But are the networks different?

Using Differential Network Analysis Software with Metabolite Data

COMETS Early Career Investigator Group Meeting: October 11, 2022

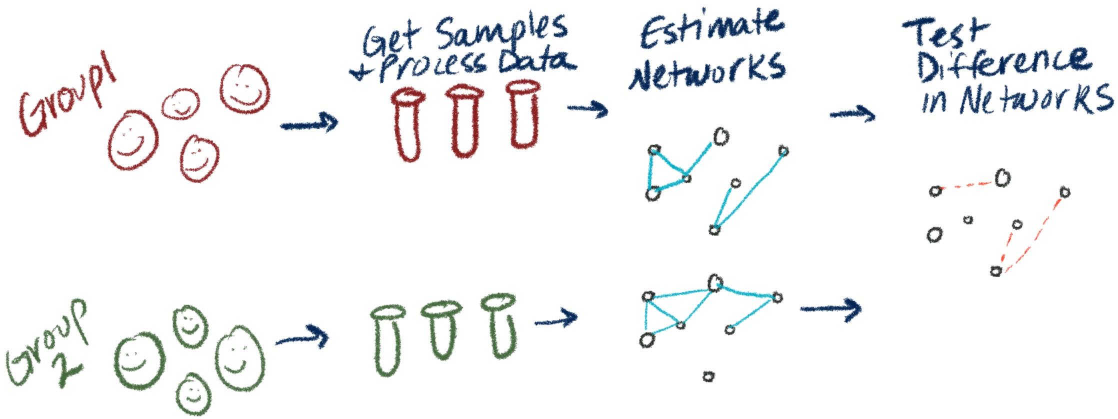
Margaret Janiczek

PhD Student, Biostatistics

University *of* Massachusetts Amherst

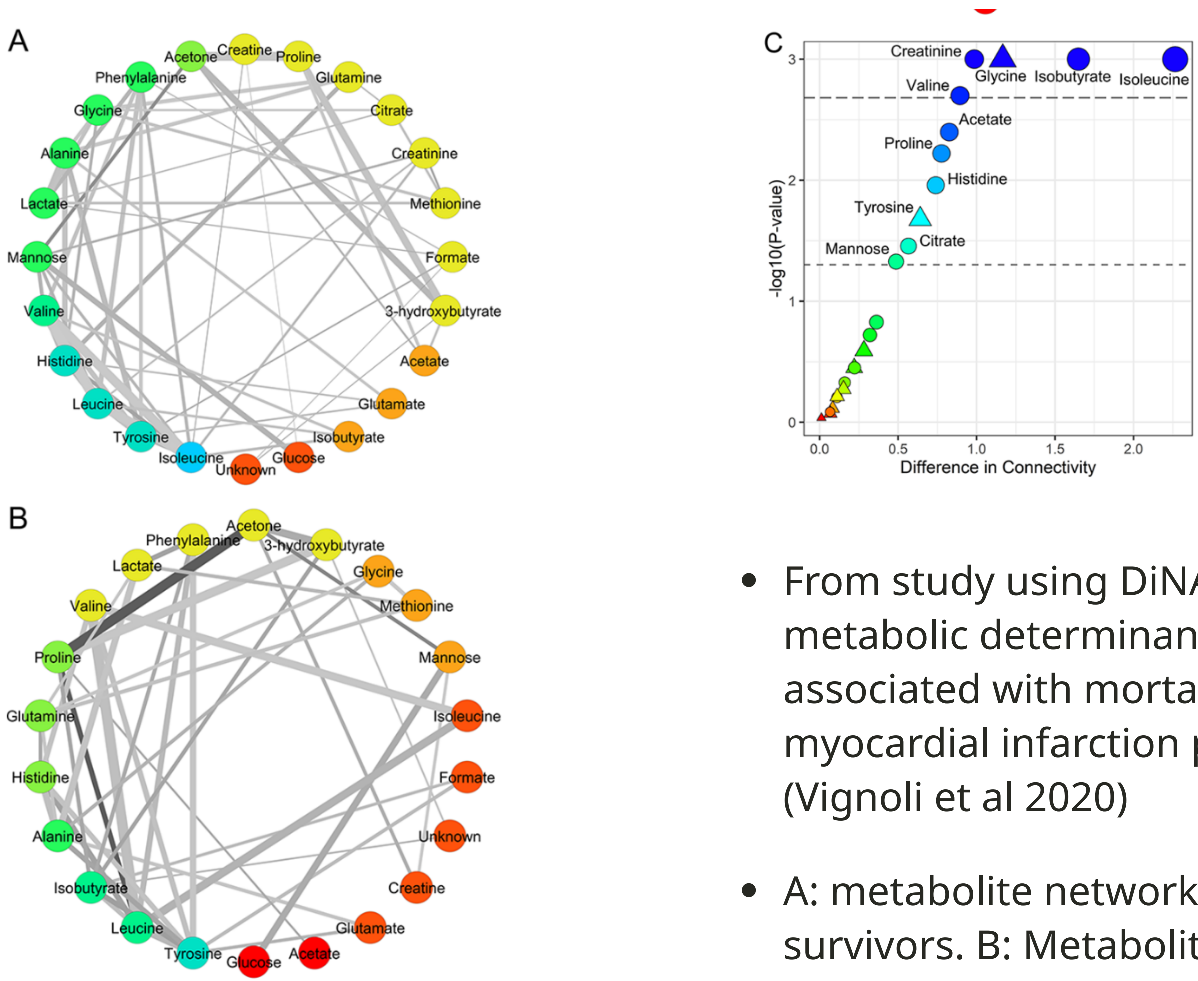
Motivation

- Identifying networks in biomedical data, and how they differ across populations, can help find drivers of disease and targets for treatment



- Certain biomedical research questions lend themselves well to network/pathway analysis
    - Data from brain scans (Alzheimer's patient scans over time)
    - Gene expression (cancer vs normal tissue)
    - Microbiome (Crohn's disease vs Healthy Control)
    - Metabolomics - any applications from the group here?

Motivation



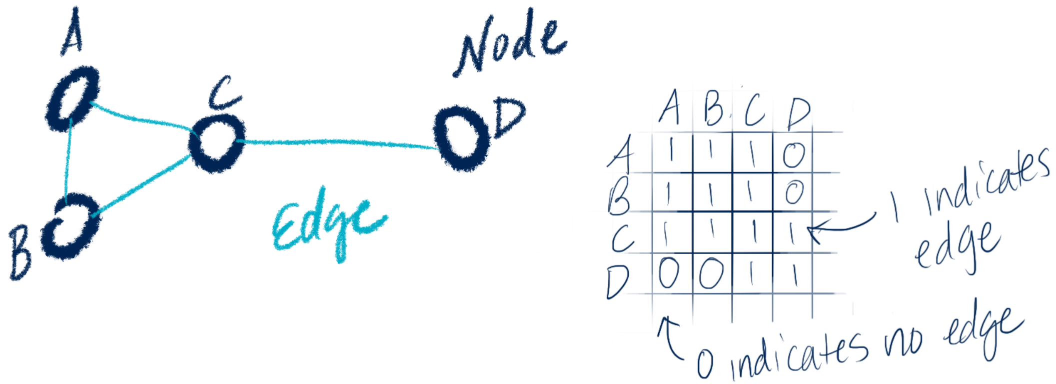
- From study using DiNA to reveal metabolic determinants associated with mortality in acute myocardial infarction patients (Vignoli et al 2020)
- A: metabolite network from survivors. B: Metabolite Network

Presentation Overview

- Background on graphical models and differential networks
- Overview of statistical landscape for differential network analysis
- Overview of available software
- Brief practical application using a few software options
- Discussion & feedback!

Background: Undirected Graphical Model

- Graphical models express connections between variables. When undirected, the connection doesn't imply any directionality.



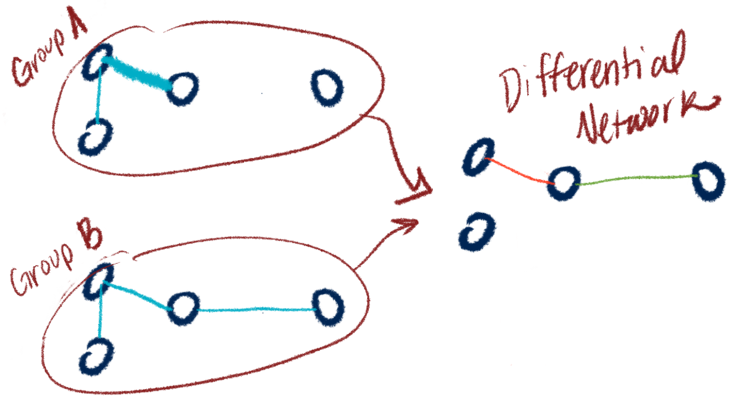
- Connected edges can be seen in a **Precision Matrix**, where anything with a zero is considered "conditionally independent"
- In this example, A and B are **conditionally independent** of D

Background: Gaussian Graphical Model

- If we can assume the data are normally distributed, the **Precision Matrix** can be estimated using the inverse of the correlation matrix!
- High dimensional data can be handled by adding shrinkage penalties which will force values down to zero.
- There are many other estimation details I won't go into here.
- See Kate Shutta's recently published tutorial on Gaussian Graphical Models for details! [Shu+22]

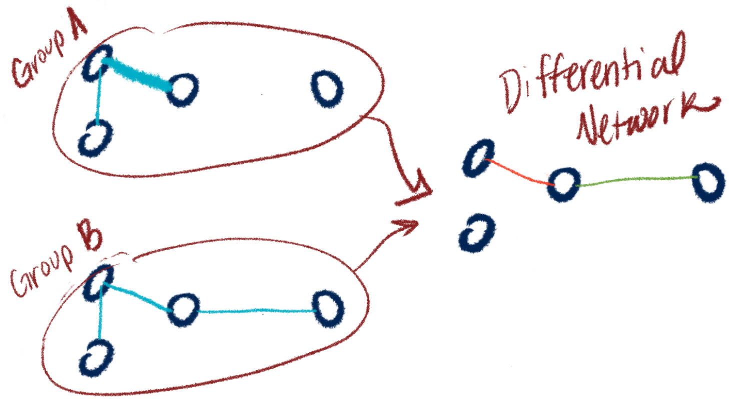
More than one graphical model

- Say you have data from two groups, like disease and healthy control.
- Say you estimate a graphical model for each group, then want to compare the resulting networks.



But are the networks different??

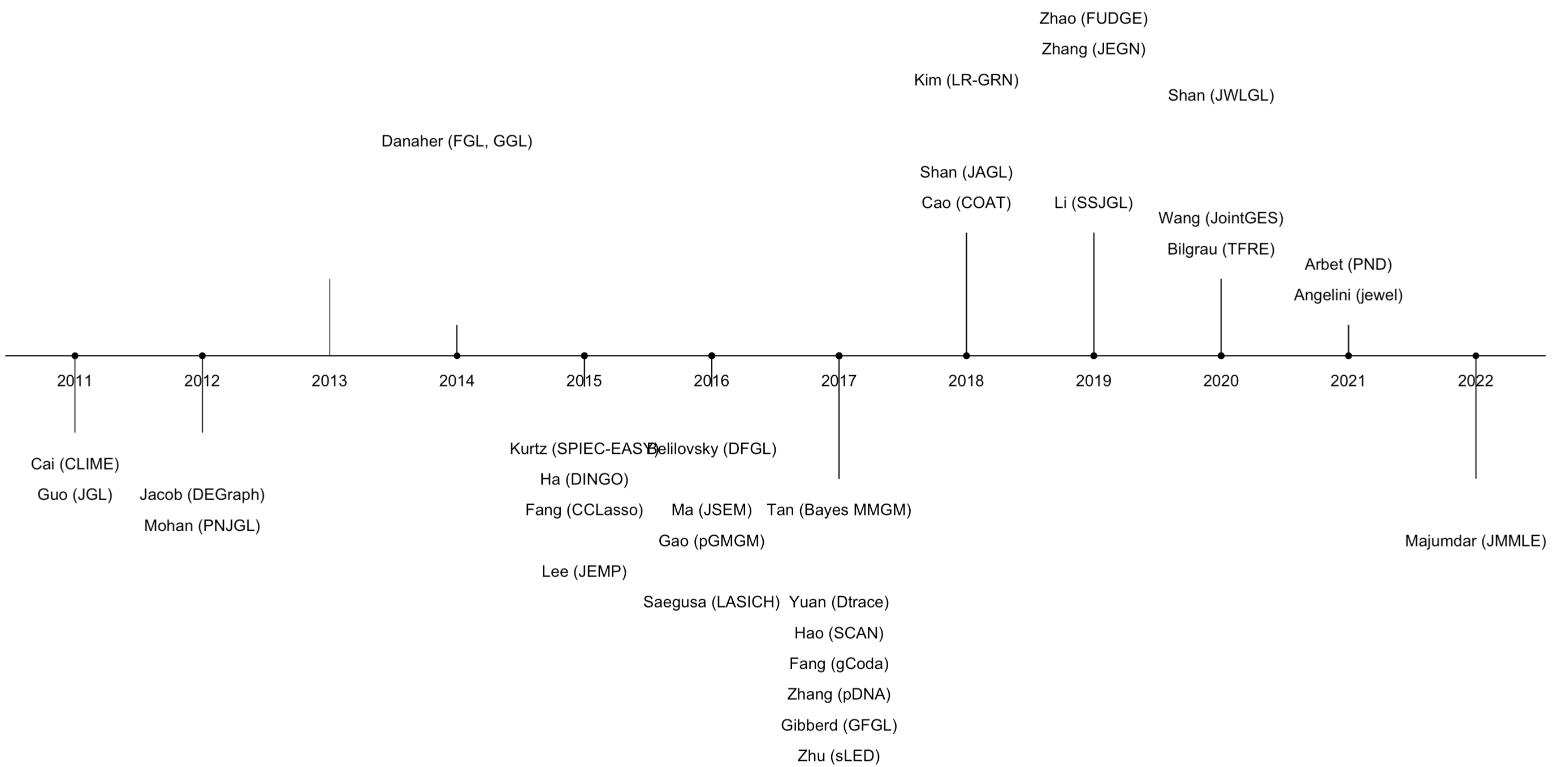
- How do you estimate them?
- How do you test the difference?
- How do you even *characterize* the difference? (edges? nodes? hubs? general structure?)
- This all falls under DIFFERENTIAL NETWORK ANALYSIS! (DiNA)



Statistical Landscape of DiNA methods

Timeline

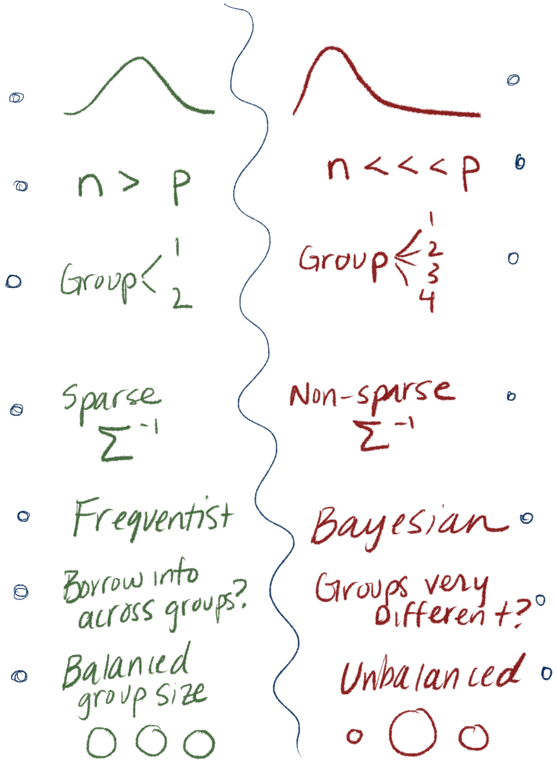
- I found 40+ methods papers on DiNA methods published in the last 10 years
- The wide variety is due to addressing many subtly different problems

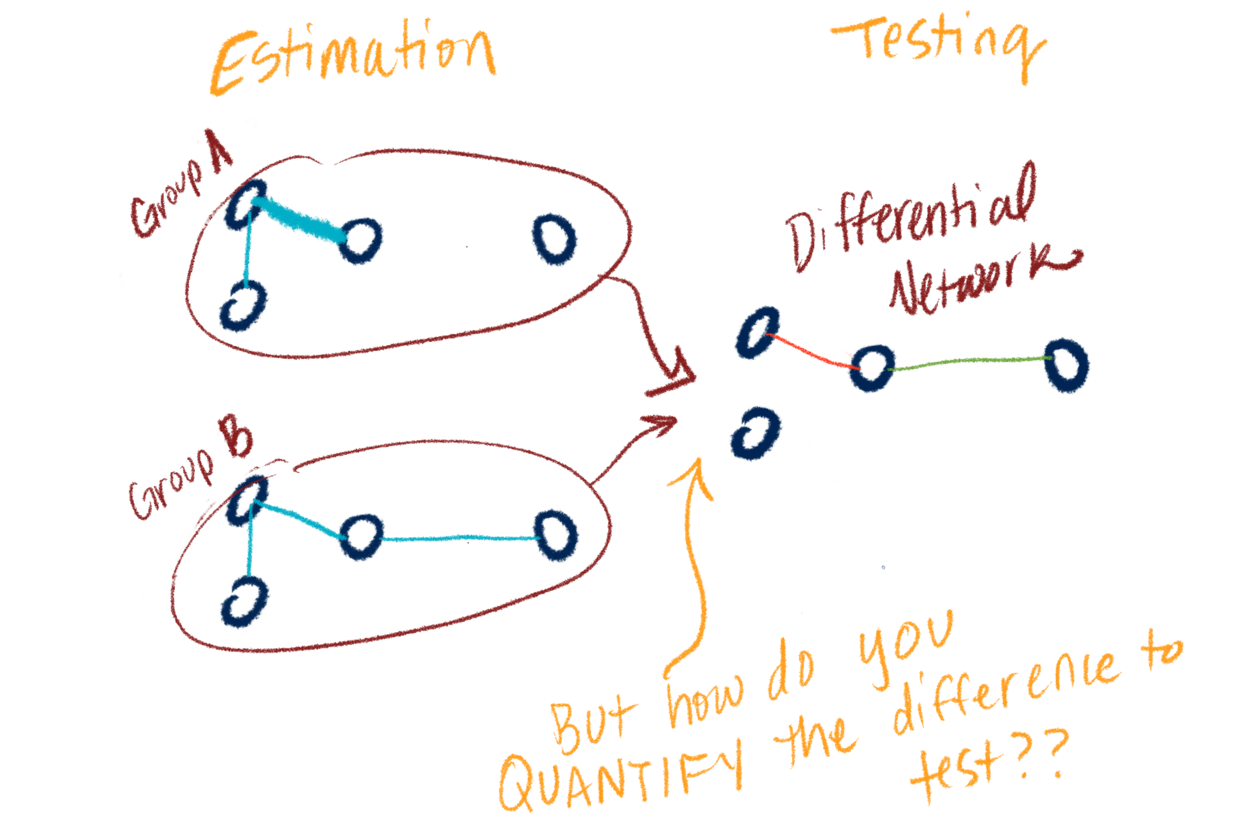


Why so many methods?

To address various data and modeling situations!

What's your data like?




Estimation
Testing
Group A
Group B
Differential Network
But how do you QUANTIFY the difference to test??

Quantifying "difference": Local Structure

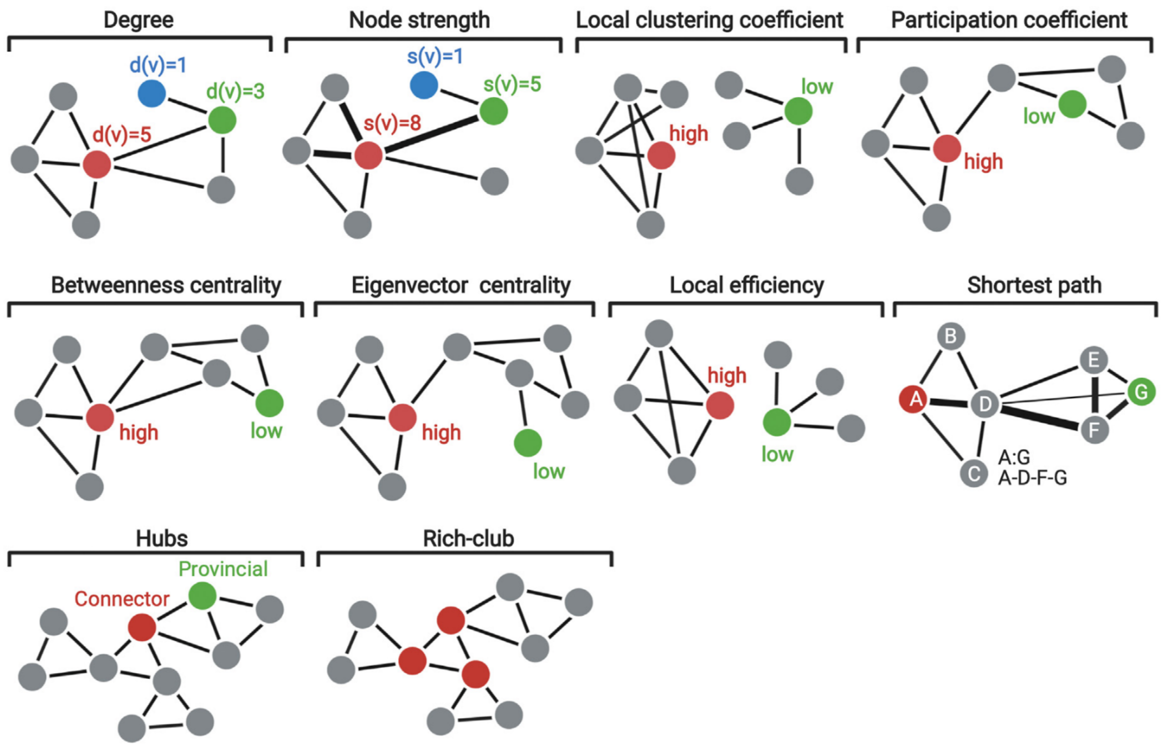


Figure 3. Illustration of local measures. Circles and connecting lines represent edges and nodes, respectively. Important nodes highlighted in red, green or blue). The route of the shortest path is shown in a weighted graph. Considering a binary graph, the shortest path then changes to A-D-G.

Quantifying "difference": Global Structure

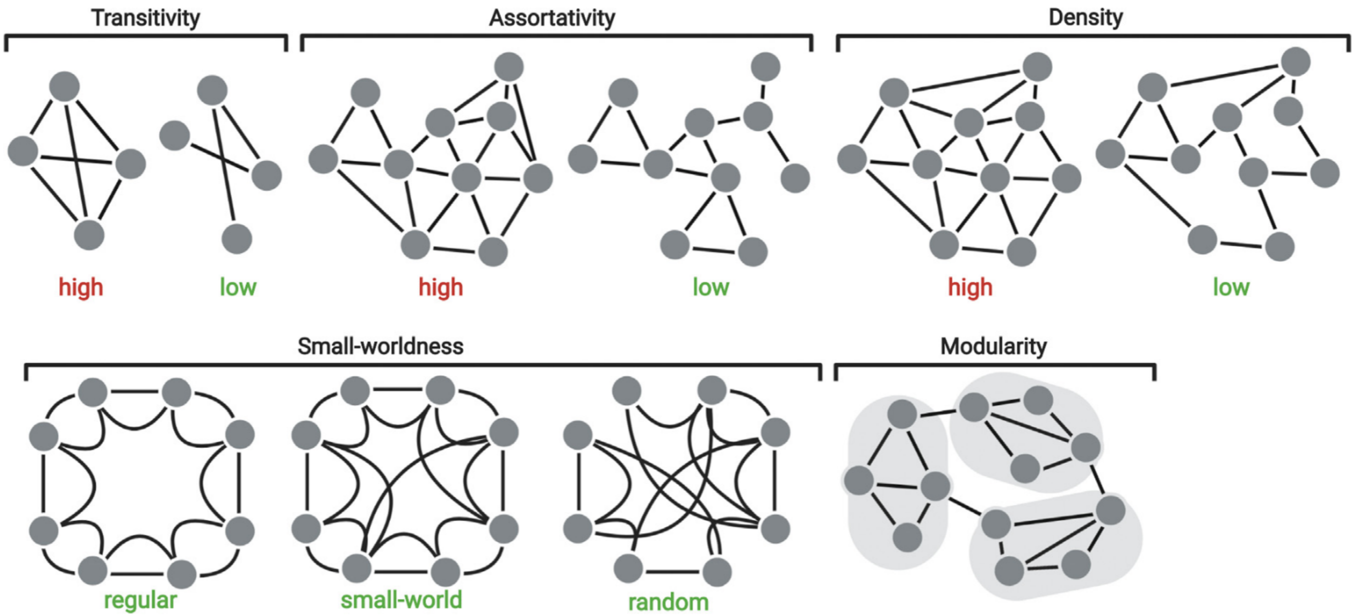


Figure 4. Illustration of global measures. Circles and connecting lines represent edges and nodes, respectively.

And finally... test the difference

For evaluating local difference:

- Can do things like test each node for local structure difference (i.e. test each metabolite for "Degree" or "Node Strength" or "Betweenness Centrality")
- Various p-value options, e.g. permutation
- Adjust for multiple testing!! Bonferoni for conservative estimate, FDR for less stringent.

For evaluating global difference:

- Can do visualization (`iGraph`) and describe global structural differences (e.g. Density, modularity)
- Can perform global hypothesis test H0: $\Sigma_1 = \Sigma_2$ vs H1: $\Sigma_1 \neq \Sigma_2$
- For low-dimensional data this is simpler (e.g. `covTestR` package)
- For high-dimensional data use method proposed by Li & Chen 2012

Software Landscape of DiNA methods

Overview DiNA software landscape

- I found 26 different R packages and 2 Python packages that implement a variety of subtly different DiNA algorithms/pipelines

Notes on software

- `JGL`, `iDingo`, `rags2ridges`, and `SpiecEasy` seem to be most popular and cited.
- I have a full tutorial for `JGL` posted on my GitHub, and Kate has one available for `iDingo`.
- Currently working making tutorials for for `rags2ridges`, `Spiec-Easy` and will work through the other available methods

Placeholder for JGL example

- JGL package runs Fused Graphical Lasso (FGL) and Group Graphical Lasso from Danaher et al 2014
- Estimates sparse covariance matrices that are *similar* across classes
- Has a lot of useful functions to analyze the networks after estimating them, for example extracting hubs, edges, degree etc.
- Graphical lasso uses L1 penalty, which encourages sparsity and as a result selects edges in the graph in the process of estimating precision matrix

Placeholder for rags2ridges example

- `rags2ridges` is great for p >>> n settings. Uses L2-penalized estimation of precision matrices
- Useful when classes are believed to share most of the same structure
- Graphical ridge uses L2 penalty which doesn't shrink things down to zero, so you select edges AFTER estimating the precision matrix. Can be useful if there's a lot of colinearity
- As a side note this package is nicely written with some fun easter eggs hidden in

Placeholder for Spiec-Easy example

Placeholder for iDingo example

Placeholder for iGraph overview

Questions for the group

- What do you find helpful in a tutorial or when identifying methods to use?
- I propose to both use simulated data (under various conditions referenced above) and several real-world datasets with all the software.

Takeaways

- DiNA has potential to be a useful tool in biomedical research
- There are many ways to customize the estimation and testing process to fit research question and data types
- However the broad landscape of methods and software and the current lack of practical applied tutorials comparing software methods seems like a barrier to widespread use
- I'm working on trying to bridge the gap between statistical methodology and applied researchers! Full tutorial forthcoming!

Questions & Comments?

Thank you!

- Dr. Raji Balasubramanian & Balasubramanian Lab
- Dr. Kate Hoff Shutta

Github: @mljaniczek
Website: mljaniczek.github.io/
Slides created via the R package **xaringan**.

References

Cai, T., W. Liu, and X. Luo (2011). "A Constrained $\ell 1$ Minimization Approach to Sparse Precision Matrix Estimation". In: *Journal of the American Statistical Association* 106.494. Publisher: Taylor & Francis _ eprint: https://doi.org/10.1198/jasa.2011.tm10155, pp. 594-607. ISSN: 0162-1459. DOI: 10.1198/jasa.2011.tm10155. URL: https://doi.org/10.1198/jasa.2011.tm10155 (visited on Aug. 18, 2022).

Danaher, P., P. Wang, and D. M. Witten (2014). "The joint graphical lasso for inverse covariance estimation across multiple classes". In: *J R Stat Soc Series B Stat Methodol* 76.2, pp. 373-397. ISSN: 1369-7412. DOI: 10.1111/rssb.12033.

Guo, J., E. Levina, G. Michailidis, et al. (2011). "Joint estimation of multiple graphical models". In: *Biometrika* 98.1, pp. 1-15. ISSN: 0006-3444. DOI: 10.1093/biomet/asq060.

Jacob, L., P. Neuvial, and S. Dudoit (2012). "More power via graph-structured tests for differential expression of gene networks". In: *The Annals of Applied Statistics* 6.2. Publisher: Institute of Mathematical Statistics, pp. 561-600. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/11-AOAS528. URL: https://projecteuclid.org/journals/annals-of-