



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

Kenny Joseph

kjoseph@cs.cmu.edu



CarnegieMellon

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

### Overview

- We're going to compare two networks from the famous work by Bernard and Killworth
- Download the data
  - Go to [casos.cs.cmu.edu](http://casos.cs.cmu.edu)
  - Tools, Models, Data -> Data -> Additional
  - Right click bkfrat 2.0.xml, save as ...
  - Load into ORA



6.15.2016

2



## Info on the Data

- A Meta-network with two networks
  - BKFRAB – records the number of times a pair of subjects were seen in conversation by an "unobtrusive" observer
  - BKFRAC - rankings made by the subjects of how frequently they interacted with other subjects in the observation week.



6.15.2016



3

## Our research question

How similar is the cognitive network to the behavioral network?

OK, go.



6.15.2016



4

## 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**



6.15.2016



5

## What is a motif?

- Partial subgraphs
- May contain only some of the edges in the large network
- Must be over-represented in the data compared to a random network
- Introduced by Uri Alon
- Also referred to as local patterns

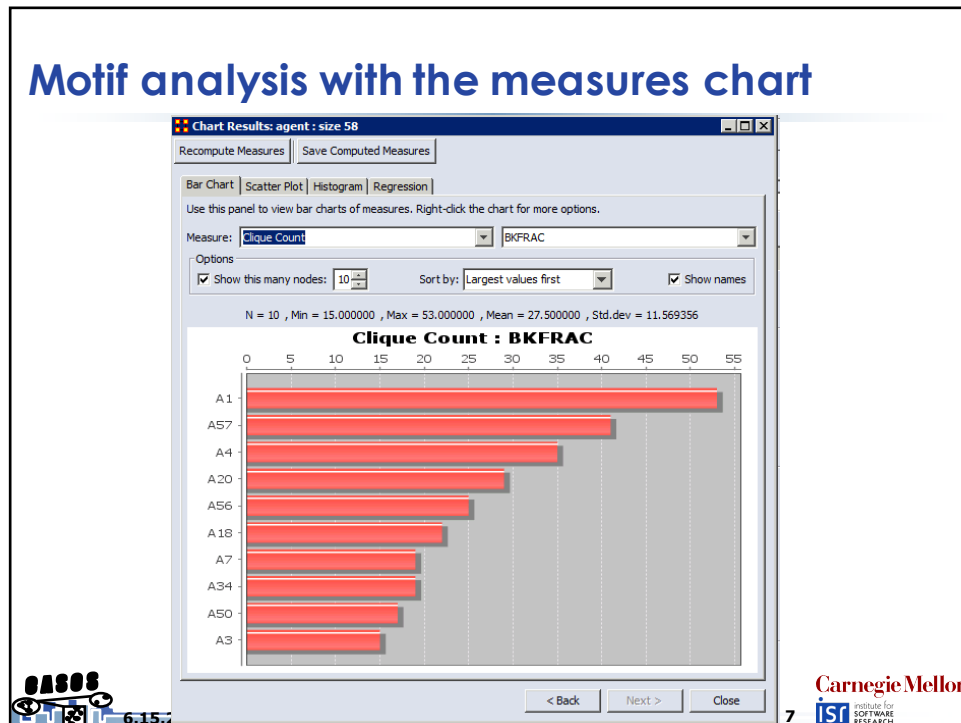


6.15.2016



6

## Motif analysis with the measures chart



## Comparing network structure

- We can also look more generally at how the entire network structures compare
- One way to do so is with distance metrics
  - Hamming Distance
  - Euclidean Distance
  - Correlations

**Carnegie Mellon**  
**ISI** Institute for Software Research

## Hamming Distance

the number of bits which differ between two binary strings:  
 $\sum |A_i - B_i|$  Alternatively, it can be calculated as  
 $\text{Union}(A_i \& B_i) - A_i$ . Either formula works for weighted  
or binary data.

Comparing two binary matrices – number of edge flips to  
make  $B = A$

Typically convert hamming to difference as a percent  
 $\text{Difference} = 100 * (\text{Max\_possible\_distance} - \text{Hamming}) /$   
 $\text{Max\_possible\_distance}$

$\text{Max\_possible\_distance} = N * N - 1$  (assuming 0 diagonals)

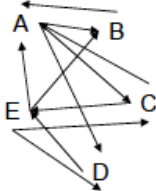
**CASOS**

January 2015 Copyright © 2015 Kathleen M. Carley - CASOS, ISR, CMU - Director

**Carnegie Mellon**  
**ISI** Institute for Software Research

## Example

1) Picture



2) Matrix

```

01110
10000
10001
00001
11110

```

3) String

```

0111010000100010000111110
0001010000100010010011100

```

4) Calculate

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

**CASOS**

January 2015 Copyright © 2015 Kathleen M. Carley - CASOS, ISR, CMU - Director

Carnegie Mellon  
IST Institute for Software Research

## Comparisons

- .25 Hamming Distance
- .75 Hamming Based Similarity (1- hamming distance)
- .60 Correlation

	A	B
density	.55	.40
average	.44	.32
min	0	0
max	1	1
Std. dev.	.51	.48

- Note: by treating a matrix as a string all standard statistics can be run

CASOS  
January 2015 Copyright © 2015 Kathleen M. Carley - CASOS, ISR, CMU - Director 7

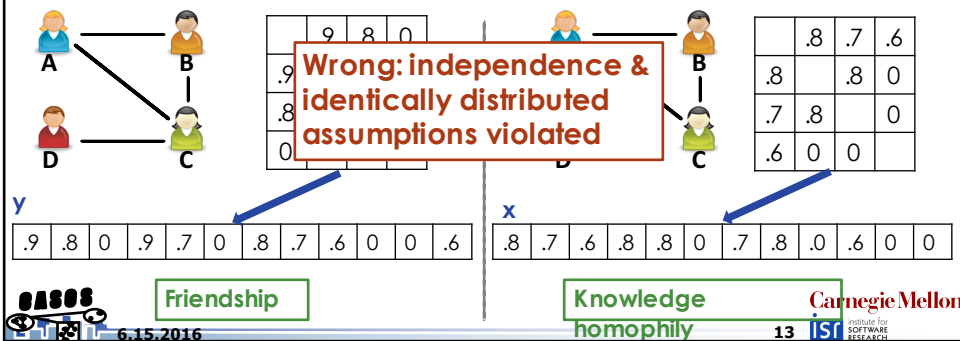
## Thinking about the correlation

- Recall our research question – how similar are these networks?
- These distance metrics answer this, but we want to know, are these correlations/distances interesting?
- Yay statistics!
  - We could run a (bootstrapped) t-test, etc..
- **But what makes this hard for networks?**

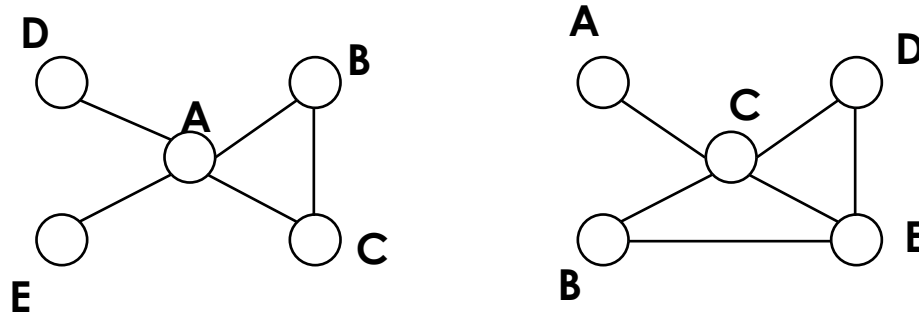
CASOS 6.15.2016 12 Carnegie Mellon  
IST Institute for Software Research

## The problem with regression/correlation

- Regression
  - Y: friendship network
  - X: knowledge homophily network
- Naïve approach
  - Write networks as vectors
  - Run OLS on vectors



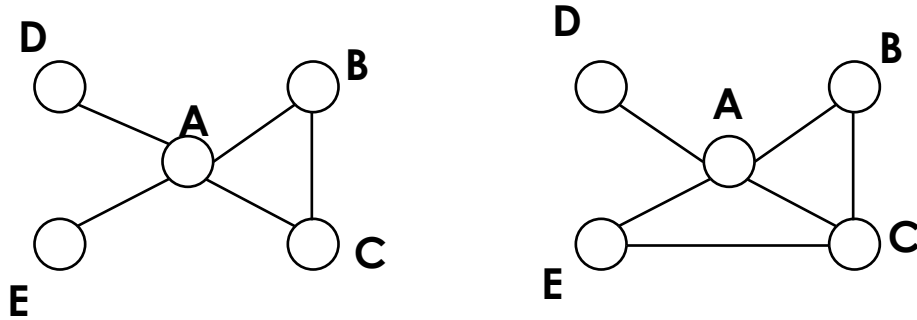
## Another way of looking at this



What is the correlation between these two networks?

**Now what is the correlation?**

## Another way of looking at this



**What about now?**



From Krackhardt, 1987

6.15.2016



15

IST Institute for Software Research

## QAP

- This is the gist of the QAP approach
  - Compute your statistic
  - Repeat for all possible namings:
    - Shuffle the names of the nodes of one of your networks
    - Recompute your statistic
  - Compute a null distribution for your statistic from these resamplings
  - Compare your statistic to this null distribution



Bootstrap-ish

6.15.2016



16

IST Institute for Software Research



## 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?



6.15.2016



17

## Statistical comparison – an example

- Now, lets compare our networks



6.15.2016



18

## What if we wanted to use both?

- We can do Multiple Regression with a QAP-based approach as well
- When doing so in ORA, interpret using the Dekker Double-Semi-Partialing approach to infer significance
- In general, be careful when working with a binary dependent variable (you're doing a linear regression)



6.15.2016



19

## Review

- Lots of tools/strategies for comparing networks
- QAP is one approach for statistical comparison
- MR-QAP can be used to extend QAP to a multiple regression setting
  - You should not use this with a binary dependent variable (let's chat)



6.15.2016



20