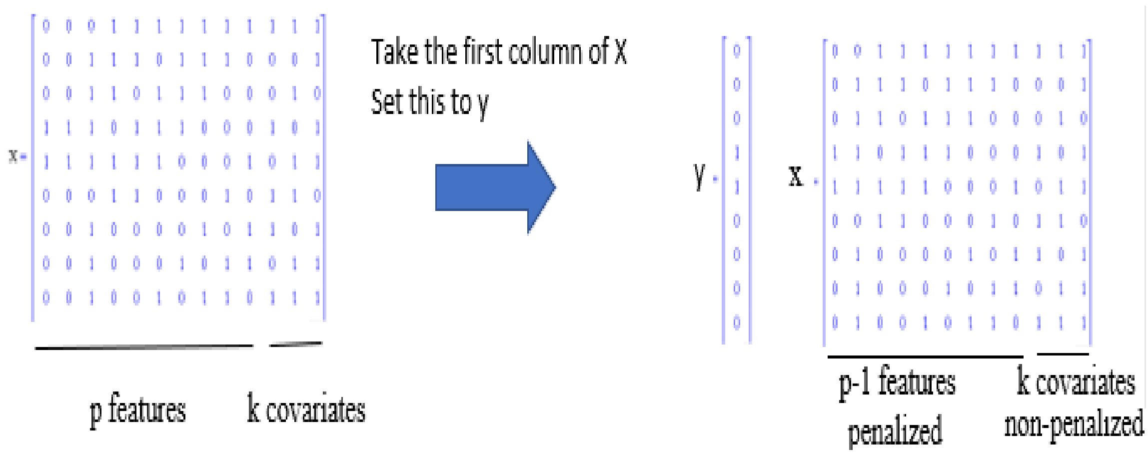
# Aim 2: Ideas
Keyword= Covariate



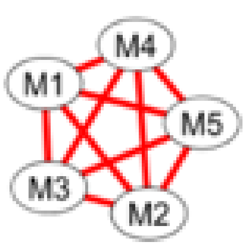
Aim 2: Problem Description
"Placental elements as covariates"



Aim2: Method – step 1
Gaussian Graphical model <-> Ising Model



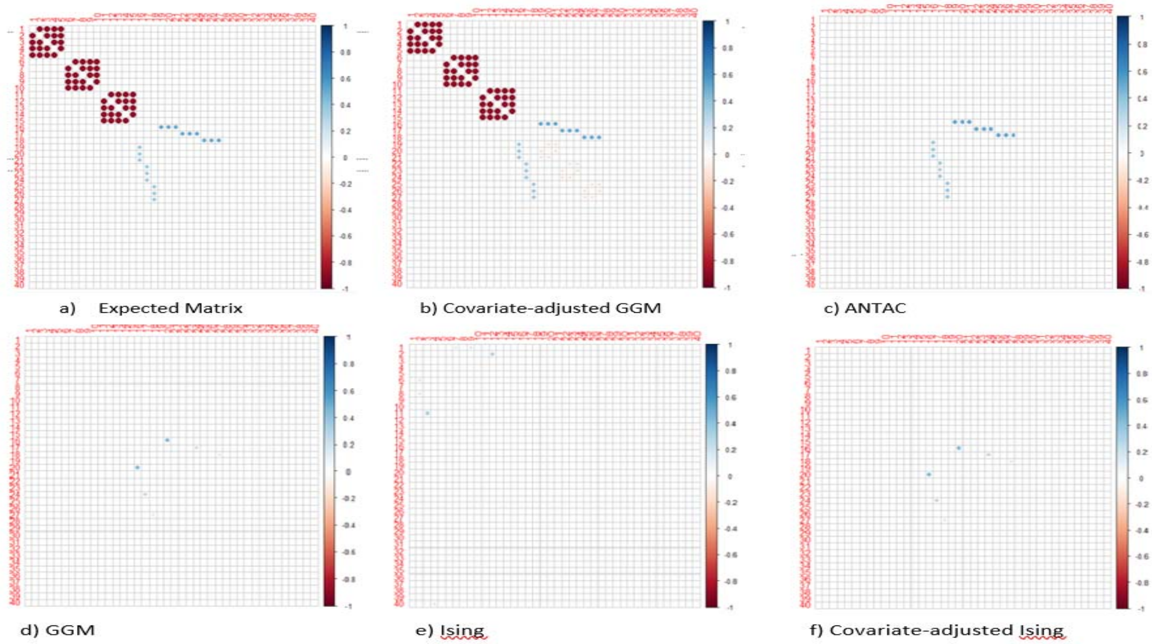
Aim2: Method – step 2
Travelling Salesman Problem on subnetworks



In this complete sub-graph, correlation values on edge indicate closeness.

To do:
Negate or get reciprocal values of weights and apply travelling salesman problem.
Then, we get the shortest path cycle

Aim2: Simulation results 1

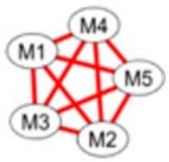


Aim2: Simulation results 2

p: features, d: degree of covariates, c: covariates, n: samples

(p, d, c, n)	Method	Adjustment effect	True Positive	True Negative
Model 1	Adjusted Ising	0%	67%	98%
(200, 10, 3,100)	ANTAC	0%	83%	99%
	Adjusted GGM	83%	100%	99%
Model 2	Adjusted Ising	11%	73%	96%
(200, 10, 3, 200)	ANTAC	0%	93%	99%
	Adjusted GGM	93%	100%	99%
Model 3	Adjusted Ising	17%	67%	98%
(200, 10, 3, 400)	ANTAC	0%	99%	96%
	Adjusted GGM	100%	99%	99%
Model 4	Adjusted Ising	0%	70%	98%
(200, 40, 3, 200)	ANTAC	0%	88%	92%
	Adjusted GGM	73%	93%	97%
Model 5	Adjusted Ising	0%	67%	98%
(200, 40, 3, 400)	ANTAC	0%	99%	97%
	Adjusted GGM	86%	96%	98%

Aim2: Simulation results 3



$$\begin{pmatrix} 0.000 & 0.269 & 0.168 & 0.201 & 0.291 \\ 0.255 & 0.000 & 0.746 & 0.669 & 0.793 \\ 0.178 & 0.743 & 0.000 & 0.826 & 0.327 \\ 0.212 & 0.624 & 0.832 & 0.000 & 0.308 \\ 0.332 & 0.783 & 0.321 & 0.342 & 0.000 \end{pmatrix}$$

M 1 -> M 3 -> M 2 -> M 5 -> M 4 -> M 1

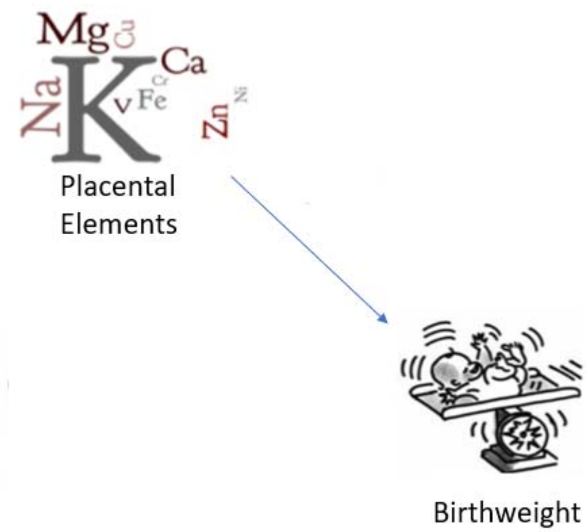
(The Shortest tour or the most correlated tour)

Aim2: Real Data application

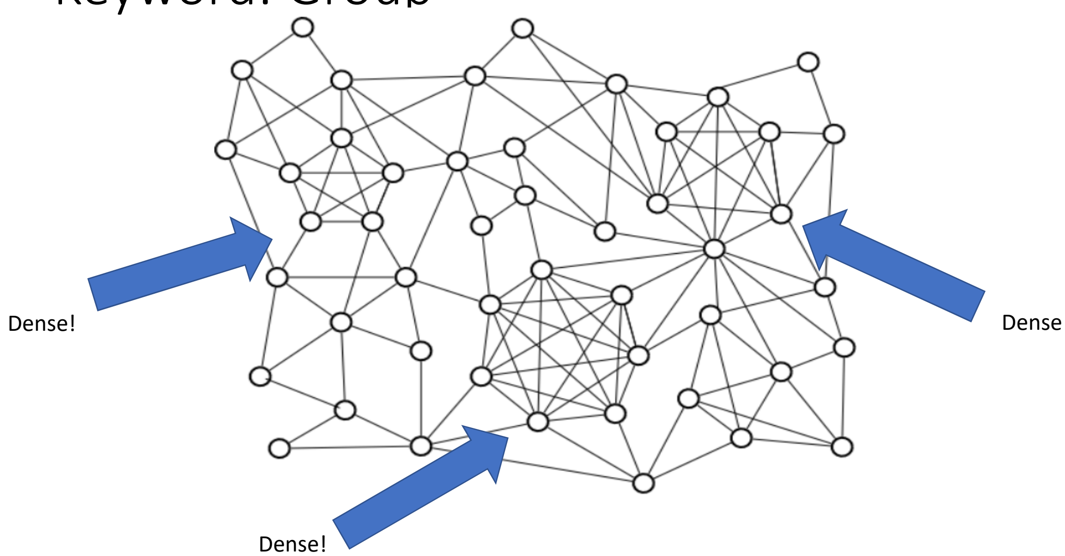
1) Real data application?
Fecal Metabolomics data by Biocrates
(Data processing is required)

2) Which metal covariates should I choose?
   a) biomedically? As, Cd, and Hg
   b) statistically? Apply Principal Component Analysis on Placental Elements Data

Aim3

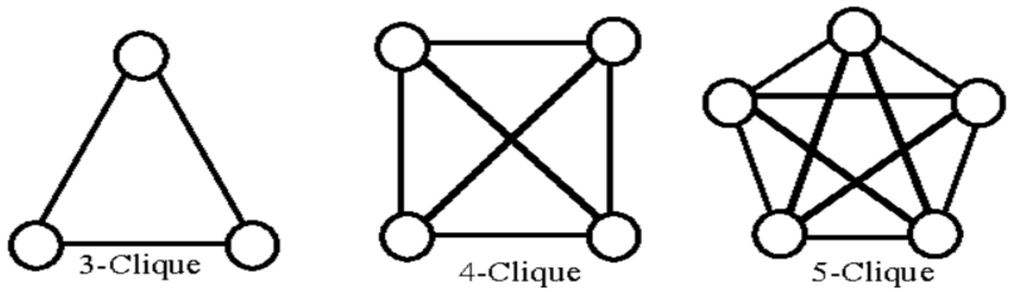


Aim 3: Ideas
Keyword: Group



Aim 3: Problem Description

Can we use dense sub-networks, possibly cliques, to estimate one vector of continuous values?



Aim3 : Method

- Step 1: Construct the network using Lasso to estimate the network after using AIC to pick tuning parameter.
- Step 2: Identify “clique”. We define clique to be a complete sub network.
- Step 3. Consider cliques as groups and apply various group Lasso methods to find the best one
- Step 4. Fit a liner regression model with cross-validation to estimate outcome y.

Aim3: Simulation Results 1

After Fitting a liner regression model with cross-validation to estimate outcome y “on 40 nodes, make three cliques of size 25, 10, 5”

Grouping Methods	Error rate ratio
grLasso: Group lasso (Yuan and Lin, 2006)	1.0012
grMCP: Group MCP (minimax concave penalty)	1
grSCAD: Group SCAD (smoothly clipped absolute deviation)	1.0136
cMCP: composite minimax concave penalty	1.0103
gel: Group exponential lasso (Breheny, 2015)	1.0054

Aim3: Simulation Results 2

Using grMCP to fit a liner regression model with cross-validation to estimate outcome y “on 40 nodes, make three cliques of size 25, 10, 5”

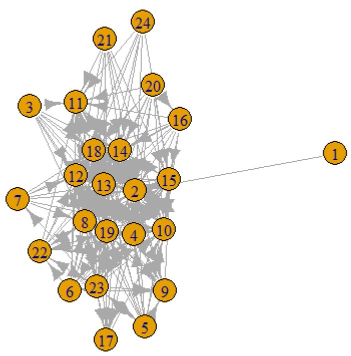
Only maximal cliques (error= 1)

Smaller cliques (eight cliques of size 5) (error=1.02)

No cliques (error=1.06)

=> Using maximal cliques gives the best results

Aim3: Real data application



Group1 (11 elements): Zn, Ca, Cu, Na, Sr, Se, P, K, Ba, <u>**Hg**</u>, Sb

Group2 (4 elements): <u>**Hg**</u>, Na, Mg, Si

Aim3: Future direction – Another Problem

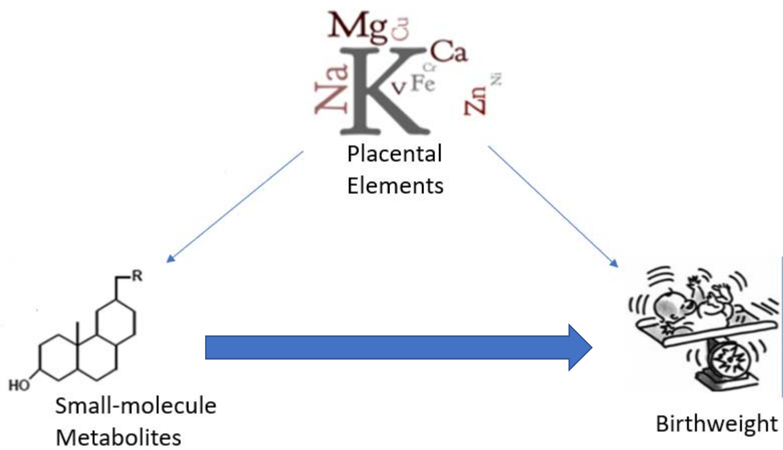
Metagenomics & Disease Status

a) Continuous data => dichotomized data

b) Gaussian Graphical Model => Ising Model

c) linear regression => Logistic Regression

Aim X: Networks, Pathways, and Mediation analysis



Aim X: Software Development

Codes for generating exposure-outcome, exposure-mediator, and mediator-outcomes were completed.

From this step, I can try two things.

1) Exposure -> Mediator -> health outcome (Shortest path analysis)

2) Exposure -> outcome

vs.

Exposure -> mediators -> outcome

(Mediation analysis)

Acknowledgements

Dissertation Committee

Dr. Jiang Gui (co-mentor)

Dr. Margaret R. Karagas (co-mentor)

Dr. Megan E. Romano

Dr. Hongzhe Li (Upenn)

Qualifying Examination Committee

Dr. Brock C. Christensen

Dr. Anne G. Hoen

Dr. Zhigang Li