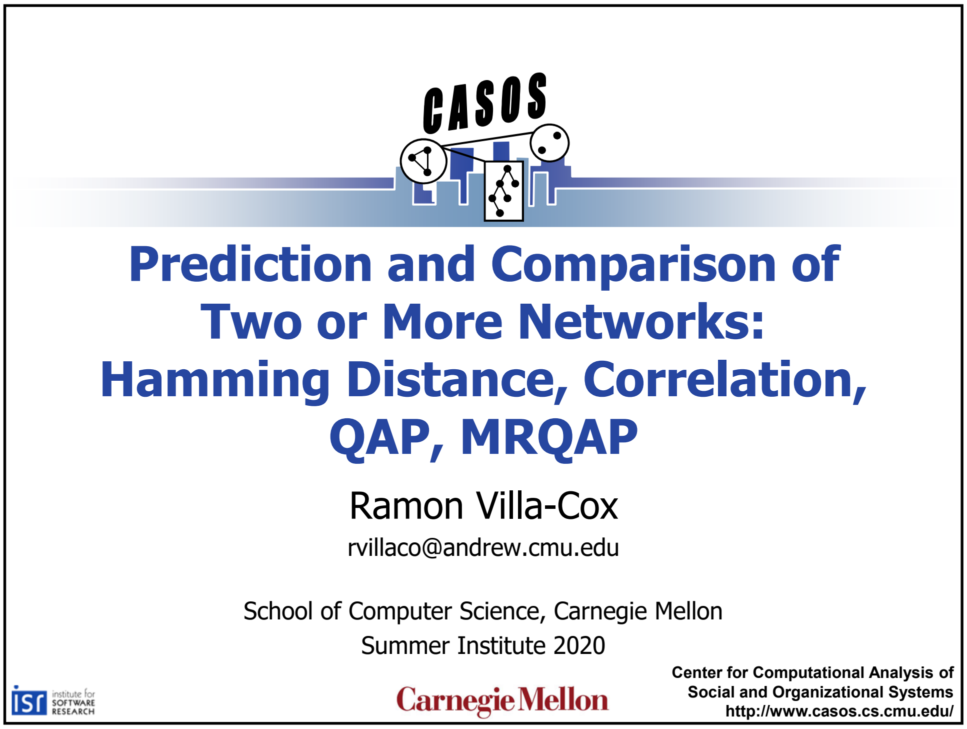


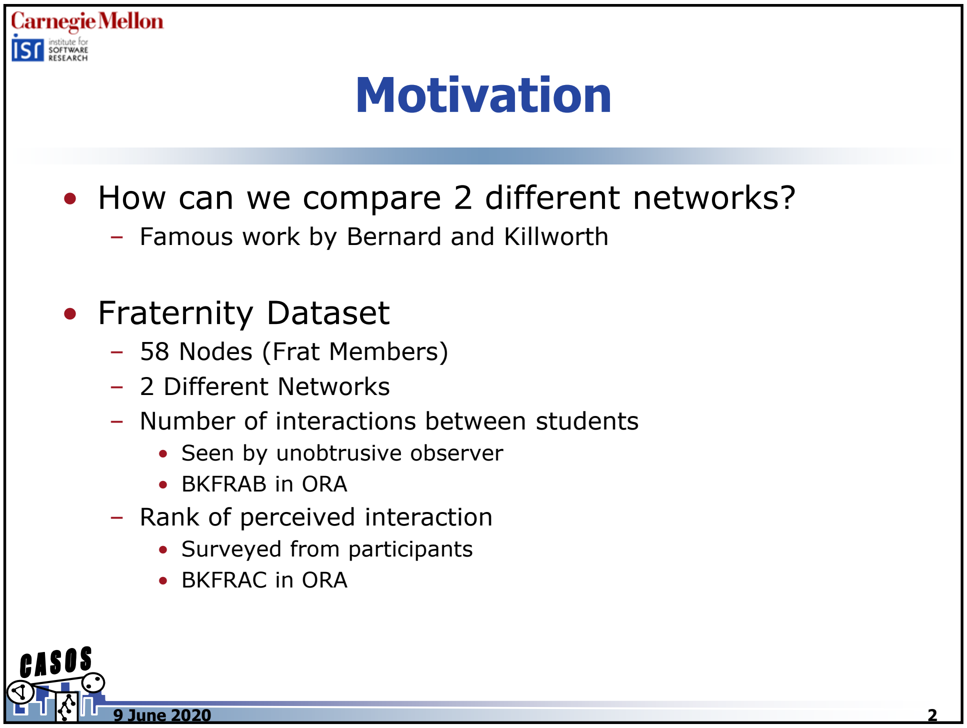
Prediction and Comparison of Two or More Networks: Hamming Distance, Correlation, QAP, MRQAP

Ramon Villa-Cox
rvillaco@andrew.cmu.edu

School of Computer Science, Carnegie Mellon

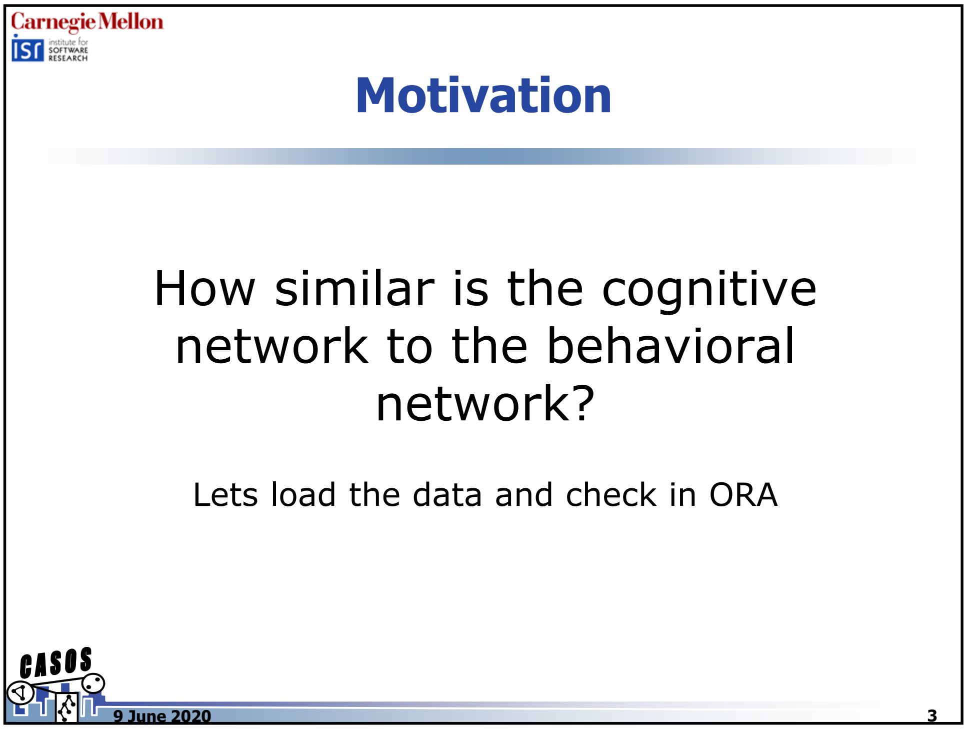
Summer Institute 2020

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/



Motivation

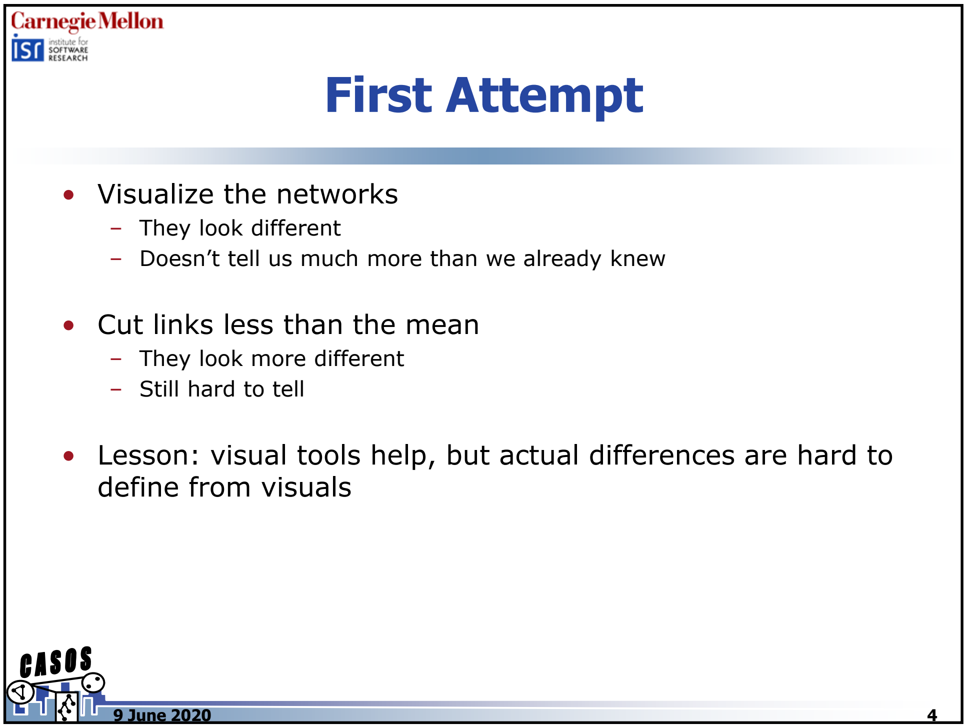
- How can we compare 2 different networks?
  - Famous work by Bernard and Killworth
- Fraternity Dataset
  - 58 Nodes (Frat Members)
  - 2 Different Networks
  - Number of interactions between students
    - Seen by unobtrusive observer
    - BKFRAB in ORA
  - Rank of perceived interaction
    - Surveyed from participants
    - BKFRAC in ORA



Motivation

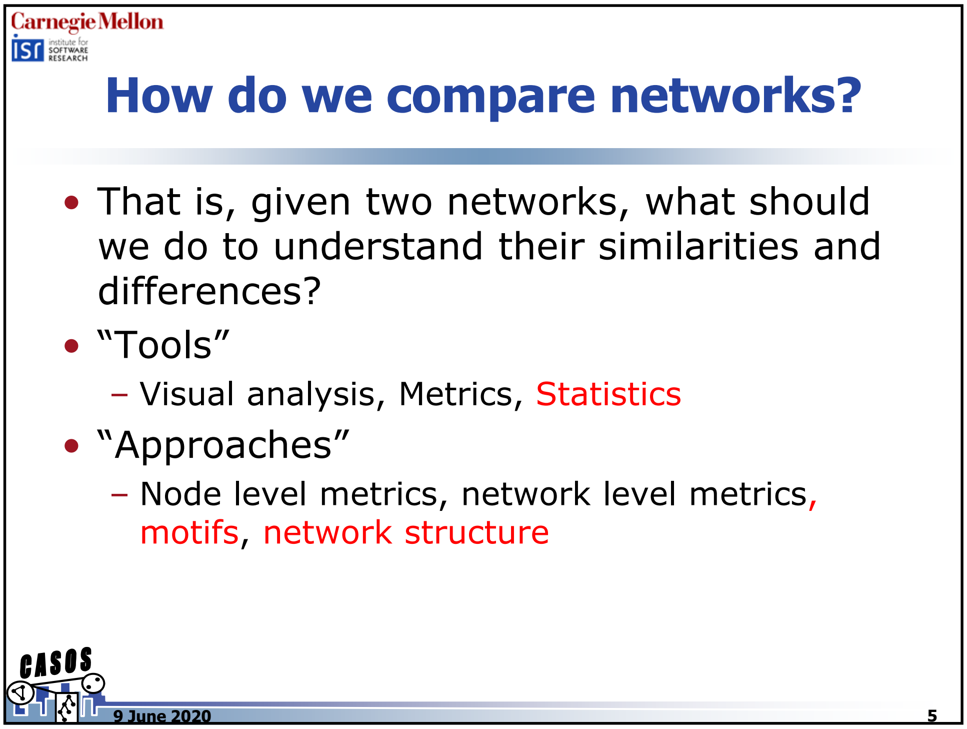
How similar is the cognitive network to the behavioral network?

Lets load the data and check in ORA



First Attempt

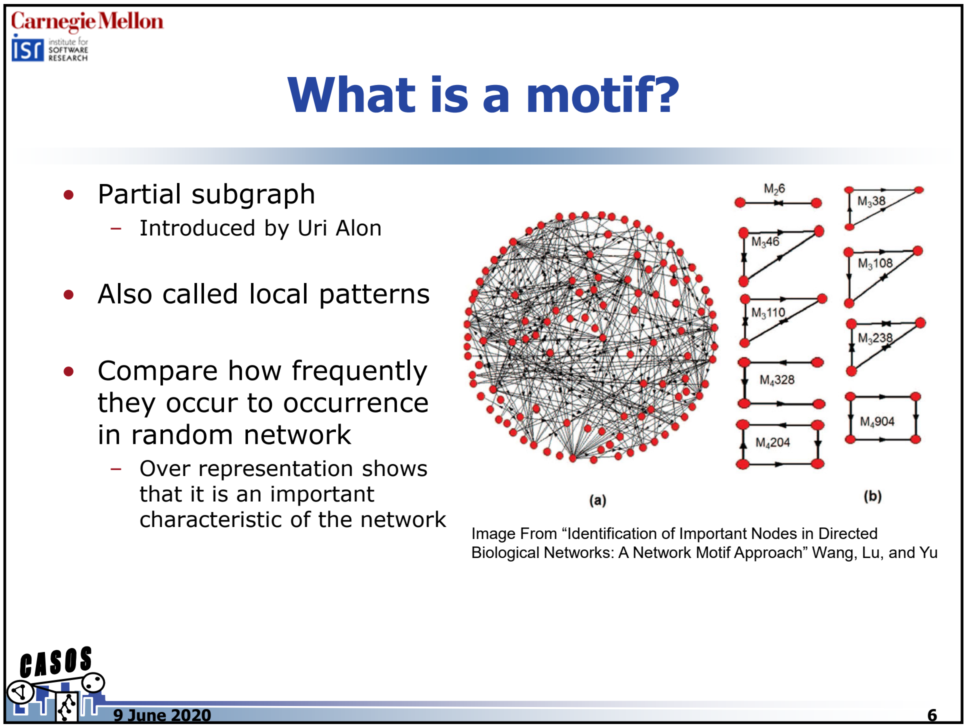
- Visualize the networks
  - They look different
  - Doesn’t tell us much more than we already knew
- Cut links less than the mean
  - They look more different
  - Still hard to tell
- Lesson: visual tools help, but actual differences are hard to define from visuals



Carnegie Mellon
institute for SOFTWARE RESEARCH

How do we compare networks?

- That is, given two networks, what should we do to understand their similarities and differences?
- “Tools”
  - Visual analysis, Metrics, Statistics
- “Approaches”
  - Node level metrics, network level metrics, motifs, network structure



Carnegie Mellon
institute for SOFTWARE RESEARCH

What is a motif?

- Partial subgraph
  - Introduced by Uri Alon
- Also called local patterns
- Compare how frequently they occur to occurrence in random network
  - Over representation shows that it is an important characteristic of the network

Image From “Identification of Important Nodes in Directed Biological Networks: A Network Motif Approach” Wang, Lu, and Yu

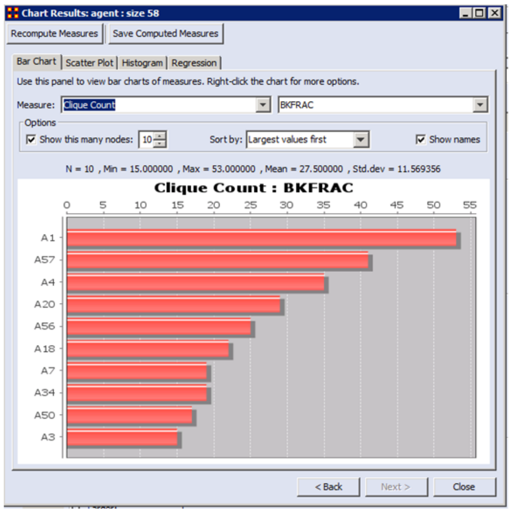
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Motifs in ORA

- Measure Charts
- All Measures
- Clique Count
- Doesn’t work for fully-connected weighted graph!
  - Have to binarize first

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Motifs in ORA



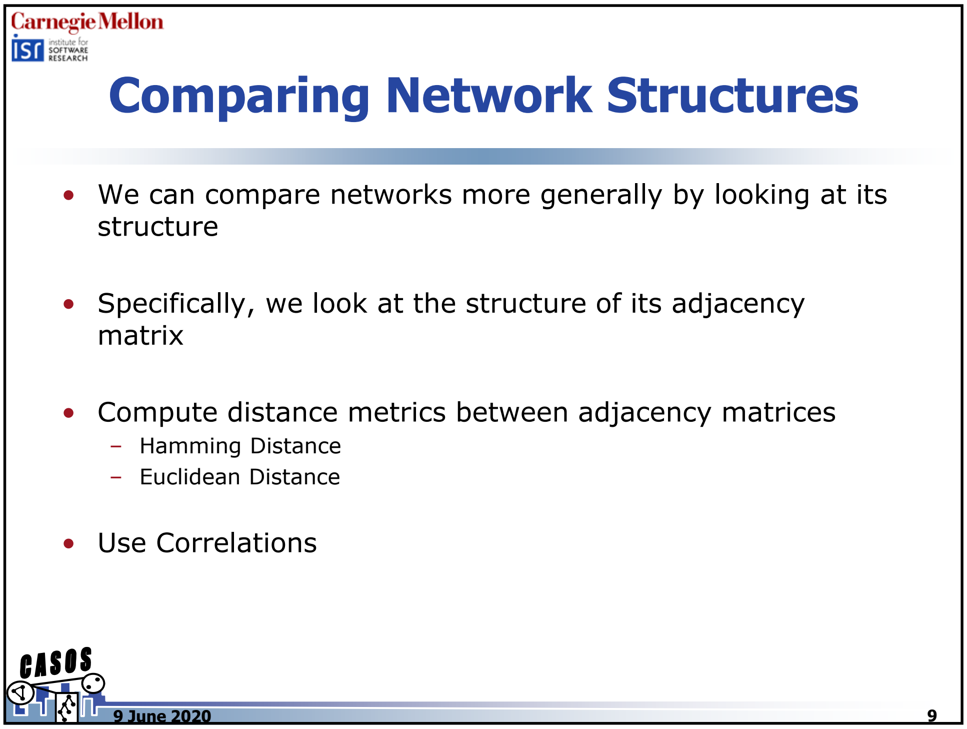
Use this panel to view bar charts of measures. Right-click the chart for more options.

Measure: Clique Count BKFRAC

Options

☑ Show this many nodes: 10 Sort by: Largest values first ☑ Show names

N = 10 , Min = 15.000000 , Max = 53.000000 , Mean = 27.500000 , Std.dev = 11.569356



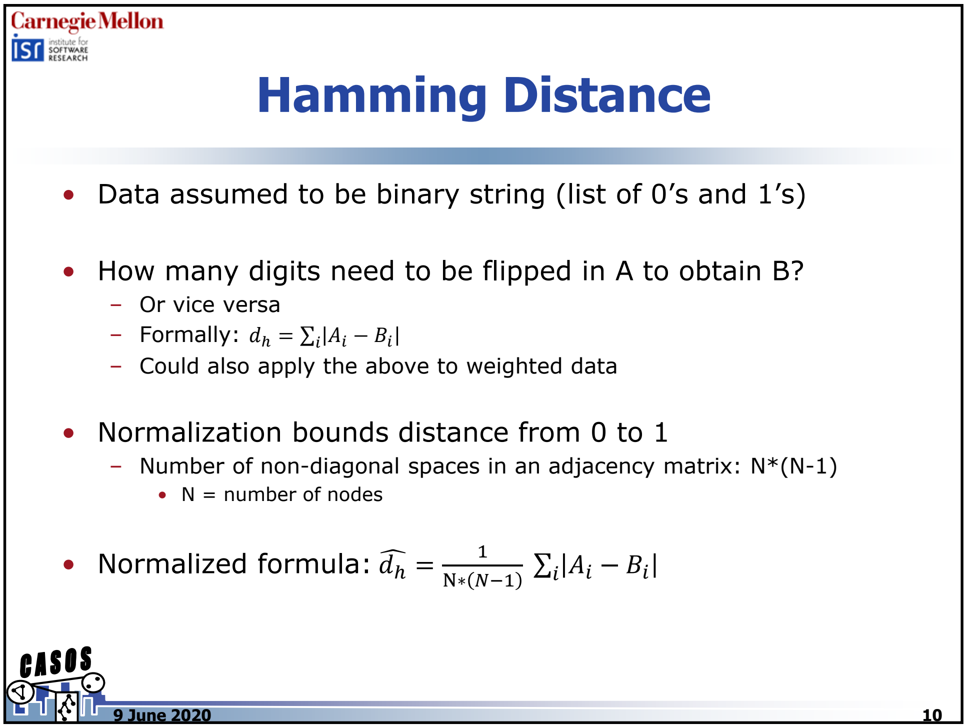
Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Comparing Network Structures

- We can compare networks more generally by looking at its structure
- Specifically, we look at the structure of its adjacency matrix
- Compute distance metrics between adjacency matrices
  - Hamming Distance
  - Euclidean Distance
- Use Correlations

CASOS



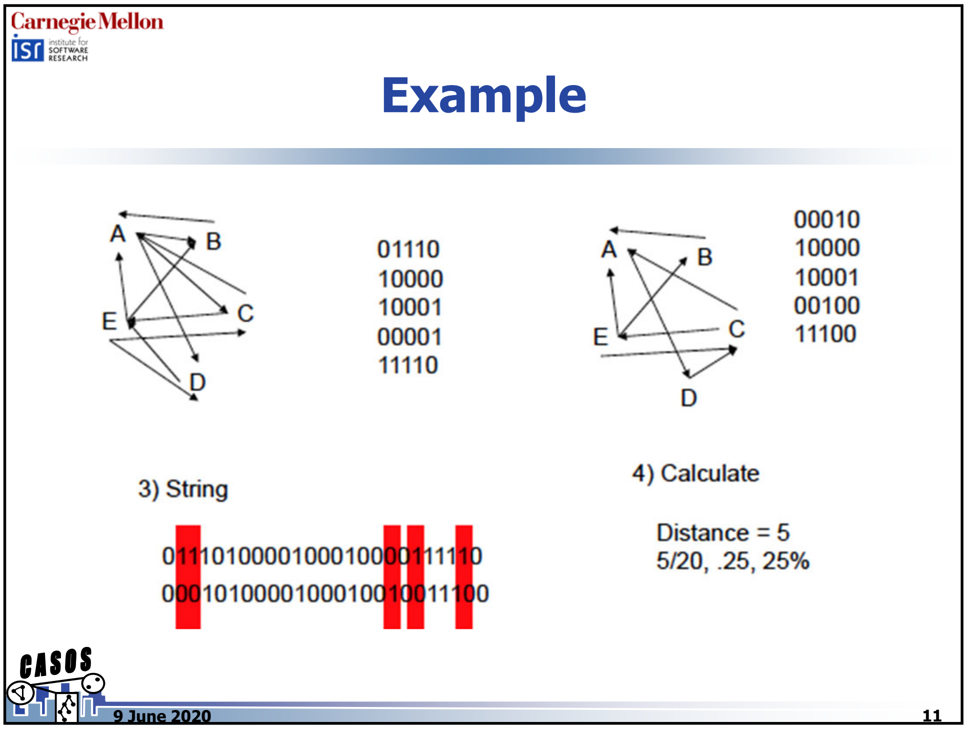
Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Hamming Distance

- Data assumed to be binary string (list of 0’s and 1’s)
- How many digits need to be flipped in A to obtain B?
  - Or vice versa
  - Formally: $d_h = \sum_i |A_i - B_i|$
  - Could also apply the above to weighted data
- Normalization bounds distance from 0 to 1
  - Number of non-diagonal spaces in an adjacency matrix: N*(N-1)
    - N = number of nodes
- Normalized formula: $\widehat{d_h} = \frac{1}{\mathrm{N}*(N-1)} \sum_i |A_i - B_i|$

CASOS



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Example

A
B
C
D
E

01110
10000
10001
00001
11110

A
B
C
D
E

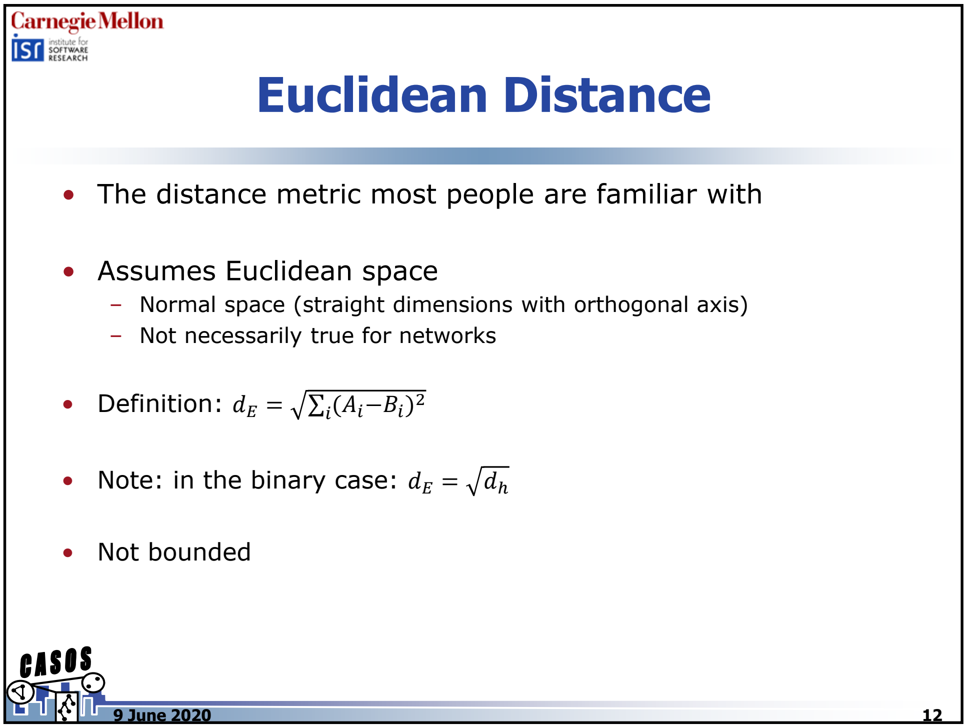
00010
10000
10001
00100
11100

3) String

0111010000100010000111110
0001010000100010010011100

4) Calculate

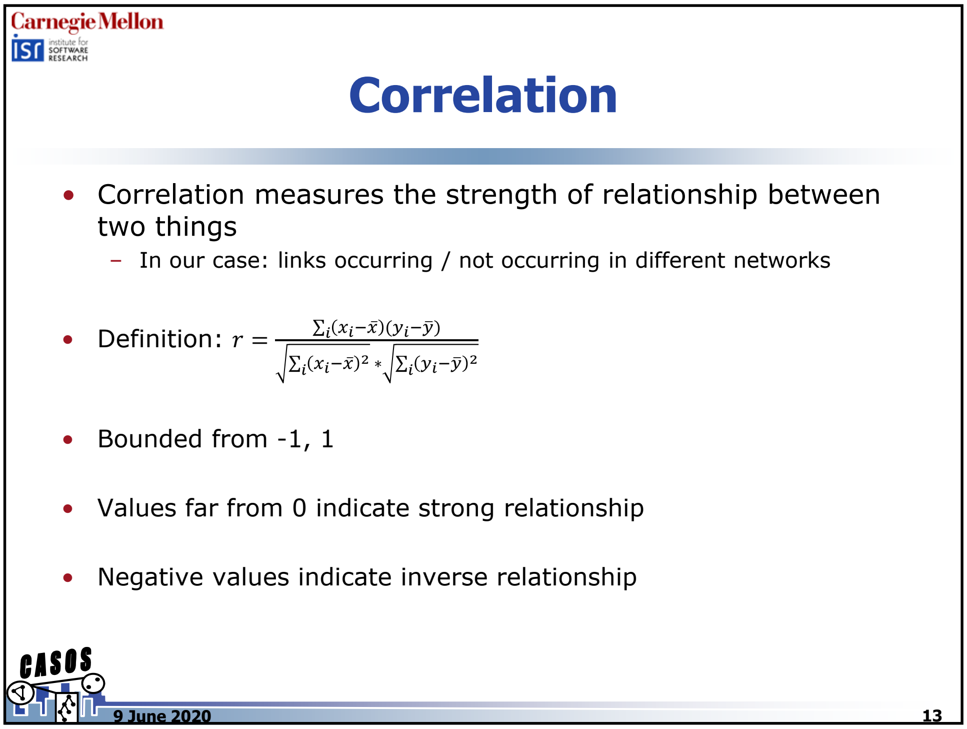
Distance = 5
5/20, .25, 25%



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Euclidean Distance

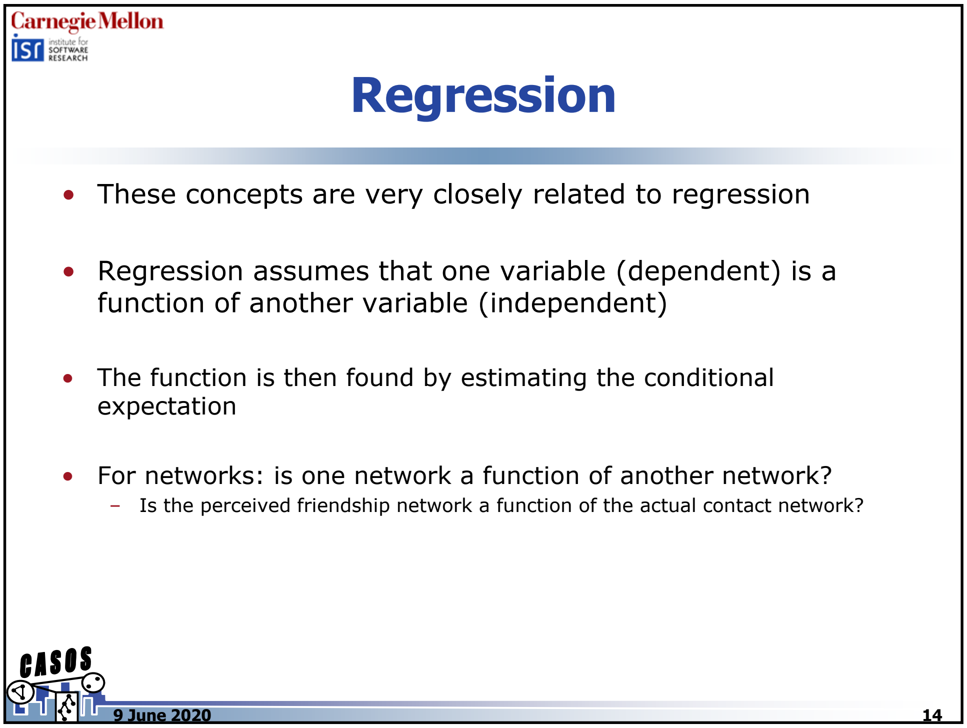
- The distance metric most people are familiar with
- Assumes Euclidean space
  - Normal space (straight dimensions with orthogonal axis)
  - Not necessarily true for networks
- Definition: $d_E = \sqrt{\sum_i (A_i - B_i)^2}$
- Note: in the binary case: $d_E = \sqrt{d_h}$
- Not bounded



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Correlation

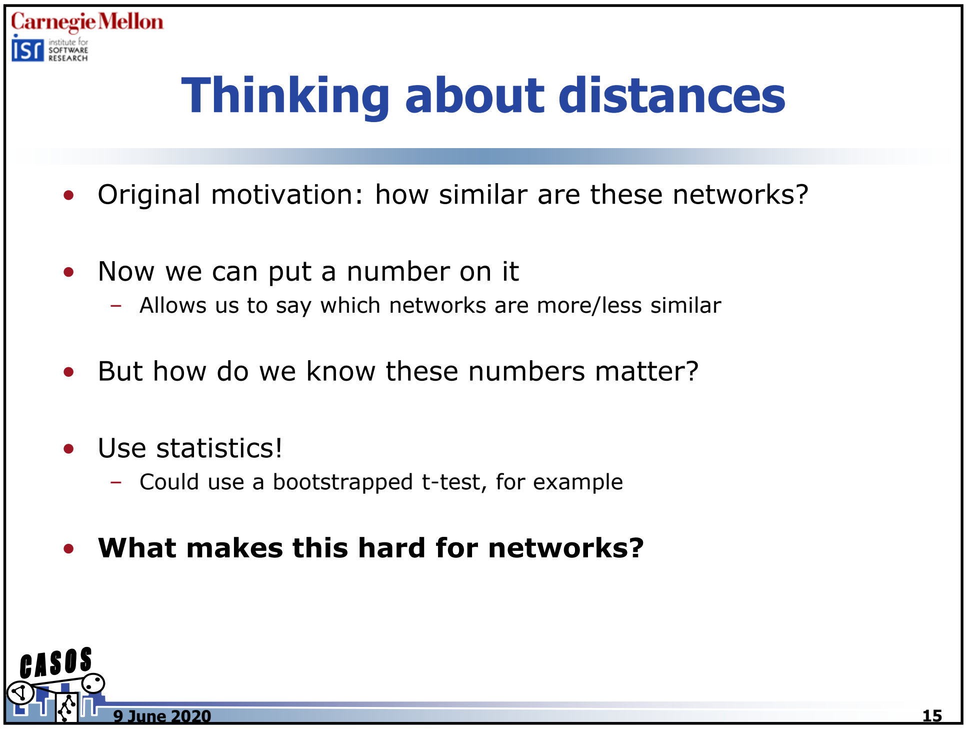
- Correlation measures the strength of relationship between two things
  - In our case: links occurring / not occurring in different networks
- Definition: $r = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2} * \sqrt{\sum_i (y_i-\bar{y})^2}}$
- Bounded from -1, 1
- Values far from 0 indicate strong relationship
- Negative values indicate inverse relationship



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Regression

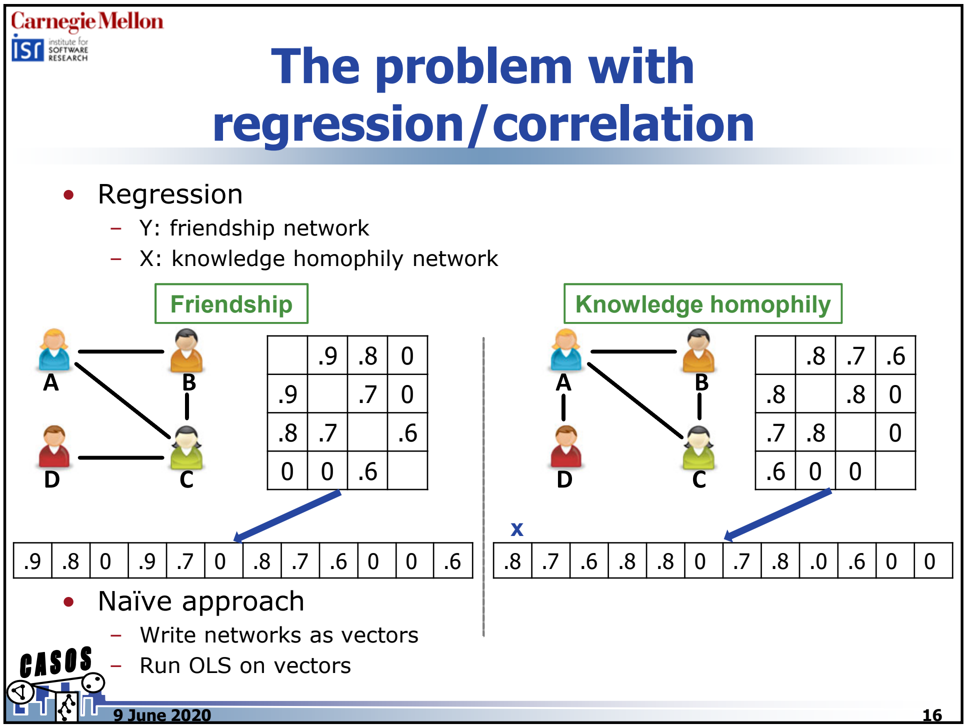
- These concepts are very closely related to regression
- Regression assumes that one variable (dependent) is a function of another variable (independent)
- The function is then found by estimating the conditional expectation
- For networks: is one network a function of another network?
  - Is the perceived friendship network a function of the actual contact network?



Carnegie Mellon

Thinking about distances

- Original motivation: how similar are these networks?
- Now we can put a number on it
  - Allows us to say which networks are more/less similar
- But how do we know these numbers matter?
- Use statistics!
  - Could use a bootstrapped t-test, for example
- **What makes this hard for networks?**



Carnegie Mellon

The problem with regression/correlation

- Regression
  - Y: friendship network
  - X: knowledge homophily network

	.9	.8	0
.9		.7	0
.8	.7		.6
0	0	.6	

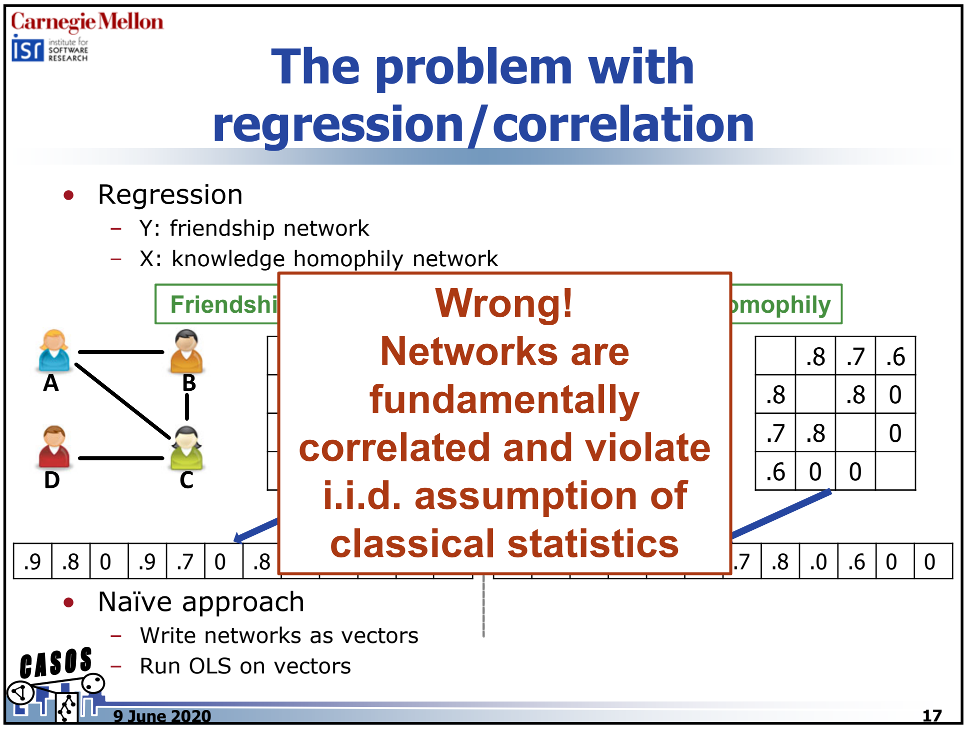
.9	.8	0	.9	.7	0	.8	.7	.6	0	0	.6

	.8	.7	.6
.8		.8	0
.7	.8		0
.6	0	0	

x

.8	.7	.6	.8	.8	0	.7	.8	.0	.6	0	0

- Naïve approach
  - Write networks as vectors
  - Run OLS on vectors



Carnegie Mellon isr institute for SOFTWARE RESEARCH

The problem with regression/correlation

- Regression
  - Y: friendship network
  - X: knowledge homophily network

Friendshi

A B D C

Wrong!
Networks are fundamentally correlated and violate i.i.d. assumption of classical statistics

omophily

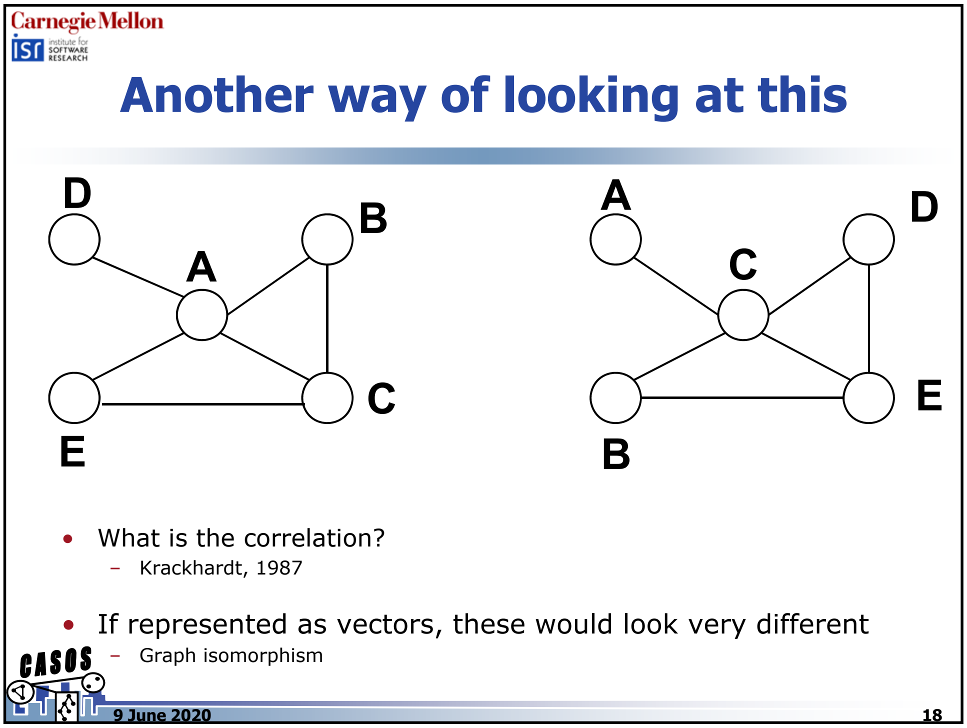
	.8	.7	.6
.8		.8	0
.7	.8		0
.6	0	0	

.9 | .8 | 0 | .9 | .7 | 0 | .8

.7 | .8 | .0 | .6 | 0 | 0

- Naïve approach
  - Write networks as vectors
  - Run OLS on vectors

CASOS

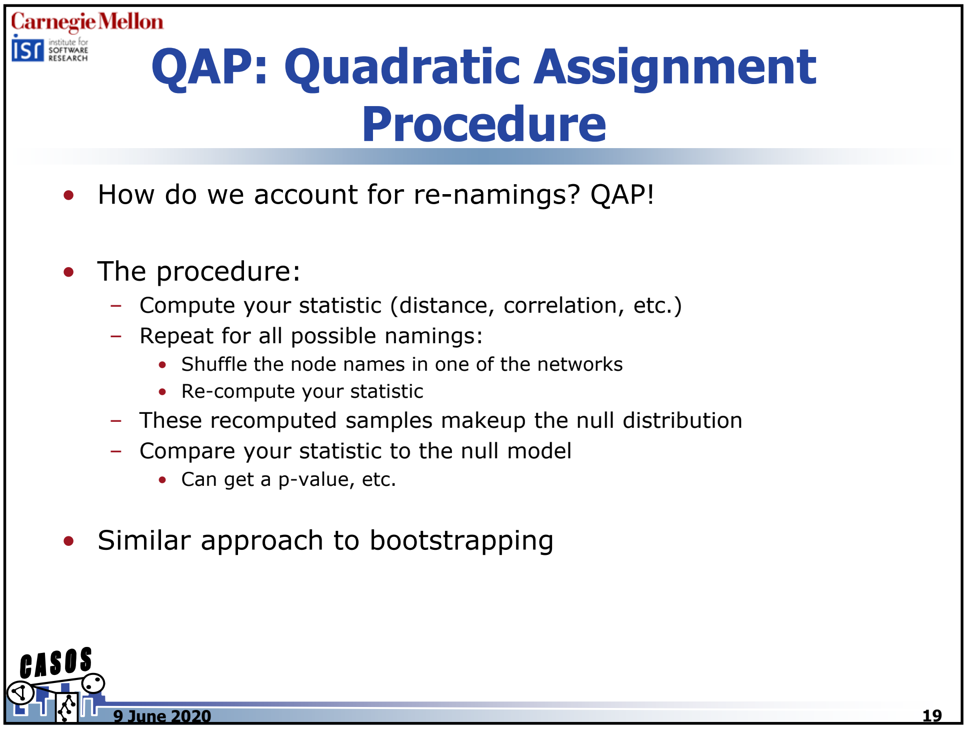


Carnegie Mellon isr institute for SOFTWARE RESEARCH

Another way of looking at this

- What is the correlation?
  - Krackhardt, 1987
- If represented as vectors, these would look very different
  - Graph isomorphism

CASOS



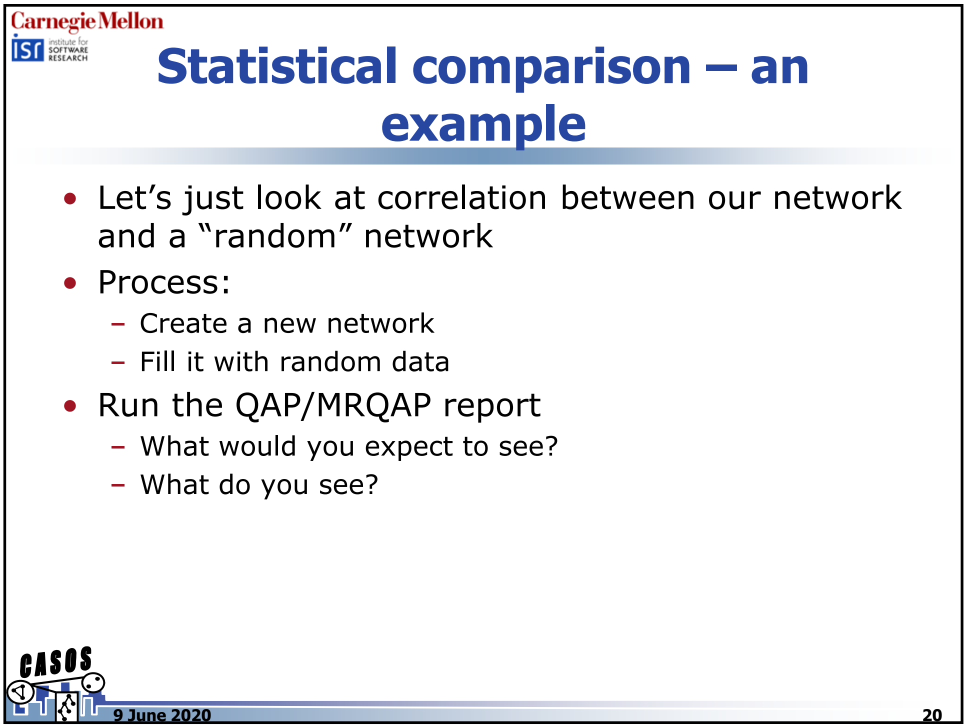
Carnegie Mellon

QAP: Quadratic Assignment Procedure

- How do we account for re-namings? QAP!
- The procedure:
  - Compute your statistic (distance, correlation, etc.)
  - Repeat for all possible namings:
    - Shuffle the node names in one of the networks
    - Re-compute your statistic
  - These recomputed samples makeup the null distribution
  - Compare your statistic to the null model
    - Can get a p-value, etc.
- Similar approach to bootstrapping

9 June 2020

19



Carnegie Mellon

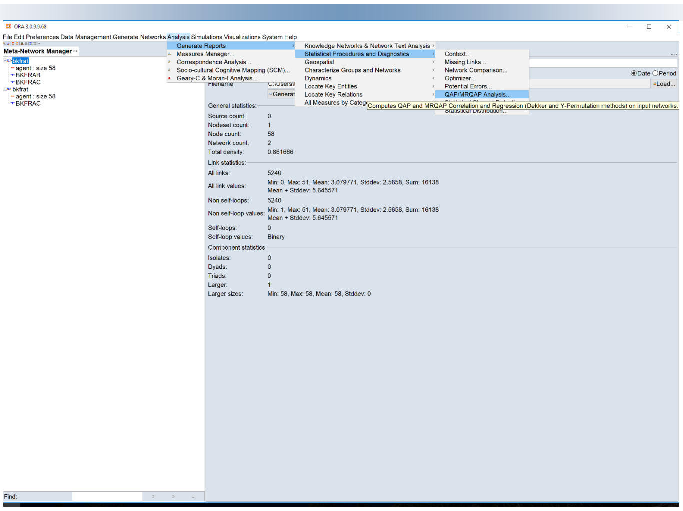
Statistical comparison – an example

- Let’s just look at correlation between our network and a “random” network
- Process:
  - Create a new network
  - Fill it with random data
- Run the QAP/MRQAP report
  - What would you expect to see?
  - What do you see?

9 June 2020

20

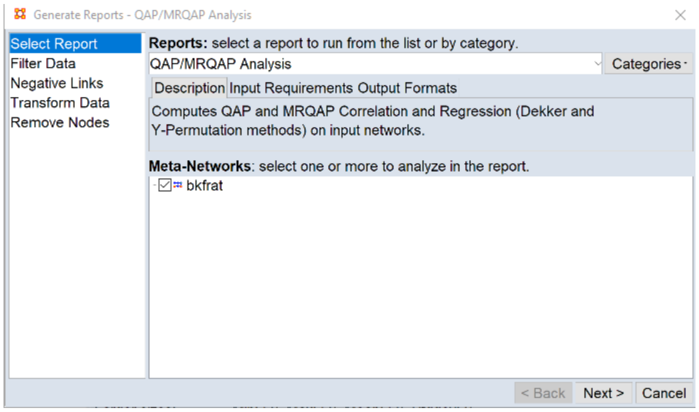
Now Lets Compare our Networks



Computes QAP and MRQAP Correlation and Regression (Dekker and Y-Permutation methods) on input networks.

```
General statistics:
Source count:        0
Nodeset count:       1
Node count:          58
Network count:       2
Total density:       0.861666
Link statistics:
All links:           5240
All link values:     Min: 0, Max: 51, Mean: 3.079771, Stddev: 2.5658, Sum: 16138
                     Mean + Stddev: 5.645571
Non self-loops:      5240
Non self-loop values: Min: 1, Max: 51, Mean: 3.079771, Stddev: 2.5658, Sum: 16138
                     Mean + Stddev: 5.645571
Self-loops:          0
Self-loop values:    Binary
Component statistics:
Isolates:            0
Dyads:               0
Triads:              0
Larger:              1
Larger sizes:        Min: 58, Max: 58, Mean: 58, Stddev: 0
```

Running QAP in ORA



Reports: select a report to run from the list or by category.

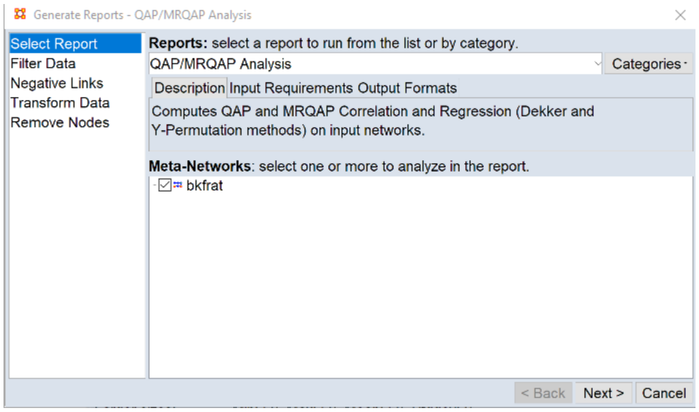
QAP/MRQAP Analysis

Computes QAP and MRQAP Correlation and Regression (Dekker and Y-Permutation methods) on input networks.

Meta-Networks: select one or more to analyze in the report.

☑ bkfrat

Running QAP in ORA



Generate Reports - QAP/MRQAP Analysis

Reports: select a report to run from the list or by category.

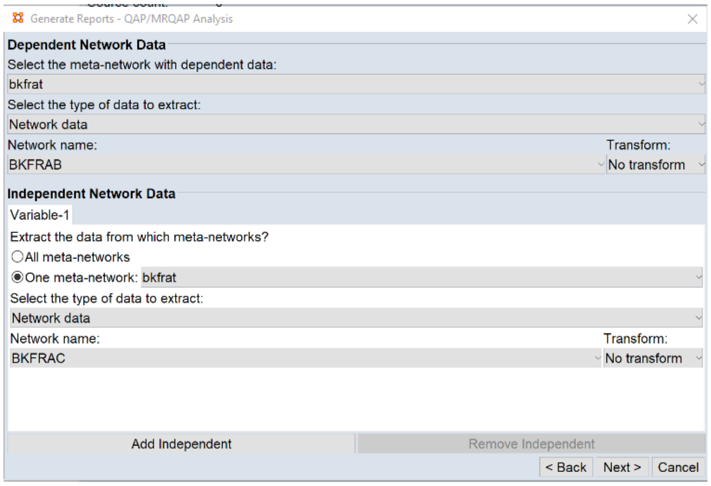
QAP/MRQAP Analysis

Computes QAP and MRQAP Correlation and Regression (Dekker and Y-Permutation methods) on input networks.

Meta-Networks: select one or more to analyze in the report.

☑ bkfrat

Running QAP in ORA



Generate Reports - QAP/MRQAP Analysis

Dependent Network Data

Select the meta-network with dependent data: bkfrat

Select the type of data to extract: Network data

Network name: BKFRAB

Transform: No transform

Independent Network Data

Variable-1

Extract the data from which meta-networks?

○ All meta-networks

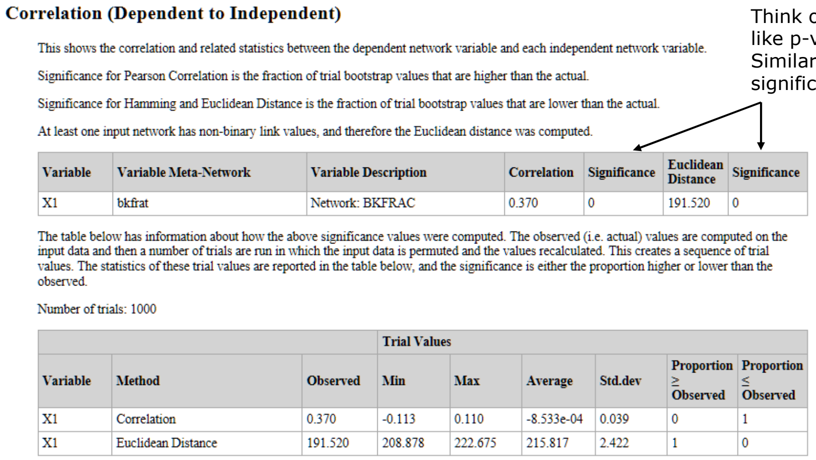
● One meta-network: bkfrat

Select the type of data to extract: Network data

Network name: BKFRAC

Transform: No transform

Running QAP in ORA



Correlation (Dependent to Independent)

This shows the correlation and related statistics between the dependent network variable and each independent network variable.

Significance for Pearson Correlation is the fraction of trial bootstrap values that are higher than the actual.

Significance for Hamming and Euclidean Distance is the fraction of trial bootstrap values that are lower than the actual.

At least one input network has non-binary link values, and therefore the Euclidean distance was computed.

Variable	Variable Meta-Network	Variable Description	Correlation	Significance	Euclidean Distance	Significance
X1	bkfrat	Network: BKFRAC	0.370	0	191.520	0

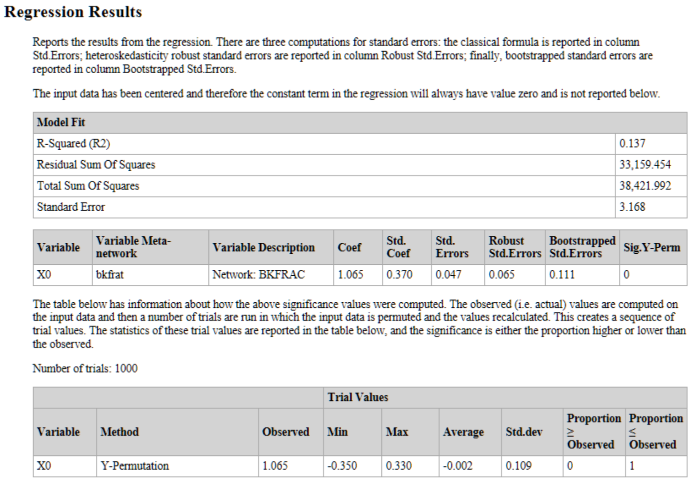
The table below has information about how the above significance values were computed. The observed (i.e. actual) values are computed on the input data and then a number of trials are run in which the input data is permuted and the values recalculated. This creates a sequence of trial values. The statistics of these trial values are reported in the table below, and the significance is either the proportion higher or lower than the observed.

Number of trials: 1000

			Trial Values					
Variable	Method	Observed	Min	Max	Average	Std.dev	Proportion ≥ Observed	Proportion ≤ Observed
X1	Correlation	0.370	-0.113	0.110	-8.533e-04	0.039	0	1
X1	Euclidean Distance	191.520	208.878	222.675	215.817	2.422	1	0

Think of these like p-values, Similarities are significant!

Running QAP in ORA



Regression Results

Reports the results from the regression. There are three computations for standard errors: the classical formula is reported in column Std.Errors; heteroskedasticity robust standard errors are reported in column Robust Std.Errors; finally, bootstrapped standard errors are reported in column Bootstrapped Std.Errors.

The input data has been centered and therefore the constant term in the regression will always have value zero and is not reported below.

Model Fit	
R-Squared (R2)	0.137
Residual Sum Of Squares	33,159.454
Total Sum Of Squares	38,421.992
Standard Error	3.168

Variable	Variable Meta-network	Variable Description	Coef	Std. Coef	Std. Errors	Robust Std.Errors	Bootstrapped Std.Errors	Sig.Y-Perm
X0	bkfrat	Network: BKFRAC	1.065	0.370	0.047	0.065	0.111	0

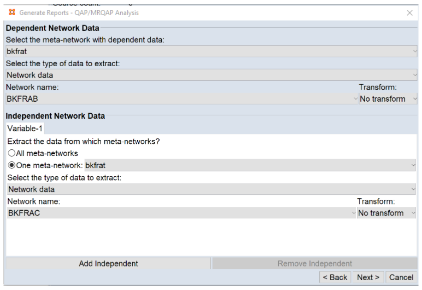
The table below has information about how the above significance values were computed. The observed (i.e. actual) values are computed on the input data and then a number of trials are run in which the input data is permuted and the values recalculated. This creates a sequence of trial values. The statistics of these trial values are reported in the table below, and the significance is either the proportion higher or lower than the observed.

Number of trials: 1000

			Trial Values					
Variable	Method	Observed	Min	Max	Average	Std.dev	Proportion ≥ Observed	Proportion ≤ Observed
X0	Y-Permutation	1.065	-0.350	0.330	-0.002	0.109	0	1

MR-QAP

- What if we want to model multiple relationships?
- Regression -> Multiple Regression
- QAP -> MR-QAP
- In ORA: “add independent” allows you to add more variables



Generate Reports - QAP/MRQAP Analysis

Dependent Network Data

Select the meta-network with dependent data: bkfrat

Select the type of data to extract: Network data

Network name: BKFRAB

Transform: No transform

Independent Network Data

Variable-1

Extract the data from which meta-networks?

○ All meta-networks

● One meta-network: bkfrat

Select the type of data to extract: Network data

Network name: BKFRAC

Transform: No transform

Recap

- Networks can be compared in a variety of ways
- Motifs allow you to see/compare “building blocks” of a network
- Distances/Correlation allow you to quantitatively find differences in network structure
- To analyze distances/correlation QAP must be used
  - Due to graph isomorphism and i.i.d. samples
- Multiple regression can also be performed using MRQAP
- Be careful with binary outcome variables!
  - Since the model is linear regression