Gaussian Graphical Models in Metabolomics

Raji Balasubramanian (UMass-Amherst) and Denise Scholtens
(Northwestern Feinberg School of Medicine)

Sunday June 23, 2019

Graphical models in medicine

Data

Introduction to network analysis in R

Gaussian Graphical Models (GGM) in R

Graphical models in medicine

Network Medicine

- **Fundamental principle**: disease module hypothesis that disease variants are connected.
- **Evidence in literature**: 10-fold increase in products of genes associated with a disorder when compared to expectation under random chance.
- **References**: Su and Clish, Metabolomics and Network Medicine, 2017; Goh, K. I., Cusick, M. E. et. al., The human disease network, 2007.

Metabolites as networks

Metabolites are naturally represented as networks:

- **Nodes**: represent individual metabolites.
- **Edges (undirected)**: denote pairwise metabolite relationships.

Example network

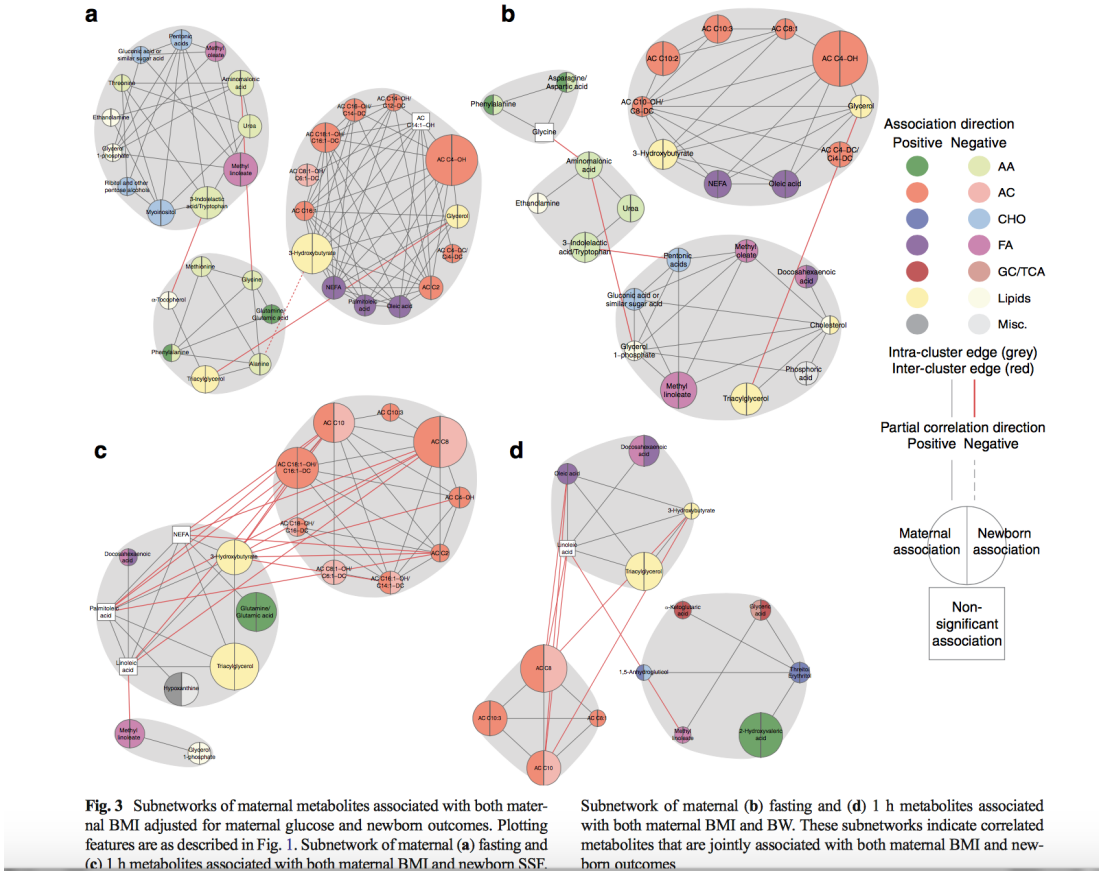


Fig. 3 Subnetworks of maternal metabolites associated with both maternal BMI adjusted for maternal glucose and newborn outcomes. Plotting features are as described in Fig. 1. Subnetwork of maternal (**a**) fasting and (**c**) 1 h metabolites associated with both maternal BMI and newborn SSF. Subnetwork of maternal (**b**) fasting and (**d**) 1 h metabolites associated with both maternal BMI and BW. These subnetworks indicate correlated metabolites that are jointly associated with both maternal BMI and newborn outcomes

Figure 1: Maternal BMI and newborn SSF associated metabolite networks from Sandler, V.,Reisetter, A. C. et. al., Diabetologia, 2017.

Correlation networks

- Correlation networks are established methods for constructing metabolite networks.
- Edges in correlation networks depict pairwise correlations between metabolite pairs.
- Networks are often created by thresholding on a correlation cut-off.

- **Recent example from literature:** A network analysis of biomarkers for Type 2 Diabetes in the Nurses Health Study. [1]

[1] Huang, T., Glass, K. et al., Diabetes, 2018.

Correlation networks

- **Drawback:** Correlations between metabolite pairs can be driven by direct and indirect relationships.
- Drivers of high correlation include shared or common enzymatic activities. [2].
- Large number of non-zero pairwise correlations are usually observed.
- Absence of an edge results from satisfying a **strong** criterion of marginal independence between metabolite pairs. [3]

---

[2] Su and Clish, Metabolomics and Network Medicine, 2017

[3] Strimmer, K., Notes on Gaussian Graphical Models. http://www.strimmerlab.org/notes/ggm.html

Gaussian graphical models (GGM)

- **Model:** Metabolites are multivariate Gaussian with mean μ and covariance matrix Σ.
- The precision (concentration) matrix $\Omega = \Sigma^{-1}$.
- If $\Omega_{jk} = 0$, then the ith metabolite is independent of the jth metabolite, given all other variables.

GGM estimation

- **Meinshausen and Buhlmann (2006)**: estimates $\Omega_{jk} = 0$ by fitting a lasso to each metabolite, using all others as predictors.
- $\hat{\Omega}_{jk} \neq 0$: if the estimated coefficients of metabolite i on j AND vice-versa are non-zero.

- **Friedman et al. (2007)**: Glasso and variants for exact maximization of the penalized log-likelihood.

Model selection

- Gaussian graphical model estimation involves a process to estimate the **optimal regularization parameter (λ)**.
- Large values of λ correspond to increasing sparsity of the resulting graph.

- Stability approach for regularization selection (StARS): uses a subsampling approach to estimate the optimal λ.
- Rotation information criterion (RIC): uses a permutation approach to estimate λ.

Correlation network versus GGM

- **Correlation network:** An edge between metabolite pairs can result from both direct AND indirect relationships.

- **GGM:** An edge exists ONLY if the metabolite pair is dependent after accounting for all other indirect relationships.

Data

HAPO Metabolomics

- **Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study** conducted during 2000 - 2006 at 15 international field centers.
- Blood samples were obtained during a 75-g oral glucose tolerance test (OGTT) between 24 and 32 weeks gestation.
- Metabolites were measured in maternal fasting and 1-h serum samples from **400** mothers in each ancestry group (Afro-Caribbean, Mexican American, Northern European, Thai).
- Mothers were sampled to span the range of maternal glucose and BMI.

HAPO Metabolomics

Data Format:

- **Column 1**: ID
- **Column 2**: Ancestry Group
- **Column 3**: Fasting glucose
- **Columns 4-54**: 51 metabolites

HAPO Metabolomics

Loading data ..

```
#PC users
#setwd("C:/Users/username/Desktop/Metabolomics Workshop 2019/")

#mac users
setwd("~/Desktop/Metabolomics Workshop 2019")
mydat <- read.csv(file = "hapo_metabolomics_2019.csv")
print(mydat[1:3,1:10])
```

```
##       id anc_gp  fpg    mt1_1     mt1_2    mt1_3    mt1_4     mt1_5
## 1 hm0001    ag3 75.6 218.2223  76.99525 19.06366 14.23091  86.75162
## 2 hm0002    ag3 84.6 292.6314 136.41320 43.14854 17.77549 120.17344
## 3 hm0003    ag4 79.2 361.1135  79.98370 22.15848 13.05497  74.75441
##      mt1_6    mt1_7
## 1 135.2109 64.00578
## 2 213.6531 91.30156
## 3 136.1587 83.67878
```

HAPO Metabolomics

Three groups of metabolites:

- Prefix **mt1**: Amino Acids (AA)
- Prefix **mt2**: Acyl carnitines (AC)
- Prefix **mt3**: Other

HAPO Metabolomics

Let's take a look at the numbers by **ancestry group**:

```
ag <- mydat[,2]
table(ag)
```

```
## ag
## ag1 ag2 ag3 ag4
## 400 400 400 400
```

HAPO Metabolomics

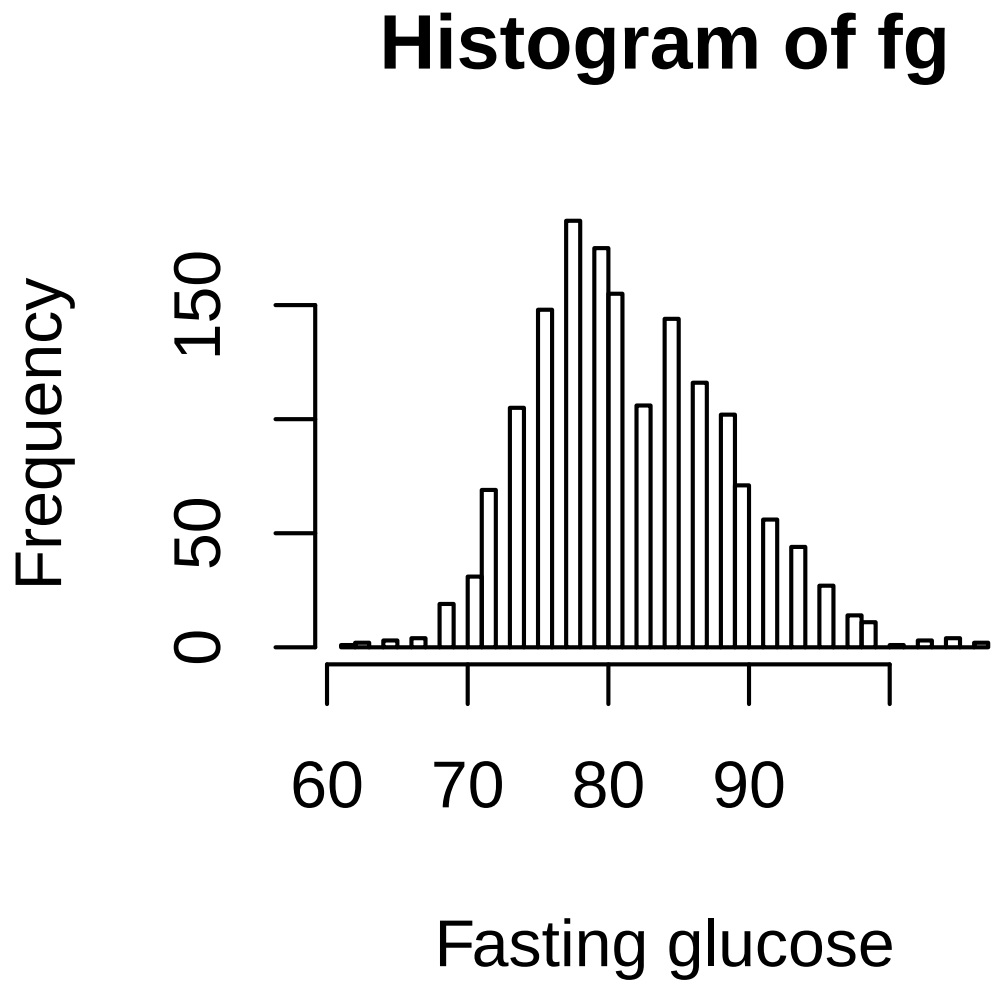
Let's take a look at the distribution of **fasting glucose**:

```
fg <- mydat[,3]
summary(fg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61.20   77.40   81.00   81.63   86.40  106.20
```

HAPO Metabolomics

Let's take a look at the distribution of **fasting glucose**:



Introduction to network analysis in R

Preliminaries

- **graph R** package: provides a way of representing graphs as a *graphNEL* object.
- **igraph R** package: also provides various tools for working with graphs.

Preliminaries

- Let's work with a small (p=6) set of metabolites sampled from the HAPO dataset.
- As an example, we start with a simple correlation network of 6 metabolites

```
mx <- mydat[,-c(1:3)]
mx.1 <- mx[ag == "ag1", c(1,2,16,17,34,35)]
cor.1 <- round(cor(mx.1, use="pairwise.complete.obs"), digits=2)

### Create an adjacency matrix using a threshold of 0.1
adj.1 <- matrix(0, nrow(cor.1), nrow(cor.1))
adj.1[abs(cor.1) > 0.1] <- 1
colnames(adj.1) <- rownames(adj.1) <- colnames(cor.1)
```

Defining network objects in R

Let p denote the number of metabolites in our network.

- **Adjacency matrix**: $p \times p$ matrix, where i, j element is 1 if there is an edge between metabolite i and metabolite j, and 0 otherwise.
- **GraphNEL object**: network object defined in the R **graph** package

```
### Adjacency matrix
print(adj.1)
```

```
##       mt1_1 mt1_2 mt2_1 mt2_2 mt3_1 mt3_2
## mt1_1     1     1     1     1     0     0
## mt1_2     1     1     0     0     0     1
## mt2_1     1     0     1     1     0     0
## mt2_2     1     0     1     1     0     0
## mt3_1     0     0     0     0     1     0
## mt3_2     0     1     0     0     0     1
```

GraphNEL R object

- Convert the adjacency matrix into a GraphNEL object using the **graph** R package.
- Extract information on the nodes and edges of the network.

```
### Converts the adjacency matrix into a graphNEL object
library(graph)
graphObj <- as(adj.1, "graphNEL")
graphObj
```

```
## A graphNEL graph with undirected edges
## Number of Nodes = 6
## Number of Edges = 11
```

```
### Extracting information about the graphNEL object
print(nodes(graphObj))
```

```
## [1] "mt1_1" "mt1_2" "mt2_1" "mt2_2" "mt3_1" "mt3_2"
```

GraphNEL R object

Extract information on the edges of the network.

```
## Printing the edges of the network
print(edges(graphObj))
```

```
## $mt1_1
## [1] "mt1_1" "mt1_2" "mt2_1" "mt2_2"
##
## $mt1_2
## [1] "mt1_1" "mt1_2" "mt3_2"
##
## $mt2_1
## [1] "mt1_1" "mt2_1" "mt2_2"
##
## $mt2_2
## [1] "mt1_1" "mt2_1" "mt2_2"
##
## $mt3_1
## [1] "mt3_1"
##
## $mt3_2
## [1] "mt1_2" "mt3_2"
```

igraph R package

We can convert an adjacency matrix to an igraph object.

```
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 3.5.2
```

```
igraph.obj <- graph.adjacency(adj.1,mode="undirected",weighted=NULL,diag=FALSE)

## Extracting nodes and edges from igraph object
V(igraph.obj)
```

```
## + 6/6 vertices, named, from bded127:
## [1] mt1_1 mt1_2 mt2_1 mt2_2 mt3_1 mt3_2
```

```
E(igraph.obj)
```

```
## + 5/5 edges from bded127 (vertex names):
## [1] mt1_1--mt1_2 mt1_1--mt2_1 mt1_1--mt2_2 mt1_2--mt3_2 mt2_1--mt2_2
```

Visualizing our network

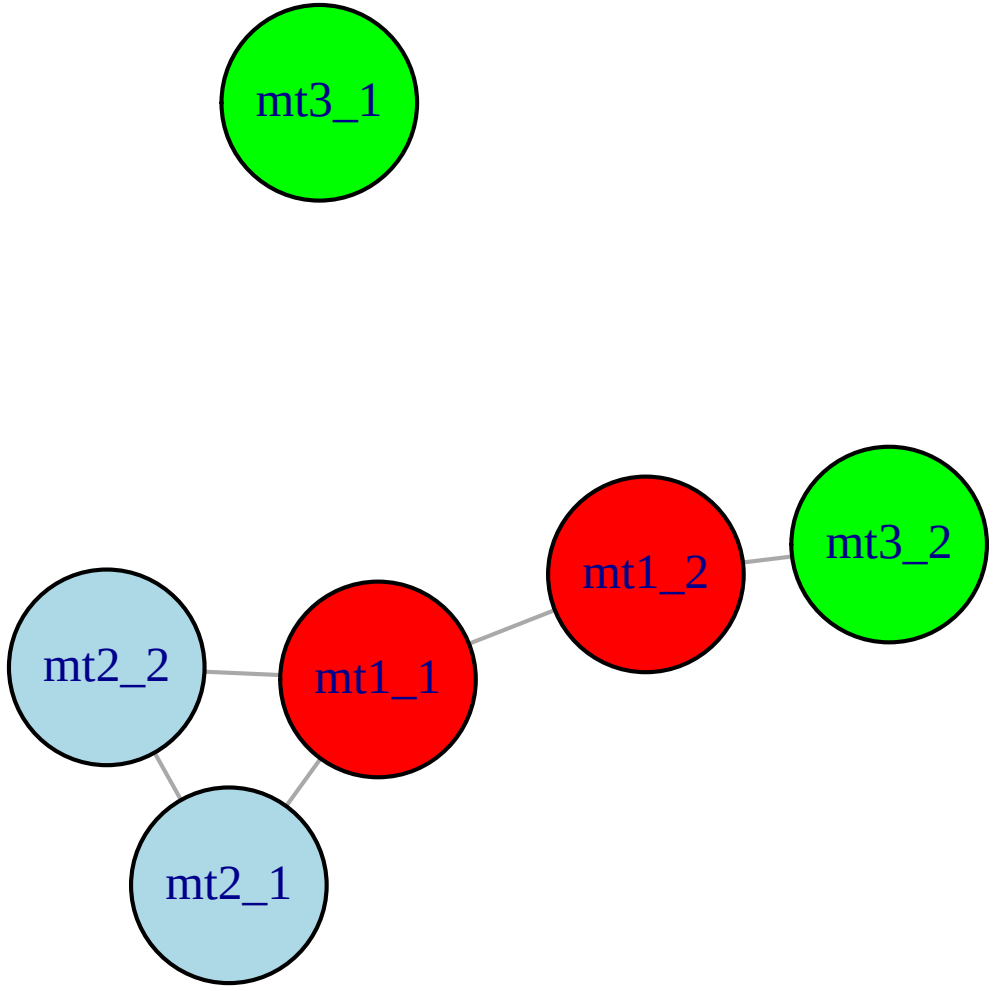
Let's assign metabolite class to each of our nodes and an associated color.

```
### Assigning attributes to the list of nodes

V(igraph.obj)$MxClass <- c(rep("AA", 2), rep("AC", 2), rep("Oth", 2))
V(igraph.obj)$color <- c(rep("red", 2), rep("light blue", 2), rep("green", 2))
V(igraph.obj)$size <- 50
V(igraph.obj)$label.cex <- 0.75
```

Visualizing our network

Visualize the network..

```
### Visualizing network
plot.igraph(igraph.obj, vertex.label = colnames(adj.1), layout = layout.fruchterman.reingold)
```



Changing node attributes

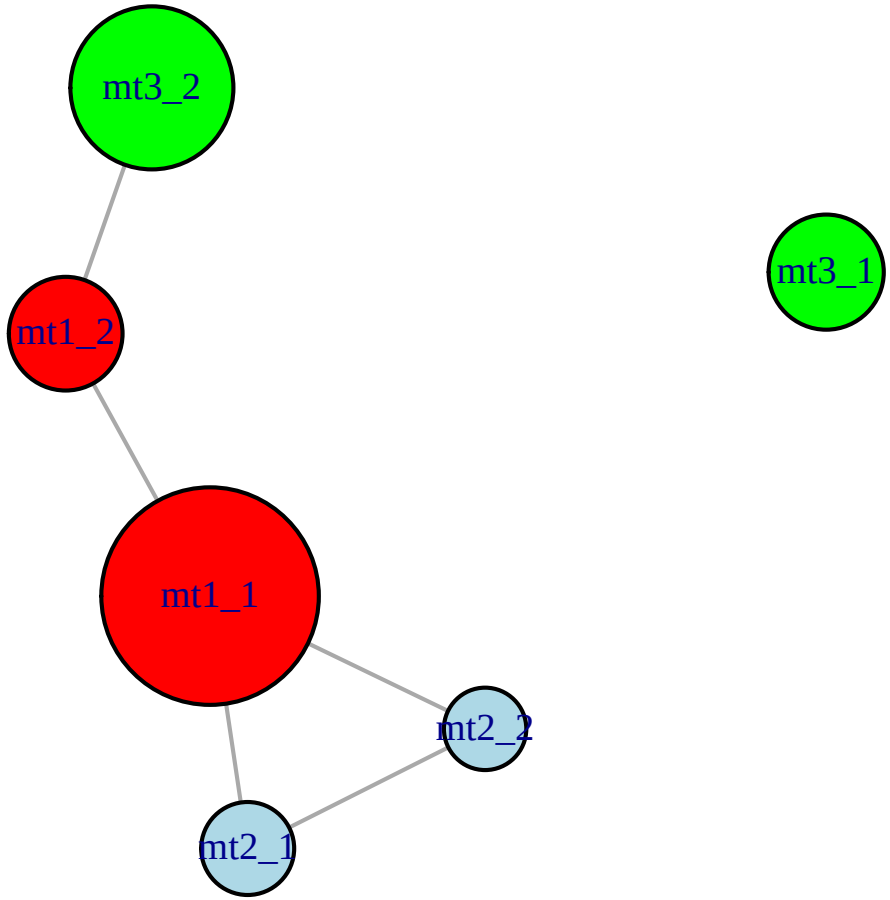
Let's change node size in proportion to significance of association with fasting glucose..

```
### Changing the node size to match the level

### of signficance with outcome (fasting glucose)

myfun <- function(metabolite, outcome) {
    mymod <- lm(outcome ~ metabolite)
    minuslogp <- -log(summary(mymod)$coef[2, 4])
    return(minuslogp)
}

fg1 <- fg[ag == "ag1"]
vals <- apply(mx.1, 2, myfun, fg1)

### scaling the node size changing the font fize of the vertex label

V(igraph.obj)$size <- vals * 3 + 20
V(igraph.obj)$label.cex <- 0.6
```

Visualizing our network

Visualize the network after changing node attributes..

```
### Visualizing network
plot.igraph(igraph.obj, vertex.label = colnames(adj.1), layout = layout.fruchterman.reingold)
```



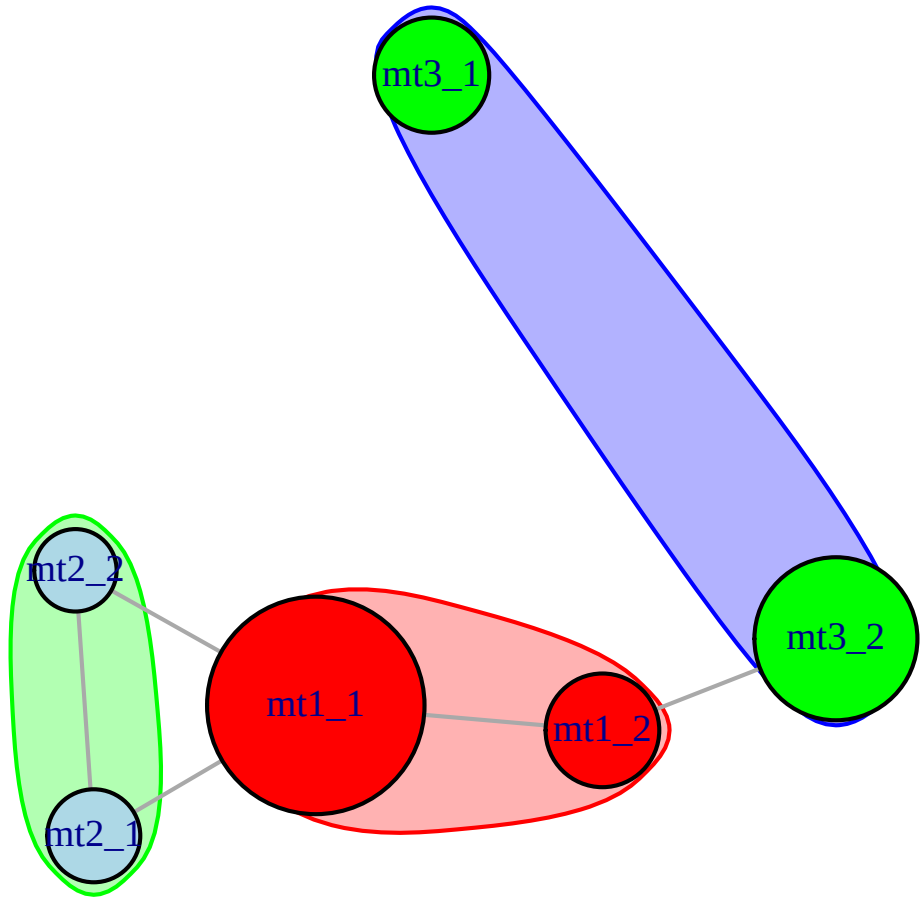
Grouping nodes

We can also visually depict metabolite classes (Amino acids, Acyl carnitines, Other) in our network ..

```
### Visualizing network with node groups
mylist <- list(c("mt1_1", "mt1_2"), c("mt2_1", "mt2_2"), c("mt3_1", "mt3_2"))
```

Grouping nodes

```
plot.igraph(igraph.obj,vertex.label=colnames(adj.1),

            layout=layout.fruchterman.reingold, mark.groups=mylist)
```



Networks in R

There are a myriad of options available for visualizing networks. For more, see help associated with plot.igraph() in the igraph package.

```
### Other layouts (Kamada-Kawai)

### For other options -- Check ?plot.igraph

l <- layout_with_kk(igraph.obj)
plot.igraph(igraph.obj, vertex.label = colnames(adj.1), layout = l, mark.groups = mylist)
```

Gaussian Graphical Models (GGM) in R

GGM in R

We illustrate estimation of the Gaussian graphical model using the R package huge.

To keep in mind:

- Missing values of metabolite levels need to be imputed prior to invoking the functions in **huge**.
- Each metabolite should be standardized to render them of unit variance.

Preliminaries

We prepare metabolite data in ancestry group ag1 for graphical model estimation.

```
### Prepping data for GGM Impute missing values Standardize

standardizeMetabolite = function(x) {
    x[x == Inf] <- NA
    x[is.na(x)] <- min(x, na.rm = T)/2
    return((x - mean(x, na.rm = T))/sd(x, na.rm = T))
}

mx.1 <- mx[ag == "ag1", ]
mx1.s <- apply(mx.1, 2, standardizeMetabolite)

summary(apply(mx1.s, 2, sd))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

GGM estimation

The key functions involved are:

- **huge:** estimates GGM over a range of penalty parameters (can be left unspecified).
- **huge.select:** implements regularization parameter selection. **Reference:** T. Zhao and H. Liu (2012). The huge Package for High-dimensional Undirected Graph Estimation in R. Journal of Machine Learning Research.

GGM estimation

Regularization parameter selection options include:

- StARS: tends to overselects edges.
- RIC: more computationally efficient, tends to underselect edges.
- **Reference**: T. Zhao and H. Liu (2012). The huge Package for High-dimensional Undirected Graph Estimation in R. Journal of Machine Learning Research.

GGM estimation

Let's estimate the GGM network for our data..

```
library(huge)
```

```
## Warning: package 'huge' was built under R version 3.5.2
```

```
### creates the GGM model object
mbModel <- huge(mx1.s, method = "mb")
```

```
## Conducting Meinshausen & Buhlmann graph estimation (mb)....done
```

```
### Optimal parameter selection using ric
mbOptRIC = huge.select(mbModel, criterion = "ric")
```

```
## Conducting rotation information criterion (ric) selection....done
## Computing the optimal graph....done
```

```
### extract the graph corresponding to optimal param
mbOptRICGraph = mbOptRIC$refit
```

GGM

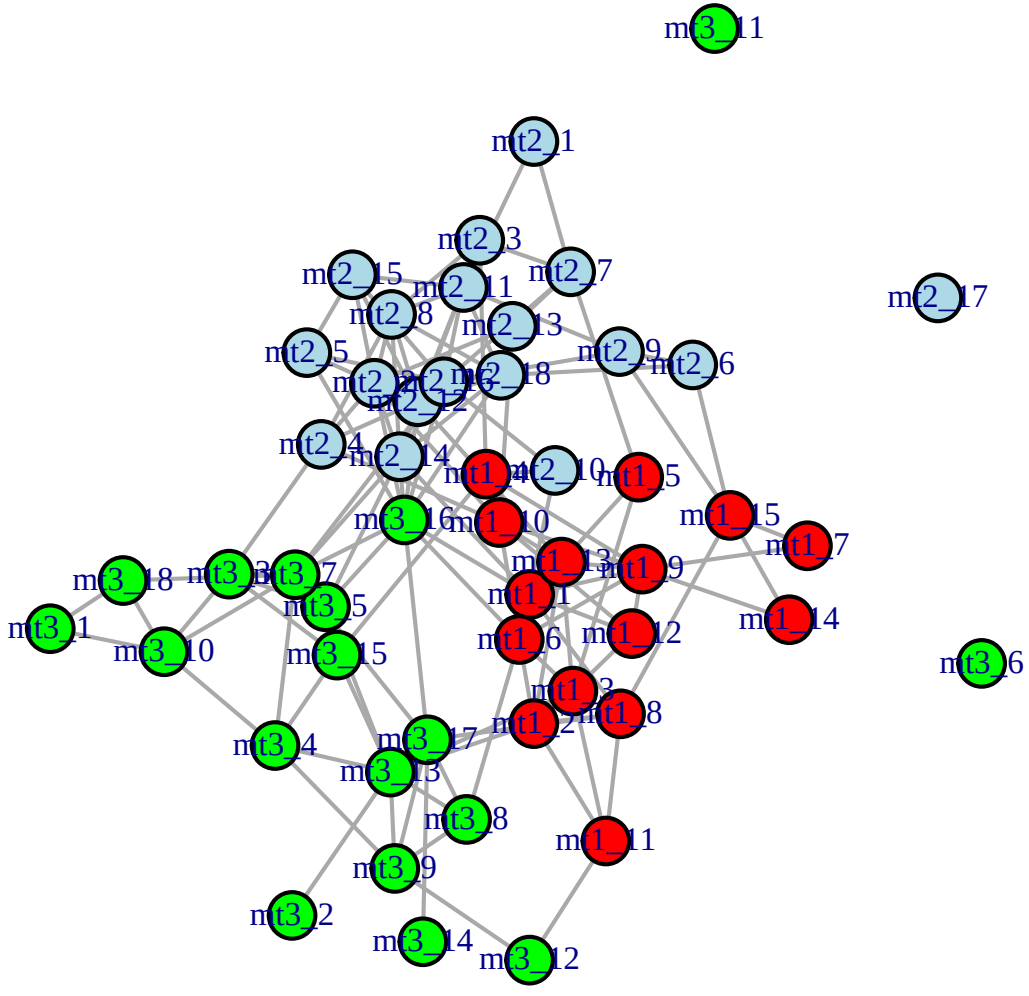
Visualize our estimated GGM ..

Let's estimate the GGM network for our data..

```
myg <- graph_from_adjacency_matrix(mbOptRICGraph, mode = "undirected")
### Assigning attributes to the list of nodes

V(myg)$MxClass <- c(rep("AA", 15), rep("AC", 18), rep("Oth", 18))
V(myg)$color <- c(rep("red", 15), rep("light blue", 18), rep("green", 18))
V(myg)$size <- 10
V(myg)$label.cex <- 0.5
```

GGM

```
### Visualizing network
plot.igraph(myg, vertex.label = colnames(mx.1), layout = layout.fruchterman.reingold)
```



Other options

- **Method:** can be changed to glasso; huge(.., method="glasso").
- **Selecting λ:** in huge.select(.., criterion="stars").
- **Relaxing Gaussian assumption:** using nonparanormal (npn) transformation; huge.npn() will return a transformed data matrix.

Next ..

Telling stories with GGMs

- Detecting communities within networks
- Differential networks
- Case studies

References

- Su, J. and Clish, C. (2018). Metabolomics and Network Medicine, Network Medicine: Complex Systems in Human Disease and Therapeutics, Harvard University Press.
- Go, KI, Cusick, ME, Valle, D, Childs B, Vidal M, Barabási AL (2007).The human disease network, PNAS, 104(21):8685-90.
- Sandler, V., Reisetter, A. C., Bain, J.R., ..., Scholtens, D.M., Lowe, W.L.Jr (2018) Associations of maternal BMI and insulin resistance with the maternal metabolome and newborn outcomes, Diabetologia, 60(3):518-530.
- Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso, Annals of Statistics, Vol. 34, No. 3, 1436-1462.
- Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9(3):432-441.
- Roeder, K., Lafferty, J., Wasserman, L., Zhao, T., Liu, H. (2012) The huge package for high-dimensional undirected graph estimation in R. Journal of Machine Learning Research, (13):1059–1062.