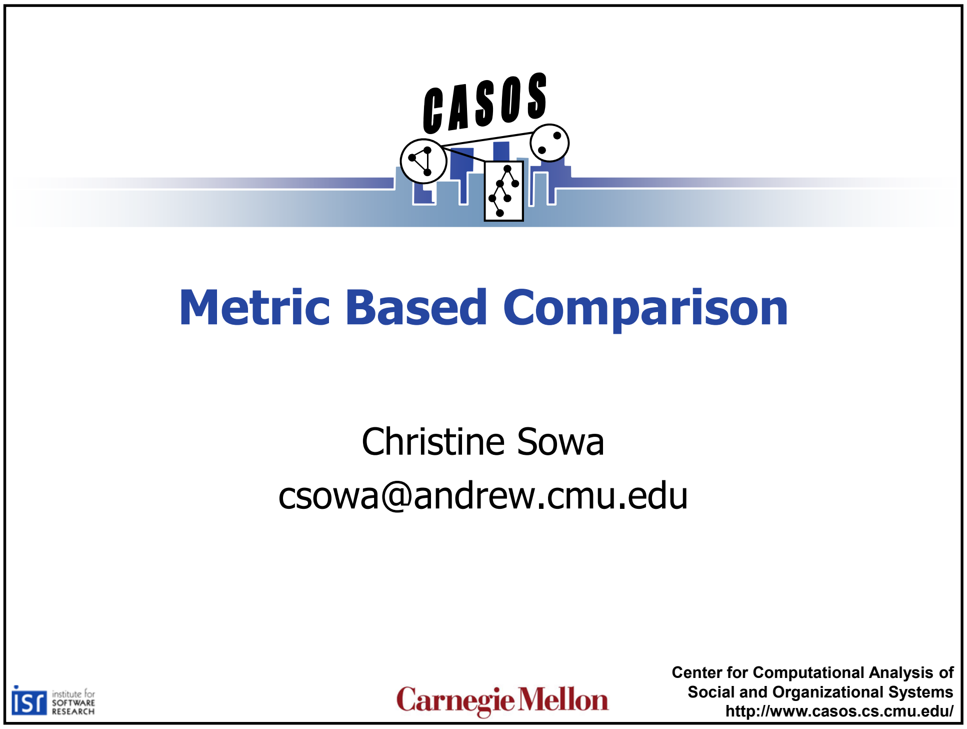
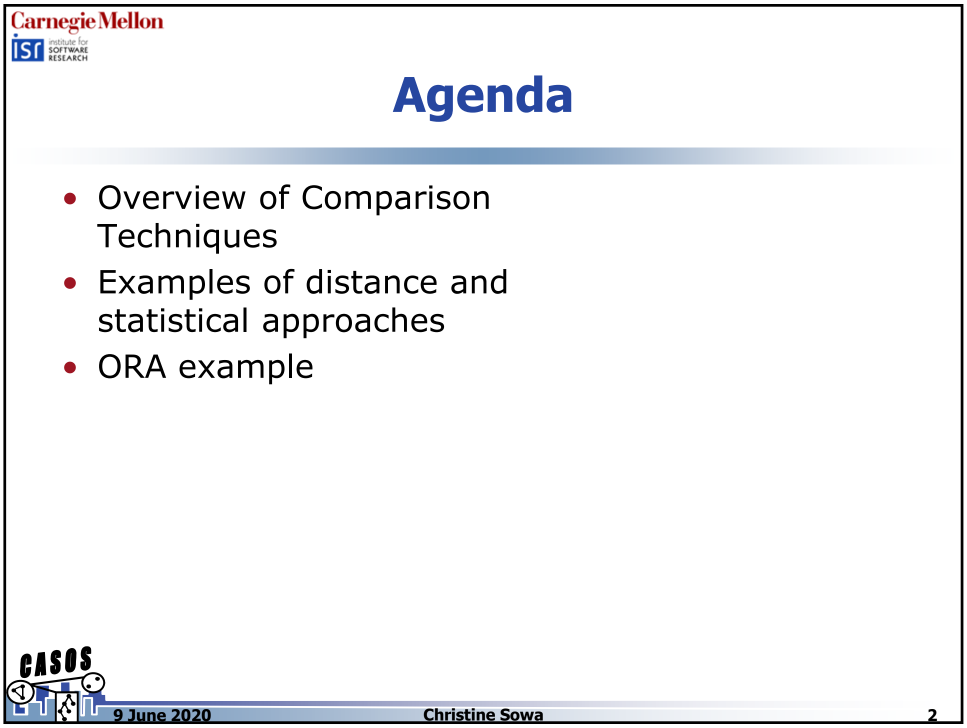


Metric Based Comparison

Christine Sowa
csowa@andrew.cmu.edu

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



Agenda

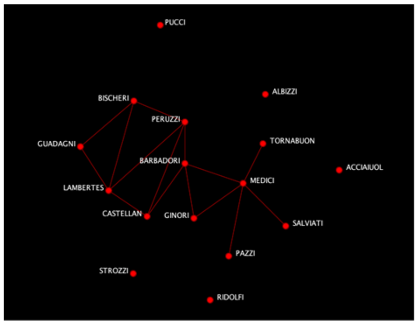
- Overview of Comparison Techniques
- Examples of distance and statistical approaches
- ORA example

Comparison Techniques

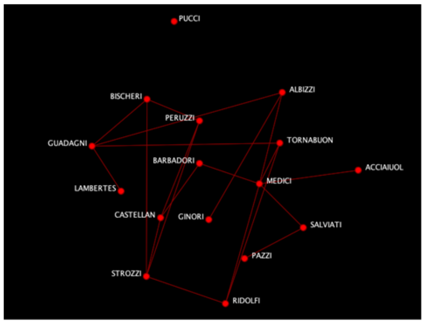
- Visualization
- Distance
- Statistical approaches
  - Summary statistics
    - Confidence intervals for measures
    - Node level statistics
  - QAP Models & MrQAP
    - Discussed in Day 3 Talks

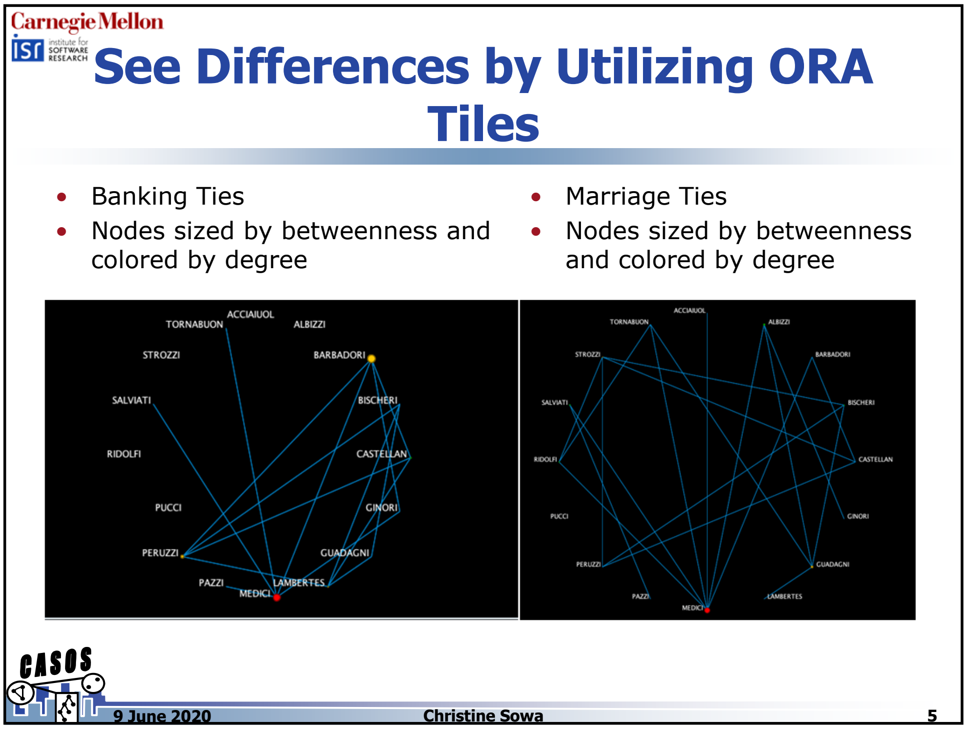
Comparison by Visualization

- Padgett Medici Banking Ties



- Padgett Medici Marriage Ties





Carnegie Mellon

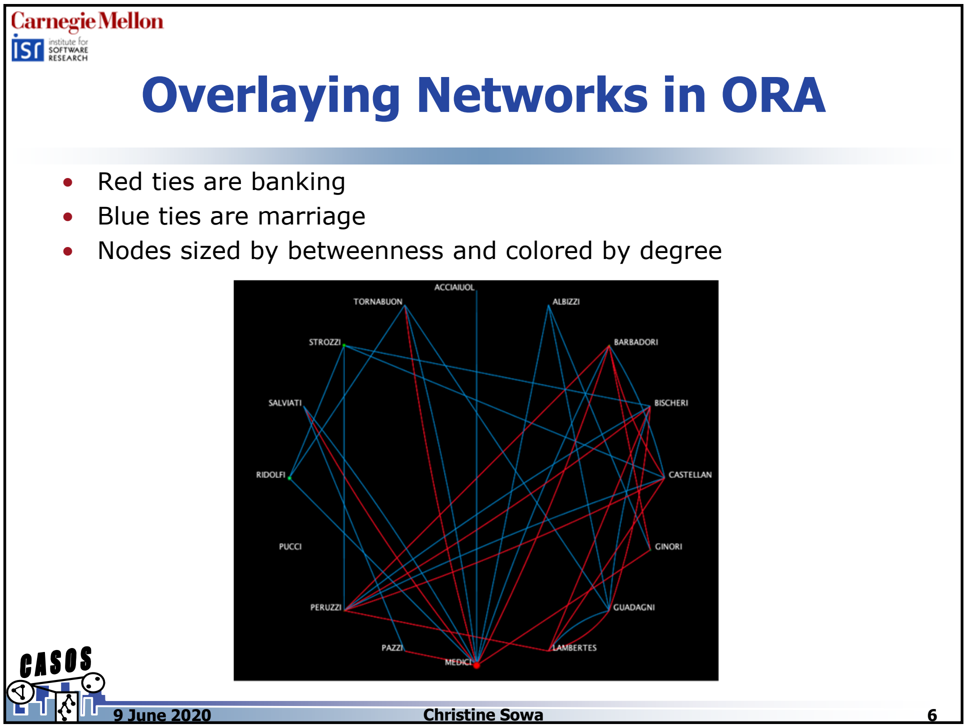
isr institute for SOFTWARE RESEARCH

See Differences by Utilizing ORA Tiles

- Banking Ties
- Nodes sized by betweenness and colored by degree

- Marriage Ties
- Nodes sized by betweenness and colored by degree

9 June 2020 Christine Sowa 5



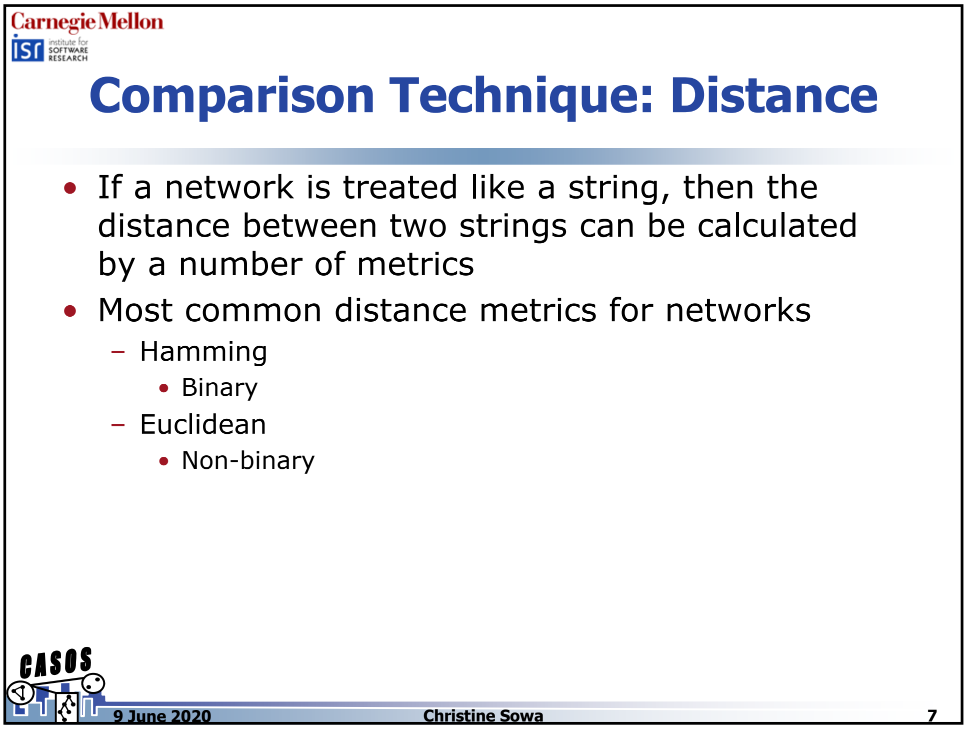
Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Overlaying Networks in ORA

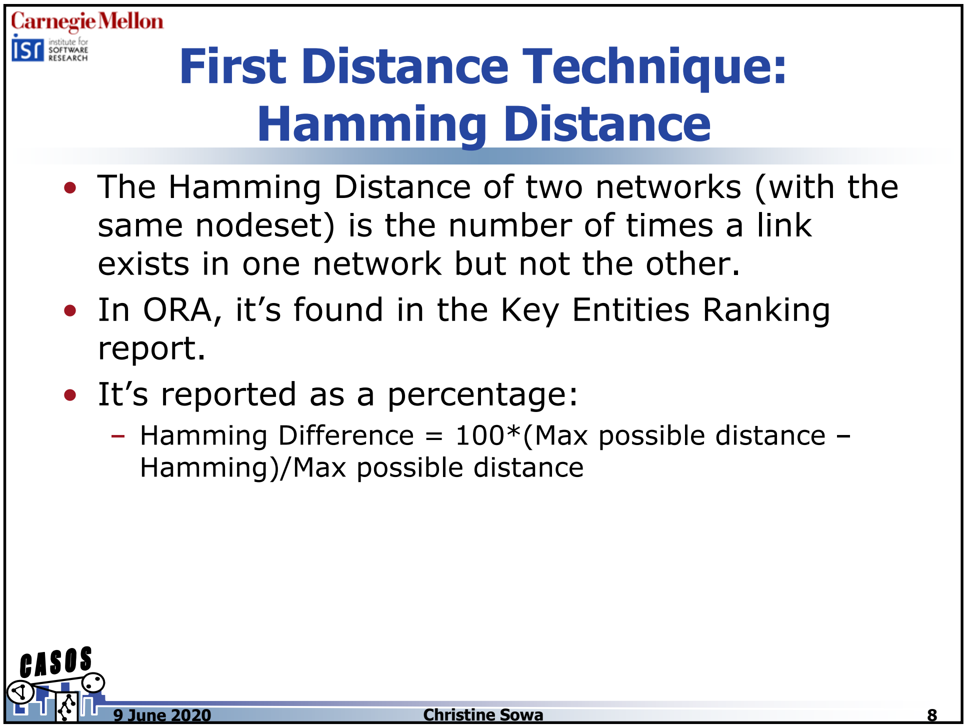
- Red ties are banking
- Blue ties are marriage
- Nodes sized by betweenness and colored by degree

9 June 2020 Christine Sowa 6



Comparison Technique: Distance

- If a network is treated like a string, then the distance between two strings can be calculated by a number of metrics
- Most common distance metrics for networks
  - Hamming
    - Binary
  - Euclidean
    - Non-binary



First Distance Technique: Hamming Distance

- The Hamming Distance of two networks (with the same nodeset) is the number of times a link exists in one network but not the other.
- In ORA, it’s found in the Key Entities Ranking report.
- It’s reported as a percentage:
  - Hamming Difference = 100*(Max possible distance – Hamming)/Max possible distance

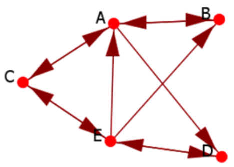
Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Hamming Distance Example

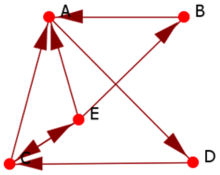
- Take the following two networks and convert them to binary strings:

Network 1



	A	B	C	D	E
A	0	1	1	1	0
B	1	0	0	0	0
C	1	0	0	0	1
D	0	0	0	0	1
E	1	1	1	1	0

Network 2



	A	B	C	D	E
A	0	0	0	1	0
B	1	0	0	0	0
C	1	0	0	0	1
D	0	0	1	0	0
E	1	1	1	0	0

0111010000100010000111110

0001010000100010010011100

- There are 5 differences between the two strings. The unnormalized hamming distance is **5**.
- Since there are 20 possible links, the normalized hamming distance is 5/20 = .2 or 20%.

CASOS

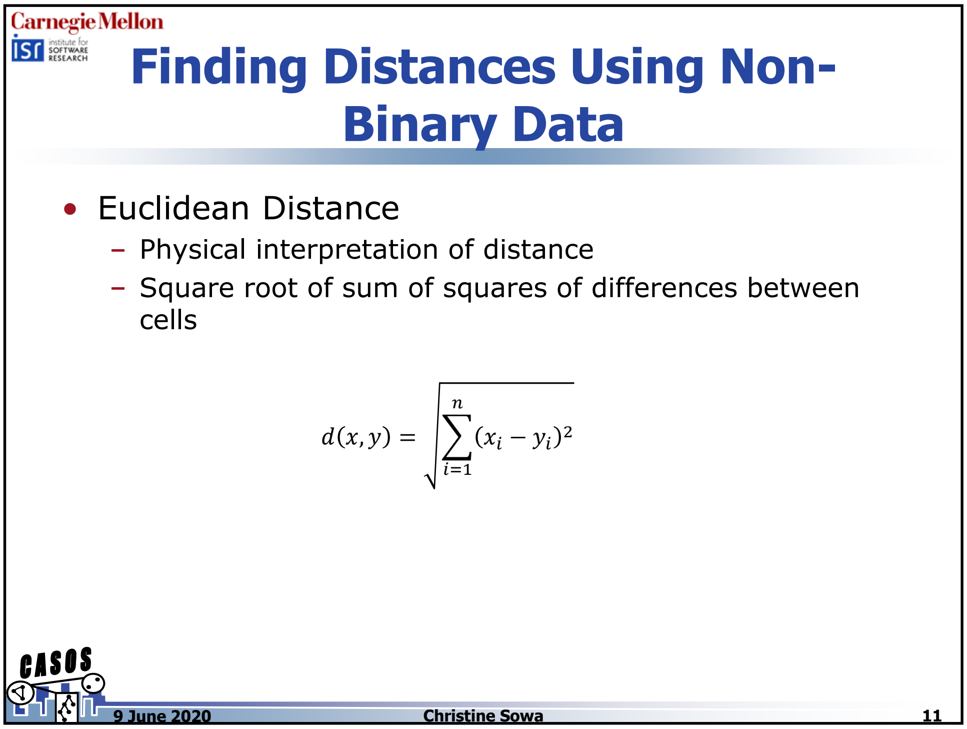
Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Variation of Hamming Distance: Damerau-Levenshtein Distance

- Damerau-Levenshtein Distance is the distance found by counting the minimum number of operations needed to transform one string into the other.
- Operations include:
  - Insertion
  - Deletion
  - Substitution of a single character
  - Transposition of two adjacent characters

CASOS

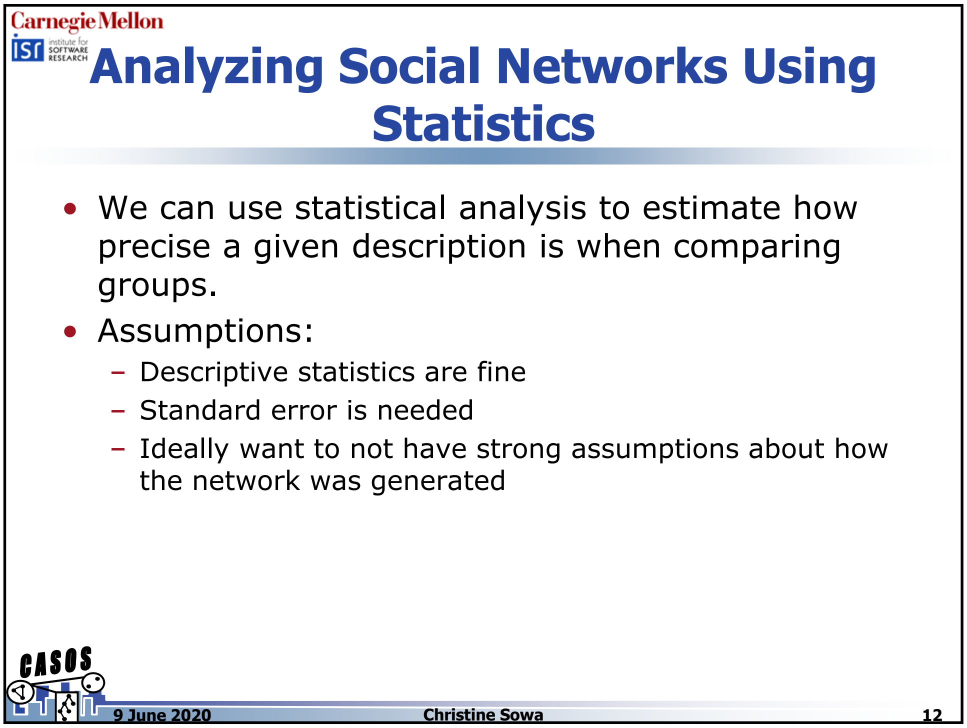


Carnegie Mellon

Finding Distances Using Non-Binary Data

- Euclidean Distance
  - Physical interpretation of distance
  - Square root of sum of squares of differences between cells

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



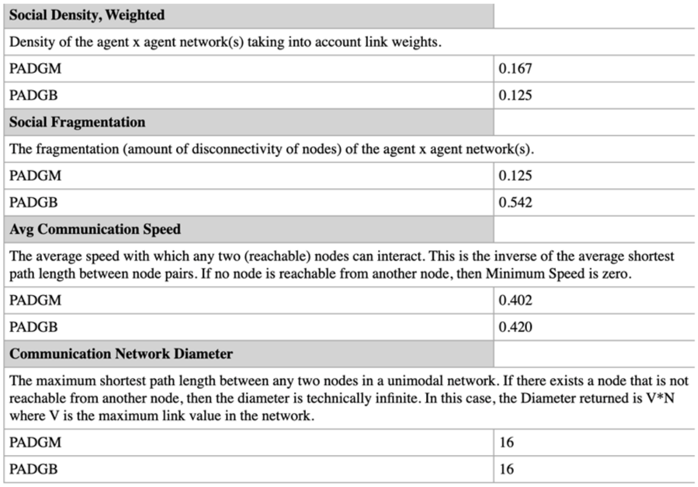
Carnegie Mellon

Analyzing Social Networks Using Statistics

- We can use statistical analysis to estimate how precise a given description is when comparing groups.
- Assumptions:
  - Descriptive statistics are fine
  - Standard error is needed
  - Ideally want to not have strong assumptions about how the network was generated

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Comparing Networks: More Florentine Families

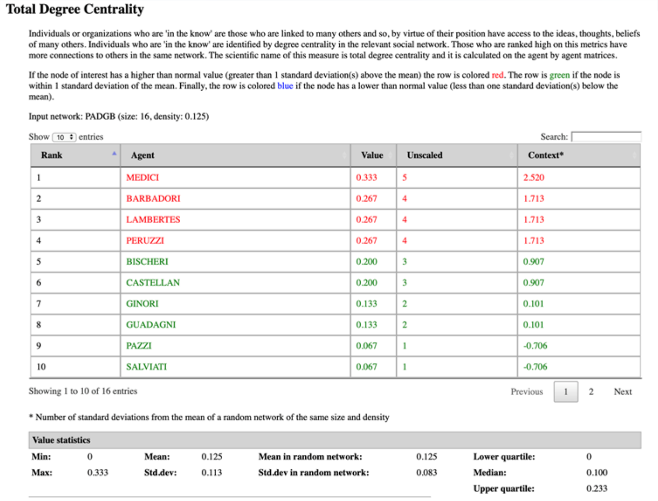


Social Density, Weighted	
Density of the agent x agent network(s) taking into account link weights.	
PADGM	0.167
PADGB	0.125
Social Fragmentation	
The fragmentation (amount of disconnectivity of nodes) of the agent x agent network(s).	
PADGM	0.125
PADGB	0.542
Avg Communication Speed	
The average speed with which any two (reachable) nodes can interact. This is the inverse of the average shortest path length between node pairs. If no node is reachable from another node, then Minimum Speed is zero.	
PADGM	0.402
PADGB	0.420
Communication Network Diameter	
The maximum shortest path length between any two nodes in a unimodal network. If there exists a node that is not reachable from another node, then the diameter is technically infinite. In this case, the Diameter returned is V*N where V is the maximum link value in the network.	
PADGM	16
PADGB	16

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Distributions of Node Level Metrics



Total Degree Centrality

Individuals or organizations who are 'in the know' are those who are linked to many others and so, by virtue of their position have access to the ideas, thoughts, beliefs of many others. Individuals who are 'in the know' are identified by degree centrality in the relevant social network. Those who are ranked high on this metrics have more connections to others in the same network. The scientific name of this measure is total degree centrality and it is calculated on the agent by agent matrices.

If the node of interest has a higher than normal value (greater than 1 standard deviation(s) above the mean) the row is colored red. The row is green if the node is within 1 standard deviation of the mean. Finally, the row is colored blue if the node has a lower than normal value (less than one standard deviation(s) below the mean).

Input network: PADGB (size: 16, density: 0.125)

Show 10 entries Search:

Rank	Agent	Value	Unscaled	Context*
1	MEDICI	0.333	5	2.520
2	BARBADORI	0.267	4	1.713
3	LAMBERTES	0.267	4	1.713
4	PERUZZI	0.267	4	1.713
5	BISCHERI	0.200	3	0.907
6	CASTELLAN	0.200	3	0.907
7	GINORI	0.133	2	0.101
8	GUADAGNI	0.133	2	0.101
9	PAZZI	0.067	1	-0.706
10	SALVIATI	0.067	1	-0.706

Showing 1 to 10 of 16 entries Previous 1 2 Next

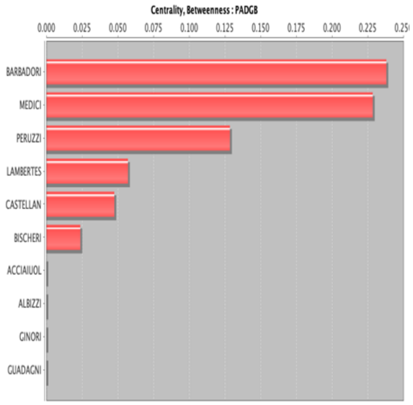
* Number of standard deviations from the mean of a random network of the same size and density

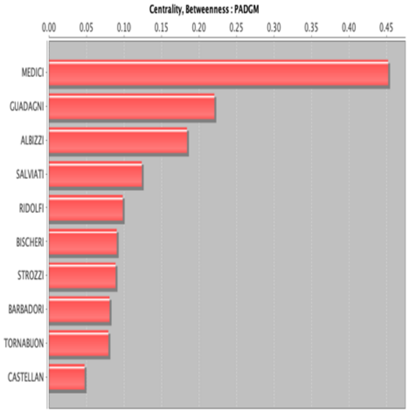
Value statistics							
Min:	0	**Mean:**	0.125	**Mean in random network:**	0.125	**Lower quartile:**	0
Max:	0.333	**Std.dev:**	0.113	**Std.dev in random network:**	0.083	**Median:**	0.100
						Upper quartile:	0.233

CASOS

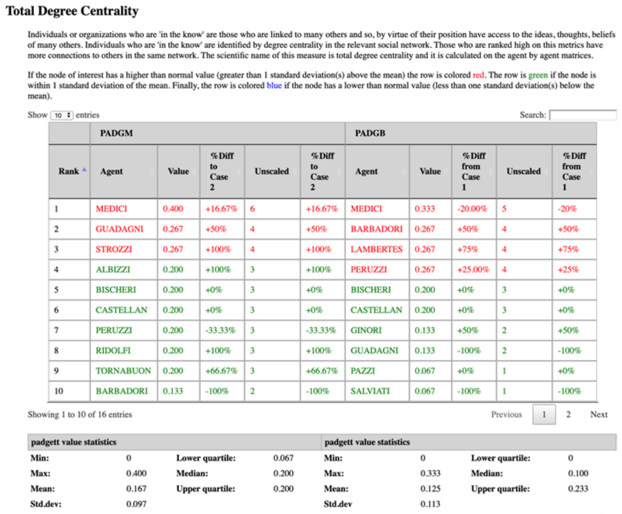
Distributions of Node Level Statistics

- Banking Network
  - Mean = .0724
  - Std. Deviation = .08898
- Marriage Network
  - Mean = .1467
  - Std. Deviation = .1133





Comparing Node Level Statistics



Total Degree Centrality

Individuals or organizations who are 'in the know' are those who are linked to many others and so, by virtue of their position have access to the ideas, thoughts, beliefs of many others. Individuals who are 'in the know' are identified by degree centrality in the relevant social network. Those who are ranked high on this metrics have more connections to others in the same network. The scientific name of this measure is total degree centrality and it is calculated on the agent by agent matrices.

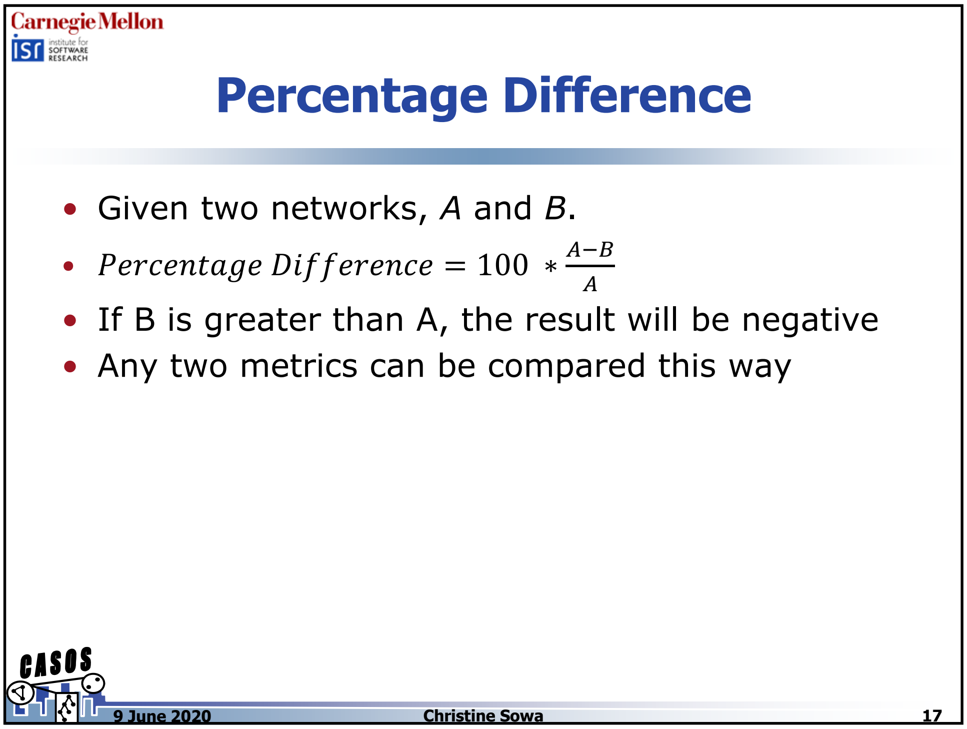
If the node of interest has a higher than normal value (greater than 1 standard deviation(s) above the mean) the row is colored red. The row is green if the node is within 1 standard deviation of the mean. Finally, the row is colored blue if the node has a lower than normal value (less than one standard deviation(s) below the mean).

Show 10 entries Search:

	PADGM					PADGB				
Rank	Agent	Value	%Diff to Case 2	Unscaled	%Diff to Case 2	Agent	Value	%Diff from Case 1	Unscaled	%Diff from Case 1
1	MEDICI	0.400	+16.67%	6	+16.67%	MEDICI	0.333	-20.00%	5	-20%
2	GUADAGNI	0.267	+50%	4	+50%	BARBADORI	0.267	+50%	4	+50%
3	STROZZI	0.267	+100%	4	+100%	LAMBERTES	0.267	+75%	4	+75%
4	ALBIZZI	0.200	+100%	3	+100%	PERUZZI	0.267	+25.00%	4	+25%
5	BISCHERI	0.200	+0%	3	+0%	BISCHERI	0.200	+0%	3	+0%
6	CASTELLAN	0.200	+0%	3	+0%	CASTELLAN	0.200	+0%	3	+0%
7	PERUZZI	0.200	-33.33%	3	-33.33%	GINORI	0.133	+50%	2	+50%
8	RIDOLFI	0.200	+100%	3	+100%	GUADAGNI	0.133	-100%	2	-100%
9	TORNABUON	0.200	+66.67%	3	+66.67%	PAZZI	0.067	+0%	1	+0%
10	BARBADORI	0.133	-100%	2	-100%	SALVIATI	0.067	-100%	1	-100%

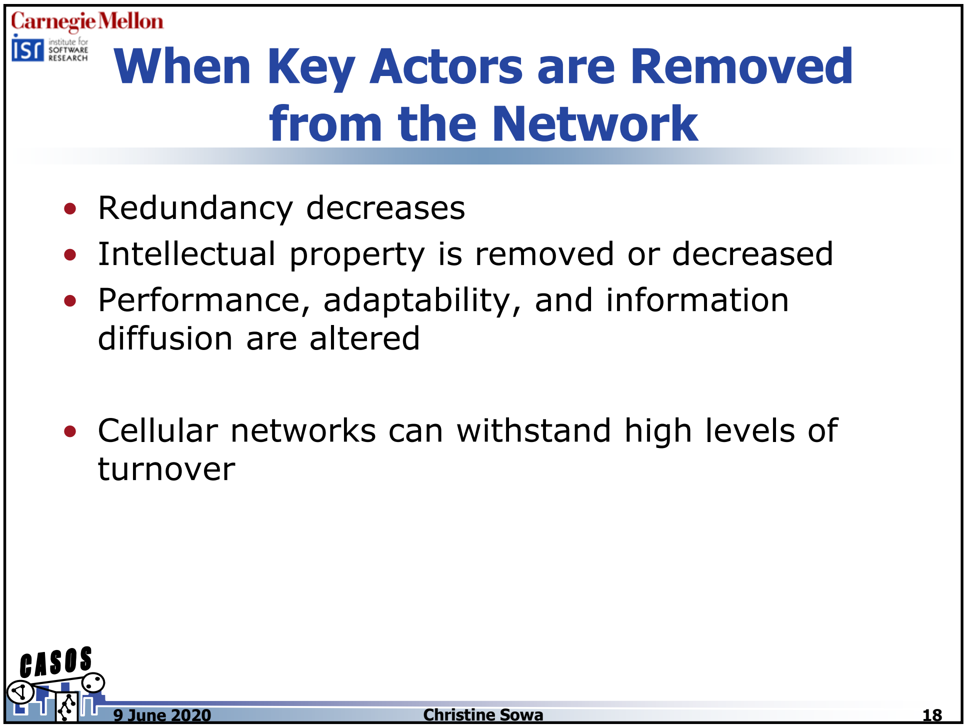
Showing 1 to 10 of 16 entries Previous 1 2 Next

padgett value statistics				padgett value statistics			
Min:	0	Lower quartile:	0.067	Min:	0	Lower quartile:	0
Max:	0.400	Median:	0.200	Max:	0.333	Median:	0.100
Mean:	0.167	Upper quartile:	0.200	Mean:	0.125	Upper quartile:	0.233
Std.dev:	0.097			Std.dev	0.113		



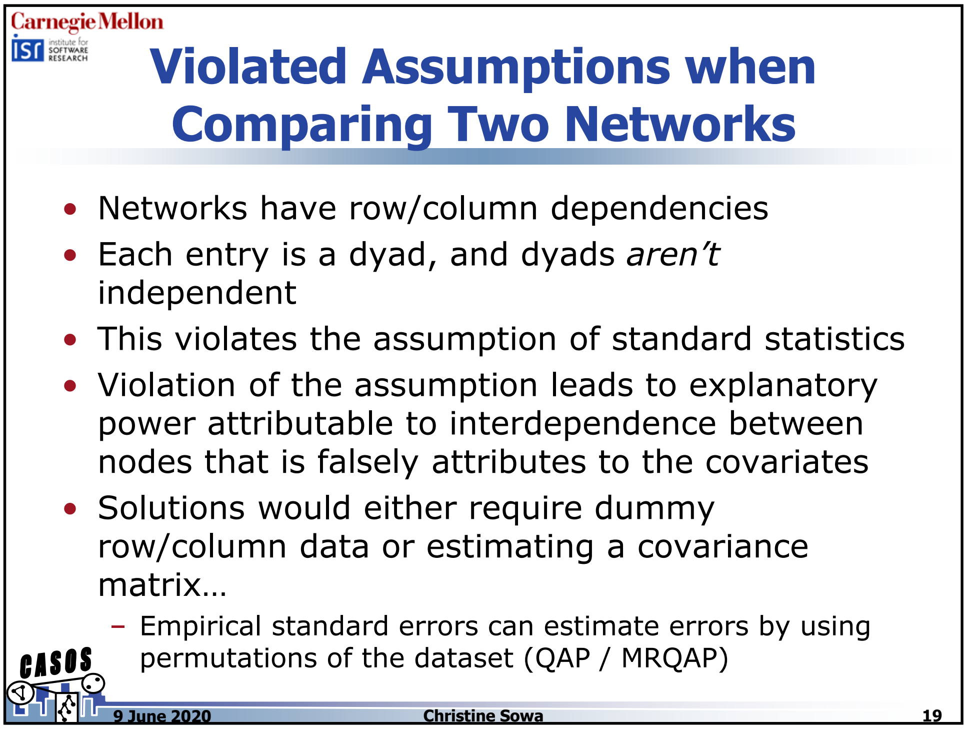
Percentage Difference

- Given two networks, *A* and *B*.
- $Percentage\ Difference = 100 \ast \frac{A-B}{A}$
- If B is greater than A, the result will be negative
- Any two metrics can be compared this way



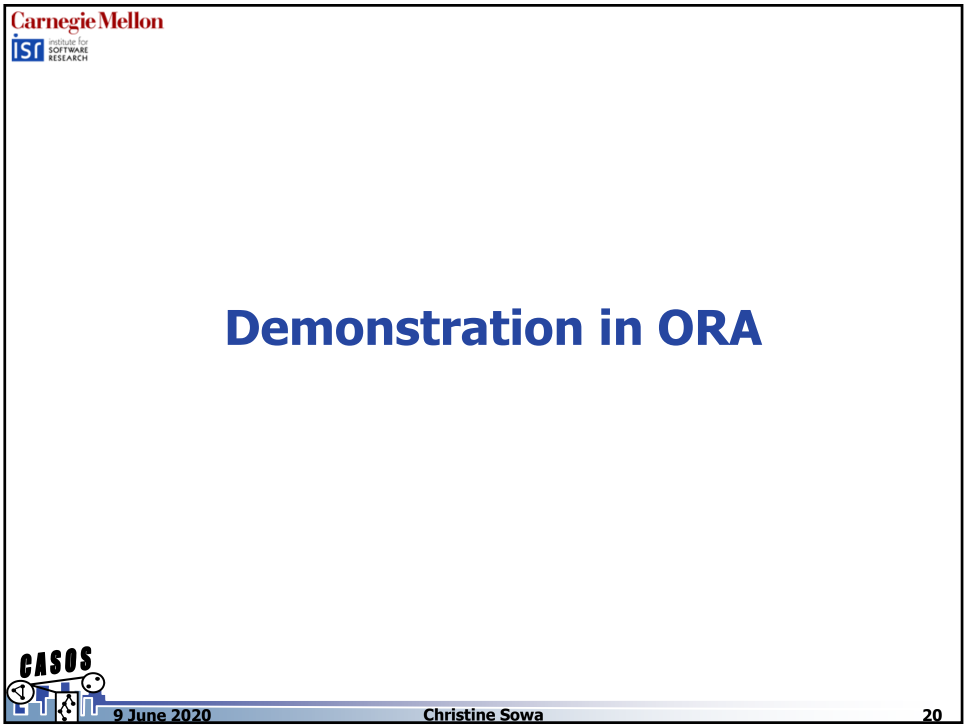
When Key Actors are Removed from the Network

- Redundancy decreases
- Intellectual property is removed or decreased
- Performance, adaptability, and information diffusion are altered
- Cellular networks can withstand high levels of turnover



Violated Assumptions when Comparing Two Networks

- Networks have row/column dependencies
- Each entry is a dyad, and dyads *aren't* independent
- This violates the assumption of standard statistics
- Violation of the assumption leads to explanatory power attributable to interdependence between nodes that is falsely attributes to the covariates
- Solutions would either require dummy row/column data or estimating a covariance matrix…
  - Empirical standard errors can estimate errors by using permutations of the dataset (QAP / MRQAP)



Demonstration in ORA