Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Pattern Recognition Letters 153 (2022) 246–253

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Central hubs prediction for bio networks by directed hypergraph - GA with validation to COVID-19 PPI☆

Sathyanarayanan Gopalakrishnan[a], Supriya Sridharan[a], Soumya Ranjan Nayak[b], Janmenjoy Nayak[c], Swaminathan Venkataraman[a,*]

[a] Department of Mathematics, School of Arts, Science, Humanities and Education, SASTRA Deemed to be University, Thanjavur, India
[b] Amity School of Engineering and Technology, Amity University, Uttar Pradesh, Noida, India
[c] Department of Computer Science, Maharaja Sriram Chandra Bhanja Deo University, Baripada, Mayurbhanj, Odisha, 757003

ARTICLE INFO

Article history:
Received 10 April 2021
Revised 3 December 2021
Accepted 22 December 2021
Available online 25 December 2021

Edited by: Maria De Marsico

2008 MSC:
05C65
05C82
68M10
90B18
90C27
90C35
91D30
92B20

Keywords:
Directed hypergraph
Centrality measures
Degree centrality
Strong tie
Weak tie
Genetic algorithm
COVID-19

ABSTRACT

Network structures have attracted much interest and have been rigorously studied in the past two decades. Researchers used many mathematical tools to represent these networks, and in recent days, hypergraphs play a vital role in this analysis. This paper presents an efficient technique to find the influential nodes using centrality measure of weighted directed hypergraph. Genetic Algorithm is exploited for tuning the weights of the node in the weighted directed hypergraph through which the characterization of the strength of the nodes, such as strong and weak ties by statistical measurements (mean, standard deviation, and quartiles) is identified effectively. Also, the proposed work is applied to various biological networks for identification of influential nodes and results shows the prominence the work over the existing measures. Furthermore, the technique has been applied to COVID-19 viral protein interactions. The proposed algorithm identified some critical human proteins that belong to the enzymes TMPRSS2, ACE2, and AT-II, which have a considerable role in hosting COVID-19 viral proteins and causes for various types of diseases. Hence these proteins can be targeted in drug design for an effective therapeutic against COVID-19.

1. Introduction

A relation that connects a group of two or more systems or people [18] forms a network. To assess a network, one may need information's such as quality of relationships, perception of co-operations, network collaborations between nodes. In recent times one of the computational concepts like the graph properties plays a significant role in analysing a network by exploring the accessibility of nodes [21]. Of which Multi-graph [3,17] can represent the complex relational data of a network and works well provided that the significant changes are made only to the existing graph analysis algorithms.

In general, since the network comprises nodes in n-ary relations, based on the literature so far, we could see Hypergraphs can handle n-ary relations more efficiently than graphs or multi-graphs. Berge [1] proposed this concept of hypergraph firstly as "a generalisation of graphs" which [3] further defined and derived the notion of the 'directed' hypergraph.

The hypergraphs [15,20] constructed from the shortest paths of the networks tend to leave out some influential nodes, which is very important in analysing the network. Granovetter [5] proposed a method to identify influential nodes by weak ties for information dissemination. To surmount the frailty for hypergraph construction, this paper constructs a directed hypergraph based on the

☆ Edited by: Maria De Marsico.

* Corresponding author.
E-mail address: swaminathan@src.sastra.edu (S. Venkataraman).

https://doi.org/10.1016/j.patrec.2021.12.015
0167-8655/

Nomenclature

ab_{ij}	the fitness of jth individual in ith generation in roulette wheel selection
$\overline{ab_{ij}}$	the average fitness of jth individual in ith generation in roulette wheel selection
$c_d^h(v_i)$	the weighted node degree centrality of a node i in H_{WDG}
cd_j	the probability for selecting jth string
$c_{sd}^h(v_i)$	the strong tie degree centrality of a node i in H_{WDG}
$c_{wd}^h(v_i)$	the weak tie degree centrality of a node i in H_{WDG}
E_{H_G}	the hyperedges of H_G
E_{HBM_G}	the hyperedges in H_{BM_G}
$E_{H_{WDG}}$	the set of all weighted directed hyperedges of the weighted directed hypergraph H_{WDG}
E_{HFM_G}	the directed hyperedges in H_{FM_G}
F	the total fitness value
f_j	the fitness value of jth individuals
$f_{\min}$	the minimum fitness
$f_{\max}$	the maximum fitness
G	the graph
H_{DG}	the directed hypergraph
HD_{HBM_G}	the head set in the minimal hypergraph H_{BM_G}
H_{WDG}	the weighted directed hypergraph
$H(E_{H_{DG}})$	the head of the directed hyperedge $E_{H_{DG}}$
H_G	the hypergraph
H_{BM_G}	the minimal B-hypergraph of H_{DG}
H_{FM_G}	the minimal F-hypergraph of H_{DG}
$M(H_{WDG})$	the number of directed hyperedges in H_{WDG}
$M(G)$	the number of edges in the graph G
$ngen$	the number of generations
$N(H_{WDG})$	the number of vertices of a weighted directed hypergraph
$N(G)$	the number of nodes in the graph G
pr_i	probability rank of the ith generation
r_{ij}	the rank of jth individual in ith generation for rank selection
$rsum_i$	sum of ranks in ith generation
$T(E_{H_{DG}})$	the tail of the directed hyperedge $E_{H_{DG}}$
$ST_{H_{WDG}}$	the set of strong tie nodes
$ST(H_{WDG})$	the range of strong tie vertices in H_{WDG}
VH_G	the set of all nodes of the hypergraph H_G
$V_{H_{DG}}$	the set of all nodes of the directed hypergraph H_{DG}
$V_{H_{WDG}}$	the set of all nodes of the weighted directed hypergraph H_{WDG}
w_{ij}	the weight of the node v_i corresponding to the directed hyperedge e_j
$WT_{H_{WDG}}$	the set of weak tie nodes
$WT(H_{WDG})$	the range of weak tie vertices in H_{WDG}
W_M	the mean of the weights
W_{q_1}	the 1st quartile of the weights
W_{q_3}	the 3st quartile of the weights
W_{QD}	the quartile deviation of the weights
W_{SD}	the standard deviation of the weights

relationship between the data. Later, degree centrality measure is employed for finding the influential nodes.

Among various centrality measures like degree, betweenness, closeness, eccentricity, cross-click, network, random walk betweenness, page rank, leverage, eigenvector, subgraph, information and many, degree centrality have high impact in analysing any networks. Also, it classifies the nodes as strong and weak ties from which weak tie is more influential so, this takes less number of nodes that are responsible for spreading of news. Hence, we use the degree centrality for predicting the influential nodes.

In this paper, initially the weight of the node is the degree centrality of a node that is, ..., the number of hyperedges incidence with the node. Thus, the same is calculated for all network nodes and fed into the Genetic Algorithm (GA) for further optimisations. After which statistical measures like mean, standard deviation and quartiles are employed to classify the nodes as strong tie and weak tie.

The proposed work is applied to Protein-Protein Interaction (PPI) networks to obtain the influential proteins. Predicting protein function is always a stumbling block in computational biology research. These proteins of PPI network help in drug target recognition, identify the role of a protein or gene, develop successful methods for treating different diseases, and provide early detection of disorders.

Recently, many researchers aims to detect the COVID-19 through images of X-rays using the concept of Cascaded Recurrent Neural Network (CRNN) [12] and ultrasound X-rays by classifying them using Multi-layers Fusion [16]. Here, we aim to detect the influential nodes which gives a promising direction in the impact of current pandemic COVID-19 and in need of designing drugs. Drug design requires the knowledge of the functionality of the COVID-19 viral protein interacting with human proteins. For this purpose, we identified some of the critical human proteins using centrality measures of the hypergraph.

These proteins belong to the enzymes - TMPRSS2, ACE2, AT-II, protein sets - IL6, cytoplasmic, cytokine storm. Some of them may cause diseases like decease chronic obstructive pulmonary disease, lower respiratory infections, blood pressure, diabetes mellitus, stroke and tuberculosis. The resultant proteins play a considerable role in COVID-19 viral interactions.

The major contributions are a state-of-the-art representation of protein interaction networks by weighted directed hypergraph and identifying influential nodes using weak tie of a weighted directed hypergraph. The degree centralities and genetic algorithm are hybridized to optimize the weights of nodes. Finally, validation of the proposed method identifies influential COVID-19 proteins from protein interactions that can be used for drug design.

Section 2 deals with basic definitions. The proposed methodology is presented in Section 3. The results of the ten biological networks and their comparison with existing graph centrality measures are presented in Section 4. Our method has been validated with the real-time pandemic COVID-19 viral-protein interactions in Section 5. The paper is concluded with a summary in the final Section.

2. Preliminaries

Some preliminary concepts on hypergraph are recalled in this section.

Let $H_G = (V_{H_G}, E_{H_G})$ be a **hypergraph** [1], where V_{H_G}, and E_{H_G} are set of all nodes and hyperedges respectively. Moreover, $V_{H_G} = \{v_i : i = 1, \ldots, n\}$ and $E_{H_G} = \{E_j : j = 1, \ldots, m\}$, with every E_j is a subset of the set V_{H_G}.

A hypergraph is a standard graph, when every hyperedge $E_i \in E_{H_G}$, satisfies $|E_i| \leq 2$ for $i = 1, 2, \ldots, m$.

A hypergraph is said to be a **directed hypergraph** (H_{DG}) [3], if every hyperarc $E_{H_{DG}} = (T(E_{H_{DG}}), H(E_{H_{DG}}))$ has a direction, where $T(E_{H_{DG}})$ is the tail of EH_G while $H(E_{H_{DG}})$ is its head.

If every node of a directed hypergraph has a weight associated with it, then the directed hypergraph is a **weighted directed hypergraph** (H_{WDG}).

The directed hyperedge or hyperarc $E_{H_{DG}} = (T(E_{H_{DG}}), H(E_{H_{DG}}))$ is said to be a **Backward hyper-arc** [3], or **B-arc**, if $|H(E_{H_{DG}})| = 1$. Similarly, the directed hyperedge or hyperarc $E_{H_{DG}} = (T(E_{H_{DG}}), H(E_{H_{DG}}))$ is said to be a **Forward hyperarc** [3], or **F-arc**, if $|T(E_{H_{DG}})| = 1$.

A directed hypergraph is said to be a *B*-graph (or *B*-hypergraph), whose hyperarcs are *B*-arcs. A directed hypergraph is said to be an *F*-graph (or *F*-hypergraph), if hyperarcs are *F*-arcs. A directed hypergraph is said to be a *BF*-graph (or *BF*-hypergraph), if hyperarcs are either *B*-arcs or *F*-arcs.

3. Proposed methodology

This section discusses the proposed technique for the identification of influential nodes in a network. It consists of following four predominant steps:

1) Construction of directed hypergraph.
2) Conversion of directed hypergraph into weighted directed hypergraph.
3) Optimizing the weights using Genetic Algorithm (GA).
4) Identifying influential nodes.

Algorithm 1 comprises the above four steps:

Algorithm 1: DHHGA.

INPUT: Network

OUTPUT: Strong and Weak Ties

Procedure: Construction of Directed Hypergraph (Network) (Algorithm 2)
Procedure: Conversion of H_{DG} into H_{WDG} (Algorithm 3)
Procedure: Genetic Algorithm (w_j) (Algorithm 4)
Procedure: Weighted Directed Hypergraph Degree Centrality [WDHDC] (Optimized w_j) (Algorithm 5)

Algorithm 2: Construction of directed hypergraph.

INPUT: Network

OUTPUT: Directed Hypergraph (H_{DG})

for $i = 1$ to n **do**
  if v_i not in $T(E_{H_{DG}})$ (by using theorem 1) **then**
    Construct $E_{H_{DG}}$ of the directed hypergraph H_{DG} with
    $T(E_{H_{DG}}) = v_i$ and
    $H(E_{H_{DG}}) = \{v_{i+1} : \text{if } v_i \text{ has a communication with } v_{i+1}\}$
  end if
end for
Return H_{DG}.

3.1. Construction of directed hypergraph

If there is a communication between v_i to $\{v_j \mid j = 1, 2, \ldots h,\ h \in \mathbb{N}\}$, then directed hyperedge $E_{H_{DG}} = (T(E_{H_{DG}}),\ H(E_{H_{DG}}))$ is constructed where $T(E_{H_{DG}}) = \{v_i\}$ and $H(E_{H_{DG}}) = \{v_j \mid j = 1, 2, \ldots h,\ h \in \mathbb{N}\}$.

Definition 3.1 (Minimal hypergraph based on. *B*-hyperarc) A directed hypergraph H_{M_G} is said to be **minimal hypergraph** (directed hyperedges are in *B*-hyperarc form or *F*-hyperarc form) with $|H_{M_G}| = k$, if there is no minimal hypergraph H_{M1_G} with $|H_{M1_G}| = p < k = |H_{M_G}|$.

Algorithm 3: Conversion of H_{DG} into H_{WDG}.

INPUT: H_{DG} from Algorithm 2

OUTPUT: Weighted Directed Hypergraph H_{WDG}

$V_{H_{WDG}} = V_{H_{DG}}$, $E_{H_{WDG}} = E_{H_{DG}}$
for $i = 1$ to n **do**
  for $j = 1$ to m **do**
    if $v_i \in e_j$ **then**
      $w_j = w_j + 1$
    end if
  end for
  Append the weight of v_i as w_j in H_{WDG}
end for
Return H_{WDG}

In general, *B*-hypergraph has $|H(E_{H_{BMG}})| = 1$ for every hyperedge, and there is no repetition in head of the hyperedge.

Suppose there are two hyperedges $((v_i, \ldots, v_k), v_j)$ and $((v_a, \ldots, v_b), v_j)$ with same $H(E_{HBM_G})$, then combine the hyperedges and regenerate it as a single hyperedge $((v_i, \ldots, v_k, v_a, \ldots, v_b), v_j)$.

Continue this until the heads of the hyperedges are distinct.

Let HD_{HBM_G} be the head set in the minimal hypergraph H_{BM_G}.

Now, add the head of each hyperedge $|E_{HBM_G}|$ of H_{BM_G} to the set HD_{HBM_G}. Thus,

$$|HD_{HBM_G}| = b < n,$$

since the heads in HD_{HBM_G} are distinct and at most $|V_{H_{DG}}|$.

Since the number of hyperedges is equal to the number of elements in the head set HD_{HBM_G} by *B*-hypergraph construction,

$$|E_{HBM_G}| = |HD_{HBM_G}|$$
$$\text{gives } |E_{HM_G}| = |HD_{HBM_G}| = b < n$$

and hence $|E_{HBM_G}| = b < n$

Similar arguments holds for *F* and *BF* hypergraphs and thus we have Theorem 1.

Theorem 1. *Let $H_{DG} = (V_{H_{DG}}, E_{H_{DG}})$ be a directed hypergraph with $|V_{H_{DG}}| = n$, then there exists a minimal hypergraph $H_{BM_G} = (V_{H_{DG}}, E_{HBM_G})$ such that every directed hyperedge E_{HBM_G} of H_{BM_G} is B-hyperarc or H_{BM_G} is a B-hypergraph. Also, there exist a minimal hypergraph $H_{FM_G} = (V_{H_{DG}}, E_{HFM_G})$ such that every hyperedge E_{HFM_G} of H_{FM_G} is F-hyperarc or H_{FM_G} is an F-hypergraph.*

3.2. Weighted node degree centrality (WNDC)

Definition 3.2. The weights of the node incidence with the corresponding hyperedge is called as weighted node degree centrality [9,13]. It is given by

$$C_d^h(v_i) = \sum_{j=1}^{m} w_j,\ \ (i = 1, 2, \ldots, n) \tag{1}$$

where w_j takes the value 1 if v_i is incident with e_j, 0 otherwise.

3.2.1. Construction of weighted node degree centrality

Initially, every vertex v_i of a directed hypergraph is assigned with a weight as its degree centrality and it is presented in Algorithm 3.

Here, the weight w_j is calculated as defined in Definition 3.2. These weights are tuned using the GA (Algorithm 4).

Algorithm 4: Genetic algorithm.

INPUT: *Genetic Algorithm*(w_j)

OUTPUT: Best solution

1: $p_t \leftarrow 0$
2: Generate populations at random ($population(p_t)$)
3: Determine the fitness values of $population(p_t)$
4: **for** $p_t = 0$ to TC (Termination Condition (TC)) **do**
5: Choose the best individuals from the groups of $population(p_r)$ using Roulette Wheel Ranking Selection
6: Apply single-point crossover to the resultant population
7: Apply uniform mutation to the resultant population
8: Evaluate optimised fitness values
9: **end for**
10: Return best solution

3.3. GA weight optimization

GA is a search heuristic, which optimizes the solution of search problems [21]. Usually, GA is a population-based search technique used in computing, with each candidate represented as fixed-length binary string chromosomes. The Roulette wheel with ranking selection, one-point crossover and a uniform mutation are the components of GA in this work. The objective function of GA is,

$$\text{(objective function) } Y = \sum_{i=1}^{n} (x_i * w_i),$$

where x_i is the initial weight of the node v_i, and w_i (weight) is the parameter that is to be maximized.

Now, Roulette wheel and ranking selection [10] methods are combined to select the best individual from the groups (of individuals) formed out the population, for the objective function. The Roulette Wheel uses,

$$\overline{ab_{i,j}} = \frac{\sum_{j=1}^{N} ab_{ij}}{N}$$

where $\overline{ab_{i,j}}$ represents the average fitness of the population for ith generation which varies from 1 to ngen. This value is used to place in the segment of roulette wheel, the bigger the value, the larger the segment and it is more probably to be selected.

$$cd_j = \frac{ab_{ij}}{\sum_{j=1}^{N} ab_{ij}}$$

where cd_j represents the probability for selecting the jth individual and ab_{ij} is the fitness value of the jth individual in the ith generation, and for ranking

$$pr_i = \frac{r_{ij}}{rsum_i},\ rsum_i = \sum_{j=1}^{N} r_{ij}$$

where i varies from 1 to *ngen* (number of generations) and j varies from 1 to N (population size).

Pioneer technique used in the crossover is the single-point crossover, and it is given as,

$$\text{(Single-point crossover) } crossover = Bas\left(\frac{\text{off spring-size}}{2}\right)$$

where *Bas* stand for Binary array of size.

We select a random gene from chromosome, lets say x_i and assign a uniform random value to it.

$$\text{(Uniformmutation)} x_i = U(a_i, b_i)$$

where $i \in [1, n]$, a_i and b_i are random integer, $U(a_i, b_i) \in [a_i, b_i]$ is a uniform random number.

The fitness value [4] F is calculated using normalized weighted sum evaluation function given by

$$\text{(Fitness)} F = \sum_{j=1}^{N} w_j \frac{f_j - f_j^{\min}}{f_j^{\max} - f_j^{\min}}.$$

where f_j is actual fitness value, $f_j^{\max}$ is the worst fitness value, $f_j^{\min}$ is the best fitness value, of jth individual.

Now, Algorithm 5 categorizes the nodes as strong ($ST_{H_{WDG}}$) and weak ($WT_{H_{WDG}}$) ties from the optimized weights of Algorithm 4.

Algorithm 5: WDHDC.

INPUT: H_{WDG} from Algorithm 3

OUTPUT: Strong ($ST_{H_{WDG}}$) and Weak ($WT_{H_{WDG}}$) Tie nodes

1: $|V_{H_{WDG}}| = n$ and $|E_{H_{EDG}}| = m$, $ST_{H_{WDG}} = \phi$ and $WT_{H_{WDG}} = \phi$,
2: Calculate the Mean (W_M) and Standard Deviation (W_{SD}) of the Weights
3: **for** $i = 1$ to n **do**
4: **if** $w_i > W_M + W_{SD}$ **then**
5: $ST_{H_{WDG}} = ST_{H_{WDG}} \cup \{v_i\}$
6: **else if** $w_i < W_M - W_{SD}$ **then**
7: $WT_{H_{WDG}} = WT_{H_{WDG}} \cup \{v_i\}$
8: **end if**
9: **end for**
10: Return the sets $WT_{H_{WDG}}$ and $ST_{H_{WDG}}$

The categories of ties based on their strength using mean and standard deviation is given as,

$$C_{sd}^h(v_i) = C_d^h(v_i) > W_M + W_{SD}, \quad \text{for strong ties,}$$
$$C_{wd}^h(v_i) = C_d^h(v_i) < W_M - W_{SD}, \quad \text{for weak ties.}$$

Here W_M stands for the mean of the weights, and W_{SD} stands for the weights' standard deviation.

Similarly, the categorization of tie strength using quartile can also be defined as follows:

$$C_{sd}^h(v_i) = C_d^h(v_i) > W_{q_3} + W_{Q_D}, \quad \text{for strong ties,}$$
$$C_{wd}^h(v_i) = C_d^h(v_i) < W_{q_1} - W_{Q_D}, \quad \text{for weak ties.}$$

Here W_{q_1}, W_{q_3} stands for 1st and 3rd quartile of the weights and W_{Q_D} stands for quartile deviation of the weights.

4. Implementation

The proposed technique is applied to the following ten biological networks [14] using Python 3.5 in Intel® Core™ i7-6700 Quad Core 3.4 GHz, 4.0 GHz system running in Ubuntu 16.4. (i) bio-WormNet-v3-benchmark, (ii) bio-DR-CX's, (iii) bio-DM-CX's, (iv) bio-HS-LC's, (v) bio-HS-CX's, (vi) bio-CE-CX's, (vii) bio-grid-fission-yeast's, (viii) bio-grid-yeast's, (ix) bio-grid-human's, (x) bio-dmela. where the networks (i)–(vi), are all a kind of WormNet network, with nodes as genes and edges as links between them and they are an integration's of all data-type-specific networks (CE-CX, CE-GN, CE-GT, CE-HT, CE-LC, CE-PG, DM-CX, DM-HT, DM-LC, DR-CX, HS-CX, HS-HT, HS-LC, SC-CC, SC-CX, SC-HT, SC-LC, SC-TS) through modified Bayesian integration. And for the remaining networks (vii)–(x), nodes are proteins and the edges are PPI.

Table 1
Range of influential nodes (Weak ties) using mean and SD, and quartiles.

H_{WDG}	$N(H_{WDG})$	$M(H_{WDG})$	Range of $WT(H_{WDG})$ nodes by mean and SD	Range of $WT(H_{WDG})$ nodes by quartiles
bio-grid-fission-yeast's	2031	2026	[400, 450]	[470, 530]
bio-WormNet-v3-benchmark	2445	2316	[490, 540]	[550, 640]
bio-DR-CX's	3289	3051	[650, 720]	[750, 850]
bio-DM-CX's	4040	3594	[820, 890]	[930, 1020]
bio-HS-LC's	4227	3391	[850, 930]	[1000, 1050]
bio-HS-CX's	4413	3975	[900, 1000]	[1000, 1150]
bio-grid-yeast's	6010	6008	[1215, 1280]	[1430, 1480]
bio-dmela	7399	6640	[1500, 1600]	[1750, 1900]
bio-grid-human's	9527	9536	[1970, 2050]	[2300, 2390]
bio-CE-CX's	16,347	14,692	[3420, 3490]	[3900, 4150]

Table 2
Comparison of number of influential nodes with other centrality measures.

G	$N(G)$	$M(G)$	DC_G	CC_G	EC_G	HC_G
bio-grid-fission-yeast's	2031	25,274	964	450	450	450
bio-WormNet-v3-benchmark	2445	78,736	2032	2295	2152	2197
bio-DR-CX's	3289	84,940	1478	1647	1344	1158
bio-DM-CX's	4040	112,688	1267	964	957	1060
bio-HS-LC's	4227	39,484	2835	1673	1753	1661
bio-HS-CX's	4413	108,818	1493	1044	1392	1037
bio-grid-yeast's	6010	313,890	3150	4791	5414	4791
bio-dmela	7399	25,571	4078	3681	6383	6370
bio-grid-human's	9527	62,364	6621	8029	8029	8029
bio-CE-CX's	16,347	762,822	7734	5081	7596	5047

Table 3
Comparison of influential proteins (Count) with our algorithm in the COVID-19 interaction data-set.

Enzyme/disease	Total number of proteins in cleaned data-set	Number of proteins obtained using our algorithm
TMPRSS2	47	47
ACE2	4	4
AT-II	11	11
Sudden cardiac attack	10	10
IL6	33	33
Cytoplasmic	1159	1159
Cytokines	3	3
Chronic obstructive pulmonary disease	2	2
Lower respiratory infections	3	3
Blood pressure	35	35
Diabetes mellitus	35	35
Stroke	23	23
Tuberculosis	18	18

4.1. Results and discussion

The influential proteins (or) genes of above ten biological networks are identified through Algorithm 1 of weak ties and presented in the Table 1 with data: the number of nodes, number of directed hyperedges in H_{WDG}, range of weak tie nodes using Mean, SD and quartile.

Various graph centrality measures like Degree Centrality (DC_G), Closeness Centrality(CC_G), Eigen Vector Centrality (EC_G) and Harmonic Centrality (HC_G) are compared with the proposed technique. Table 2 presents the influential nodes of the above explained ten biological networks.

The minimum number of influential nodes are to be derived which are responsible to maximize the influence to entire network. It is apparent from the values tabulated, our proposed work yields the minimum number of influential nodes both in mean, SD and quartile when comparing with the other centrality measures expect the bio-grid-fission-yeast's. In bio-grid-fission-yeast's network the number of the influential nodes using quartile is greater than the existing centralities.

Fig. 1 illustrates the count of edges of graph and hyperedges, it is apparent that number of hyperedges is much lesser than the number of edges of graph. Figs. 2 and 3 depicts the comparison of degree centralities of the proposed work based on mean-standard deviation, and quartiles-quartile deviation respectively, with the graph based centralities.

5. COVID-19 validation

In this section, we intend to validate our technique for COVID-19 protein-protein interaction. On Dec 8, 2019, the coronavirus (COVID-19) had identified in the seafood market in the Wuhan city of China. Coronavirus is one of a kind belonging to severe acute respiratory syndrome (SARS) virus. The world health organization (WHO) declared coronavirus as a pandemic.

Coronavirus (SARS-COV-2 or COVID-19) is one of the family members of Coronaviridae and order Nidovirales. This family contains two subfamilies, namely, Coronavirnae and Torovirinae. The Coronavirnae are classified into four categories:

- Alphacoronavirus - which consists of human coronavirus (HCOV)
- Betacoronavirus - which includes of human coronavirus (HCOV) with the SARS-COV-2 virus.
- Gammacoronavirus - which includes the viruses of bird and whales.

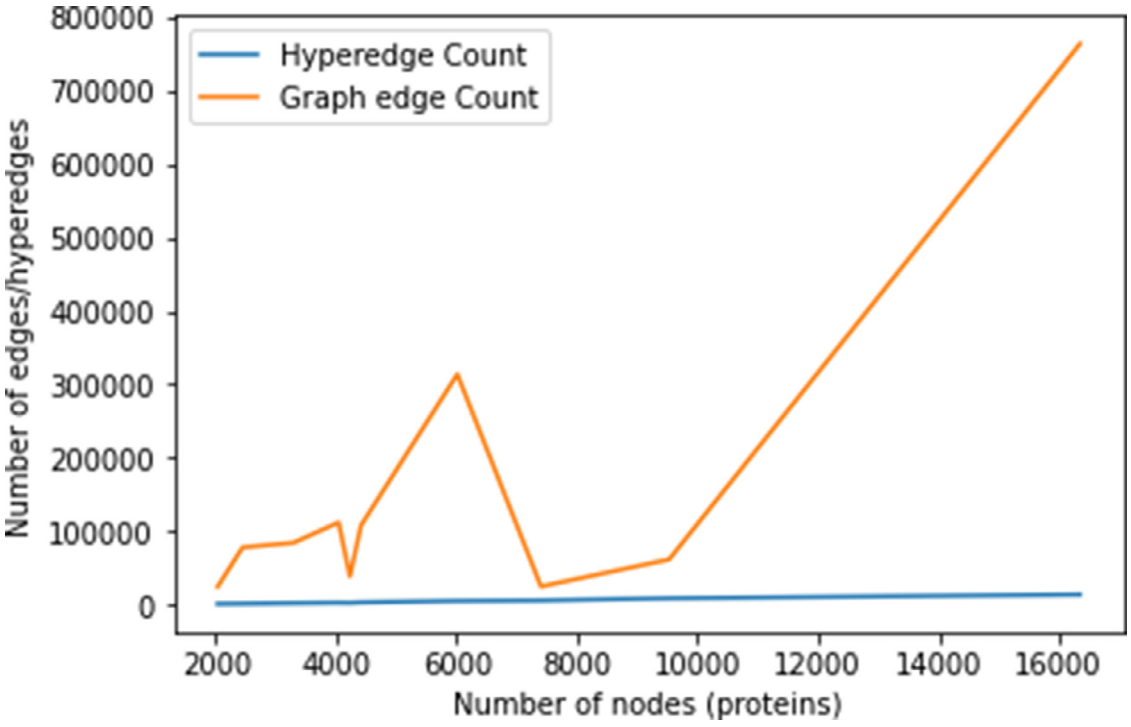

Fig. 1. Comparison of edge and hyperedge count.

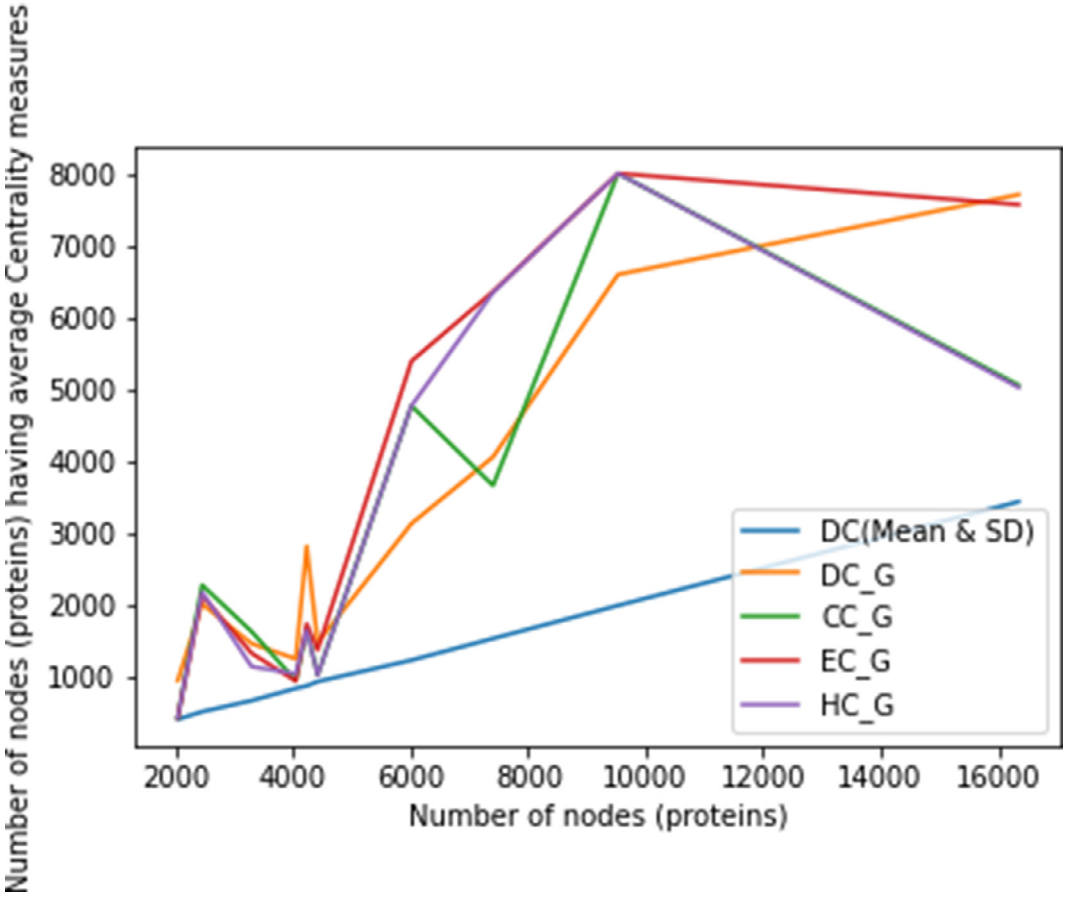

Fig. 2. Comparison of degree centrality of graph with hypergraph (Using mean and SD).

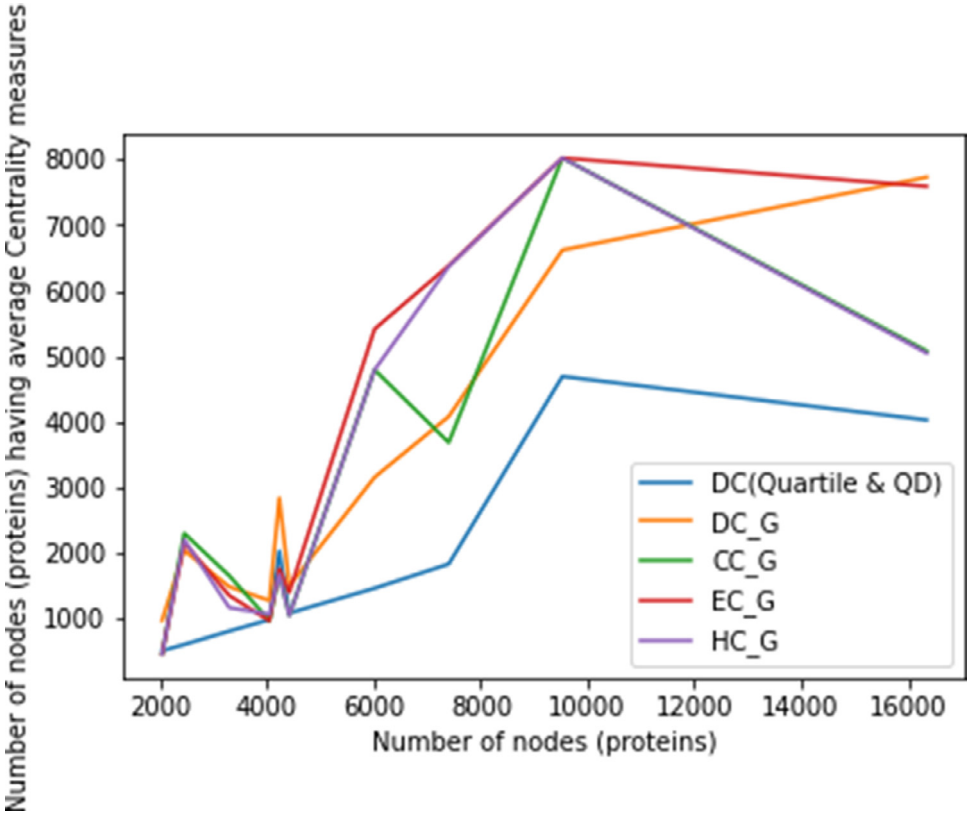

Fig. 3. Comparison of degree centrality of graph with hypergraph (Using quartiles).

- Deltacoronavirus - which consists of viruses which are isolated from birds and pigs.

The COVID-19 is Betacoronavirus together with the impact of viruses, namely, middle SARS viruses and pathogenic viruses. From the biological laboratory results [2,6,11,19] some crucial proteins have been identified, that plays a vital role in the protein-protein interactions(PPI's) of COVID-19 with the human body. And, some of these essential proteins belong to the enzymes TMPRSS2 (Transmembrane protease, serine 2), ACE2 (Angiotensin - Converting Enzyme 2), and AT2 (Angiotensin II).

The COVID-19 protein-protein interactions (PPI's) from [7] has been constructed as directed hypergraph. Here the nodes are the proteins and the directed hyperedge is constructed if there is an interaction between viral protein with human host proteins and human host protein with human proteins. Now, the directed hypergraph is transformed to weighted directed hypergraph by assigning the weights as the number of PPI's. These weights are tuned using GA and the classification of weak tie proteins is summarized in Table 3.

Proteins in TMPRSS2, ACE2, AT-II enzymes are 47, 4 and 11, respectively, in the cleaned SARS COVID II and human interactome data-set [8]. These proteins act as a major cause of various disease. We had also identified the proteins which cause the cytoplasmic, cytokine storm, chronic obstructive pulmonary disease, lower respiratory infections, blood pressure, diabetes mellitus, stroke, tuberculosis.

6. Conclusion

In this work, hypergraph is being exploited as a more powerful tool that reduces the complexity considerably compared to graphs as the weighted directed hypergraph of any network has fewer directed hyperedges. The influential nodes of the network are obtained by weak ties of degree centrality. The weights of the nodes are tuned using GA is employed by combining the Roulette wheel and ranking selection. The empirical results obtained from the computation show that proposed work perform better than other graph based centrality measures. Also, obtained critical proteins which play an influential role in COVID-19 viral interactions. These proteins may be a direct or indirect host of the COVID-19 viral protein and useful in drug design. For big data the elapsed time proliferates in identifying the influential nodes by the proposed technique. In the future, a suitable dimensionality reduction scheme will be introduced along with a congenial evolutionary algorithm to handle big data efficiently. Also, the expected protein interactome will be verified by protein docking based on the different mathematical modelling.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Sathyanarayanan Gopalakrishnan and Swaminathan Venkataraman thank the Department of Science and Technology - Fund Improvement of S&T Infrastructure in Universities and Higher Educational Institutions Government of India (SR/FST/MSI - 107/2015) for carrying out this research work and, the author Supriya Sridharan wishes to express sincere thanks to the INSPIRE fellowship (DST/INSPIRE Fellowship/2019/IF190271) for their financial support.

References

[1] C. Berge, Hypergraphs North Holland mathematical library, 1989,

[2] B.J. De Witt, E.A. Garrison, H.C. Champion, P.J. Kadowitz, L-163,491 is a partial angiotensin at1 receptor agonist in the hindquarters vascular bed of the cat, Eur. J. Pharmacol. 404 (1-2) (2000) 213–219.
[3] G. Gallo, G. Longo, S. Pallottino, S. Nguyen, Directed hypergraphs and applications, Discrete Appl. Math. 42 (1993) 177–201, doi:10.1016/0166-218X(93)90045-P.
[4] M. Gen, R. Cheng, Genetic Algorithms and Engineering Optimization, vol. 7, John Wiley & Sons, 1999.
[5] M.S. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) 1360–1380.
[6] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T.S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, et al., SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor, Cell 181 (2) (2020) 271–280.
[7] INTACT contributors, Coronavirus proteins interactions data-set, 2020a, https://www.ebi.ac.uk/intact/.
[8] INTACT contributors, Enzymes data-set, 2020b, https://www.uniprot.org/uniprot/?query=tmprss2(/ace2/at-II)&sort=score.
[9] K. Kapoor, D. Sharma, J. Srivastava, Weighted node degree centrality for hypergraphs, in: 2013 IEEE 2nd Network Science Workshop (NSW), IEEE, 2013, pp. 152–155.
[10] R. Kumar, et al., Blending roulette wheel selection & rank selection in genetic algorithms, Int. J. Mach. Learn. Comput. 2 (4) (2012) 365–370.
[11] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, et al., Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, Nature 581 (7807) (2020) 215–220.
[12] G. Muhammad, M.S. Hossain, COVID-19 and non-COVID-19 classification using multi-layers fusion from lung ultrasound images, Inf. Fusion 72 (2021) 80–88.
[13] T. Opsahl, F. Agneessens, J. Skvoretz, Node centrality in weighted networks: generalizing degree and shortest paths, Soc. Netw. 32 (2010) 245–251, doi:10.1016/j.socnet.2010.03.006.
[14] R.A. Rossi, N.K. Ahmed, The network data repository with interactive graph analytics and visualization, 2015, http://networkrepository.com.
[15] S. Roy, B. Ravindran, Measuring network centrality using hypergraphs, in: Proceedings of the Second ACM IKDD Conference on Data Sciences, 2015, pp. 59–68.
[16] K. Shankar, E. Perumal, V.G. Díaz, P. Tiwari, D. Gupta, A.K.J. Saudagar, K. Muhammad, An optimal cascaded recurrent neural network for intelligent COVID-19 detection using chest X-ray images, Appl. Soft Comput. 113 (2021) 107878.
[17] R. Vagnetti, M.C. Pino, F. Masedu, S. Peretti, I. Le Donne, R. Rossi, M. Valenti, M. Mazza, Exploring the social cognition network in young adults with autism spectrum disorder using graph analysis, Brain Behav. 10 (3) (2020) e01524.
[18] Wikipedia contributors, Social network—Wikipedia, the free encyclopedia, 2004, [Online; accessed 22-February-2020], https://en.wikipedia.org/wiki/Socialnetwork.
[19] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2, Science 367 (6485) (2020) 1444–1448.
[20] Y. Yoshida, Almost linear-time algorithms for adaptive betweenness centrality using hypergraph sketches, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1416–1425.
[21] S. Zhang, S. Cui, Z. Ding, Hypergraph spectral analysis and processing in 3D point cloud, arXiv preprint arXiv:2001.02384 (2020).