Focal Structure Analysis

Mustafa Alassad

PhD Student

Advisor : Dr. Nitin Agarwal

EIT – UALR

Fall 2018

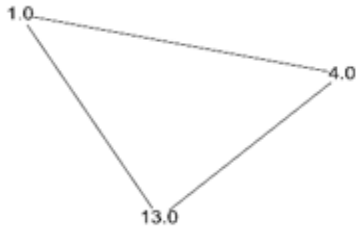
Outlines

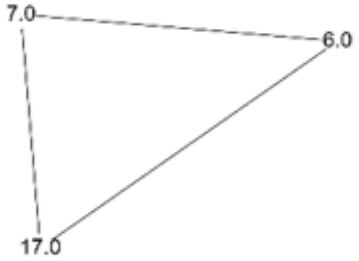
- Introduction;
- Problem Definition;
- Methodology;
- Solution strategy;
- Results and evaluation; and
- Conclusion

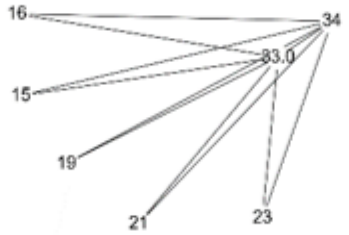
Introduction

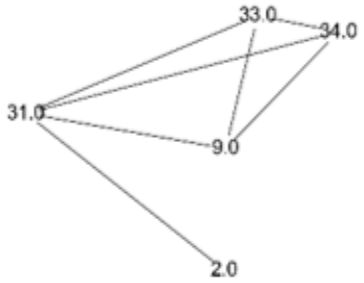
- Social networks are exposed to deep research in last few years.
- Communities, individuals, interactions, groups, etc.
- Methods are proposed for communities detections, and others are proposed to study the individuals actions.
- Modular comm., dense comm. Robust comm. Balanced comm. And hierarchical comm. are used to detect communities in graph; and
- Node degree, node similarity, and node reachability are member based community detection algorithms.
- But here we would like to mix between community based algorithms and members based algorithms to produce focal structures.

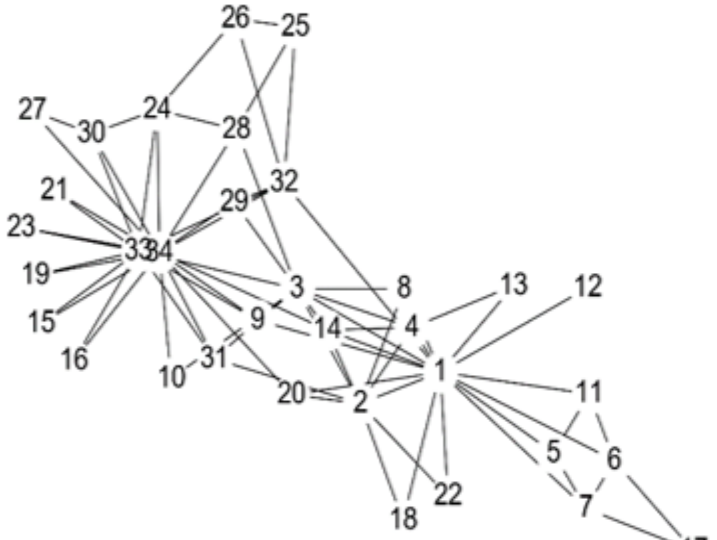
Zachary's karate Club Focal Structures

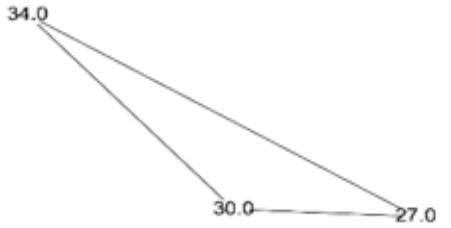


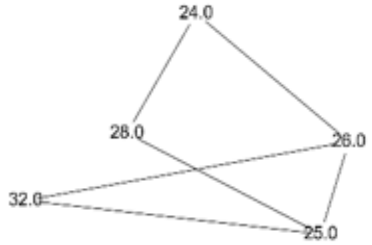


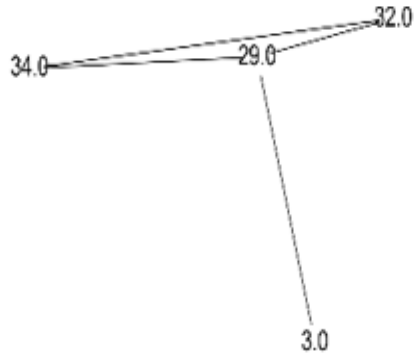


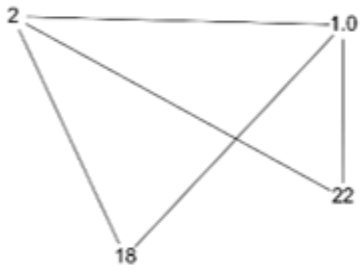


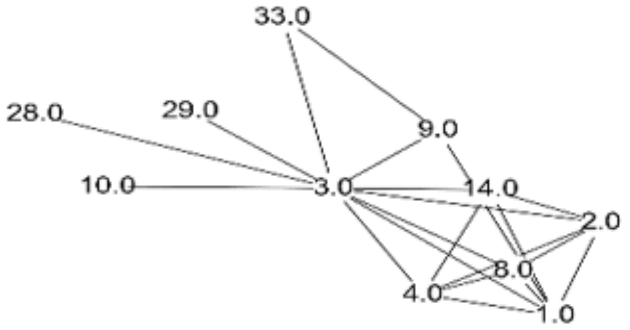


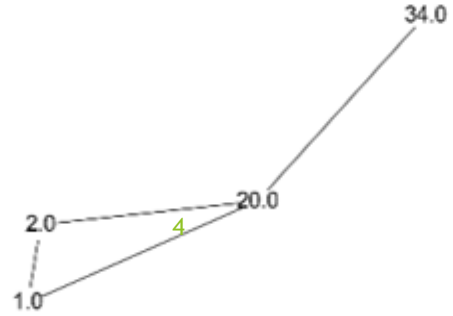












Problem Definition

- Focal structure definition
  - Focal structure in social network is defined to be the main set of individuals who may be in charge of organizing events, protest, or leading citizen engagements efforts.
  - In systems it is the complexity and modularity metrics.
  - In science it is the higher-order connectivity patterns.

Identifying the influential set of individuals instead of a set of influential individuals

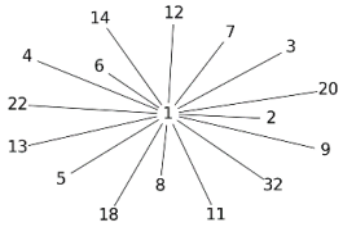
- Focal structure characteristics
  - Focal structure is a set of vertices at least three nodes and edges;
  - Focal structure needs to consist a triad or close to a complete graph;
  - Focal structure should produce high centrality, and jointly maximize graph modularity;
  - Can include members action in different groups or organizations;
  - By acting together may increase the graph modularity; and
  - They are unique, small, cannot be a subgraph, and different from regular communities.

Methodology

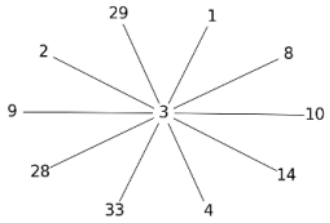
1. Model's First Level- Local Communities

1.1 Degree Centrality :

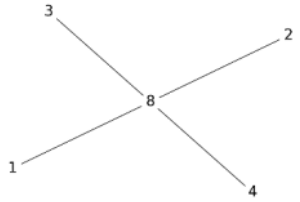
- One of the simplest and best known centrality measures is the degree centrality. It is counting the sum of edges occurrence upon a specified node and recognized that every edge is a walk of length 1.
- Dropping one neighbor nodes.



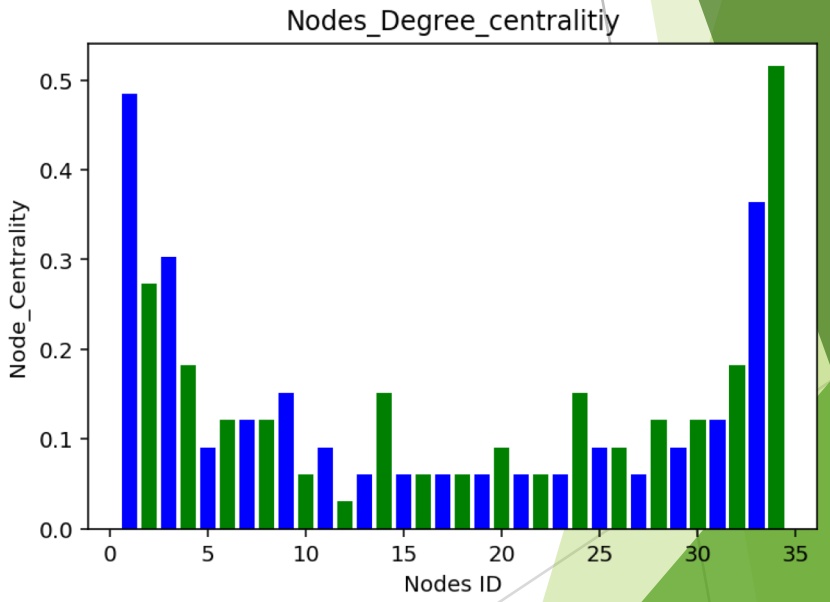
a) Community 1 *(16, 0.48)*



b) Community 3 (*10,0.3*).



c) Community 8 *(4 , 0.12)*.



Centrality degree measures of Zachary's karate club

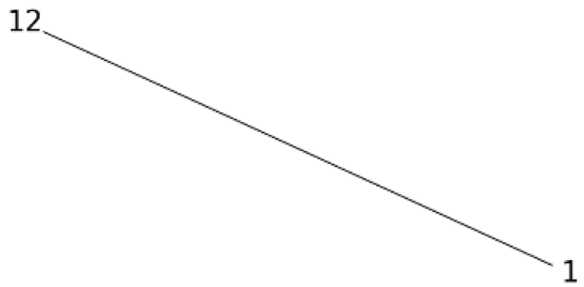
Methodology

1.2 Communities average Clustering Coefficient

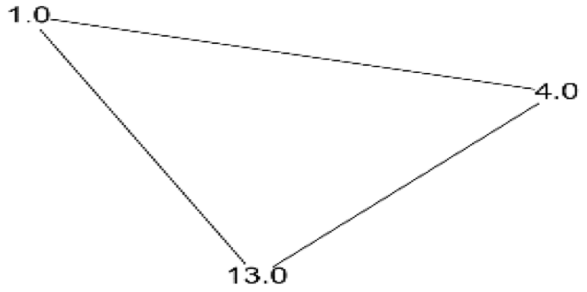
In social networks clustering coefficient is considered to study the individual's linking behavior with other individuals.

Transitivity is when a friend of my friend is my friend

- Global Clustering Coefficient " is for the network", and
- Local Clustering Coefficient " is for a node".



$\mathcal{AC}$ =0



$\mathcal{AC}$ =1

Zachary's Network Clustering Coefficient measures

Methodology

2 Model's Second Level- Global Communities

2.1 Modularity Values

$$Q = \frac{1}{2m}\left(\sum_{i,j} A_{ij} - \frac{d_i d_j}{2m}\right)\delta(C_i, C_j)$$

$$B = A_{ij} - \frac{\boldsymbol{d}\boldsymbol{d}^T}{2m}$$

d= is the degree vector for all nodes. $\boldsymbol{d} \in \mathbb{R}^{n\times 1}$

- Finding the assignment matrix Δ that maximizes Q is an NP-hard problem.
  - Good approximation, K-means clustering, K cut, Cliques, Similarity matrix, strong community and weak community, and community centrality probability.

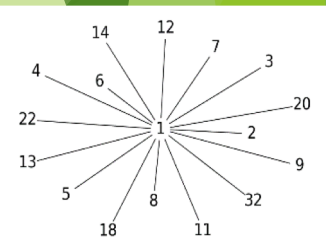
$$Q = \frac{1}{2m} Tr(\Delta^T B \Delta)$$

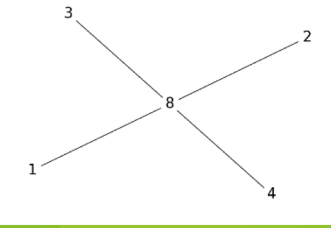
Δ is indicator function (partition membership) $\Delta \in \mathbb{R}^{n\times k}$

Similar to the spectral clustering matrix formulation, the modularity formulation explained below needs vectors as inputs.

$C_{v1}^T = [0,1,1,1,1,1,1,1,1,0,1,1,1,1,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0]$

$C_{v8}^T = [1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]$





Methodology

2.2 Stitching Similar Focal Structures

- **Jaccard similarity;**
- **Building similarity matrix;**
- **Tunable threshold;and**
- **Connected subgraph.**

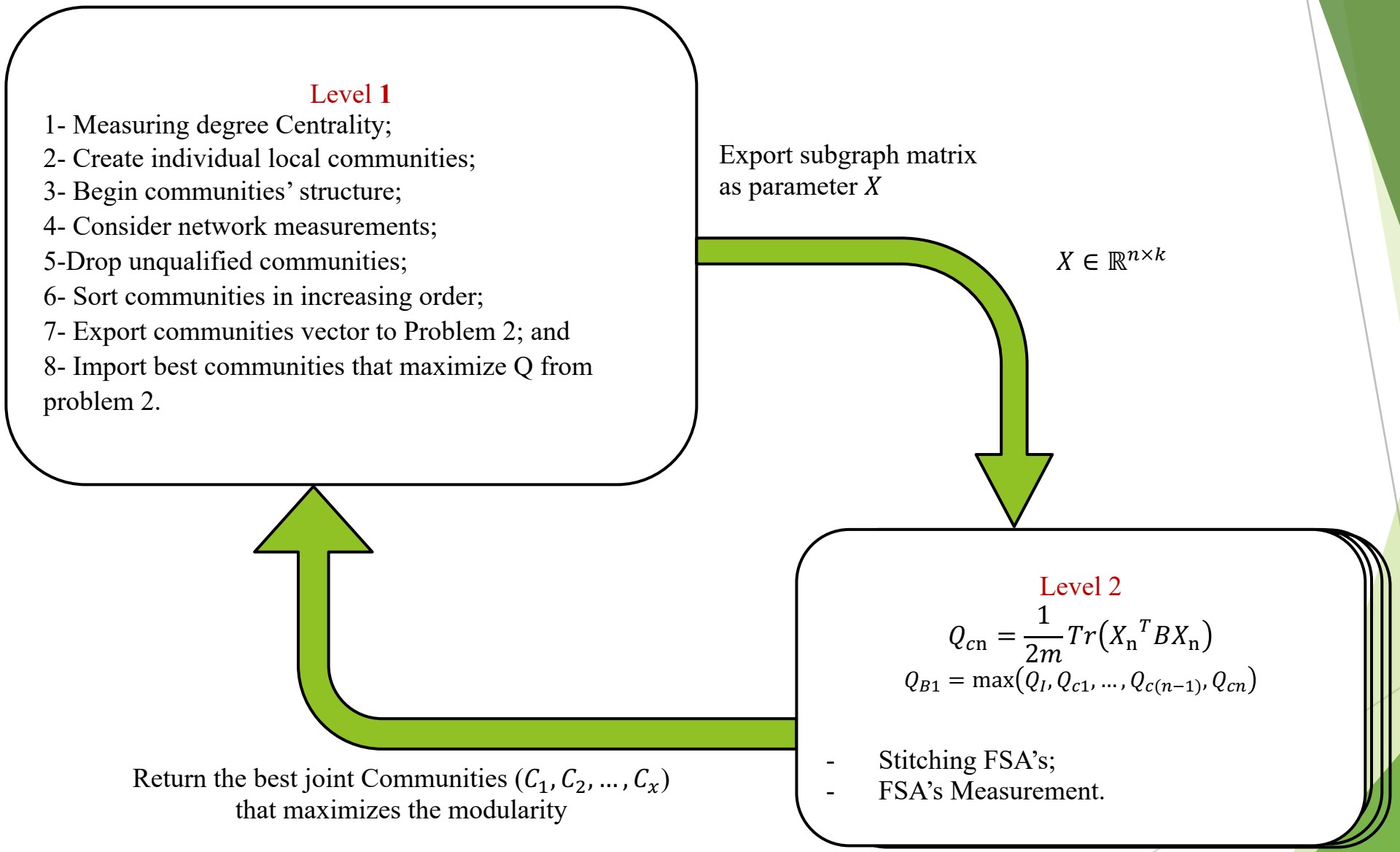
$$S = \begin{bmatrix} 0 & S_{12} & \cdots & S_{1n} \\ S_{21} & 0 & \cdots & S_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ S_{n1} & S_{n2} & \cdots & 0 \end{bmatrix}_{n \times n}$$

$$n = |FSA|$$

2.3 Focal Structure quality Measurements

- **Average Clustering Coefficient($\mathcal{AC}_F$);**
- **Average degree centrality ($\mathcal{ADC}_F$) ;**
- **Average shortest paths ($\mathcal{A\ell}_F$);**
- **Focal structure Diameter($\mathcal{D}_F$); and**
- **Focal structure Density ($\mathcal{DN}_F$).**

Solution Strategy



Results and Evaluation

Since the model consist of two level solution, maximizing modularity-centrality "MMC", the results should be measure in two levels too, as following:

1 Modularity Experiments and Evaluation

Best modularity values founded by applying famous algorithms on real-world data sets [16-27]

Network Name	$\|E\|$	$\|V\|$	**Modularity values**							
			LAG	**G-N**	**S-A**	**Ncut**	**FN**	**Mincut**	**CMDR**	MMC
Karate	78	34	0.42	0.40	0.42	0.34	0.253	0.23	0.417	0.417
Dolphin	159	62	0.52	0.52	0.52	0.37	0.372	0.37	0.52	0.509
LesMis	254	77	0.56	0.54	0.56	-	-	-	-	0.544

Method	**Karate**	**Dolphin**
Mincut[25]	0.008	0.01
FMM/GA[26]	1295	3783
FN[24]	0.031	0.078
Ncut[22]	0.021	0.027
CPP[27]	100.93	256.821
MMC	69.41	132

Running time (second) for algorithms on two real-world data sets.

Results and Evaluation

2 Focal Structure Experiments and Evaluation

All produced results were compared with Fatih's et al. results based on the online Focal Structure tools.

	MMC $\|FSA\|$= 11 Groups					Fatih's et al $\|FSA\|$= 5 Groups				
	$\mathcal{AC}_F$	$\mathcal{DN}_F$	$\mathcal{D}_F$	$\mathcal{Al}_F$	$\mathcal{ADC}_F$	$\mathcal{AC}_F$	$\mathcal{DN}_F$	$\mathcal{D}_F$	$\mathcal{Al}_F$	$\mathcal{ADC}_F$
Max	1.00	1.00	2.00	1.62	0.25	0.867	1.000	3.000	1.711	0.191
Min	0.33	0.38	1.00	1.00	0.13	0.000	0.378	1.000	1.000	0.167
Average	0.751	0.746	1.727	1.254	0.205	0.261	0.696	1.800	1.322	0.179
Abbrev.	$k\mu_1$	$k\lambda_1$	$k\xi_1$	$k\varsigma_1$	$k\psi_1$	$k\mu_2$	$k\lambda_2$	$k\xi_2$	$k\varsigma_2$	$k\psi_2$

Zachary's karate club results from two approaches. In this experiment, the MMC has more robust focal structures with respect to mean of average clustering coefficient and mean of average centrality. Fatih's et. al approach suffered from a chain FSAs, where some of them have $\mathcal{AC}_F = 0$. In other hand, MMC explored more FSAs including individuals acting in different groups, and this is the adequate the real world networks and people behaviors, but in the other approach, each individual was part of only one FSA

Also, to support results' evaluation, we applied two hypothesis for measuring the difference in the means of five applied metrics as following:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

Where μ_1 are the mean value of any FSA's $\mathcal{AC}_F$or ($\mathcal{DN}_F$, $\mathcal{D}_F$, $\mathcal{A\ell}_F$, and $\mathcal{ADC}_F$) and μ_2 is representing the same meaning for Fatih's et al.

MMC	Fatih's et. al	H_0	H_a	T-test	Accepted hypotheses	Rejected hypotheses
$k\mu_1$	$k\mu_2$	$k\mu_1 = k\mu_2$	$k\mu_1 > k\mu_2$	0.022	H_a	H_0
$k\lambda_1$	$k\lambda_2$	$k\lambda_1 = k\lambda_2$	$\lambda_1 > \lambda_2$	0.369	H_0	H_a
$k\xi_1$	$k\xi_2$	$k\xi_1 = k\xi_2$	$k\xi_1 > k\xi$	0.431	H_0	H_a
$k\varsigma_1$	$k\varsigma_2$	$k\varsigma_1 = k\varsigma_2$	$k\varsigma_1 > k\varsigma$	0.337	H_0	H_a
$k\psi_1$	$k\psi_2$	$k\psi_1 = k\psi_2$	$k\psi_1 > k\psi$	0.030	H_a	H_0

The two hypotheses are to measure were the mean of the FSA's metrics produced by two method are equal (H_0), or the alternate hypotheses is assuming that the MMC has improved the FSA values (H_a). the results show that the null hypotheses were rejected at the $\mathcal{AC}_F$, and $\mathcal{ADC}_F$, which means that the MMC's $\mathcal{AC}_F$, and $\mathcal{ADC}_F$ are higher that the other approach's FSAs. Also, this means that the MMC's FSAs are have stronger relationships and have more influential that other approach. Also, the test accepted the null hypotheses for the other metrics.

Dolphin social network results from two approaches. In this experiment, the MMC has more robust focal structures with respect to mean of average clustering coefficient , mean of average centrality values, and mean of density values as shown in Table 10. The Fatih's et. al approach suffered from a chain FSAs, where some of them have $\mathcal{AC}_F = 0$. In other hand, MMC explored more FSAs including individuals acting in different groups, and this is the adequate the real world networks and people behaviors, but in the other approach, each individual was part of only one FSA

	Current instances Number of FSAs $\|FSA\|$= 19 Groups					Fatih's et al instances Number of FSAs = 9 Groups				
	$\mathcal{AC}_F$	$\mathcal{DN}_F$	$\mathcal{D}_F$	$\mathcal{A\ell}_F$	$\mathcal{ADC}_F$	$\mathcal{AC}_F$	$\mathcal{DN}_F$	$\mathcal{D}_F$	$\mathcal{A\ell}_F$	$\mathcal{ADC}_F$
Max	0.84	0.73	2.00	1.60	0.13	1.00	1.00	3.00	1.82	0.13
Min	0.43	0.40	2.00	1.27	0.09	0.00	0.36	1.00	1.00	0.10
Average	0.63	0.58	2.00	1.42	0.11	0.42	0.72	2.00	1.34	0.11
Abbrev.	$D\mu_1$	$D\lambda_1$	$D\xi_1$	$D\varsigma_1$	$D\psi_1$	$D\mu_2$	$D\lambda_2$	$D\xi_2$	$D\varsigma_2$	$D\psi_2$

The two hypotheses are to measure were the mean of the FSA's metrics produced by two method are equal (H_0), or the alternate hypotheses is assuming that the MMC has improved the FSA values(H_a). the results show that the null hypotheses were rejected at the $\mathcal{AC}_F$,$\mathcal{DN}_F$, and $\mathcal{ADC}_F$, which means that the MMC's $\mathcal{AC}_F$,$\mathcal{DN}_F$, and $\mathcal{ADC}_F$ are higher that the other approach. Also, this means that the MMC's FSAs are have stronger relationships, the dolphins have more interaction with each other when the test rejected the null hypotheses of the density mean, and the have more centrality that other approach's FSAs. Also, the test accepted the null hypotheses for the other metrics. As results, the MMC improved three important metrics by accepting the alternate hypotheses and showed an equal means for $\mathcal{D}_F$ and $\mathcal{A\ell}_F$ and rejecting the (H_a).

MMC	Fatih's et. al	H_0	H_a	T-test	Accepted hypotheses	Rejected hypotheses
$D\mu_1$	$D\mu_2$	$D\mu_1 = D\mu_2$	$D\mu_1 > D\mu_2$	0.0716	H_a	H_0
$D\lambda_1$	$D\lambda_2$	$D\lambda_1 = D\lambda_2$	$D\lambda_1 > D\lambda_2$	0.0932	H_a	H_0
$D\xi_1$	$D\xi_2$	$D\xi_1 = D\xi_2$	$D\xi_1 > D\xi$	0.5	H_0	H_a
$D\varsigma_1$	$D\varsigma_2$	$D\varsigma_1 = D\varsigma_2$	$D\varsigma_1 > D\varsigma$	0.25	H_0	H_a
$D\psi_1$	$D\psi_2$	$D\psi_1 = D\psi_2$	$D\psi_1 > D\psi_2$	0.095	H_a	H_0

Saudi_Arabian women's 2014 network (Oct 09), the results from two approaches. In this experiment, the MMC has huge difference in the in the mean of average clustering coefficient, and the other approach. The Fatih's et. al approach has a chain FSAs, where some of them have $\mathcal{AC}_F = 0$. Also, there is a slight improvement in the degree centrality values, as show in the Table 14.In other hand, MMC explored more FSAs including individuals acting in different groups, and this is the adequate the real world networks and people behaviors, but in the other approach, each individual was part of only one FSA

	Current instances Number of FSAs $\|FSA\|$= 18 Groups					Fatih's et al instances Number of FSAs =9 Groups				
	$\mathcal{AC}_F$	$\mathcal{DN}_F$	$\mathcal{D}_F$	$\mathcal{Al}_F$	$\mathcal{ADC}_F$	$\mathcal{AC}_F$	$\mathcal{DN}_F$	$\mathcal{D}_F$	$\mathcal{Al}_F$	$\mathcal{ADC}_F$
Max	1.00	1.00	5.00	2.36	0.12	0.58	1.00	3.00	2.00	0.15
Min	0.20	0.05	1.00	1.00	0.03	0.00	0.29	1.00	1.00	0.04
Average	0.68	0.64	2.00	1.44	0.09	0.13	0.67	2.00	1.39	0.07
Abbrev.	$S\mu_1$	$S\lambda_1$	$S\xi_1$	$S\varsigma_1$	$S\psi_1$	$S\mu_2$	$S\lambda_2$	$S\xi_2$	$S\varsigma_2$	$S\psi_2$

The two hypotheses are to measure were the mean of the FSA's metrics produced by two method if they are equal (H_0), or the alternate hypotheses is assuming that the MMC has improved the FSA values(H_a). the results show that the null hypotheses were rejected at the $\mathcal{AC}_F$, and $\mathcal{ADC}_F$, which means that the MMC's $\mathcal{AC}_F$, and $\mathcal{ADC}_F$ are higher than the other approach. Also, this means that the MMC's FSAs are have stronger relationships, and they have more centrality that other approach's FSAs. Also, the test accepted the null hypotheses for the other metrics. As results, the MMC improved two important metrics by accepting the alternate hypotheses and showed an equal means for $\mathcal{DN}_\mathrm{F}$, $\mathcal{D}_F$ and $\mathcal{Al}_F$ and rejecting the (H_a).

MMC	Fatih's et. al	H_0	H_a	T-test	Accepted hypotheses	Rejected hypotheses
$S\mu_1$	$S\mu_2$	$S\mu_1 = S\mu_2$	$S\mu_1 > S\mu_2$	0.0000418	H_a	H_0
$S\lambda_1$	$S\lambda_2$	$S\lambda_1 = S\lambda_2$	$S\lambda_1 > S\lambda_2$	0.383	H_0	H_a
$S\xi_1$	$S\xi_2$	$S\xi_1 = S\xi_2$	$S\xi_1 > S\xi_2$	0.5	H_0	H_a
$S\varsigma_1$	$S\varsigma_2$	$S\varsigma_1 = S\varsigma_2$	$S\varsigma_1 > S\varsigma_2$	0.364	H_0	H_a
$S\psi_1$	$S\psi_2$	$S\psi_1 = S\psi_2$	$S\psi_1 > S\psi_2$	0.0413	H_a	H_0

Conclusion

- The model showed very promising results at the modularity maximization level,
- The model identified the influential sets of individuals.
- The model does not required additional information from the user, no number of groups requested initially, and no nonlinearity.
- The model was able to identify more communities than other method.
- The focal structure has higher quality with respect to degree centrality and average clustering coefficient.
- The model run time is acceptable; and
- The model was able to solve the problem defined in this research.