Clustering Social Networks to Discover Topologies

Merve Celen (Univ. of Texas at Austin)
Satyabrata Pradhan (Infosys Labs, Hyderabad)
Radha Krishna Pisipati (Infosys Labs, Hyderabad)

Agenda

- Social Networks
- Motivation
- Project Overview
- Topology Score Determination
- Clustering
- Experimentation
- Future Work

Social Networks

- 202 major active social networking websites
- 17 virtual communities with more than 100M users

Facebook

- More than 750M active users
- Average user has 130 friends and is connected to 80 community pages, groups and events

Windows Live Messenger

- 330M users

Twitter

- Around 200M users
- More than 200M tweets each day

LinkedIn

- 100M users

MySpace

- 50M users

Motivation

- Airline discount offers
- Spread of diseases
- Spread of smoking, obesity, etc.
- Music, movie, etc. download recommendations
- Word-of-mouth marketing, viral ad campaigns
  - Can reach millions in very short amount of time
  - Impossible to market the product to every individual
  - Critical to recognize key influencers and communities to promote the product or service

Related Work

- Girvan and Newman (2002), *Community Structure in Social and Biological Networks*
  - Proposed a method for clustering social and biological networks using betweenness centrality to find cluster boundaries.
  - Focus instead on those edges that are least central, the edges that are most ‘‘between’’ communities.
- Mishra et al (2007), *Clustering Social Networks*
  - A deterministic algorithm for discovering overlapping clusters in social networks
  - Assumes that each cluster has a champion and there is a sufficiently large gap between internal density and external sparsity

Project Overview

- Aim of the project:
  - Cluster a social network using topology discovery
- Project mainly involves:
  - Application of graph theoretical approaches for discovering topologies
  - Devise novel clustering approaches to derive meaningful insights from social networks
- Challenges:
  - Social networks are huge
  - Requires long computation time
  - No known optimal method to discover topologies

Easy to find the key influencers

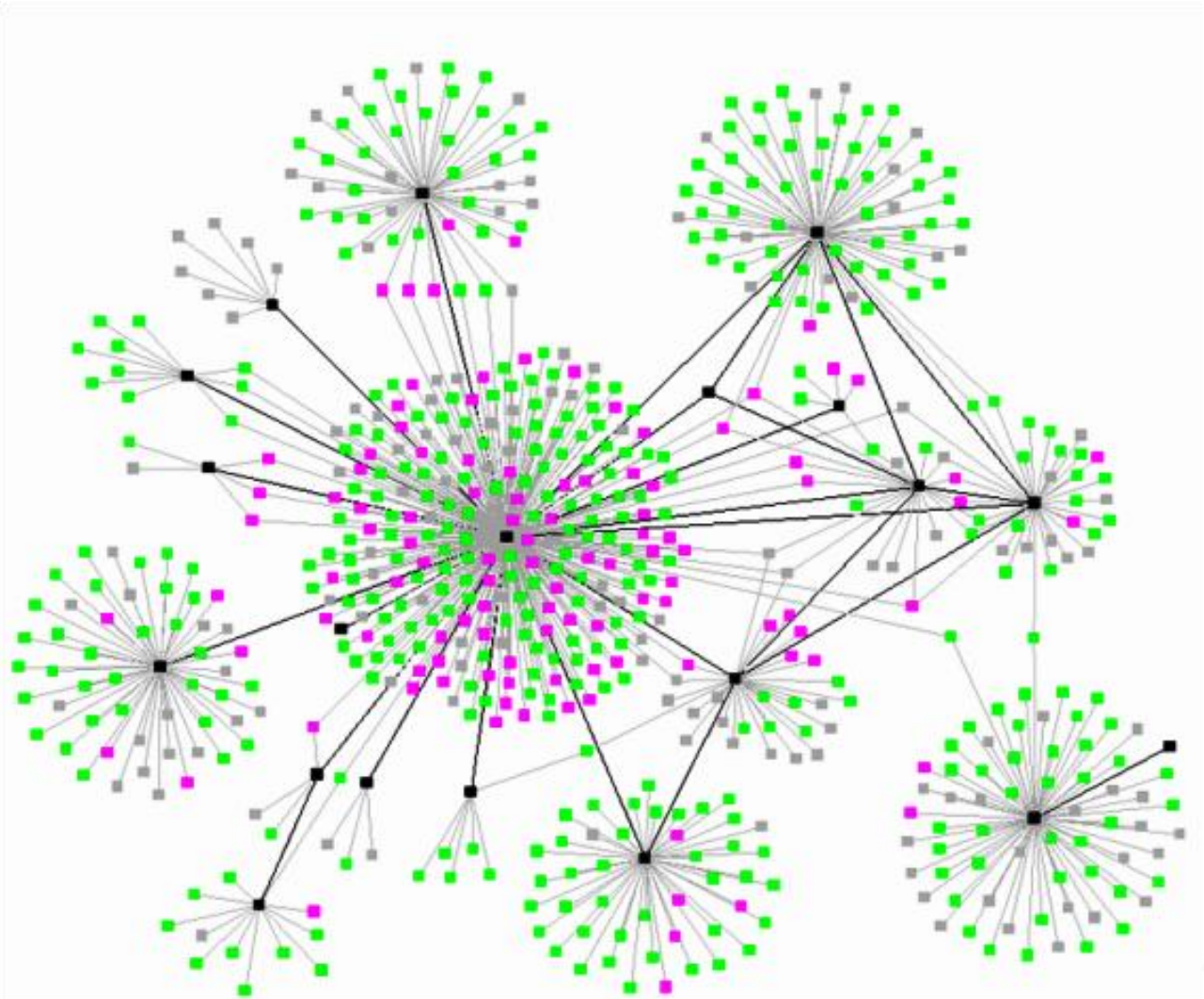


Photo credit:
prblog.typepad.com

Facebook network of just one average user

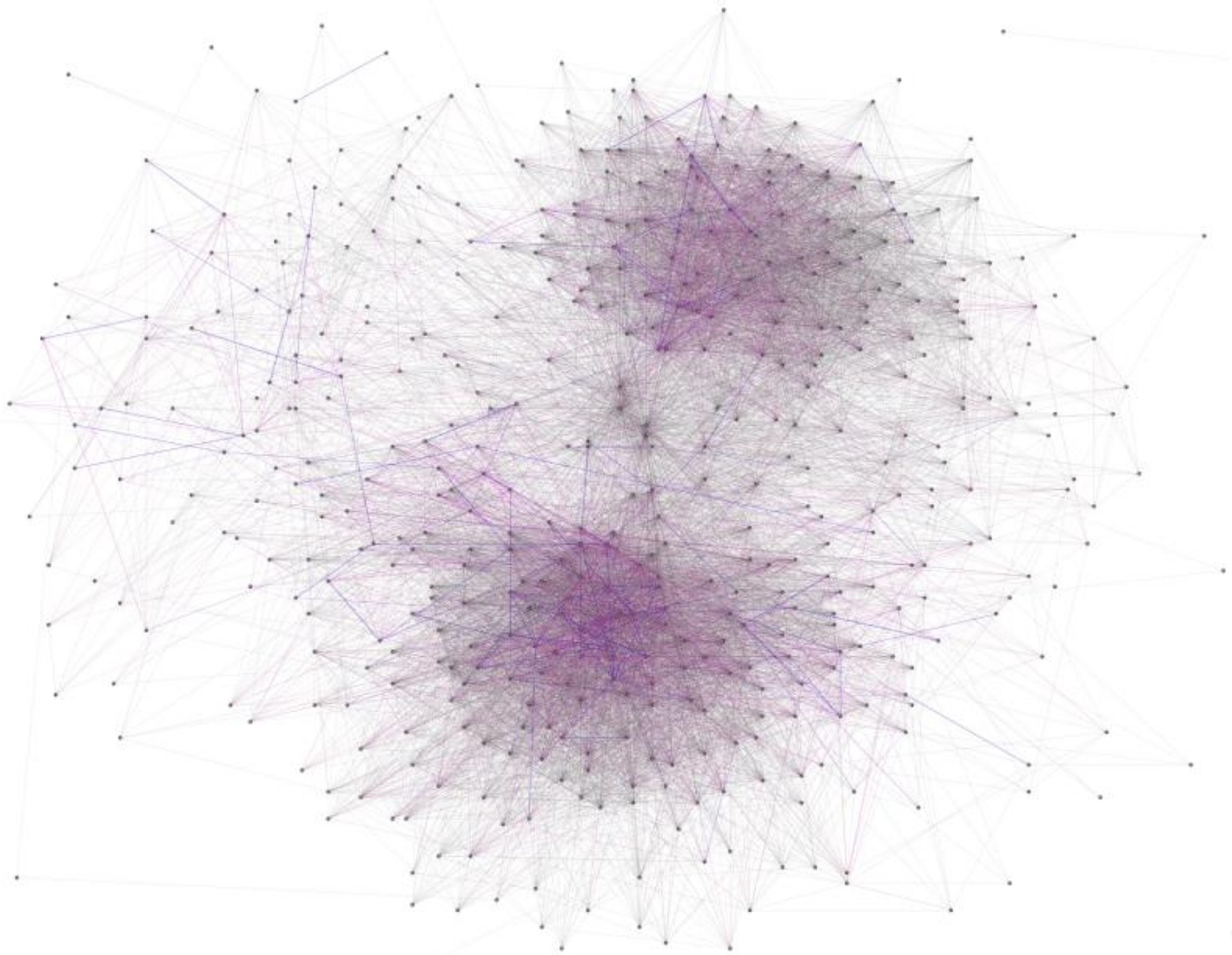


Photo credit: afrolegs.com

Very hard to determine key influencers and communities

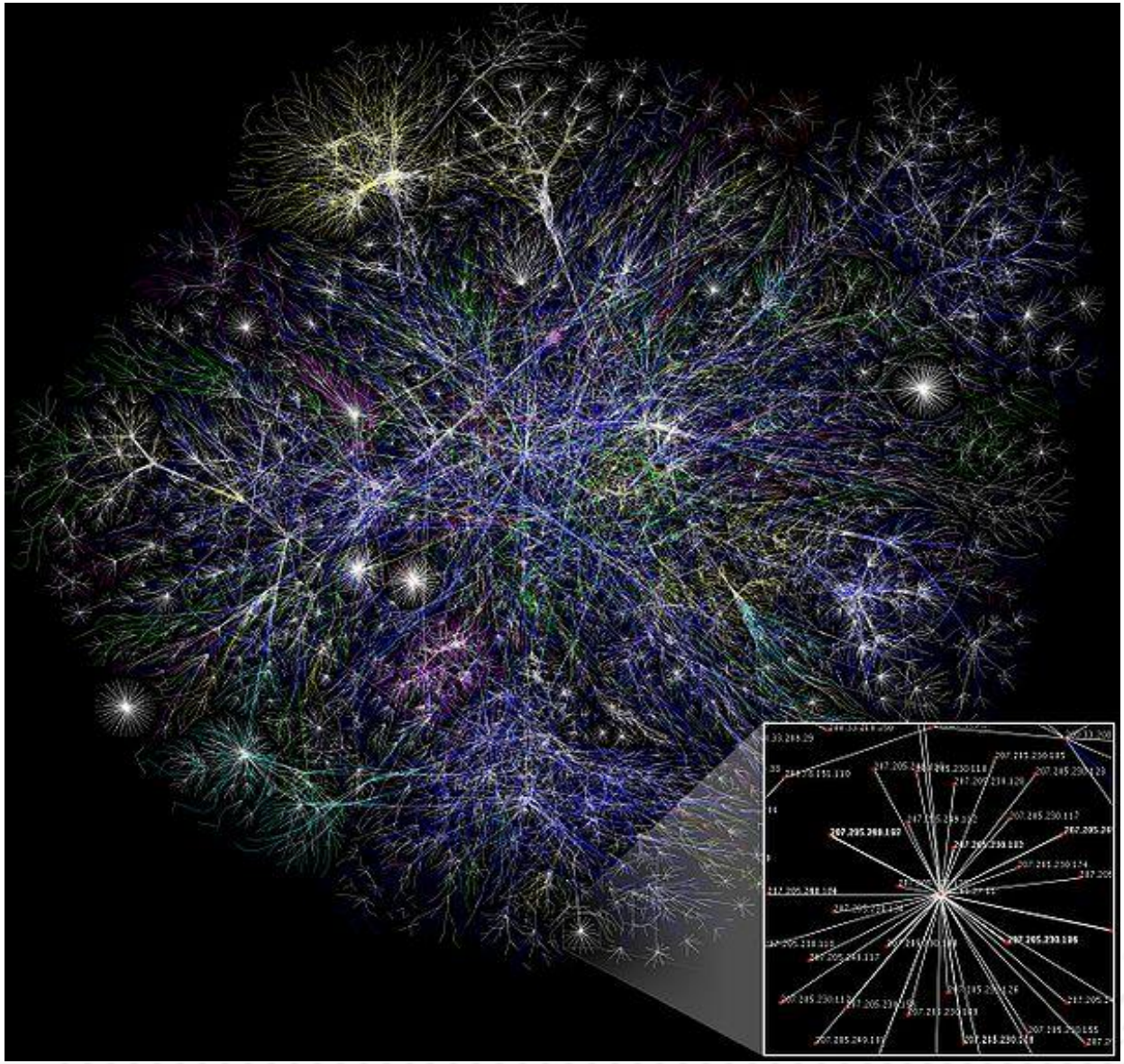


Photo credit:
en.wikipedia.org

Project Overview (continued)

- Main parts of the project:

1. Topology score determination
    - Determine the topology scores for vertices
    - Determine top ranking (more influencing) vertices

2. Clustering
    - Cluster the network using calculated topology scores
    - Determine the extent to which the clusters grow

Topology Score Determination

1. Measures in social network analysis:

- *Betweenness:*

The importance of a vertex in retaining connectivity among distant vertices of the network.

$$b(v) = \sum_{s \in V, t \in V, s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- *Closeness:*

The efficiency of each vertex (individual) in spreading information to all the other vertices which are reachable in the network.

$$cl(s) = \frac{\sum_{t \in V \setminus s} d_{st}}{|V| - 1}$$

- *Degree:*

The number of connections a vertex has with its neighbors.

$$d(s) = |N_s|$$

- *Clustering Coefficient:*

Indicates how close the neighbors of a vertex are in forming a clique.

$$cc(s) = \frac{2 \cdot E_s}{|N_s|(|N_s| - 1)}$$

- *Eccentricity:*

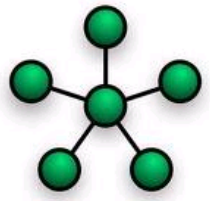
The maximum shortest path length that is possible from a vertex to all its reachable vertices in the network.

$$e(s) = \max_{t \in V} d_{st}$$

Topology Score Determination (continued)

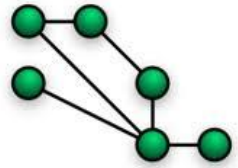
2. Finding top centrality vertices of a topology:

- *Star Topology:*
  - Has the most influencing vertex at the center of the star network.
  - We assign less weight to vertices which connect members who are directly connected among themselves.



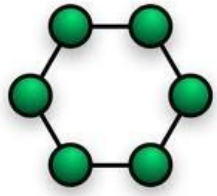
Star

- *Mesh Topology:*
  - Nearly all edges would be present among all the members of lattice or mesh.



Mesh

- *Ring Topology:*
  - A failure in one of the core ring network vertices may result in disruption of the information flow.



Ring

Photo credit:
en.wikipedia.org

Clustering

Proposed methodology is composed of two main parts:

1. Edge weight determination:

   The social network in consideration is converted into an undirected weighted graph, where weights of the links between vertices are calculated with respect to the degree of the interaction between those vertices.

2. Clustering Algorithm:

   It is applied on the undirected weighted graph of the network to find the clusters.

Edge weight determination for DBLP data

We have assumed that authors can be linked to each other by either co-authoring a paper or by giving reference to a paper of the other one

a_i , a_j : two authors

w_{ij} = the weight of the link between a_i and a_j

d_{ij} = the distance of the link between a_i and a_j

w_{ij}^{C} = the weight of the link between a_i and a_j due to coauthorship

w_{ij}^{R} = the weight of the link between a_i and a_j due to references

$$\delta_i^k = \begin{cases} 1 \text{ if } a_i \text{ is an author of paper } k \\ 0 \text{ otherwise} \end{cases}$$

n_k = number of authors of paper k

$$r_{ij}^k = \begin{cases} 1 \text{ if } a_i \text{ gives reference to } a_j \text{ in paper } k \\ 0 \text{ otherwise} \end{cases}$$

R_i^k = number of references a_i uses in paper k

Edge weight determination for DBLP data (continued)

Co-authorship weight ----------> $w_{ij}^{C} = \sum_k \frac{\delta_i^k \cdot \delta_j^k}{n_k - 1}$

Citation weight ----------> $w_{ij}^{R} = \begin{cases} 0 & \text{if} \quad \sum_k r_{ij}^k = 0 \wedge \sum_l r_{ji}^l = 0 \\ \frac{\left[\left(\sum_k r_{ij}^k\right)+1\right]\left[\left(\sum_l r_{ji}^l\right)+1\right]}{\left(\sum_k R_i^k\right)\left(\sum_l R_j^l\right)} & \text{otherwise} \end{cases}$

Link weight ----------> $w_{ij} = w_{ij}^{C} + w_{ij}^{R}$

Link distance ----------> $d_{ij} = \frac{1}{w_{ij}}$

Clustering Algorithm

Parameters:

D_c: maximum allowed distance of the furthermost neighbor to c

$ne(c)$: neighborhood of node c

d_{cv}: shortest distance between nodes c and v

R_θ: minimum required number of overlapping elements to be in the same cluster

$$R_\theta = \theta \times min(|ne(c)|, |ne(v)|) \text{ where } \theta \in [0,1]$$

$t(v)$: topology score of node v

T: threshold topology score

Experimentation

Parameter Selection:

- D_c: *selected as the average distance of node c to all other nodes*
- $\theta \in [\theta^0 - 0.3, \theta^0 + 0.3]$ where θ^0 is the average clustering coefficient
- T: selected as the topology score at the 10th percentile

Parameter Properties

- Theta (θ) ↑
- Cluster # ↓
- Uncovered Node # ↓
- Overlapping Node # ↓
- Multi-Overlapping Node# ↑ (*ratio of nodes overlapped more than twice to # overlapping nodes*)

Experimentation (continued)

Sampling 1:

- 94 nodes, 167 edges
- 5 connected components

Star Topology Results:

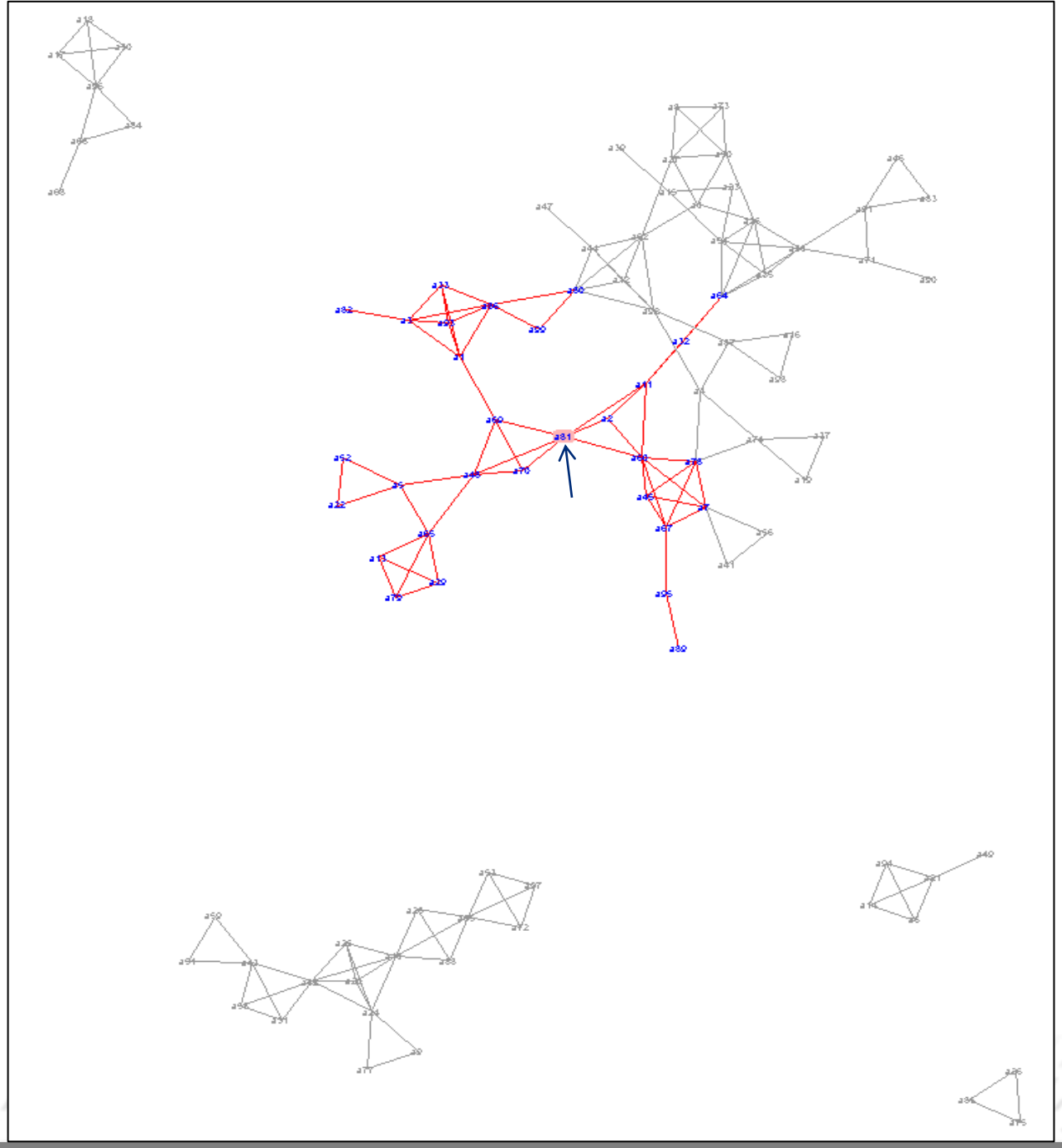
Theta	# of Clusters	# of Left Node	# of Overlap Node	# of Total Overlap	Cluster Sizes
0.803	6	10	19	19	(6,30,4,11,2,50)

Mesh Topology Results:

Theta	# of Clusters	# of Left Node	# of Overlap Node	# of Total Overlap	Cluster Sizes
0.653	9	7	16	16	(17,16,10,19,4, 19,4,11,4)

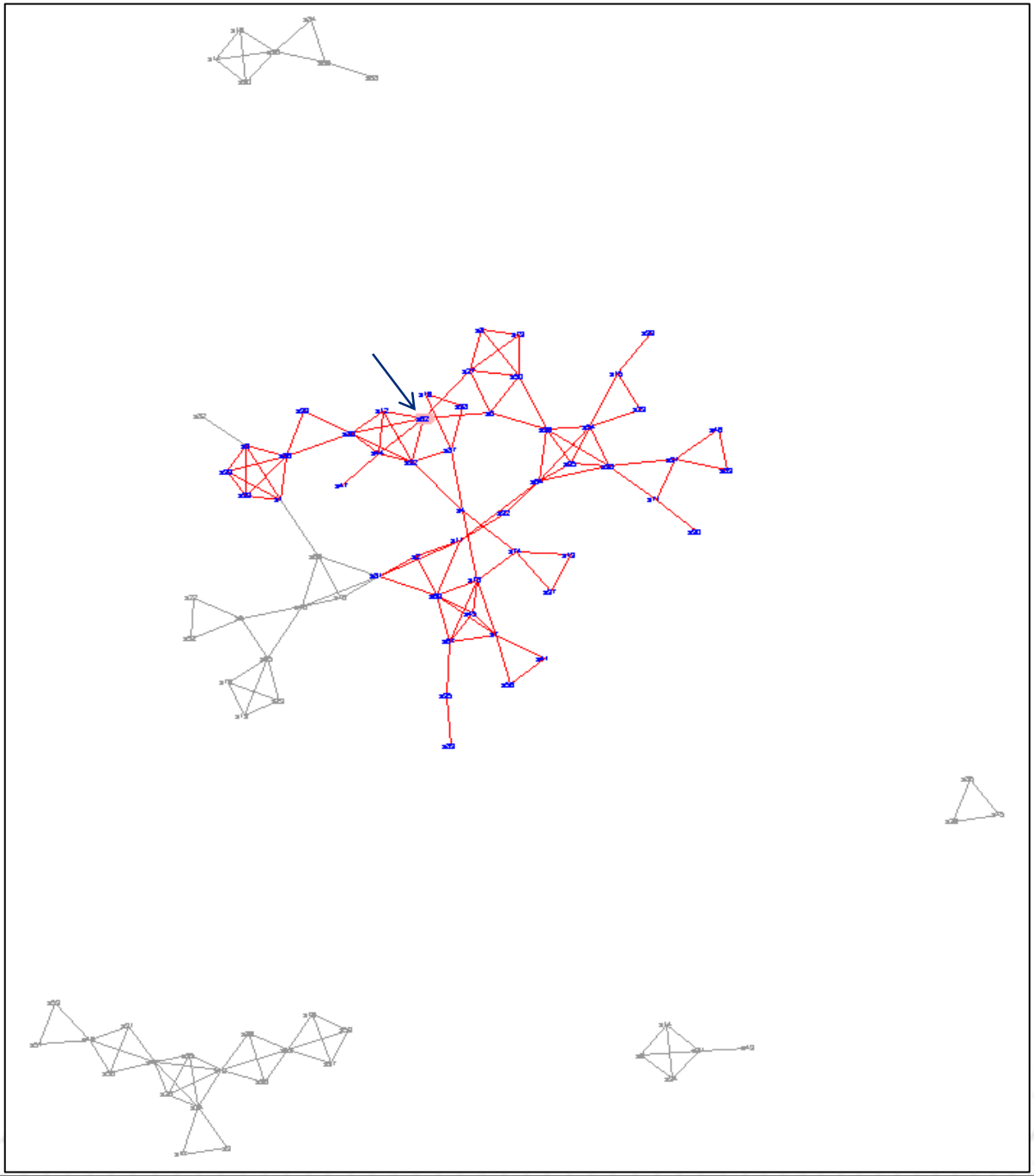
Star topology:

Cluster 1



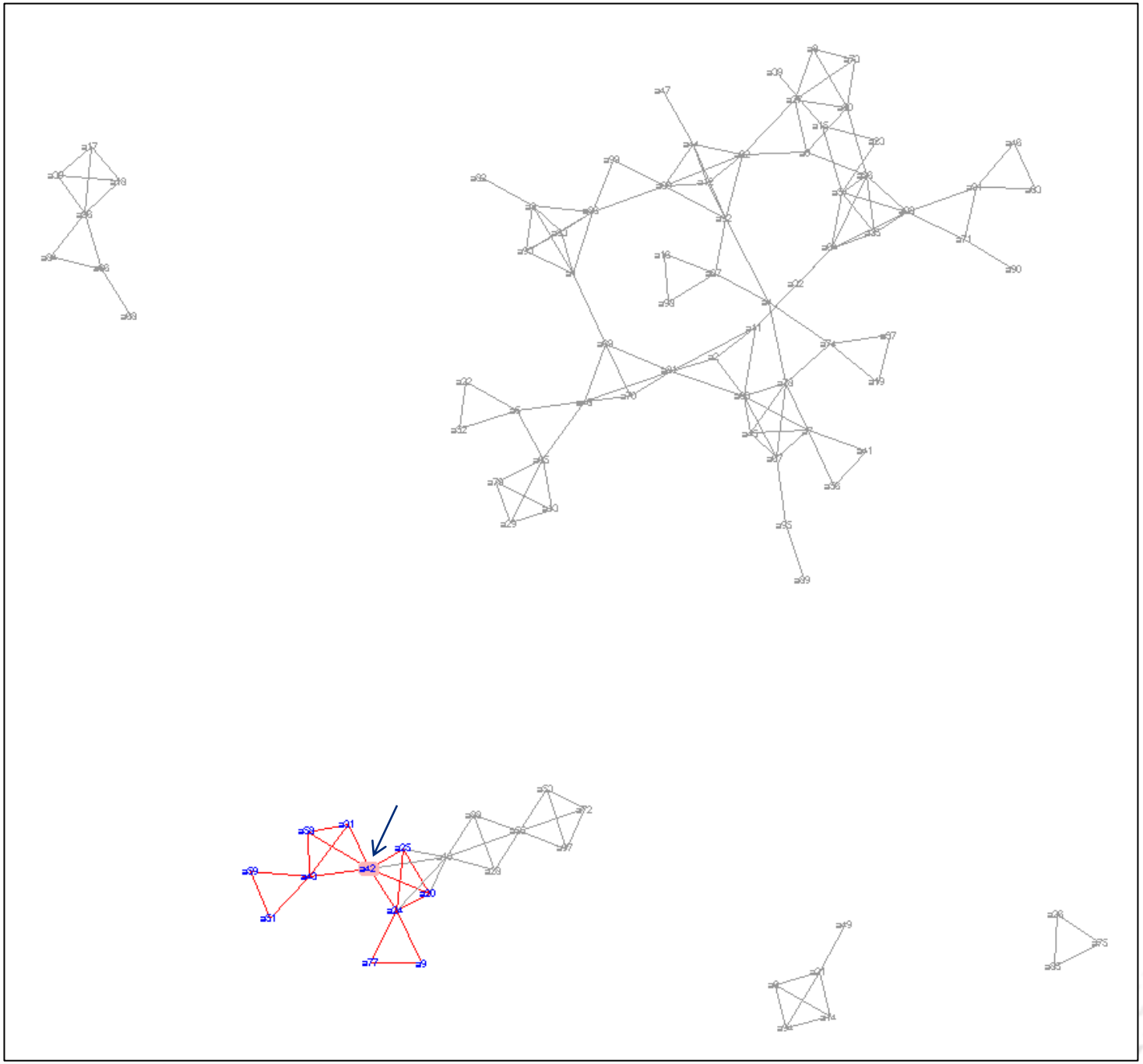
Star topology:

Cluster 2



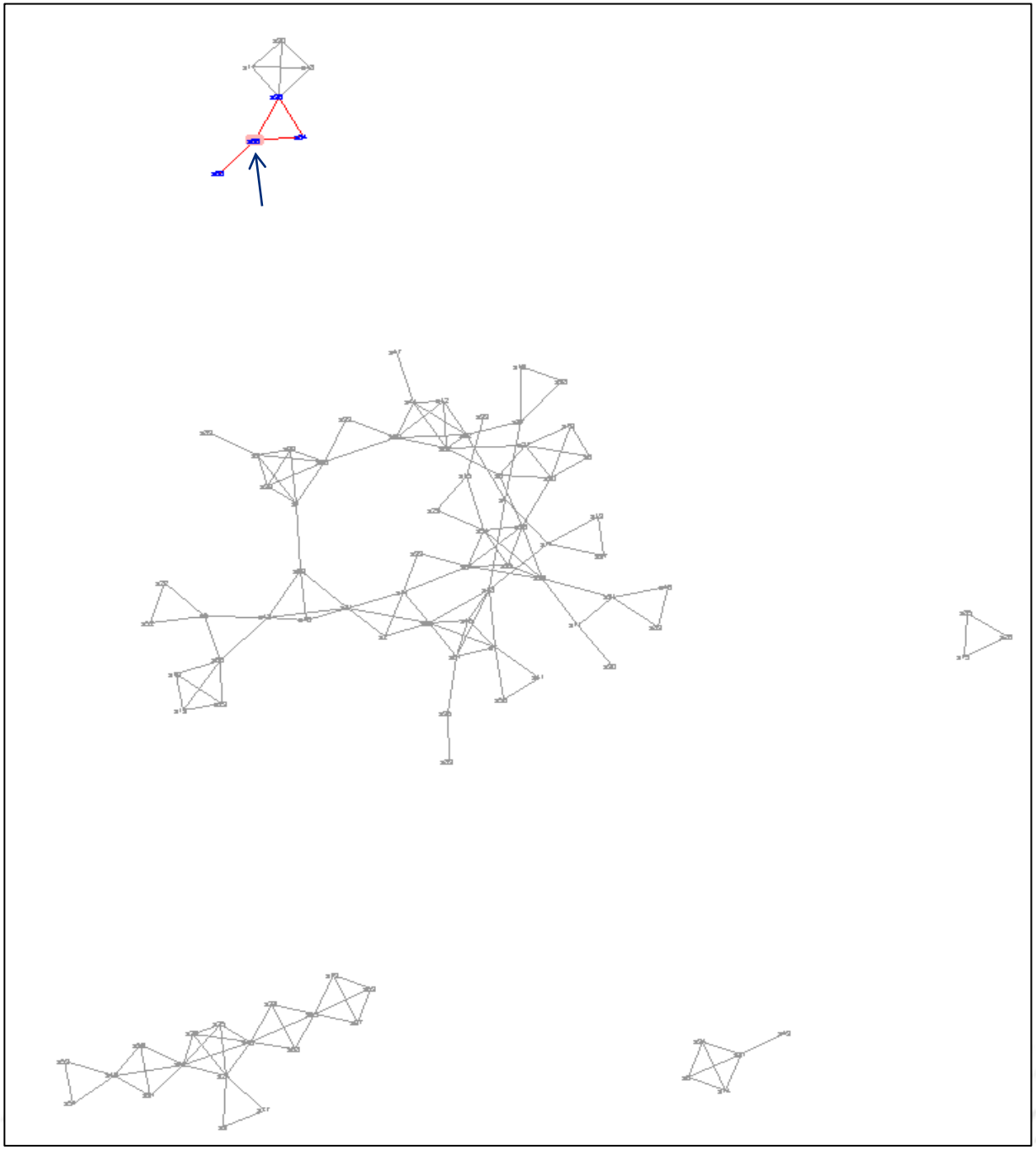
Star topology:

Cluster 3



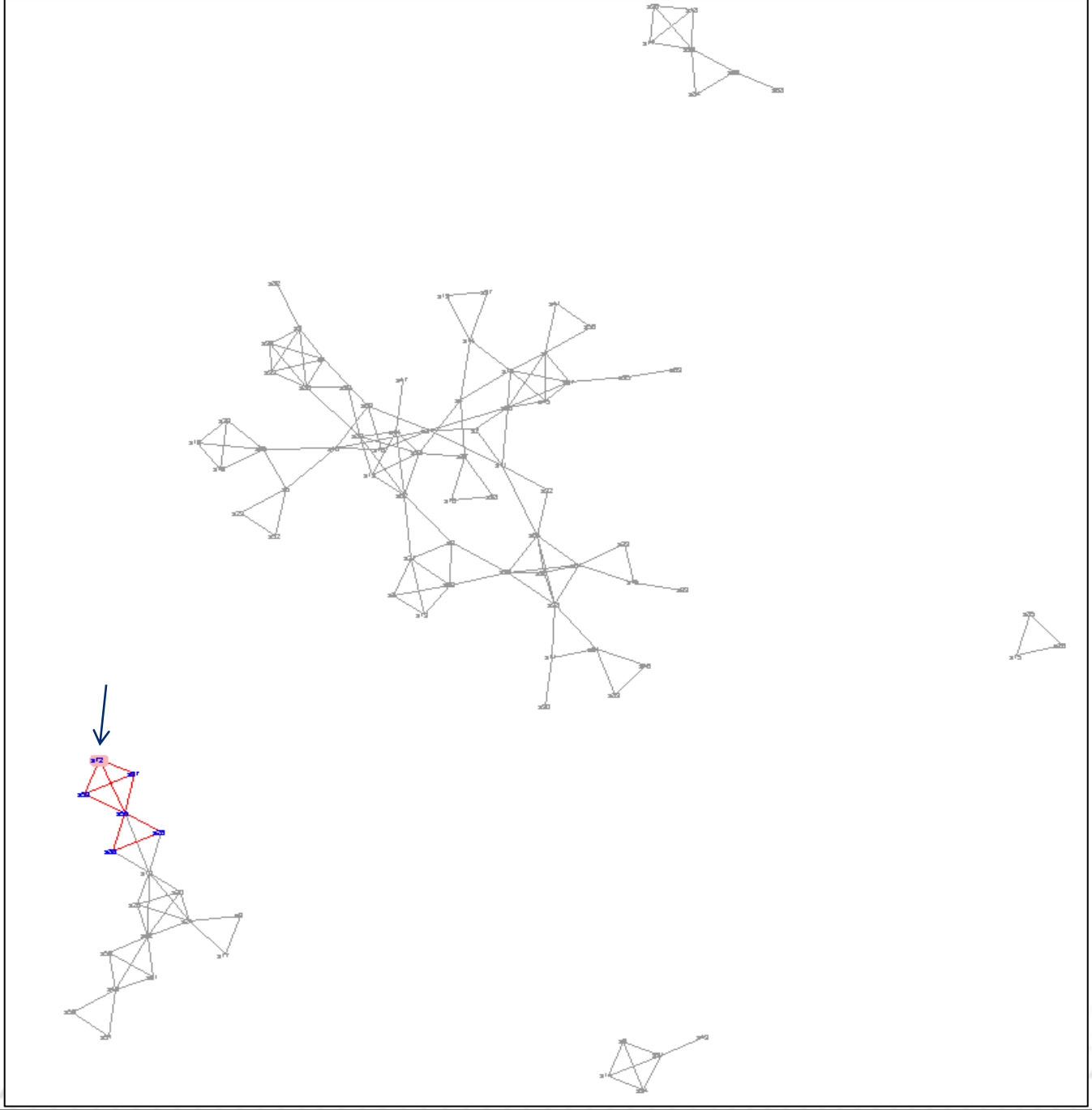
Star topology:

Cluster 4



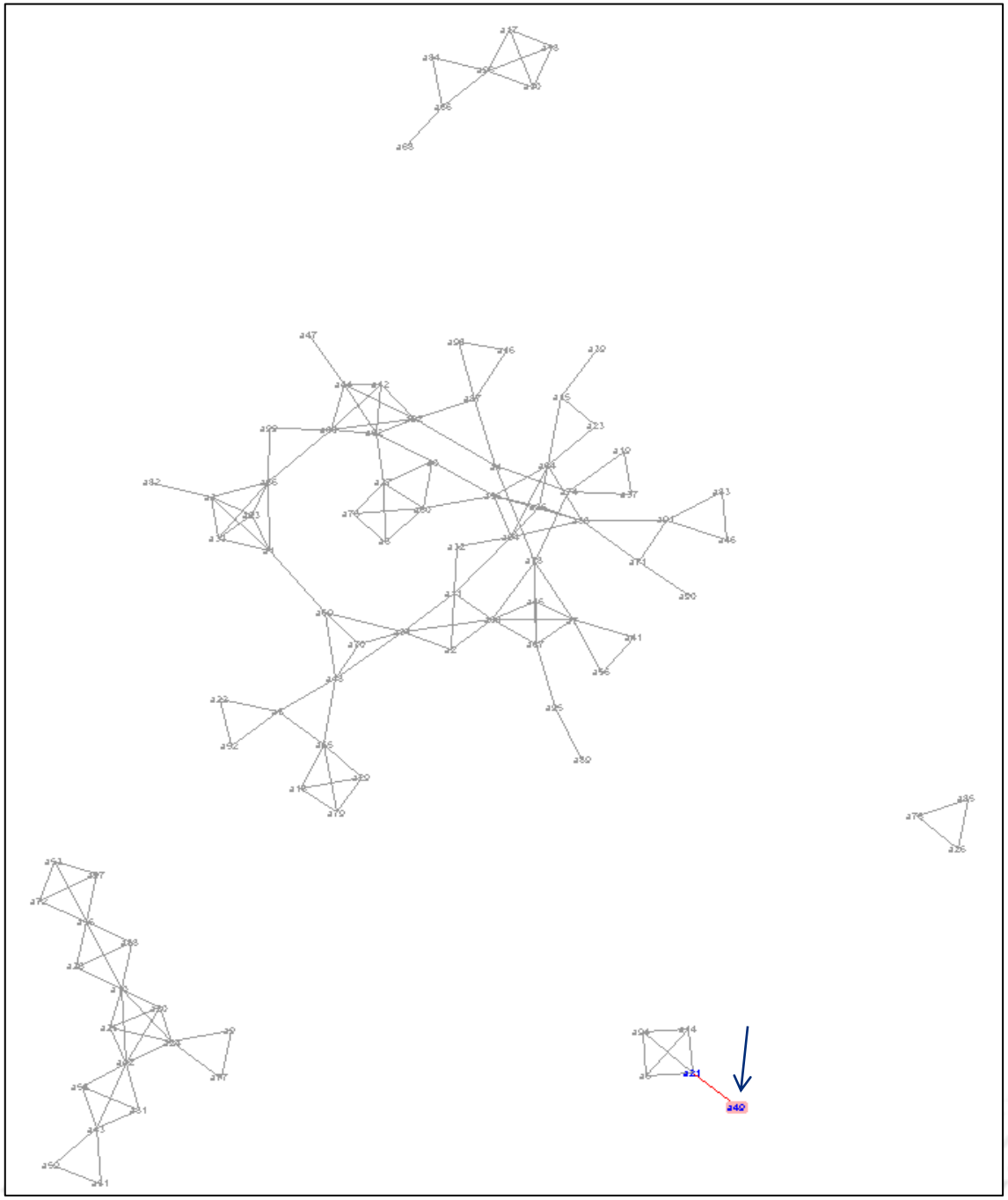
Star topology:

Cluster 5



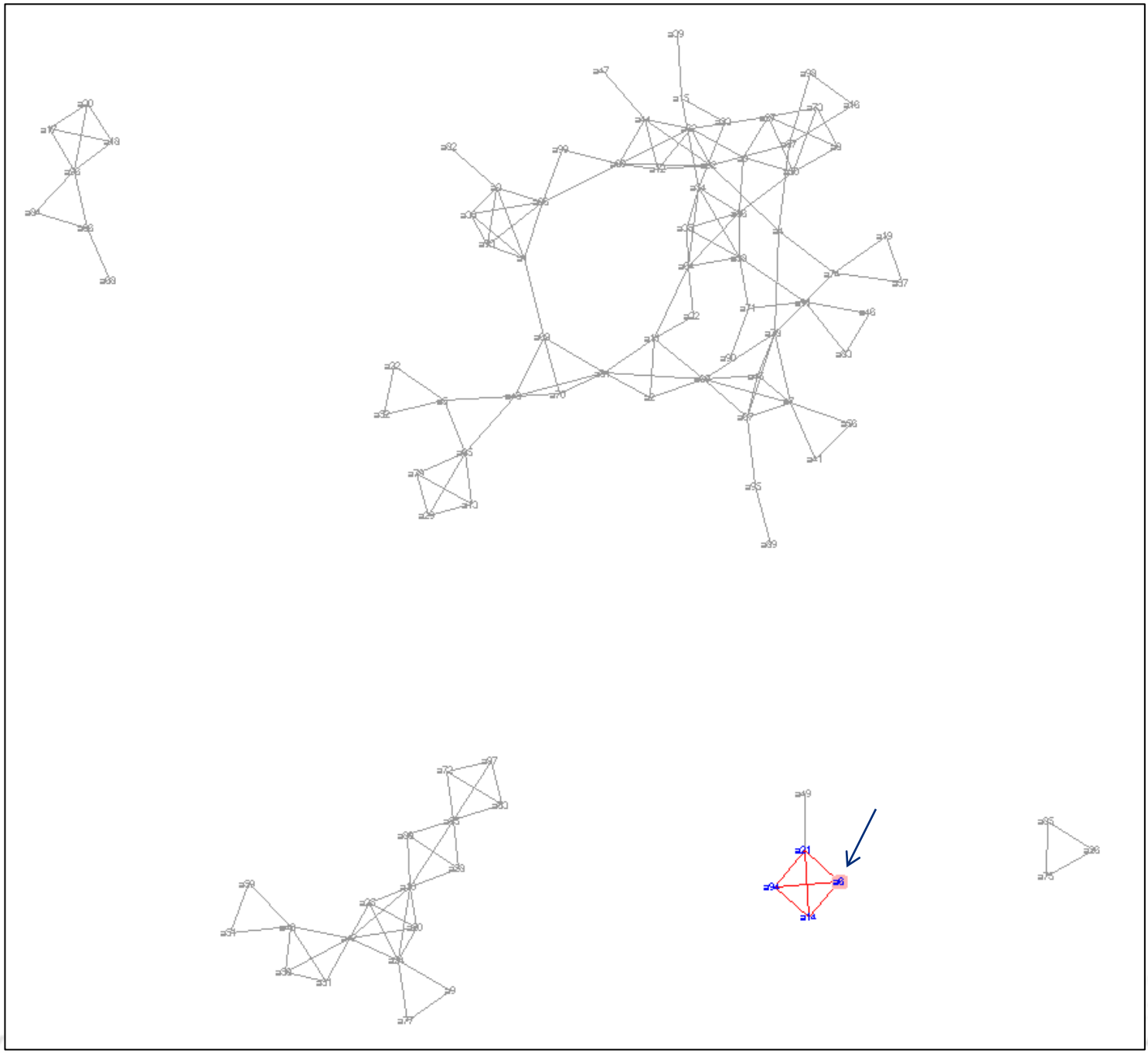
Star topology:

Cluster 6



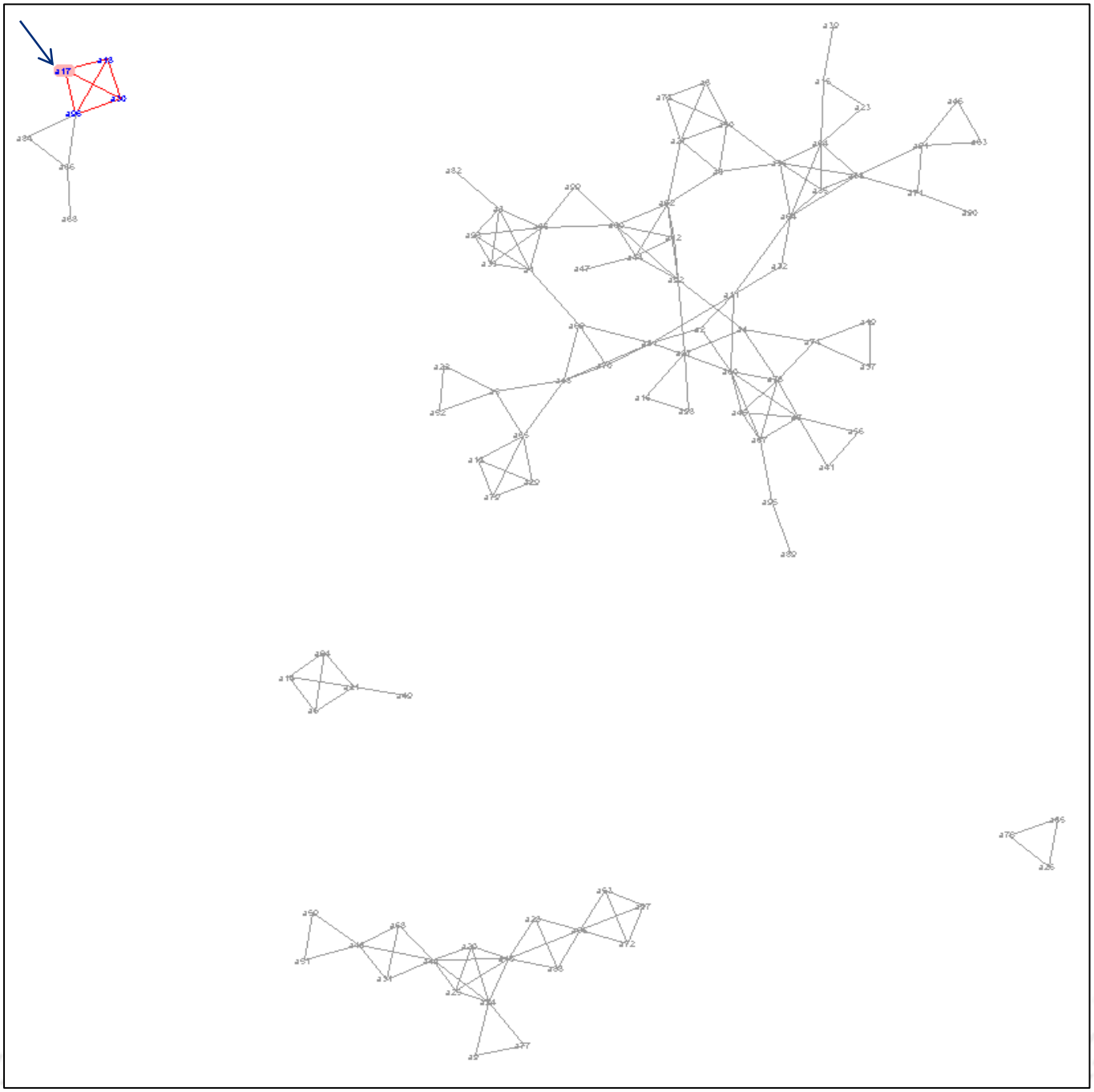
Mesh topology:

Cluster 1



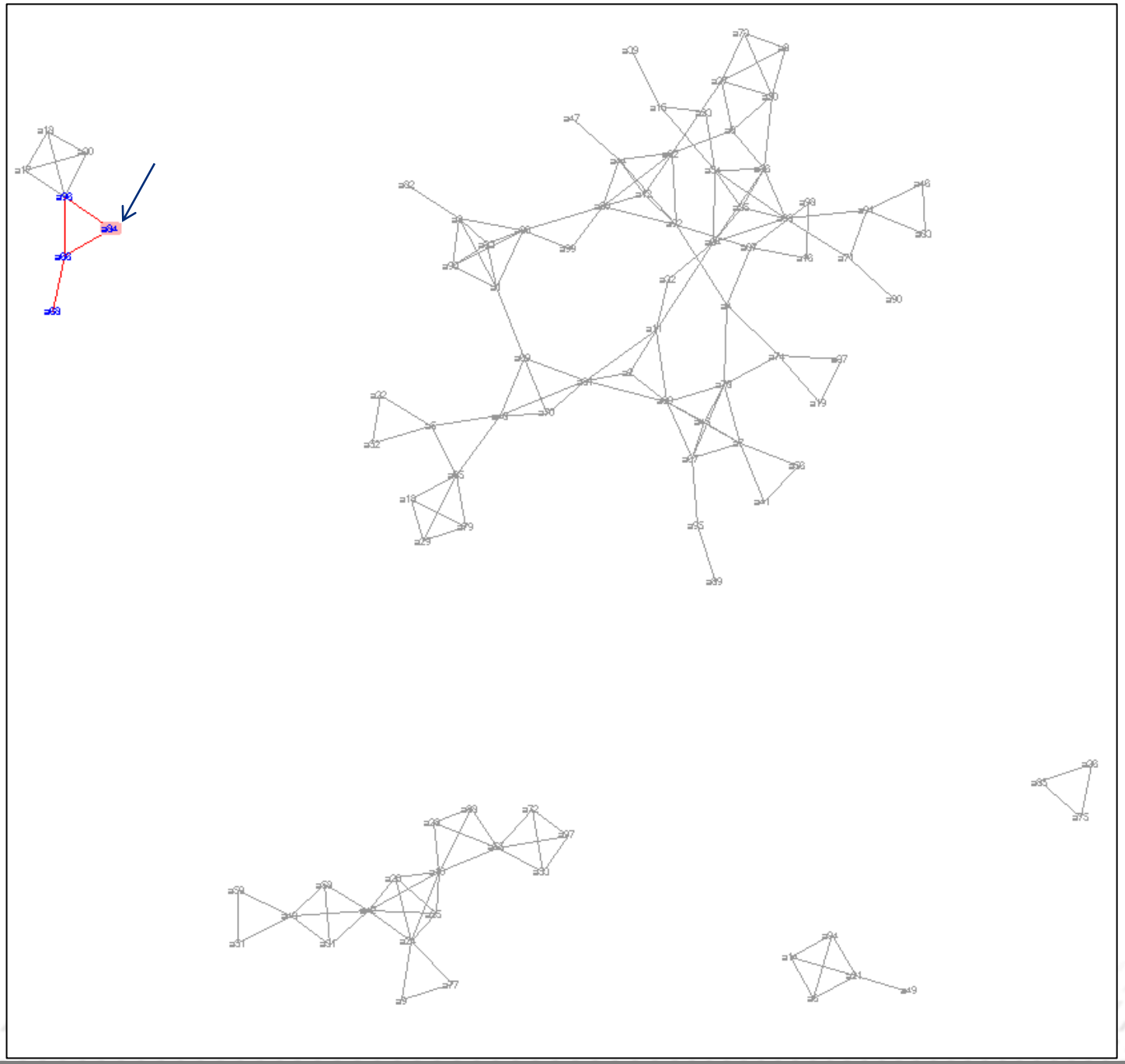
Mesh topology:

Cluster 2



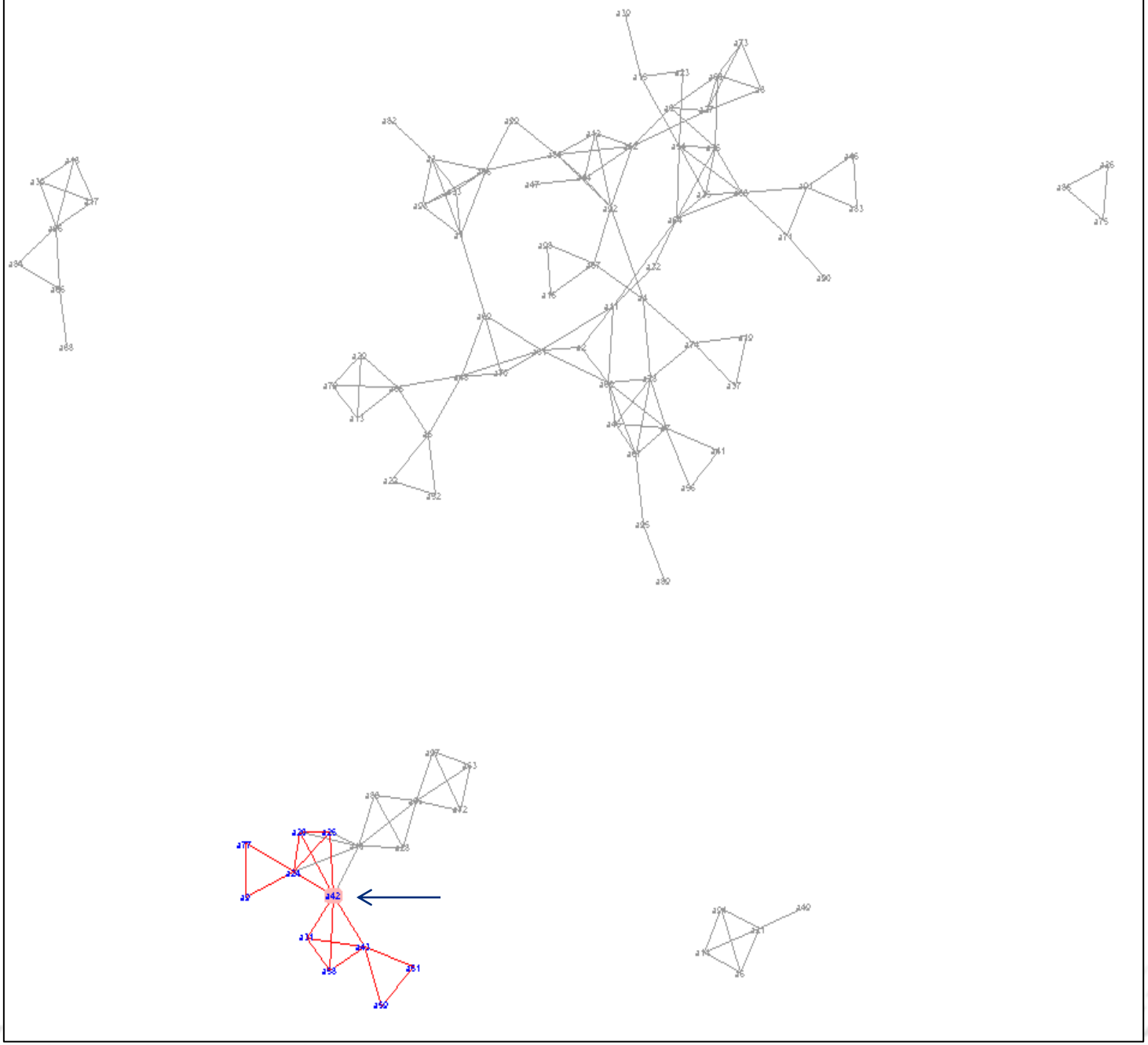
Mesh topology:

Cluster 3



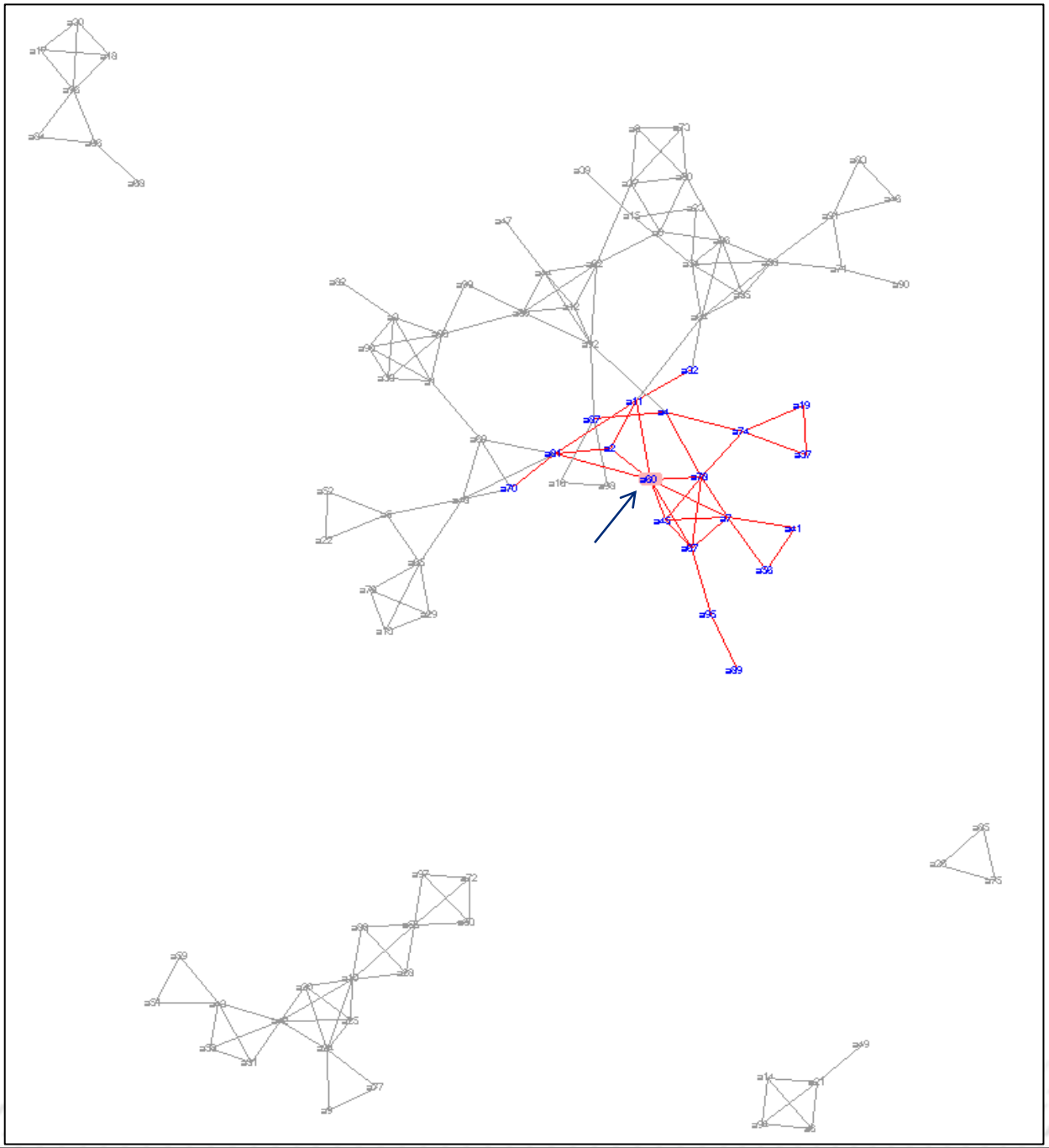
Mesh topology:

Cluster 4



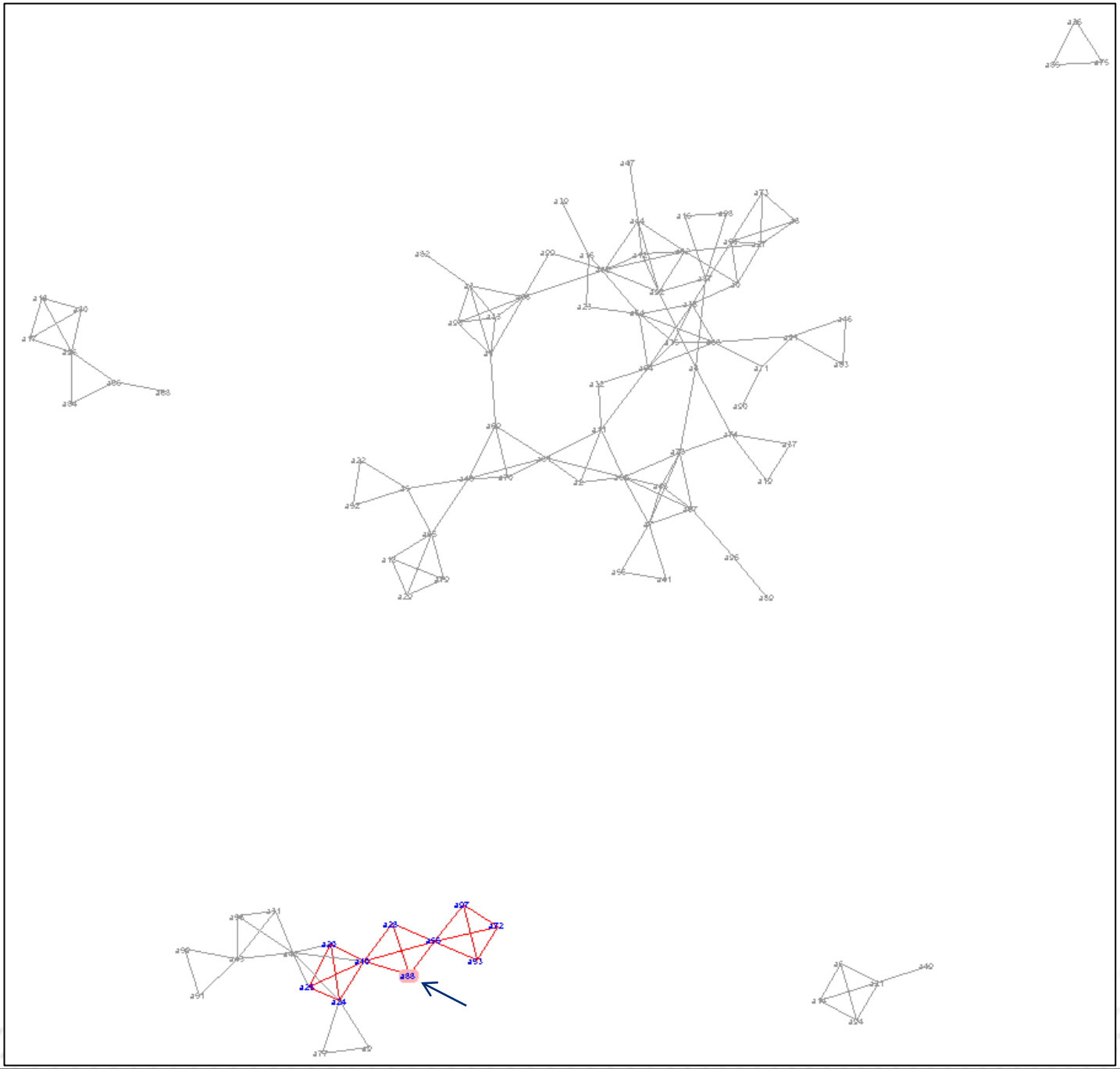
Mesh topology:

Cluster 5



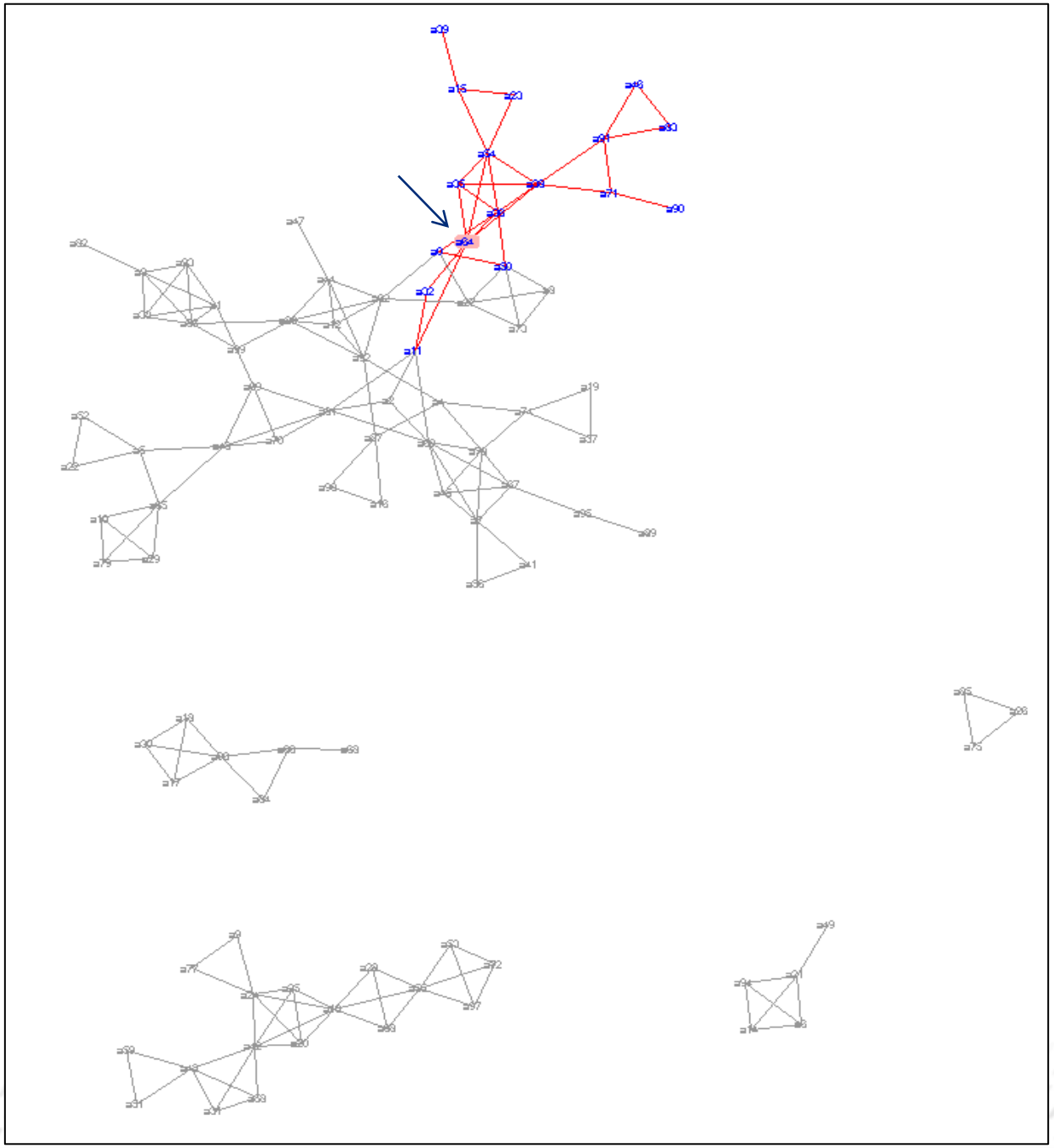
Mesh topology:

Cluster 6



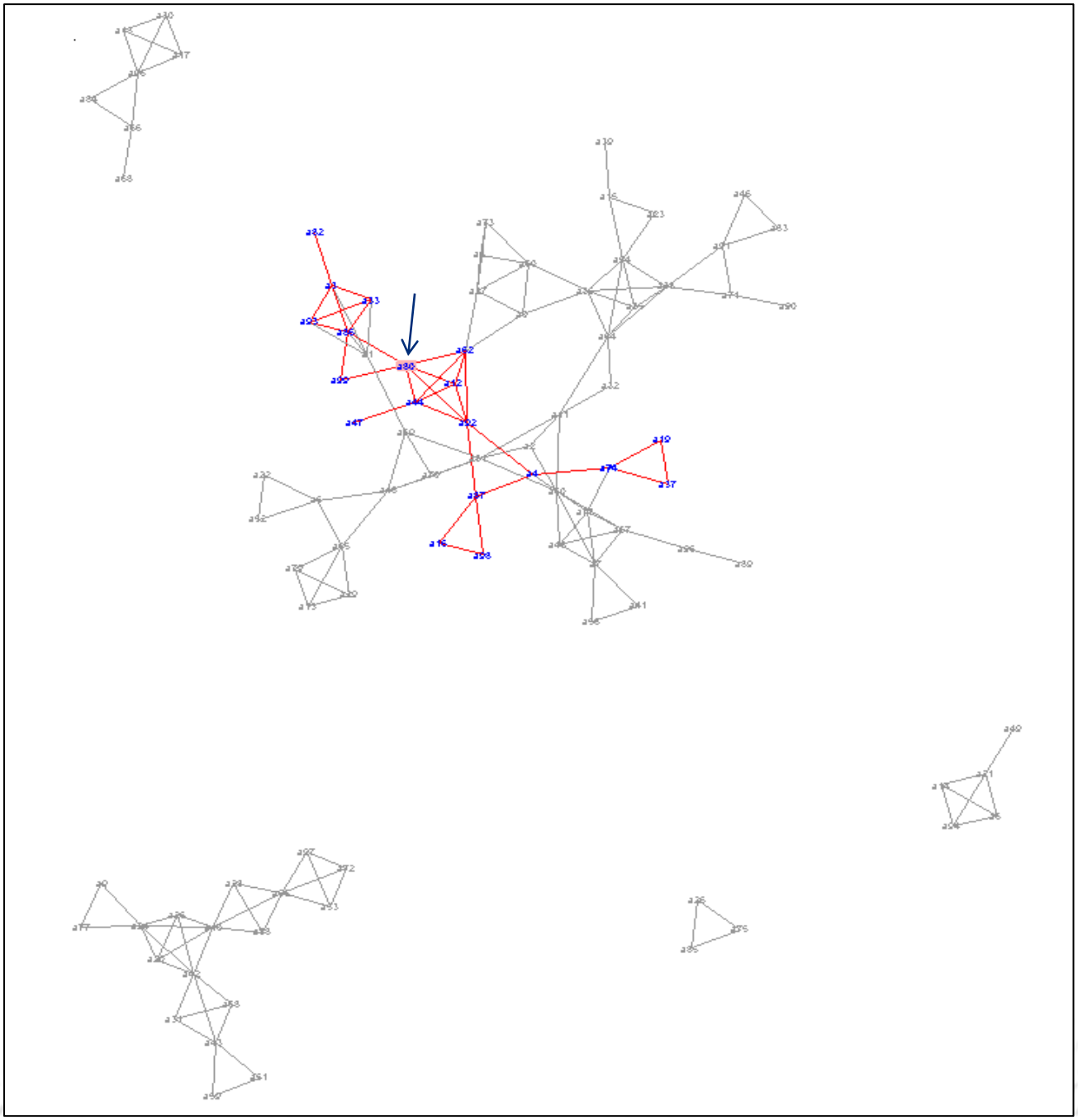
Mesh topology:

Cluster 7



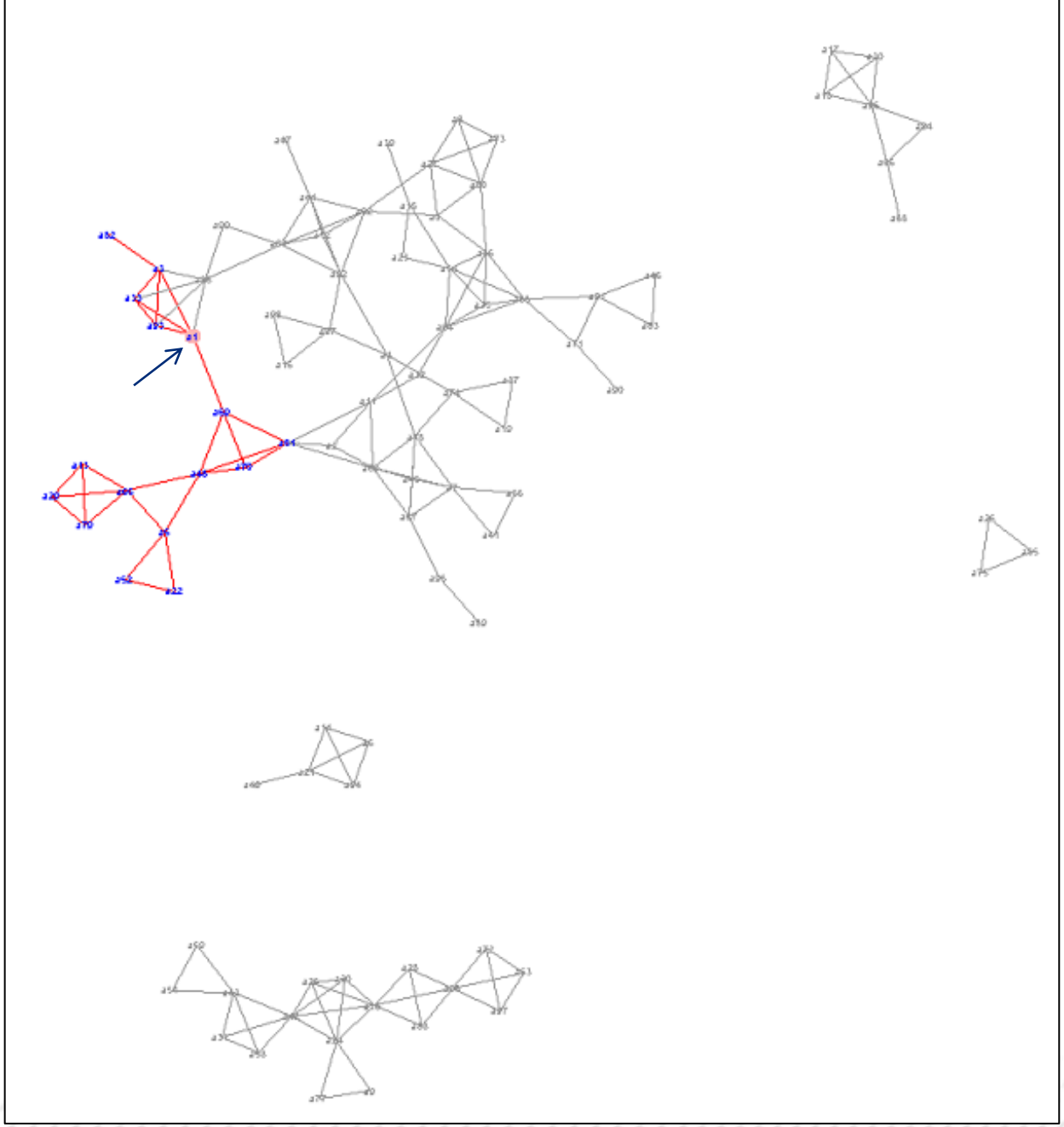
Mesh topology:

Cluster 8



Mesh topology:

Cluster 9



Experimentation (continued)

Sampling 2:

- 100 nodes, 232 edges
- 6 connected components

Star Topology Results:

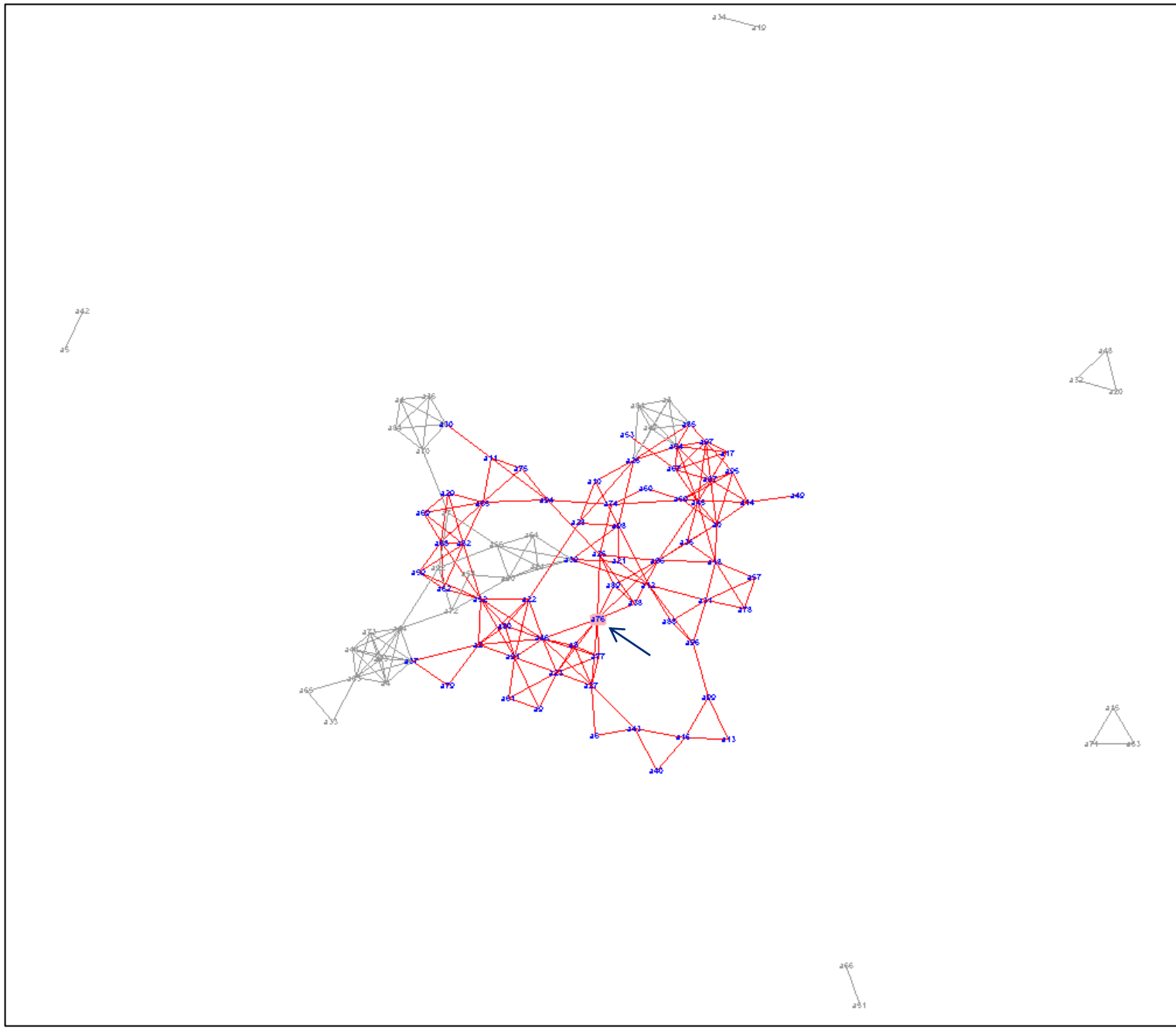
Theta	# of Clusters	# of Left Node	# of Overlap Node	# of Total Overlap	Cluster Sizes
0.639	2	15	28	28	(48,65)

Mesh Topology Results:

Theta	# of Clusters	# of Left Node	# of Overlap Node	# of Total Overlap	Cluster Sizes
0.789	3	16	20	20	(45,15,44)

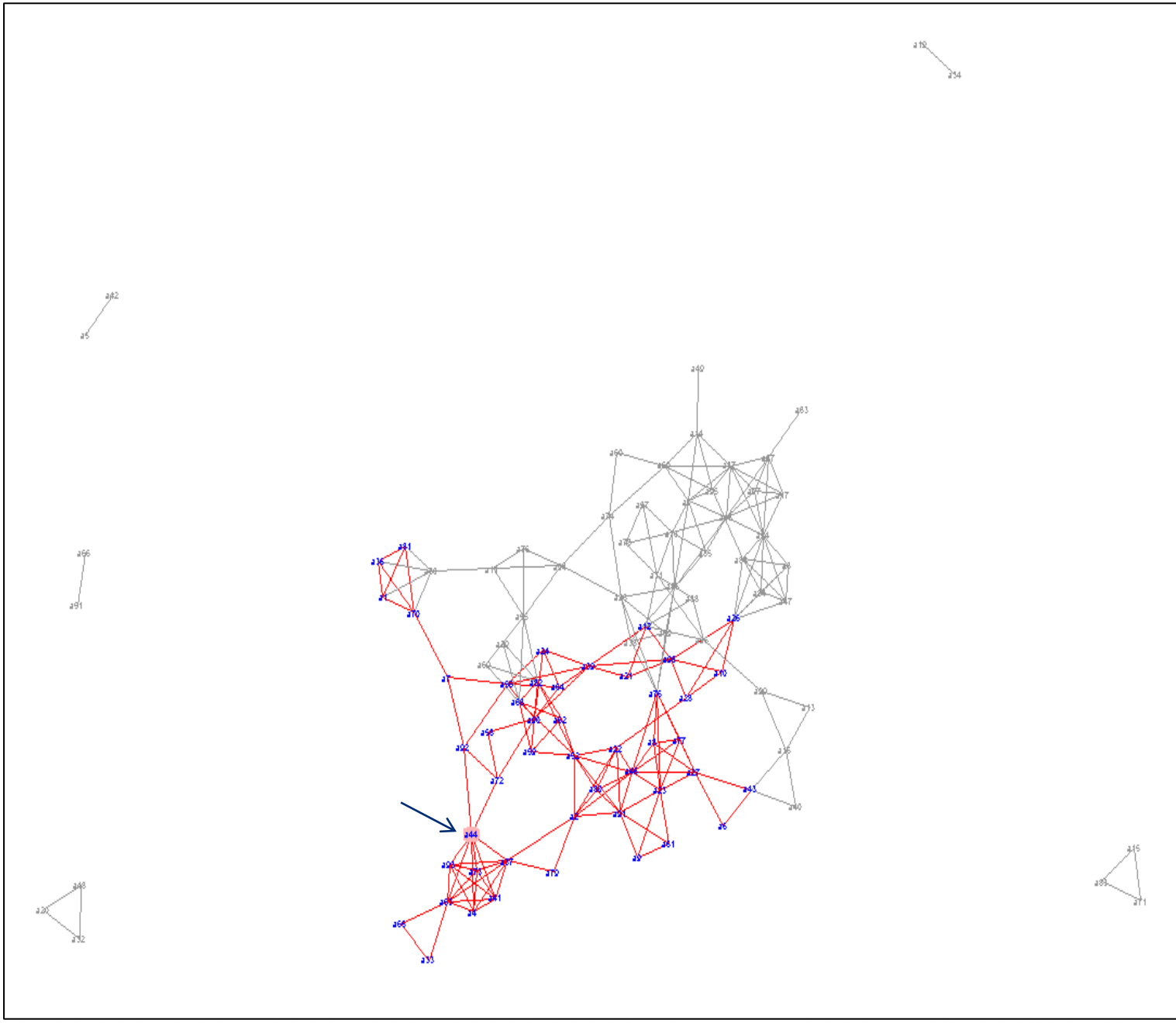
Star topology:

Cluster 1



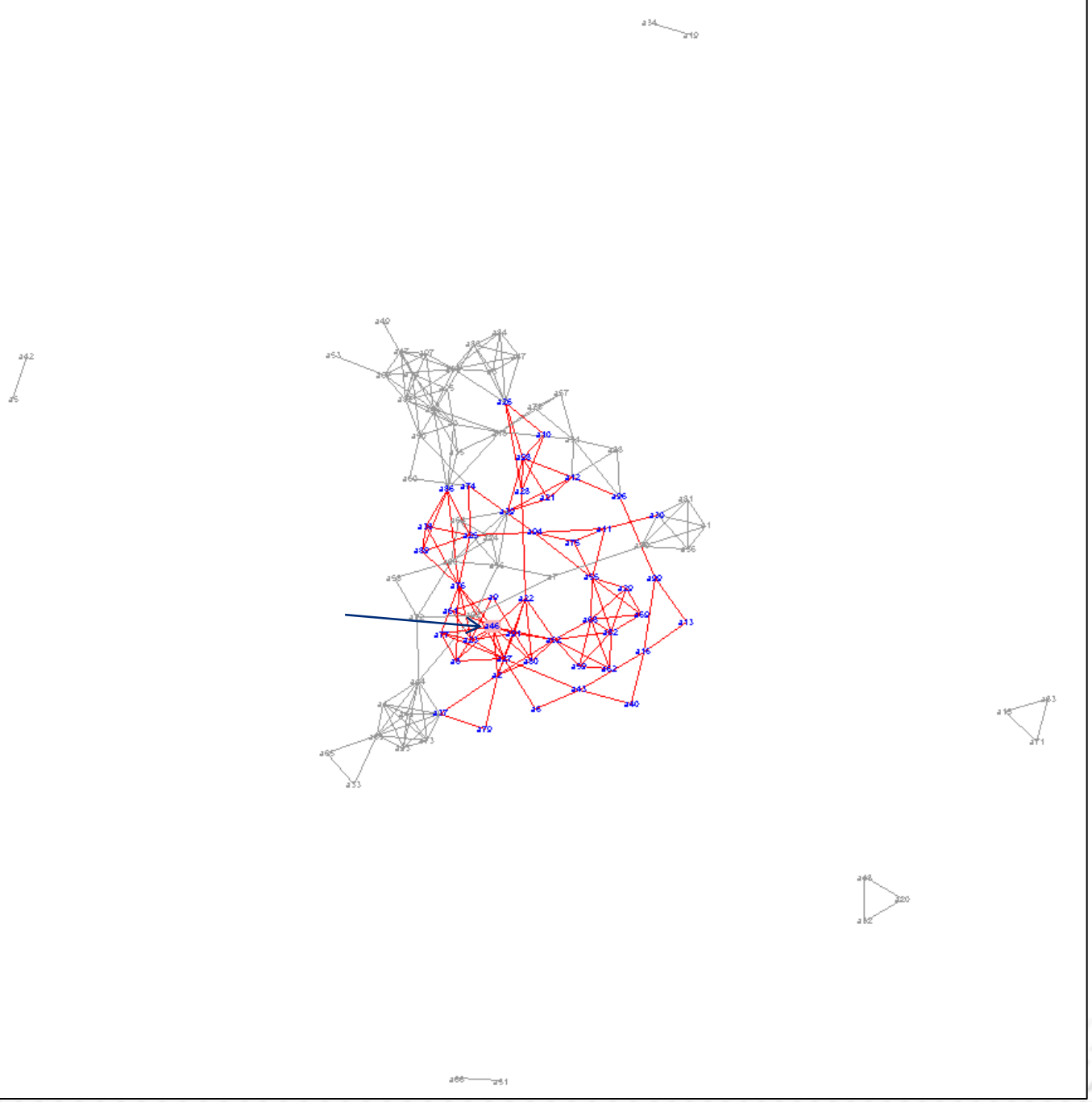
Star topology:

Cluster 2



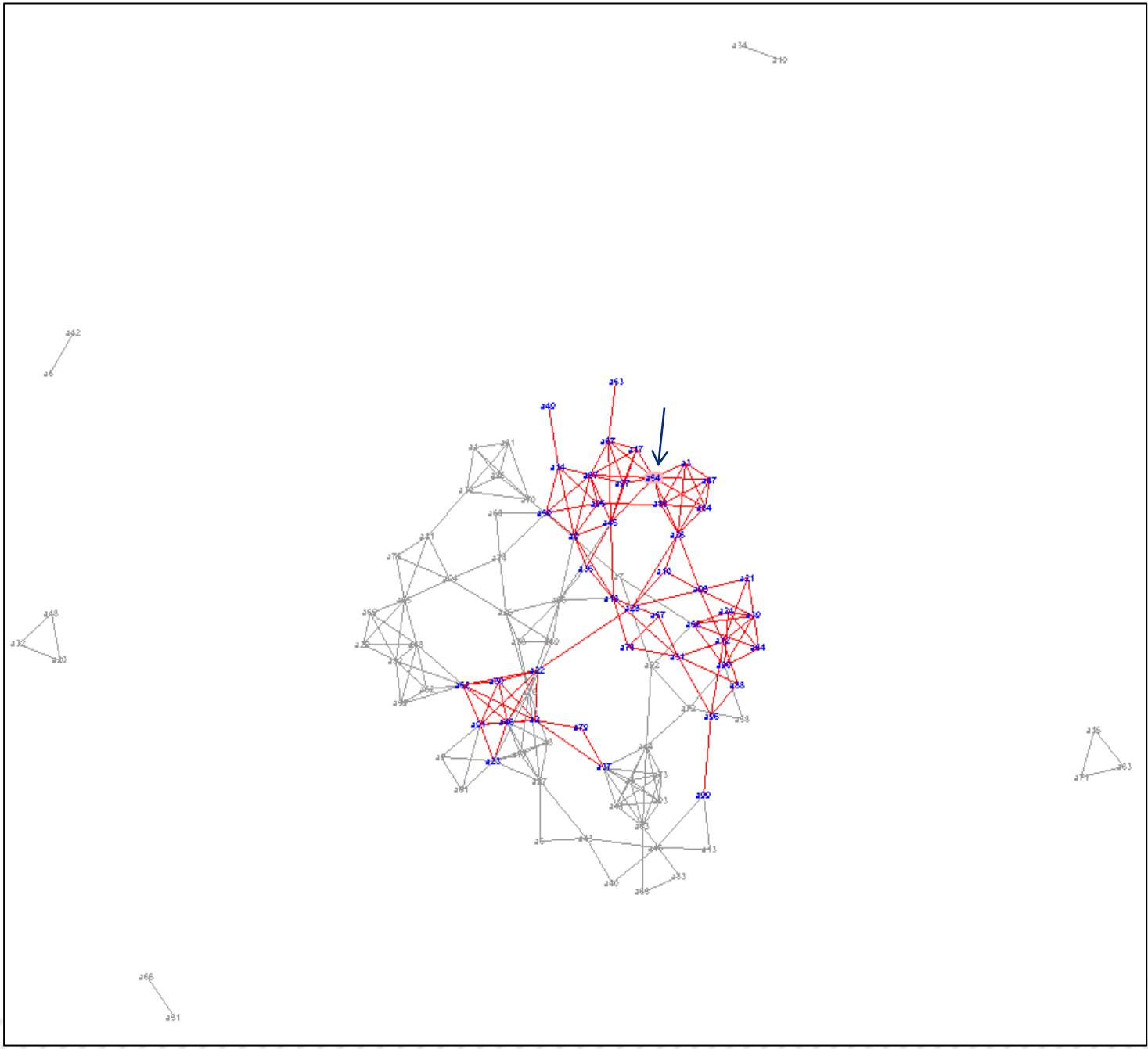
Mesh topology:

Cluster 1



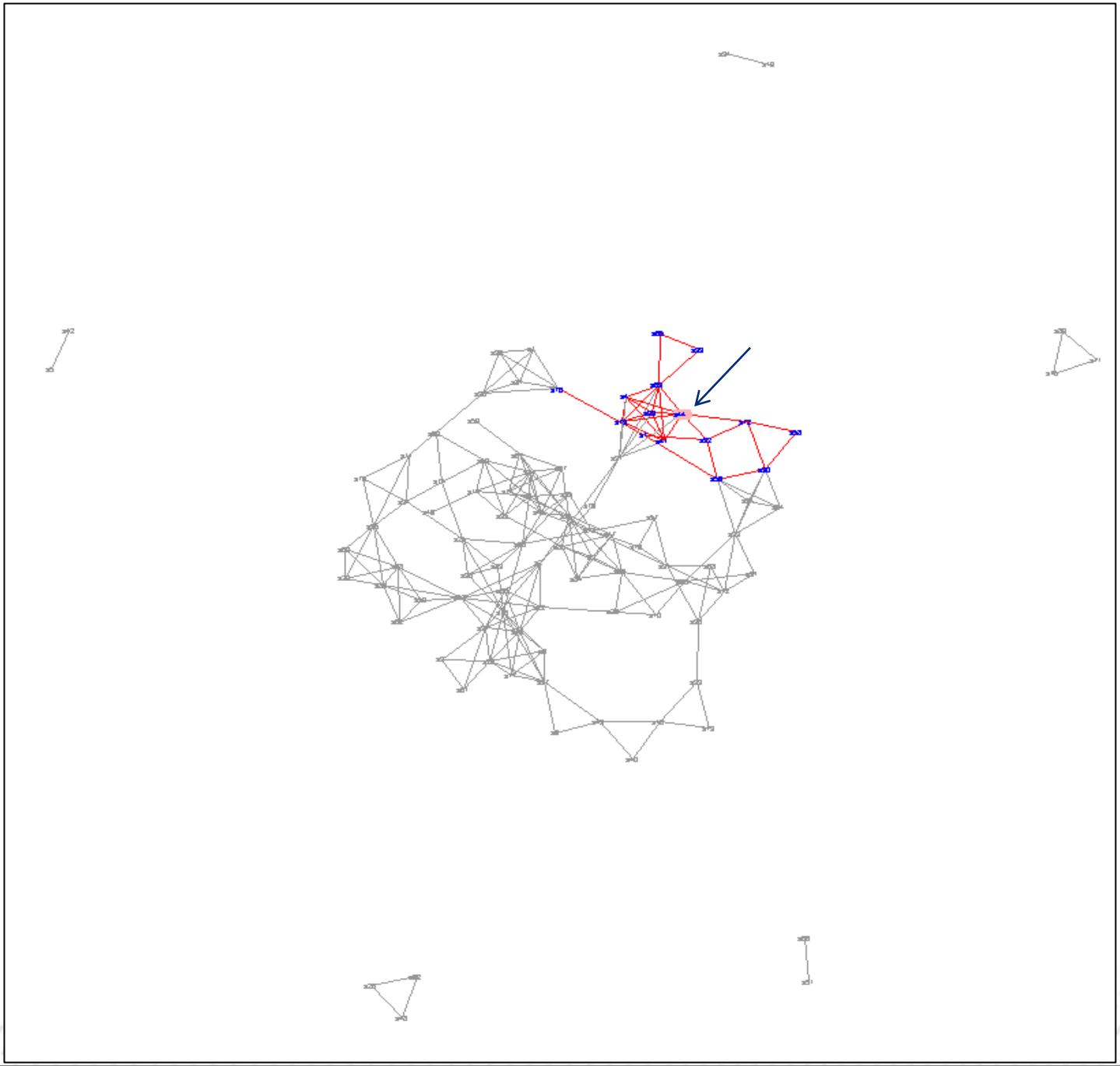
Mesh topology:

Cluster 2



Mesh topology:

Cluster 3



Experimentation (continued)

Sampling 3:

- 99 nodes, 329 edges
- 1 connected component

Star Topology Results:

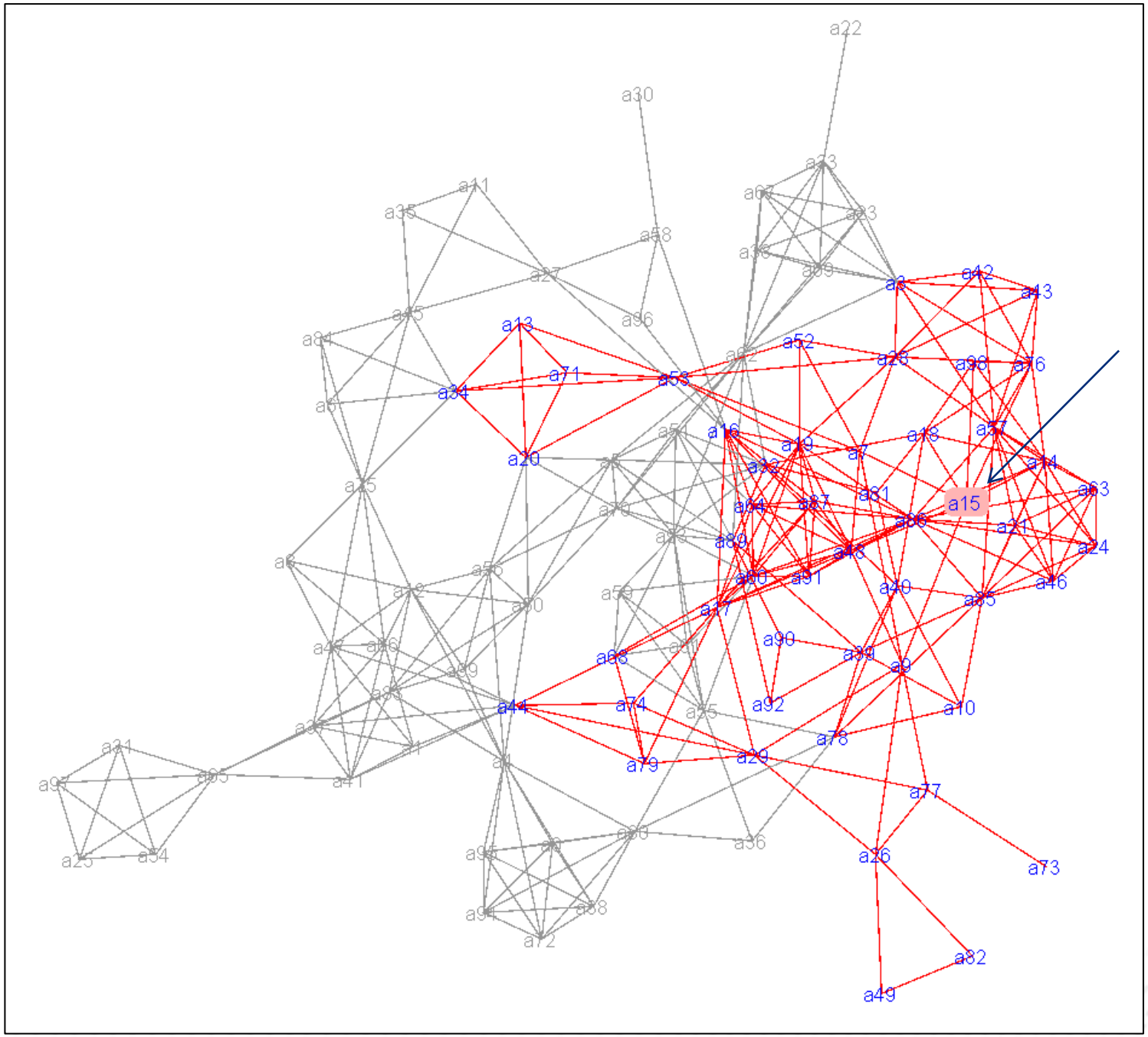
Theta	# of Clusters	# of Left Node	# of Overlap Node	# of Total Overlap	Cluster Sizes
0.926	4	7	19	19	(8,43,3,51)

Mesh Topology Results:

Theta	# of Clusters	# of Left Node	# of Overlap Node	# of Total Overlap	Cluster Sizes
0.626	2	1	48	48	(91,55)

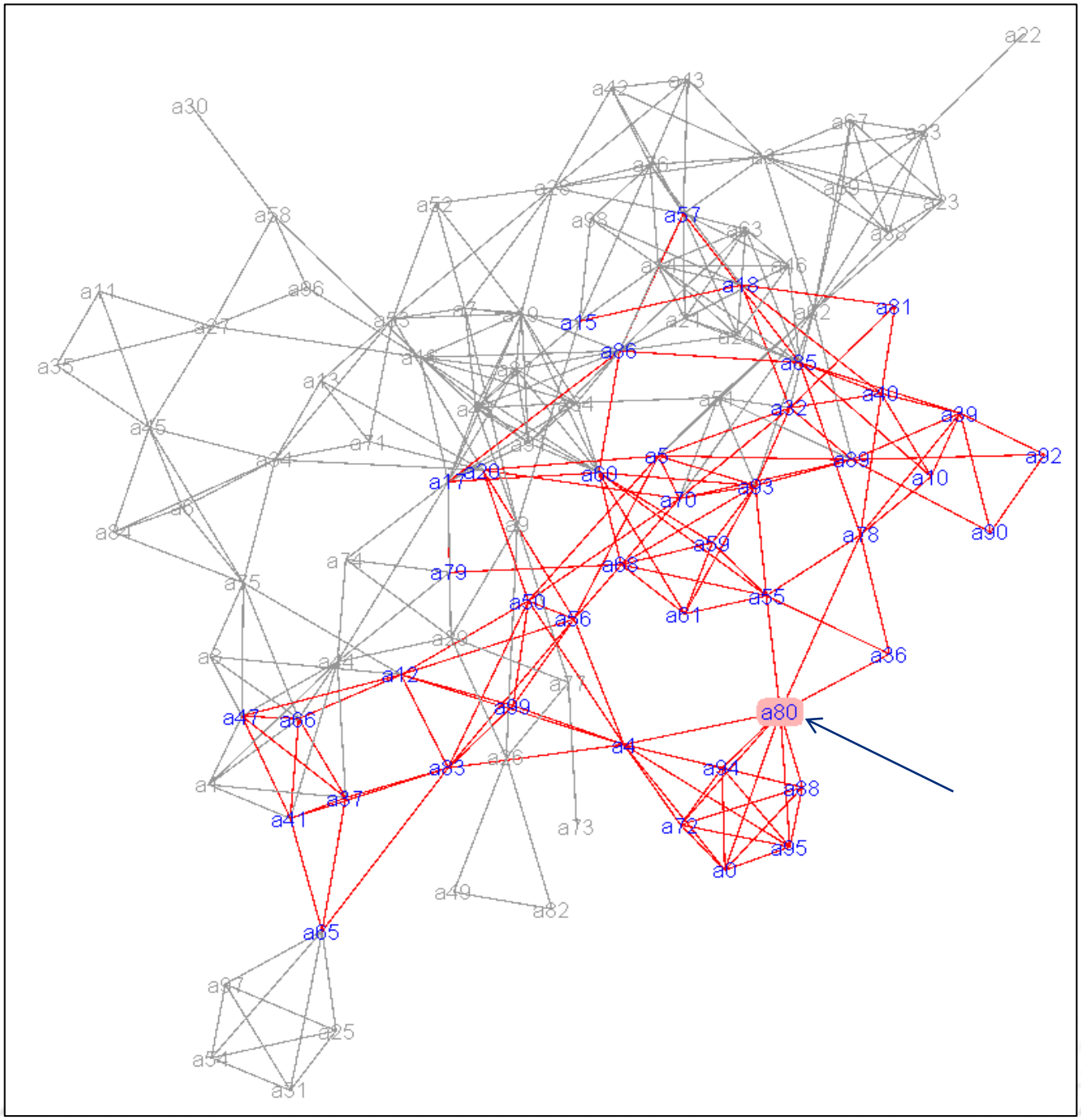
Star topology:

Cluster 1



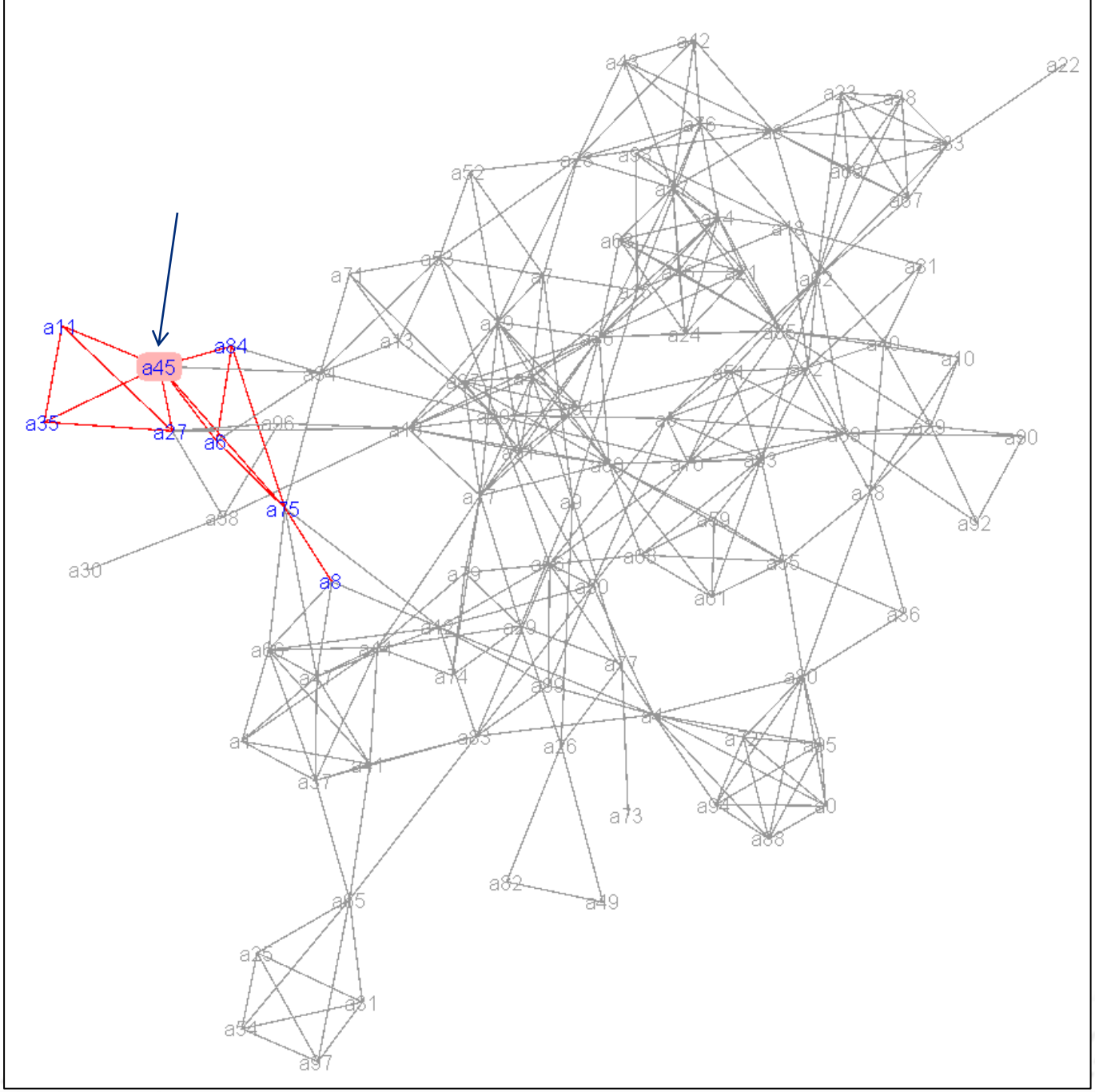
Star topology:

Cluster 2



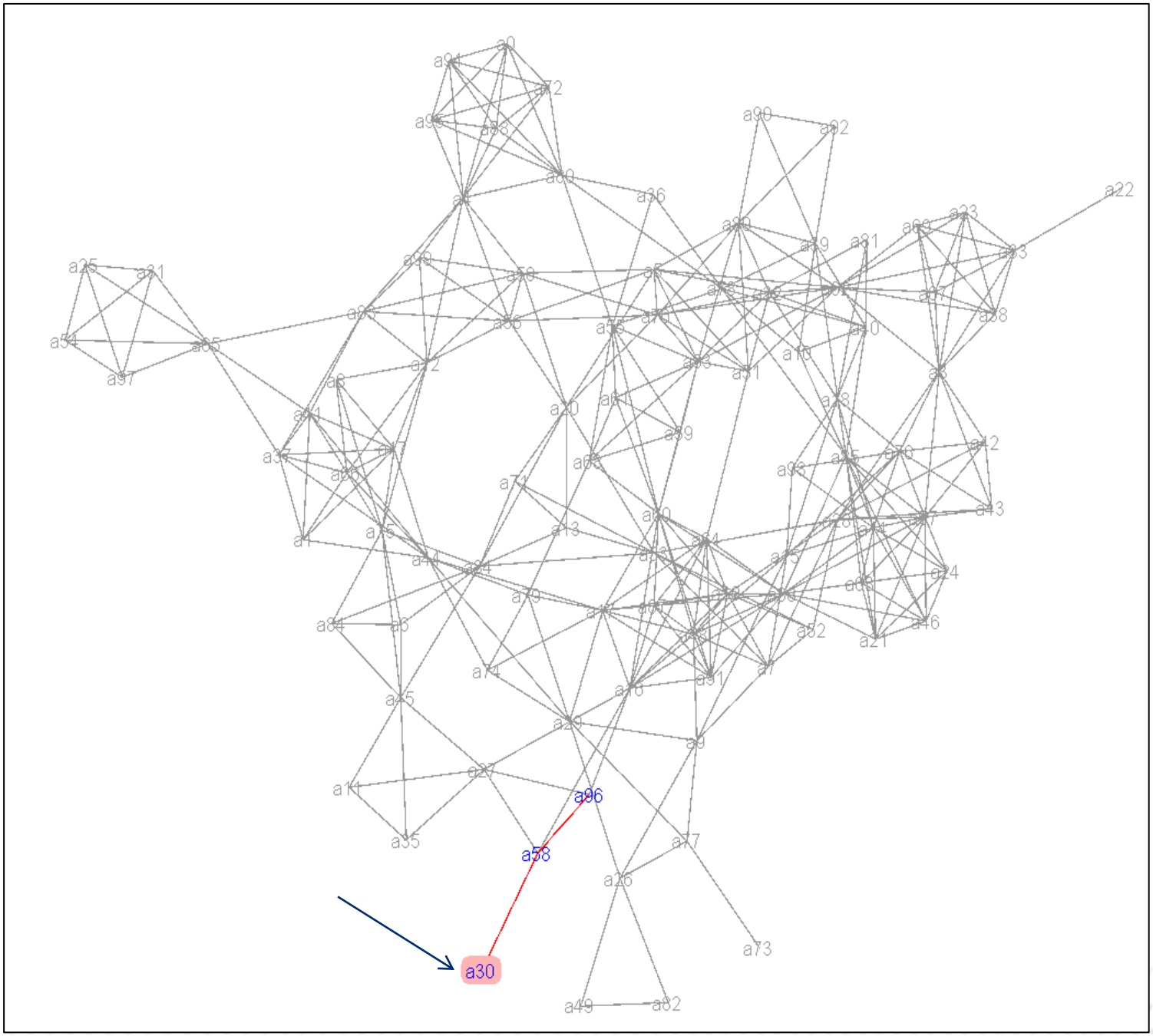
Star topology:

Cluster 3



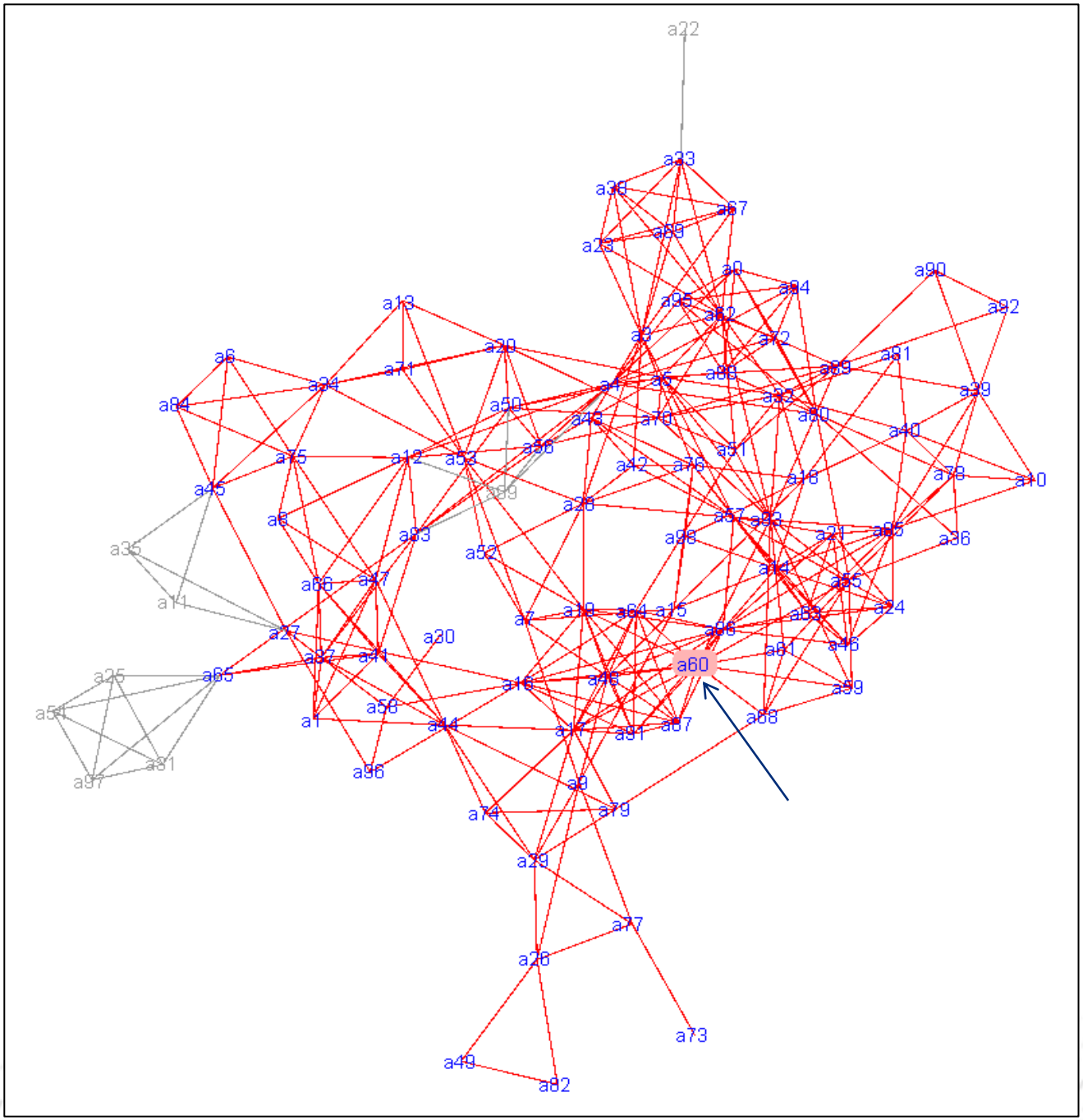
Star topology:

Cluster 4



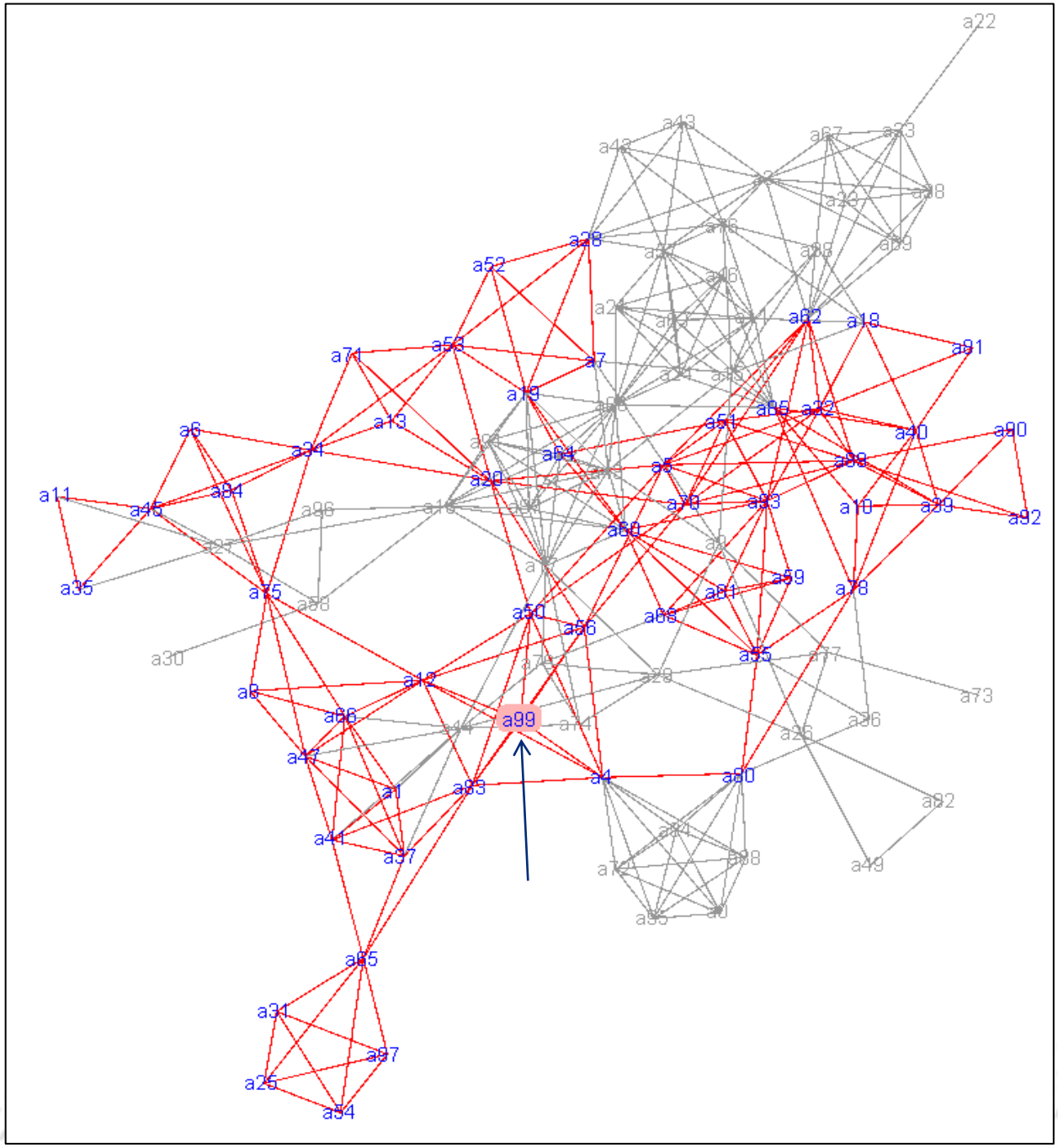
Mesh topology:

Cluster 1



Mesh topology:

Cluster 2



Future Work

- Fine tune the parameters used in the algorithm.
- Improve cluster acceptation/rejection criteria
- Relook into the requirement of possible different strategies of clustering for different topologies.
- Compare the results of the proposed methodology with the results of other clustering algorithms.
- Experiment the methodology on full DBLP dataset and on other social network databases such as Facebook/twitter.

Infosys®

THANK YOU