


# Topology & Groups, Motifs, Communities



Prof. Kathleen M. Carley

[kathleen.carley@cs.cmu.edu](mailto:kathleen.carley@cs.cmu.edu)




**Carnegie Mellon**

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



# Topology



6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU



Carnegie Mellon  
IST Institute for Software Research

## What is Topology?

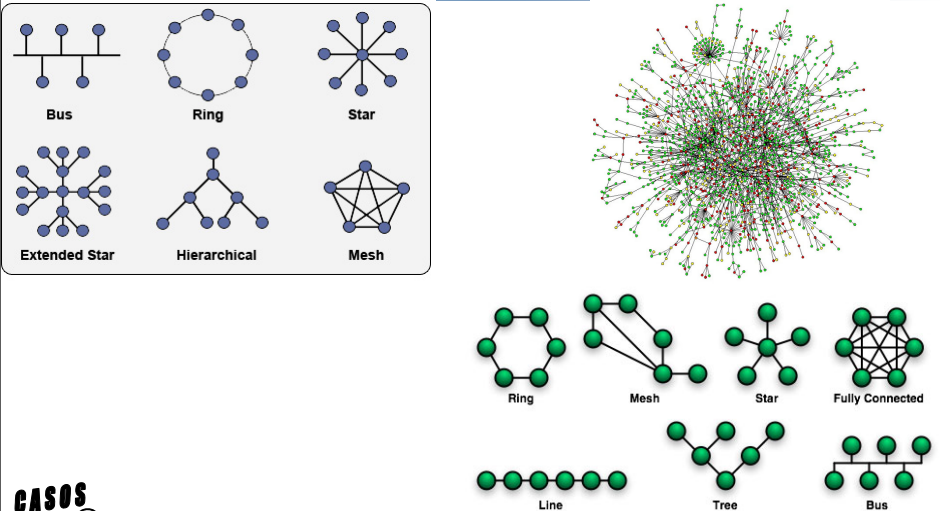
- The shape of the network
  - Overall structure = pattern of nodes and relations
  - Overall structure as a whole
  - Overall structure as the sum of the parts
- Approaches
  - Organization theory approach
  - Engineering – computer network (related to the original experimental psychology approach)
  - Component approach – statistics and mathematics
  - Stylized forms approach – physics, network science

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 3

Carnegie Mellon  
IST Institute for Software Research

## Images Networks



Bus Ring Star

Extended Star Hierarchical Mesh

Ring Mesh Star Fully Connected

Line Tree Bus

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 4

Carnegie Mellon  
ISR Institute for SOFTWARE RESEARCH

## Common Topologies

<p>Social Networks Network Science</p> <ul style="list-style-type: none"> <li>• Erdos-Renyi, i.e., random graph</li> <li>• Scale Free</li> <li>• Lattice</li> <li>• Small World (Lattice Ring)</li> <li>• Cellular</li> <li>• Core-periphery</li> <li>• Fully connected</li> </ul>	<p>Organization Theory</p> <ul style="list-style-type: none"> <li>• Hierarchy</li> <li>• Flat hierarchy</li> <li>• Matrix</li> <li>• Team</li> </ul> <p>Router &amp; Engineering</p> <ul style="list-style-type: none"> <li>• Bus</li> <li>• Ring</li> <li>• Star</li> <li>• Extended Star</li> <li>• Hierachy (Tree)</li> <li>• Mesh</li> <li>• Line</li> <li>• (Fully-Connected)</li> </ul>
--	---

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU

Carnegie Mellon  
ISR Institute for SOFTWARE RESEARCH

## Why is Knowing the Topology Important?

- Topology places limits on the empirical range of node level metrics and graph level metrics
- Topology impacts the distribution of node level metrics
- Knowing the topology gives you a high level view of what is going on

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 6



Carnegie Mellon  
IST Institute for Software Research

## Measuring Topology

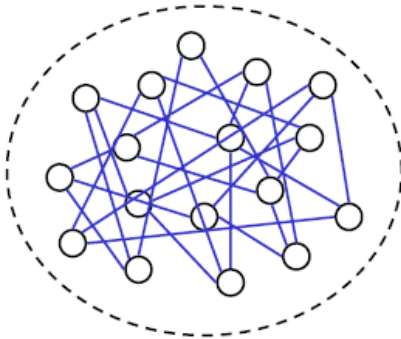
- Graph level indicators
  - Measures s.t. there is one measure for the graph
  - May or may not be indicative of topology
  - Key metrics
    - Density
    - Size
    - Limit what topology is possible
    - Always provide these
- Measured by graph level indicators
  - Hierarchy
  - Centralization
  - Clustering coefficient
  - Degree distribution

**CASOS** But there is not a single metric for each topology

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 7

Carnegie Mellon  
IST Institute for Software Research

## Random networks (Erdos-Renyi, '60)



A random graph with  $p = 3/N = 0.18$

- in a *random graph* each pair of nodes is connected with probability  $p$
- *LOW average path length*:  
 $L \approx \ln N / \ln \langle k \rangle \sim \ln N$  for  $N \gg 1$   
 (because the entire network can be covered in about  $\langle k \rangle^L$  steps:  $N \sim \langle k \rangle^L$ )
- *LOW clustering coefficient* (if sparse):  
 $C = p = \langle k \rangle / N \ll 1$  for  $p \ll 1$   
 (because the probability of 2 neighbors being connected is  $p$ , by definition)
- *PEAK (Poisson) degree distribution* (low value):  
 $\langle k \rangle \approx pN, \quad P(k) \approx \delta(k - pN)$

**CASOS**

6/7/2020 Slide by Kraemer & Barabasi, Bonabeau (SciAm'03) 8

Carnegie Mellon  
ISR Institute for SOFTWARE RESEARCH

## Random Networks

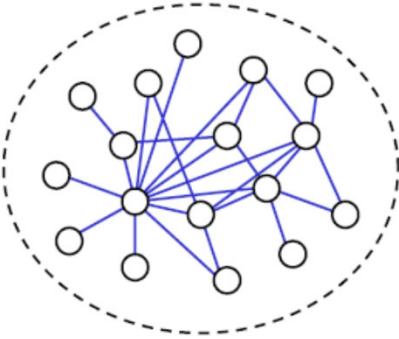
- Most common form studied
- Statistical tests to decide if your network is random
- Easy to generate
- Good mathematical properties
- Very different than real world networks

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 9

Carnegie Mellon  
ISR Institute for SOFTWARE RESEARCH

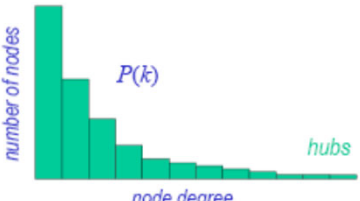
## Scale-free networks



A schematic scale-free network

- in a *scale-free network* the degree distribution follows a **POWER-LAW**:  

$$P(k) \sim k^{-\gamma}$$
- there exists a small number of highly connected nodes, called *hubs* (tail of the distribution)
- the great majority of nodes have few connections (head of the distribution)



number of nodes

$P(k)$

node degree

hubs

CASOS

6/7/2020 Slide by Kraemer & Barabasi, Bonabeau (SciAm'03) 10

Carnegie Mellon  
ISR Institute for Software Research

## Properties of Scale Free Networks

- A small number of nodes contribute heavily to connectivity.
  - These nodes are called hubs.
- Any two nodes, even in a very large network, can be connected via few other intermediary nodes.
- A power law has a characteristic (constant) exponent (dimension).
  - Regardless of size ... the dimension stays the same.
  - Thus the term "scale-free".
- Scale-free networks are "self similar".
  - Any part of the network is statistically similar to the whole network.
  - Self similarity is the key feature of fractals.
- Scale-free networks are "robust".
  - It can operate with the random removal of a few nodes.
  - Connectivity failure occurs when a hub is removed.
- Scale-free networks tend to promote high speed transfer of information or energy.
  - Hubs have a combination of high global connectivity with highly developed local clustering.
  - This leads to rapid information diffusion.

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 11

Carnegie Mellon  
ISR Institute for Software Research

## As Opposed to Random Network

- In a random network, each node contributes approximately the same to the overall connectivity of the network.
- Any two nodes are not guaranteed to connect.
- There is no characteristic (constant) exponent (dimension).
- Random networks are "self similar". *\* debated*
  - Any part of the network is statistically similar to the whole network.
  - Self similarity is the key feature of fractals.
- Random networks are "robust".
  - It can operate with the random removal of a few nodes.
  - Connectivity failure occurs when a hub is removed.
- Information tends to move slowly in a random network.

CASOS

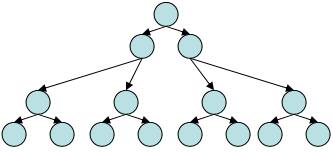
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 12



Carnegie Mellon  
IST Institute for Software Research

## Hierarchy

- Unified chain of command
- Breadth
- Depth
- Information flows up
  - With information loss
- Decisions and commands flow down
- Information compressed as it goes up
- Consequent cap on performance



CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 13

Carnegie Mellon  
IST Institute for Software Research

## Measures of Hierarchy

- Krackhardt Hierarchy
- Breadth
- Depth
- Centralization (based on degree)
- Distribution for Clustering Coefficient

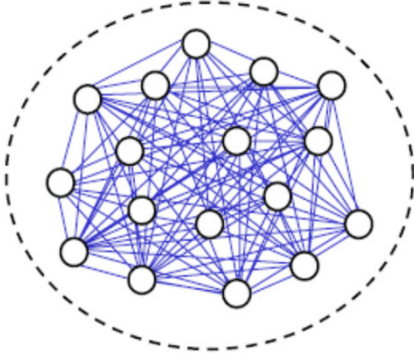
CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 14



Carnegie Mellon  
IST Institute for Software Research

## Regular Networks – Fully Connected



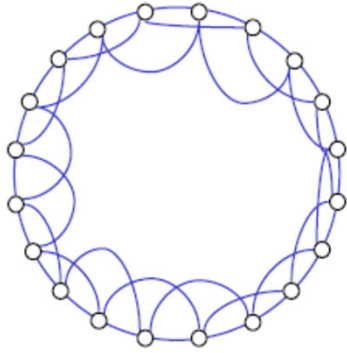
A fully connected network

- in a *fully (globally) connected* network, each node is connected to all other nodes
- fully connected networks have the **LOWEST** path length and diameter:  
 $L = D = 1$
- the **HIGHEST** clustering coefficient:  
 $C = 1$
- and a **PEAK** degree distribution (at the largest possible constant):  
 $k_d = N-1, \quad P(k) = \delta(k - N+1)$
- also the highest number of edges:  
 $E = N(N-1)/2 \sim N^2$

CASOS 6/7/2020 Slide by Kraemer & Barabasi, Bonabeau (SciAm'03) 15

Carnegie Mellon  
IST Institute for Software Research

## Regular networks – Lattice: ring world



A ring lattice with  $K = 4$

- in a *ring lattice*, nodes are laid out on a circle and connected to their  $K$  nearest neighbors, with  $K \ll N$
- **HIGH** average path length:  
 $L \approx N/2K \sim N$  for  $N \gg 1$   
(mean between closest node  $l = 1$  and antipode node  $l = N/K$ )
- **HIGH** clustering coefficient:  
 $C \approx 0.75$  for  $K \gg 1$   
(mean between center with  $K$  edges and farthest neighbors with  $K/2$  edges)
- **PEAK** degree distribution (low value):  
 $k_d = K, \quad P(k) = \delta(k - K)$

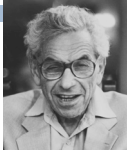


CASOS 6/7/2020 Slide by Kraemer & Barabasi, Bonabeau (SciAm'03) 16



Carnegie Mellon  
IST Institute for Software Research

## Small World

- What is your Erdos Number?  
– <http://xkcd.com/599/>
- Six degrees of Kevin Bacon?
- Stanley Milgram - Small world experiment

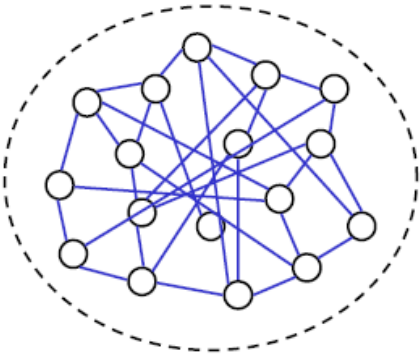




CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 17

Carnegie Mellon  
IST Institute for Software Research

## Small-world networks (Watts-Strogatz, '98)



A Watts-Strogatz small-world network

- a network with *small-world EFFECT* is ANY large network that has a low average path length:  
 $L \ll N$  for  $N \gg 1$
- famous "6 degrees of separation"
- the *Watts-Strogatz (WS) small-world MODEL* is a hybrid network between a regular lattice and a random graph
- WS networks have both the **LOW average path length** of random graphs:  
 $L \sim \ln N$  for  $N \gg 1$
- and the **HIGH clustering coefficient** of regular lattices:  
 $C \approx 0.75$  for  $K \gg 1$

CASOS

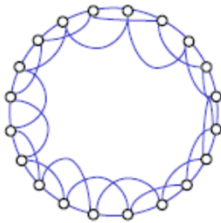
6/7/2020 Slide by Kraemer & Barabasi, Bonabeau (SciAm'03) 18

Carnegie Mellon  
IST Institute for Software Research

## Small-world networks

**Ring Lattice**

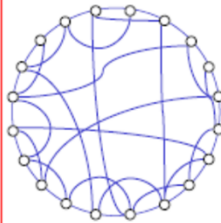
- large world
- well clustered



$p = 0$  (order)

**Watts-Strogatz (1998)**


- small world
- well clustered



$0 < p < 1$


**Random graph**

- small world
- poorly clustered



$p = 1$  (disorder)

➤ the WS model consists in gradually rewiring a regular lattice into a random graph, with a probability  $p$  that an original lattice edge will be reassigned at random



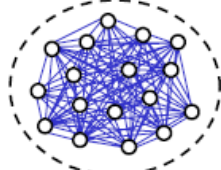
6/7/2020

Slide by Kraemer & Barabasi, Bonabeau (SciAm'03)

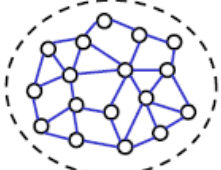
19

Carnegie Mellon  
IST Institute for Software Research

## Small-world networks



full,  $\langle k \rangle = 16$

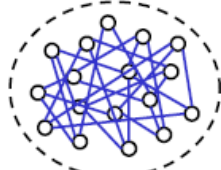


lattice,  $\langle k \rangle = 3$

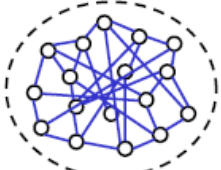
➤ on the other hand, the WS model still has a **PEAK (Poisson) degree distribution** (uniform connectivity)

➤ in that sense, it belongs to the same family of **exponential networks**:

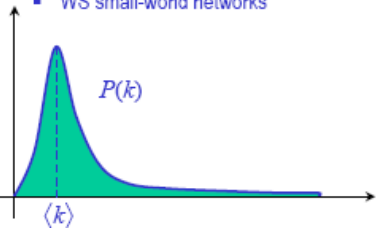
- fully connected networks
- lattices
- random graphs
- WS small-world networks



random,  $\langle k \rangle = 3$




WS small-world,  $\langle k \rangle = 3$



$P(k)$

$\langle k \rangle$



6/7/2020

Slide by Kraemer & Barabasi, Bonabeau (SciAm'03)


20



Carnegie Mellon  
IST Institute for Software Research


## Other Common Topologies

- Core-Periphery
  - A network where there is a substructure that has a set of members that are very densely connected, and then a set of others that are connected to only a few of the core members
- Cellular
  - A network where there are a set of substructures such that each substructure is densely connected and each of these substructures is connected to only one or two other substructures. Most members only have connections within there cell.

 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 21

Carnegie Mellon  
IST Institute for Software Research

# Groups Motifs Community Detection

 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 22



Carnegie Mellon  
IST Institute for Software Research

## Three Topics

- What is a group?
- How do you assess groups?
- How do you find groups?

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 23

Carnegie Mellon  
IST Institute for Software Research

## What is a Group? Community? Motif?

- Any number of entities considered as a unit
- Nominal group – “named” collective e.g., nurses
- Collection of entities with features in common
- Small Group
  - 3-15 members
  - Able to communicate freely & openly with all group members
  - Norms
  - Roles
  - Common purpose
- Community: A set of connected nodes with something in common
- Motif: Predefined pattern

(a) Community without overlapping

(b) Community B & C are overlapped

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 24



Carnegie Mellon  
IST Institute for Software Research

## Why “group” Nodes? Grouping Provides a Useful Summary

- Find communities that are likely to
  - Share attributes
  - Share information / beliefs
  - Experience the same future influences
  - Have similar goals / strategies in selecting links
- Use observed member traits to predict unobserved.
- Find unique individuals (local leaders, spanners, etc.)

CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 25

Carnegie Mellon  
IST Institute for Software Research

## Why are Groups Important? Structuralism as Social Phenomenon

- Similar nodes have similar outcomes
  - If two nodes occupy the same position, then they will get the same results, even if unconnected to each other
    - Even if only connected to similar others – cohesion
    - Only if connected to same others – equivalence
- Networks with similar structures will have similar outcomes
  - Similar structures = similar topology
  - E.g., Similarly structured teams will have similar performance outcomes
- Members of group will have similar outcomes
  - Ideas, attitudes, illnesses, behaviors
  - Due to interpersonal transmission
    - Transference
    - Influence / persuasion
    - Co-construction of beliefs & practices
      - As in communities of practice

CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 26



Carnegie Mellon  
IST Institute for Software Research

## 4 conceptual reasons for why groups matter

- Cohesion
  - Because the nodes have the same kind of position – relations to same type of other nodes
  - Network region might contain cohesive subgroups
- Equivalence
  - Because the nodes have the same linkages – relations to the same other nodes
- Distinction
  - Because the nodes are different from other nodes around them, anomalies
- Similarity
  - Because the nodes have the same kind of features

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 27

Carnegie Mellon  
IST Institute for Software Research

## How Do I Know that the Group I Found Makes Sense?

- Recognizable
  - Are members similar on some dimension?
    - Statistical members on attributes or links in rest of meta-network
  - Are groups distinctive?
    - Ties, lack of ties, or patterns of ties are different
- Coverage
  - Are the members correct?
    - Optimal clustering/breaking
    - Comparison of results of grouping algorithms
- Theoretically sound
  - Does algorithm generate groups that meet the theoretical criteria?

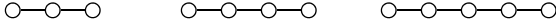
CASOS

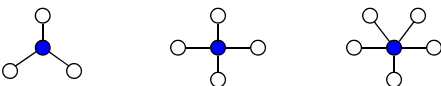
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 28

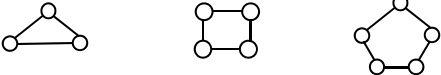


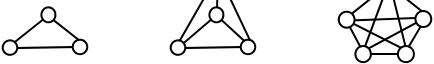
Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH


## Illustrative Motifs


**Paths** 

**Stars** 

**Cycles** 

**Complete Graphs** 


**Bipartite Graphs** 

CASOS  6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 29

Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## Groups

- Set of nodes that meet some criteria – a node set
- Goal is to extract these automatically based on node properties (such as – how they are connected)
- Finding groups is pattern analysis
- 2 types of approaches mechanistically
  - Bottom up – combine
    - E.g., Clustering nodes
    - E.g., Cluster “dyads” or “links”
  - Top down – split entire set into subsets
    - E.g. break up groups (Concor)
    - E.g. segregate set of links
- 2 types of approaches based on need
  - Locate members, locate anomalies
  - Break the network (locate components, sub-cells, ...), segregate links

CASOS  6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 30



Carnegie Mellon  
 Institute for SOFTWARE RESEARCH

## Groups & Positions

- Groups - Cliques, clusters, components, cores, circles, etc.
- Subgraph - any collection of points selected from a whole graph of a network.
  - examples - random selection, males and females, people who smoke, etc.
- Goal:
  - discovers the underlying structure
  - using a criterion find the largest sub-graph possible that maintains this criterion
  - the sub-graph is maximal

CASOS  
 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 31

Carnegie Mellon  
 Institute for SOFTWARE RESEARCH

## Community or Group Detection Methods

Non Overlapping	Overlapping
• Components	• Clique Percolation
• Minimum-cut method	• FOG
• Hierarchical clustering	• K-core
• Girvan-Newman algorithm	
• Modularity maximization	
• Louvain method	

CASOS  
 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU





Carnegie Mellon  
IST Institute for Software Research

## Finding Groups

- Components
  - Isolates
- Nominal Groups
- Group Identification Algorithms
- Community Detection Algorithms/ Optimization Algorithms
- Similarity Based Algorithms
- Issues:
  - number of groups/communities within the network is typically unknown
  - groups are often of unequal size and density

CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 33

Carnegie Mellon  
IST Institute for Software Research

## Terminology: Components

- A subgraph  $S$  of a graph  $G$  is a component if  $S$  is maximal and connected
- If  $G$  is a digraph, then
  - $S$  is a weak component if it is a component of the underlying (undirected) graph
  - $S$  is a strong component if for all dyads  $u, v$  in  $S$ , there is a path from  $u$  to  $v$
- Finding components is the first step in analysis of large graphs
  - Analyze each component separately, or discard very small components

CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 34

Carnegie Mellon  
IST Institute for Software Research

## Isolates

- A node not connected to any other nodes in a network
- Each isolate is its own component
- Dealing with isolates
  - Delete them
    - Often used with large networks
  - Lump into their own group
    - Often used when issues of cohesion need to be addressed
  - Leave each as their own component
    - Often used with small networks

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 35

Carnegie Mellon  
IST Institute for Software Research

## Terminology: K-Cores Clique algorithm

- A maximal subgraph  $S$  such that for all  $u$  in  $S$ ,  $\alpha(u,S) \geq k$ 
  - each point is adjacent to  $k$  other points
  - $S=A$  is 1-core & 2-core; B and C 3-core
  - There is no 4-core or higher
- All nodes in a  $k$ -core have a degree greater than or equal to  $k$ .
- Finds large regions within which cohesive subgroups may be found
- Identifies fault lines across which cohesive subgroups do not span

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 36

Carnegie Mellon  
IST Institute for Software Research

## Terminology: K-Cores

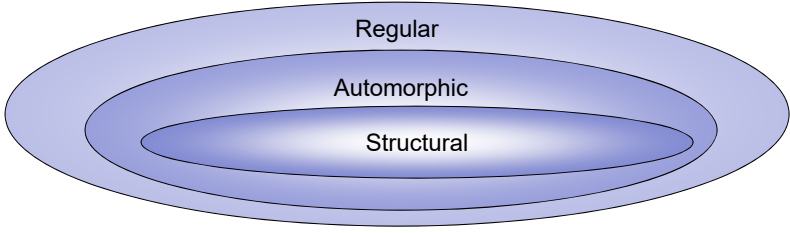
- An area within a graph of high cohesion.
  - Dense heterogeneous groups
- 1k-core is a component
  - every node has one connection
  - Every node is connected to every other node by some path
- 2k-core drops all nodes of degree one, then finds the connected components
- The higher the  $k$ , the higher the core's density
- K-core collapse - process of increasing  $k$  until the core collapses
  - The point where the greatest number of nodes drops out.
  - The pattern of the core collapse indicates the degree of clumpiness in the core.

CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 37

Carnegie Mellon  
IST Institute for Software Research

## Groups and Equivalences

- Many grouping mechanisms are based on equivalences
- Common ones:
  - Structural
  - Regular
  - Automorphic \*At least as defined in JMS paper in 1994.
- These are subsets



CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 38



Carnegie Mellon  
IST Institute for Software Research

## Terminology: Equivalence

- An equivalence is just the relation  $E$  induced by a partition
- Is any relation that satisfies 3 conditions:
  - Transitivity:  $(a,b), (b,c) \in E$  implies  $(a,c) \in E$
  - Symmetricity:  $(a,b) \in E$  iff  $(b,a) \in E$
  - Reflexivity:  $(a,a) \in E$

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 39

Carnegie Mellon  
IST Institute for Software Research

## Structural Equivalence

- A coloration  $C$  is structurally equivalent
  - if  $C(u)=C(v)$  iff  $N(u)=N(v)$
  - $N(u) = N(v)$  iff  $N^i(u)=N^i(v)$  and  $N^o(u)=N^o(v)$
- In other words – two nodes are structurally equivalent if they are connected to the exact same set of others

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 46



Carnegie Mellon  
ISR Institute for Software Research

## Structural Equivalence

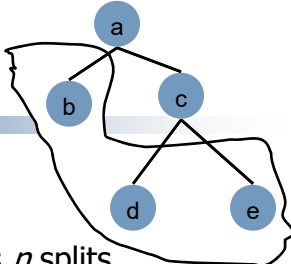
- Structurally indistinguishable
  - Same degree, centrality, belong to same number of cliques, etc.
  - Only the label on the node can distinguish it from those equivalent to it.
  - Perfectly substitutable: same contacts, resources
- Face the same social environment
  - Similar forces affecting them – same influencers
  - On average, hear things equally early, influenced similarly, have similar things to cope with

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 41

Carnegie Mellon  
ISR Institute for Software Research

## CONCOR



- Works by splitting groups
- Specify number of splits
- Recursively splits partitions, user selects  $n$  splits.
  - $n$  splits  $\rightarrow 2^n$  groups
- At each split, divides nodes based on maximum correlation in outgoing connections.
- Builds a hierarchical decomposition
- Calculates correlation between each pair of rows/columns
  - Then the correlation of the correlations ...
  - Repeats until reaches “stableness”
  - Then splits the nodes into two groups based on the correlation

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 42



Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## CONCOR

- Finds ZERO blocks
- Issues –
  - First correlation does most of the work
  - Heuristic approach
  - Located groups are “cliques” and often only regularly equivalent
- PRO: Only commonly used algorithm detects relaxed structural equivalence. (except arguable PCA)
- CON: Top down splitting of nodes imposes structure
- CON: Requires user to choose a power of 2 for the number of groups.

CASOS  
6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 43

Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## Girvan Newman’s method (partition the nodes)

- The Girvan–Newman algorithm detects communities by progressively removing edges (with high betweenness centrality) from the original network.
- These edges are believed to connect communities
- Algorithm stops when there are no edges between the identified communities.

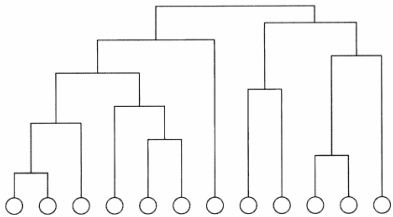


Fig. 2. An example of a small hierarchical clustering tree. The circles at the bottom represent the vertices in the network, and the tree shows the order in which they join together to form communities for a given definition of the weight  $W_{ij}$  of connections between vertex pairs.

<http://www.istor.org/stable/pdf/3058918.pdf>


CASOS  
Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 44



Carnegie Mellon  
IST Institute for Software Research

## Newman-Girvan


- Detects groups
  - “community structure”
  - A community consists of a subset of nodes within which the node-node connections are dense, and the edges to nodes in other communities are less dense
- Procedure:
  - Calculate betweenness of all existing edges in the network
  - Remove edge with the highest betweenness is removed
  - Recalculate betweenness of all edges affected by the removal
  - Repeat until no edges remain
- Procedure to find optimal grouping
- Fast
- Groups sometimes difficult to interpret

 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 45

Carnegie Mellon  
IST Institute for Software Research

## Newman-Girvan

- NG takes divisive approach
- Finds edge (link) with highest betweenness
- Removes that link
- Calculates community groups
- Repeats process (finds edge with largest betweenness, deletes it, calculates communities)
- At each step need to calculate index of fit

 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 46



Carnegie Mellon  
ISR Institute for SOFTWARE RESEARCH

## Newman-Girvan

- Modularity
  - Consider a matrix  $e$  ( $k \times k$ ) in which elements indicate the fraction of edges in the 2 groups
  - Trace of matrix is the sum of the main diagonal
  - High values of the trace indicate a good partition of the network (because it would indicate all the links are within communities).
- Row and column sums indicate cross group links
- Modularity is the sum of the difference between the diagonal and the off-diagonal elements
- Higher the number the more partitioned the network is

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 47

Carnegie Mellon  
ISR Institute for SOFTWARE RESEARCH

## How Good is the Grouping? Modularity

- Modularity is defined as:  
 $Q = \# \text{ edges within communities} - \text{expected } \# \text{ edge of a null model network (same size)}$   
 Where “expected” come from a “null model” to compare our network against: networks with the same  $n$  and  $m$ , where edges are placed at random

$$Q = \frac{1}{2m} \sum_{C \in \mathcal{P}} \sum_{i, j \in C} [A_{ij} - P_{ij}] \quad \begin{cases} \text{if } P_{ij} = \langle k \rangle^2 / 2, & \text{then } Q \equiv Q_{\text{unif}} \\ \text{if } P_{ij} = k_i k_j / 2m, & \text{then } Q \equiv Q_{\text{conf}}. \end{cases}$$

- A scale value between -1 and 1 that measures the density of edges inside communities to edges outside communities
- Larger values of  $Q$  indicating stronger community structure.
- Goal: assign nodes to community to maximize  $Q$

CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 48





Carnegie Mellon  
IST Institute for Software Research

## Louvain method (partition the nodes)

- Goal: optimize modularity → theoretically results in the best possible grouping of the nodes of a given network (it depends on the function of the network, the reason behind clustering)
- The Louvain Method of community detection:
  - find small communities by optimizing modularity locally on all nodes,
  - then each small community is grouped into one node
  - then the first step is repeated
- Visualization: <https://www.youtube.com/watch?v=dGa-TXpoPz8>

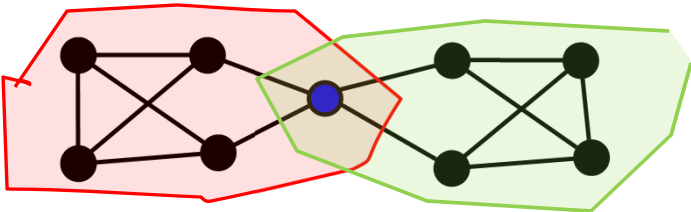
CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 49

Carnegie Mellon  
IST Institute for Software Research

## What is FOG?

- Fuzzy, Overlapping Groups
  - Multiple group memberships
  - Varying strength of membership
  - No arbitrary assignments on boundary spanners
    - Reveals details of interstitial roles
- Designed for Link Data or Network Data
- Generative model (rather than pattern matching)



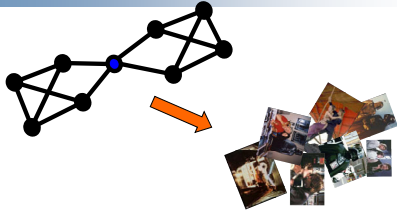
CASOS

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 50

**Carnegie Mellon**  
**ISR** Institute for Software Research


## Sampling Link Data From Networks

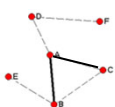
- Random tree
- Models iterative interaction
  - Informal gathering
  - Spread of rumor or info

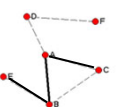


Tree

Link

  
 {A,B}

  
 {A,B,C}

  
 {A,B,C,E}

**CASOS** 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 51

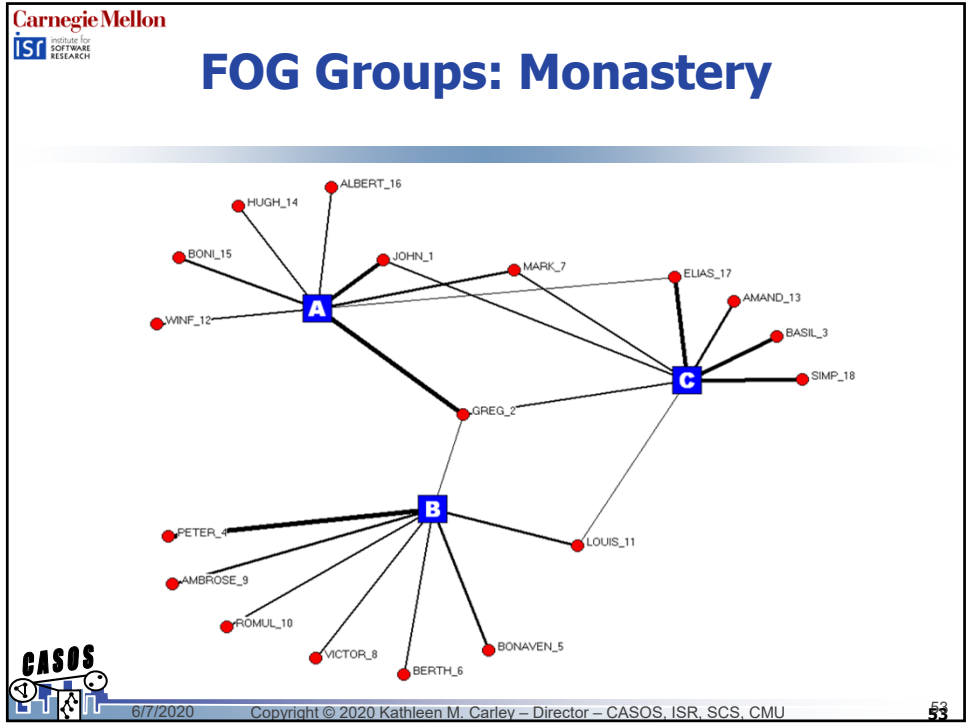
**Carnegie Mellon**  
**ISR** Institute for Software Research

## FOG Algorithms

Algorithm	Based On	Pros	Cons
H-FOG	Hierarchical Clustering	<ul style="list-style-type: none"> <li>• Nested Groups</li> <li>• Run once; explore tree to determine # of groups.</li> </ul>	Scales poorly $O(n^4)$
k-FOG	K-Means	Scales well	Must guess # of groups, $k$
$\alpha$ -FOG	Dirichlet Process	Fast, Does not require guessing number of groups ( $\alpha$ parameter is expected concentration)	Data-hungry

**CASOS** 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 52





**Carnegie Mellon**  
**ISR** Institute for SOFTWARE RESEARCH

## Sampson's Monastery

	10	5	9	6	4	11	8	12	1	2	14	15	7	16	13	17	18
ROMUL	10	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
BONAVEN	5	0	0	2	0	3	3	0	0	0	1	0	0	0	0	0	0
AMBROSE	9	0	1	0	0	2	0	2	1	2	1	0	0	0	0	0	0
BERTH	6	0	1	3	0	4	2	0	0	0	1	0	0	0	0	0	0
PETER	4	3	1	0	4	0	4	0	0	0	0	0	0	0	0	0	0
LOUIS	11	0	3	2	0	2	0	2	0	0	1	0	0	1	0	0	0
VICTOR	8	0	1	2	3	4	2	0	0	1	0	0	0	0	0	0	0
WINF	12	0	0	0	0	0	0	0	3	3	1	0	2	0	0	0	0
JOHN	1	0	1	0	0	0	0	1	4	0	1	2	0	1	0	0	1
GREG	2	0	1	0	0	0	0	1	3	4	0	0	0	3	0	0	0
HUGH	14	0	0	0	0	0	0	0	3	4	3	0	4	0	1	0	0
BONI	15	0	0	0	0	0	0	0	1	2	4	3	0	2	0	0	0
MARK	7	0	0	0	0	0	0	0	3	0	4	0	2	0	4	0	0
ALBERT	16	0	0	0	0	0	0	0	1	0	4	0	4	4	0	0	0
AMAND	13	0	4	0	0	0	3	0	0	0	0	0	0	0	0	0	1
BASIL	3	0	0	0	0	0	0	0	0	4	0	0	0	0	4	0	4
ELIAS	17	0	0	0	0	0	0	0	0	0	3	0	0	0	1	3	0
SIMP	18	0	0	0	0	0	0	0	1	4	0	0	0	0	0	3	4

Surveys: Sampson, 1968                      Collation: Breiger, 1975

CASOS 6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 54



Carnegie Mellon  
 Institute for Software Research

## Block Modeling

- A block model is a reduced form representation such that nodes are divided into a set of mutually exclusive groups
- The resulting groups can then be analyzed as a network such that
  - The group's connection to itself is the density of the connections among members
  - For each pair of groups, the inter-group connection is the density of the connections of group 1 (row) to group 2 (column)
  - The resulting block matrix can be turned into a binary matrix by simply comparing the level of connections in the block to the overall density of the original matrix such that there if the value of the cell is  $\geq$  to the overall density then we replace it with a 1, else 0

55

Carnegie Mellon  
 Institute for Software Research

## Example

	A	B	C	D	E	F	G	H	I	J
A	0	1	1							
B	1	<b>G1</b>		1						
C	1		0					1		
D	1	1		0						
E										
F										
G	1						0	1		1
H		1					1		1	
I									<b>G3</b>	1
J	1			1			1			0

	G1	G2	G3
G1	.58	0	.06
G2	0	1	0
G3	.25	0	.58

	G1	G2	G3
G1	1	0	0
G2	0	1	0
G3	1	0	1

Density =  $21/90 = .22$


56



Carnegie Mellon  
 Institute for SOFTWARE RESEARCH

## Common Blockmodels

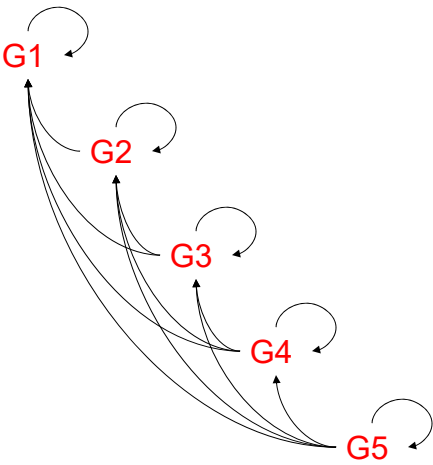
- Completely connected  $\longrightarrow$   $\begin{matrix} 11 \\ 11 \end{matrix}$
- Opposing groups  $\longrightarrow$   $\begin{matrix} 10 \\ 01 \end{matrix}$
- Supporters and Supporting  $\longrightarrow$   $\begin{matrix} 01 \\ 10 \end{matrix}$
- Central Core  $\longrightarrow$   $\begin{matrix} 10 \\ 00 \end{matrix}$
- Hierarchy  $\longrightarrow$   $\begin{matrix} 10 \\ 10 \end{matrix}$
- Core with Outreach  $\longrightarrow$   $\begin{matrix} 11 \\ 00 \end{matrix}$
- Core-periphery  $\longrightarrow$   $\begin{matrix} 11 \\ 10 \end{matrix}$
- Isolates  $\longrightarrow$   $\begin{matrix} 00 \\ 00 \end{matrix}$


CASOS  6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 57

Carnegie Mellon  
 Institute for SOFTWARE RESEARCH

## Illustrative Hierarchy

	G1	G2	G3	G4	G5
G1	1	0	0	0	0
G2	1	1	0	0	0
G3	1	1	1	0	0
G4	1	1	1	1	0
G5	1	1	1	1	1



CASOS  6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 58



Carnegie Mellon  
 ISI Institute for SOFTWARE RESEARCH

## Illustrative Alternative Hierarchy

	G1	G2	G3	G4	G5
G1	1	0	0	0	0
G2	1	1	0	0	0
G3	0	1	1	0	0
G4	0	0	1	1	0
G5	0	0	0	1	1

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 59

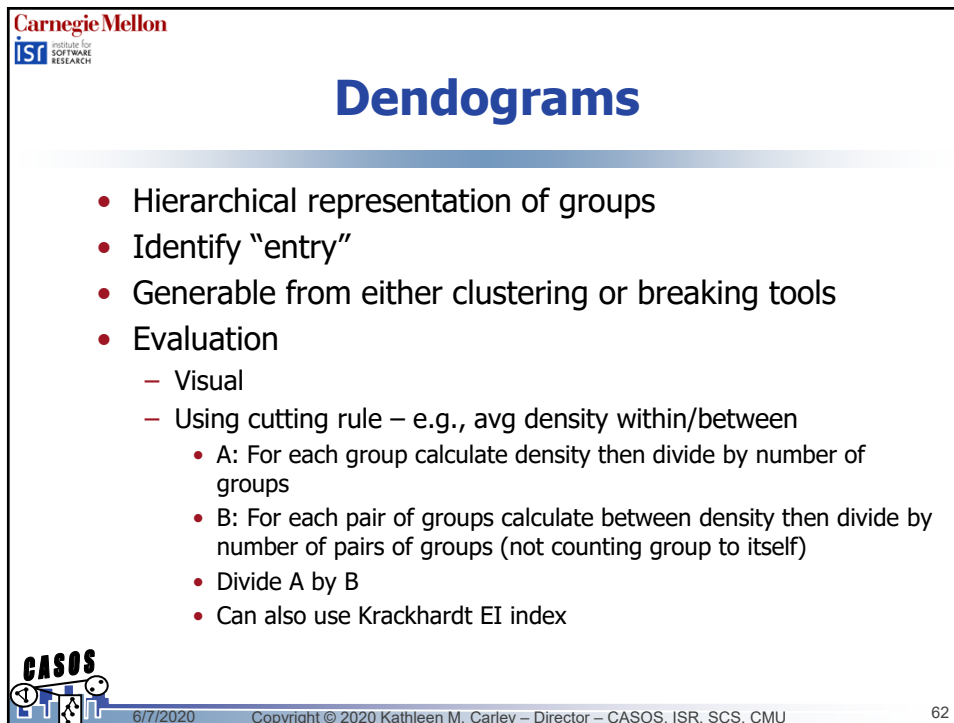
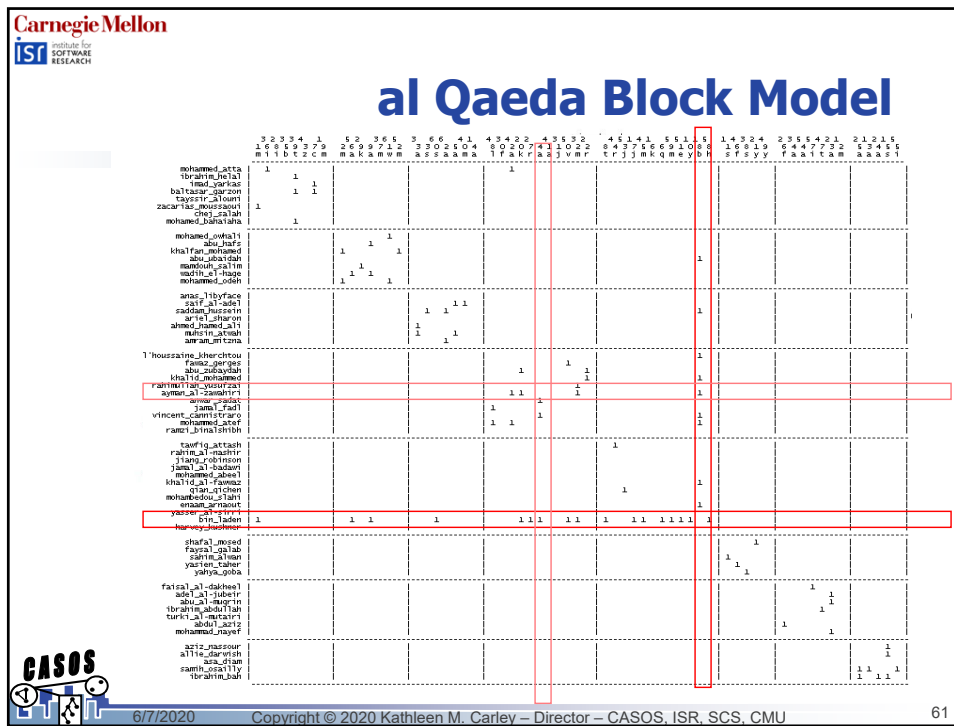
Carnegie Mellon  
 ISI Institute for SOFTWARE RESEARCH


## ORA Demonstration al Qaeda 2000 network

- Hawaswi
- Fakhiri
- Khachrou
- Haroun
- OmerSheikh
- Nashin
- Hanjour
- Omani
- Tballi
- Badawi
- Ouso
- AlAzdi
- Johani
- Dandasi
- Ahdal
- Bandar
- Muqin
- Reshoud
- Mujeli
- Zaidan
- Aul
- Ruzhud
- Ersoz
- Tugluoglu
- Ilhan
- Zinedine
- Akar
- Dumont
- Haouari
- Boukhari
- Maroni
- Beneli
- Bensakhria
- Jastar
- Kadri
- Benahmed

6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU 60







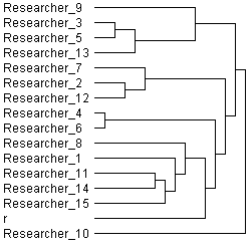
## Dendograms

**Hierarchical Clustering Diagram**


```

Level      1      1      1 1 1 1 1
-----
          9 3 5 3 7 2 2 4 6 8 1 1 4 5 6 0
-----
-0.05     . . . . . XXXX . . . . .
-0.00     . XXX . . . . XXX . . . . .
0.039     . XXX . . XXX XXX . . . . .
0.083     . XXXXX . XXX XXX . . . . .
0.122     . XXXXX XXXXX XXX . . . . .
0.160     . XXXXX XXXXX XXX . . XXX . . .
0.229     . XXXXX XXXXX XXX . . XXXXX . .
0.264     . XXXXX XXXXX XXX . XXXXXXXX . .
0.290     . XXXXX XXXXX XXX XXXXXXXXXXXX . .
0.309     XXXXXXXX XXXXX XXX XXXXXXXXXXXX . .
0.321     XXXXXXXX XXXXX XXX XXXXXXXXXXXX . .
0.284     XXXXXXXX XXXXX XXXXXXXXXXXX . .
0.204     XXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXX . .
0.000     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX . .
0         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX . .
                
```


**Dendrogram**



Group	Size	Members
1	6	1 8 11 14 15 16
2	3	2 7 12
3	4	3 5 9 13
4	2	4 6
5	1	10




Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU



## Summary

- **Why Group?**
  - Reconstruct “real” groups
  - Find individuals who might be or act similarly
  - Find individuals who have unusual community ties/
- **CONCOR: Structural Similarity**
  - Finds groups with similar roles in network, even if dispersed
- **Newman-Girvan/Louvain: Cohesive Communities**
  - Finds unusually dense clusters, even in large networks
  - If big data use Louvain – very fast
- **FOG: Fuzzy, Overlapping Groups**
  - Gives better understanding of individuals spanning groups
  - Analyzes network data or raw link data



6/7/2020 Copyright © 2020 Kathleen M. Carley – Director – CASOS, ISR, SCS, CMU

