Carnegie Mellon

CASOS

Robustness Analysis

Terrill L. Frantz
14 June 2006
CASOS Summer Institute

Center for Computational Analysis of Social and Organizational Systems

http://www.casos.cs.cmu.edu/

Carnegie Mellon

Contents

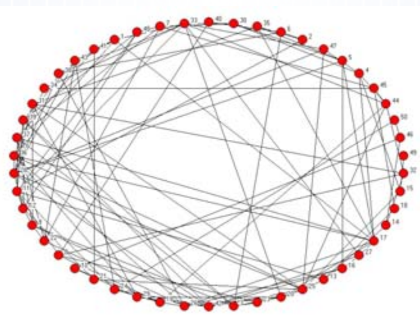
- Network Topology Basics
- ORA Activity – Generate Random Networks
- Robustness Of Centrality Measures

Network Topology

- Uniform Random
- Small-World
- Scale-Free
- Core-Periphery
- Cellular

Uniform Random

Spring Embedded Layout



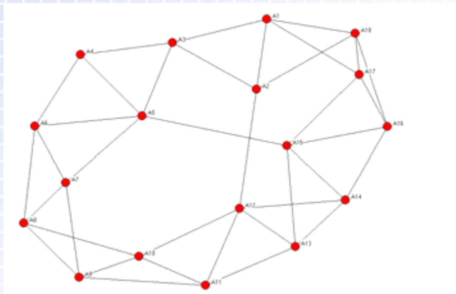
Circular Layout

Existence of a tie is a function of a fixed probability.

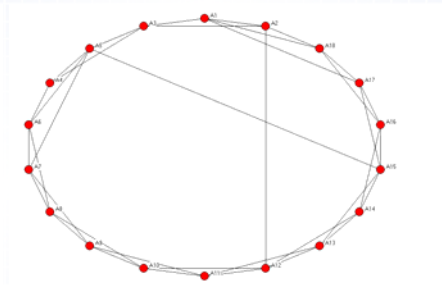
50 nodes, 10% density, undirected

Carnegie Mellon

Small-World



Spring Embedded Layout



Circular Layout

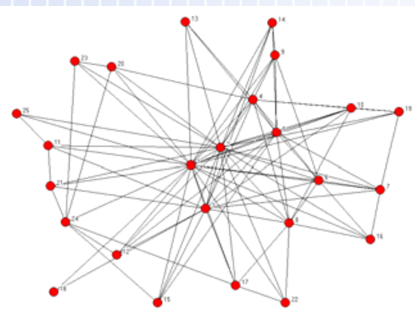
Nodes are tied to n-immediate neighbors; then a random few are re-wired to other nodes, selected at random.

50 nodes, 10% density, undirected

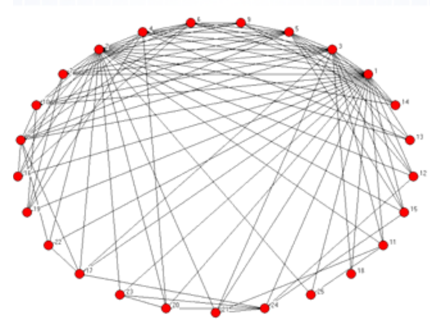
CASOS

Carnegie Mellon

Scale-Free



Spring Embedded Layout



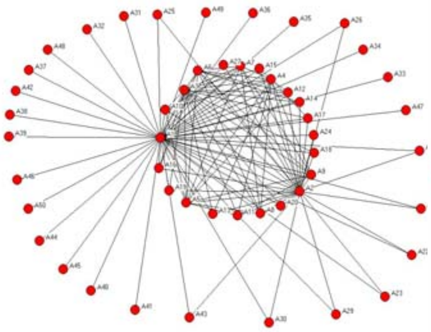
Circular Layout

Network is grown; preferential-attachment for ties to others with high degree centrality.

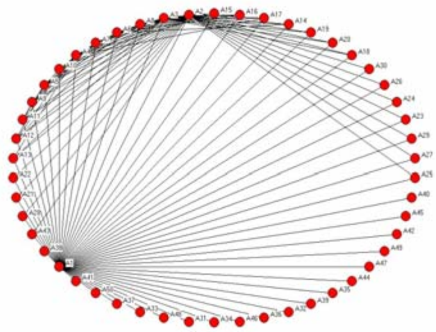
50 nodes, 10% density, undirected

CASOS

Core-Periphery



Spring Embedded Layout



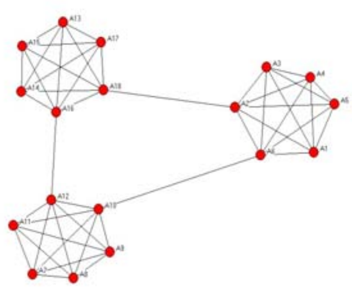
Circular Layout

Nodes in the core subset are highly connected to one another; nodes in periphery connected to few in the core.

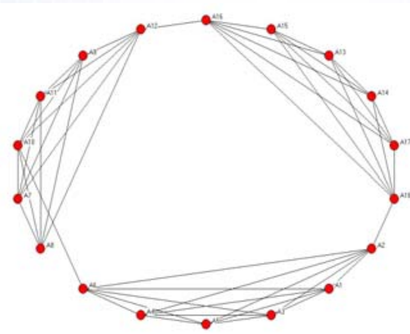
50 nodes, 10% density, undirected

Carnegie Mellon

Cellular



Spring Embedded Layout



Circular Layout

Cell subgroups are highly connected within; cells are loosely connected.

50 nodes, 10% density, undirected

Activity

Task:

Conduct a comparative analysis of network measures for a pair of topologies.

Process:

Working in a group, use ORA to conduct a social network analysis comparing an example <topology-of-your-choice> network with an example random network.

ROBUSTNESS EXPERIMENTS

Carnegie Mellon

Problem

Social-network data is prone to error…

Documented nodes and ties may be quite mistaken

Carnegie Mellon

Motivation

- Explore the robustness of social network measures
- Given that missing and inaccurate data is the norm, then how reliable are the measures?
- What caveats should we assign to interpreting the measures?
- Can the bounds and robustness of measures be quantified under conditions of data error?

Borgatti, Carley, & Krackhardt (2005)

Findings:

- Measure accuracy declines predictably with increasing error
- Four centrality measures have similar robustness pattern and levels
- Type of the error (node / edge) has little affect on the robustness
- Increased density reduces accuracy for all kinds of error, except edge addition where accuracy increased.

Our Research Question

How is network topology related to the robustness of measures of centrality?

Carnegie Mellon

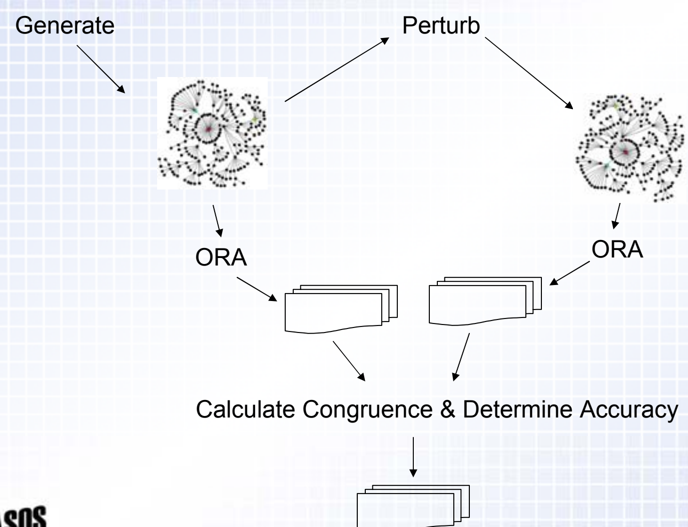
Methodology

- Run O(2,400) controlled experiments under varied scenarios to generate data. In each experiment:
  - Select sample networks O(10,000) randomly from a bounded population of candidates (TRUE NETWORK - *T*)
  - Independently perturb each sample (OBSERVED NETWORK - *O*)
  - Collect the respective measures for each *T*, *O pair.*
- Compute descriptive statistics of measures each experiment
- Aggregate data from independent experiments into various configurations for analysis

CASOS

Carnegie Mellon

Experiments



CASOS

Factorial Design

- Each experiment/scenario* has 5 independent attributes:
  - Topology (3) (Uniform Random, Cellular, Core-periphery)
  - Node Size(4): 10, 25, 50, 100
  - Density(8): 1%, 2%, 5%, 10%, 30%, 50%, 70%, 90%
  - Introduced Error:
    - Type(5): EdgeRemove, EdgeAdd, NodeRemove, NodeAdd, NodeAlias
    - Level(5): 1%, 5%, 10%, 25%, 50%
- Factorial Design Experiment
  - 3x4x8x5x5 = 2,400 experiments
- * Except for topology, experiment replicates Borgatti, Carley, & Krackhardt (forthcoming).

Experiment Design

- Each experiment trial (*n*=2,400x10,000) has 4 graph-level-measure responses:
  - Betweenness Centrality
  - Closeness Centrality
  - Degree Centrality
  - Eigenvector Centrality

Carnegie Mellon

Experiment Design

- Each of the 7 graph-level-measure responses is represented by 6 criteria:
  - Top 1 – Proportion Rank 1 in True is Rank 1 in Observed
  - Top 3 – Proportion Rank 1 in True is Rank 1-3 in Observed
  - Top 10% - Proportion Rank 1 in True is Rank Top 10% in Observed
  - Intersection R-Squared – Pearson correlation between T & O
  - Union R-Squared – Pearson correlation between T & O
  - Overlap – (Num nodes in T top 10% and O top 10%) / Num in either 10%
    - i.e., Overlap = |NT ∩ NO | / | NT U NO |
- Each criteria is reflected in 4 statistics:
  - Minimum
  - Maximum
  - Mean
  - Standard Deviation

CASOS

Carnegie Mellon

Experiment Design

For example, each experiment (*n*=2,400) results in 7 of these unique charts:

Uniform Topology; 100 Nodes; 50% Density; Edge Remove; 10% error (*n*=10,000)

Betweenness Centrality

	Minimum	Maximum	Mean	Std. Dev.
Top 1	0.000	1.000	0.484	0.500
Top 3	0.000	1.000	0.795	0.404
Top 10%	0.000	1.000	0.983	0.129
Intersection R-Squared	0.804	0.953	0.899	0.020
Union R-Squared	0.804	0.953	0.899	0.020
Overlap	0.111	1.000	0.529	0.135

CASOS

Carnegie Mellon

Current Experiments

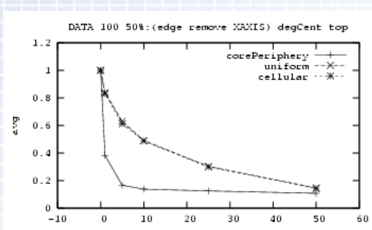
Run 3 full experiments (Uniform, Cellular, Core-periphery)

- 33.6 million graphs produced
- 840,000 data values to analyze

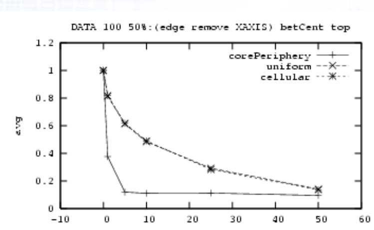
CASOS

Carnegie Mellon

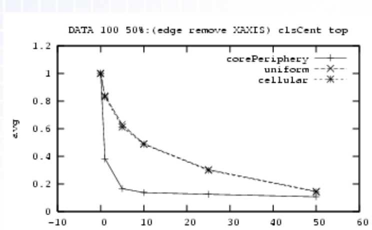
Four Centrality Measures Similar



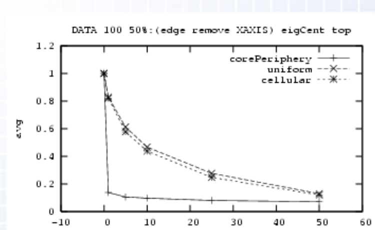
DEGREE



BETWEENNESS



CLOSENESS

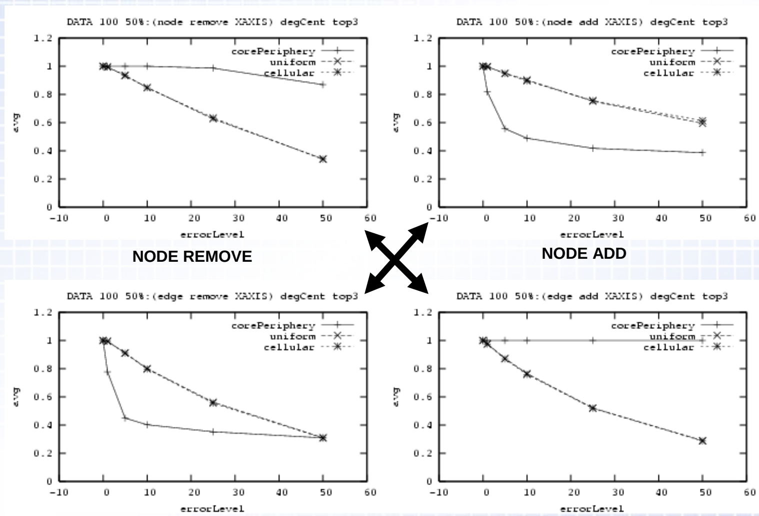


EIGENVECTOR

CASOS

Carnegie Mellon

Uniform vs. Core-Periphery



NODE REMOVE

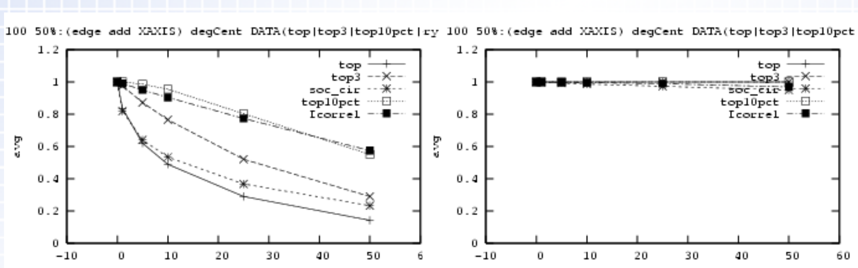
NODE ADD

EDGE REMOVE

EDGE REMOVE

Carnegie Mellon

Uniform vs. Core-Periphery

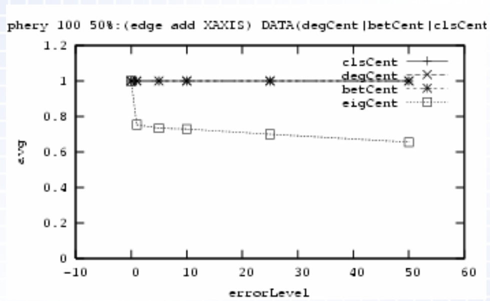


UNIFORM

CORE PERIPHERY

Carnegie Mellon

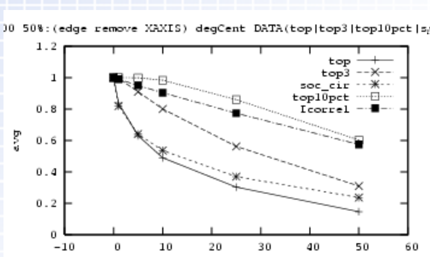
Core Periphery



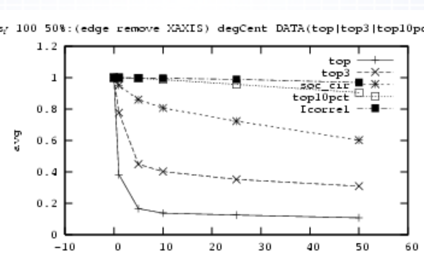
CASOS

Carnegie Mellon

Uniform vs. Core-Periphery



UNIFORM

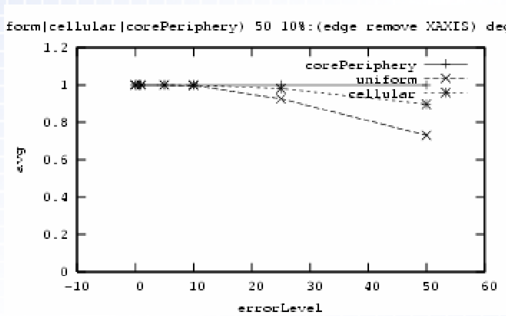


CORE PERIPHERY

CASOS

Carnegie Mellon

Cellular



Carnegie Mellon

Findings

1. Accuracy profile for all four centrality measures is similar
2. Core-Periphery to uniform/cellular is more robust in case of node-remove and edge-add; less robust in cases of node-add and edge-remove.
3. Core-Periphery edge-add has very high (near 1) robustness across all error levels
4. Core-Periphery less robust to smaller levels of error.

