Evolution of protein-protein interaction networks from influential seeding

RA Keerthan * **J Mahesh** **

* *IIT Madras, Chennai, India (e-mail: ch17b078@smail.iitm.ac.in)*
** *IIT Madras, Chennai, India (e-mail: ch17b049@smail.iitm.ac.in)*

Abstract: Modelling of protein-protein interaction networks (PPIN) has been under the spotlight for several decades owing to its colossal impact on the understanding of several biological processes and mechanisms. An important component of growth-based PPIN models is the *seed* graph from which the model is grown. On the same lines, (1) proposed that choosing the "right" *seed* graph is essential for obtaining an accurate model of a PPIN. The authors chose a *seed* graph that included a fully connected sub-graph of maximum size in the PPIN. In this paper, we attempt to select the nodes of *seed* graph as the minimal set of high influence nodes (2) present in PPIN. Following this, we adopt the duplicate and diverge algorithm for proteome evolution (3) to model the PPIN. For evaluation of the model, the *network-centrality matrix*(4) is constructed for each network. The dimensionality-reduction technique of Principal Component Analysis (PCA) is applied on the *network-centrality matrix*. Through PCA, we obtain the set of vectors along which the centrality measures vary the most. We compare the distribution formed by the first two Principal Components (PCs) of the original network against the generated networks using the two-sample KS Test. We believe that PCA will allow us to combine different centrality measures while comparing the networks.

Keywords: Protein-protein interaction networks, seed graph, proteome evolution, Principal Component Analysis

1. INTRODUCTION

The study of evolution of protein-protein interaction networks (PPIN) has been a key research area in network biology owing to its importance in understanding several biological processes. One of the widely recognized approaches to approximate the evolution of PPIN is the duplicate-diverge model that was proposed by (3). However, (3) did not discuss the choice of the right seed network for network growth. To study the impact of the choice of seed network, (1) proposed to choose two cliques of maximum sizes present in the original PPIN data. We propose an alternative method of constructing the seed using the most influential nodes as found by the Collective Influence algorithm discussed in (2).

In addition to the endeavour of building a model, evaluating the model in the context of how closely it resembles the original network is not an easy task. With numerous metrics, statistics and centrality measures describing the properties of a network, it is insufficient to choose a single such statistic to numerically estimate the similarity between the generated and original networks. The relative significance of these statistics is context-dependent and usually unknown. Therefore, we attempt to combine multiple centrality measures using PCA to evaluate the closeness of the generated network with the original network.

1.1 Objectives

The objectives of this paper are:

- ➔ To implement the Duplicate-Diverge algorithm of Network Growth and apply it to model the yeast PPIN (CCSB-Y2H) and worm interactome (WI-2007) datasets.
- ➔ Implement the Collective Influence algorithm to identify the most influential nodes in a network, and following that, use these nodes to construct a seed network for growth.
- ➔ Compare the performance of using the Maximum Size Cliques Seed and the Most Influential Nodes Seed on the network growth model by applying PCA-based dimensionality reduction on the network-centrality matrix.

The rest of the paper is organized as follows: section 2 will brief about relevant existing methods, section 3 will discuss our contributions, section 4 will detail the experimentation results and finally we conclude with section 5.

2. METHODS

2.1 Duplicate-Diverge model

The Duplicate-Diverge algorithm for growth of a network model, as proposed by (3), is as follows:

- Select a node at random and duplicate it preserving its connectivity.
- With probability δ, remove an edge emanating from the newly formed node.
- With probability α, form an edge between newly formed node and rest of the nodes.

The duplication step involves creation of a new node with same structure as that of the original node. Note that the creation and deletion of links are with respect to the newly formed node. As inferred from the above steps, there are two parameters δ, the *new node edge removal probability* and α, the *new node edge addition probability.*

The algorithm shall be run until the desired network size is achieved. As mentioned in (1), one of the drawbacks of this growth model is that it produces many singletons (nodes without edges), unlike in known PPINs where singletons are rarely observed. In cases when the parameter α is close to 0 and the parameter δ is higher than 0.5, singletons are especially prominent. The modification to address this issue, as was proposed in (1), is to delete any singleton nodes if they form after any step.

2.2 Clique based seeding

A seed graph is defined as the graph from which network evolution takes place. With this definition, the seed graph can also be viewed as the core of PPIN. As stated by (1), an important observation while obtaining an optimal seed is that the duplicate and diverge model will not produce large cliques. Therefore, (1) proposed to set the seed graph to the top two largest cliques present in the PPIN. However, the time complexity of finding a clique in a network with n nodes is $O(3^{n/3})$ (8) whereas finding influential nodes as described in subsection 2.3 has the time complexity of $O(nlogn)$ (7).

2.3 Finding influential nodes

Identification of important nodes have become pivotal in modern day network analysis. In social networks, identification of influential nodes is helpful for tailoring marketing strategies, which could potentially boost business. In the context of biological networks, identification of influential nodes plays a crucial role in many applications such as drug target identification and large scale information diffusion. We adopt the approach presented in (2) to identify the *minimal set* of influential nodes whose removal will cause the whole network to collapse. The authors of (2) used an influence maximization formulation to identify the minimal set of influential nodes. The idea is to remove nodes in decreasing order of *Collective Influence*, denoted as **CI**, until the network breaks down.

$$\mathbf{CI}_L(i) = (k(i) - 1) \sum_{j \in Ball(i,L)} (k(j) - 1) \qquad (1)$$

where $Ball(i, L)$ is the collection of all nodes that are L shortest path away from node i, $k(i)$ denotes the degree of node i. The event of collapse of a network is algebraically identified when the largest eigenvalue of the *coupling matrix* equals unity. For more details on the same, refer to (2) and (7).

3. INFLUENTIAL SEEDING AND PCA-CENTRIC EVALUATION

In this section, we will discuss in detail the key contributions of the present work. Our first contribution is the determination of N_{seed} number of influential nodes as the seed graph while maintaining the inter-node connectivity, if any. That is, the subgraph formed by these influential nodes is taken as the seed. Let total number of nodes in the graph be N and total number of influential nodes be N_{inf}. We define N_{seed} to be around $min(1.5\%$ of N, $N_{inf})$. In the case where N_{seed} equals 1.5% of N, top $0.015N$ influential nodes were selected out of N_{inf} influential nodes.This definition of N_{seed} has no particular mathematical relevance but rather is defined in such a way that we have a reasonable amount of nodes in the seed graph to start developing the PPIN. Once we are set with the seed graph, we will follow the Duplicate-Diverge model discussed in subsection 2.1 to simulate the evolution of PPIN.

The choice of metric for evaluation of the generated network is essential to corroborate the efficiency of the proposed approach. However, in most cases, choice of such metric remains a dilemma. To help us resolve this, (4) proposed a PCA based method to identify important centrality measures by constructing a *network-centrality* matrix, a matrix with number of rows and columns equal to number of nodes and centrality measures used respectively. In the present work, we slightly modified their technique by projecting the network-centrality matrix along the top two principal components and using the reduced matrix for comparison of networks. This idea was inspired by the ability of principal components to capture latent relationships between the centrality measures present in the network-centrality matrix. In this work, we chose Degree Centrality, Betweenness Centrality and Closeness Centrality as the measures while constructing the network-centrality matrix. In section 4, we will look in detail how the comparison is made between the principal components of the network-centrality matrices corresponding to different networks.

4. RESULTS AND DISCUSSION

All computation that is presented in this work were implemented in a Jupyter Notebook using Python 3.7 with the assistance of the Python library NetworkX. We used yeast PPIN [CCSB-Y2H (5)] and worm interactome [WI-2007 (6)] datasets for testing the model. The CCSB-Y2H dataset consisted of 1280 proteins and 1810 interactions (inclusive of self-loops). WI-2007 dataset consisted of 1498 proteins and 1817 interactions (inclusive of self-loops).

4.1 Parameter Estimation

The performance of the model is governed by the two parameters: α and δ. These parameters were estimated using a grid-search algorithm that minimised the deviations observed in two key properties: number of edges and the degree distribution of the generated network. The deviation was calculated with respect to the original network. This ensured that the best fit parameters represented a model that best corresponds to the original network. The objective function used for this purpose is shown in Equation 2:

$$L(G_m, G_o) = (\epsilon + |E_m - Eo|) \times \frac{\left(\frac{1}{Z_{deg}+1} + Z_{deg} + 1\right)}{2} \qquad (2)$$

Where,

- G_m, G_o stand for the model-generated graph and original network respectively.
- E_m, E_o stand for the number of edges in their respective graphs.
- ϵ is a non-negative number to make objective function non-zero. Our chosen value for ϵ was 1.
- Z_{deg} is the KL Divergence (KLD) between the degree centrality distributions of the original graph and the network model.

The objective function consists of an edge difference term multiplied by a KLD factor term. The fit of degree centrality distribution is encoded by the use of KL Divergence which is defined as:

$$H(p,q) = \sum p(x) log\left(\frac{p(x)}{q(x)}\right)$$

Where $p(x)$ and $q(x)$ corresponds to degree centrality distributions of generated and original network respectively. The edge difference has a minimal value of 0, when the number of edges of the generated graphs are equal to that of the original network, whereas the KLD factor has a minimum value of 1, at 0 KL Divergence, which is achieved when the degree distribution of the original and the generated networks are identical. This objective function will ensure that the optimal network will minimise the relative difference in number of edges generated while also minimising the deviation between the degree distributions of the original and the generated network.

Direct Estimation: An alternative approach of estimating α and β is proposed by (1). It is based on the assumption that the average degree of the network asymptotically approaches a fixed value. Under the approximation that this asymptotic average degree is equal to the observed average degree, a relation between the parameters of the network is obtained.

$$a = \frac{2\delta}{1 - Pr_s - 2\alpha} \tag{3}$$

Where a is the asymptotic average degree of the network. Pr_s is the asymptotic probability that a newly added node ends up as a singleton (before it is removed). For a network with N number of nodes, the expression of Pr_s may be obtained.

$$Pr_s = \sum_{k=1}^{N} \frac{n_{d=k}}{N} \delta^k (1 - \frac{\alpha}{N})^{N-k} \tag{4}$$

This expression does not depend on N asymptotically, as the value $\frac{n_{d=k}}{N}$ vanishes for large values of k. Therefore, Equation 3 is used to obtain a relation between α and δ by setting the value of a to the average degree of the original network. This relation, along with the bounds on the values of α and δ, may be used in an optimization problem with a suitable objective function (in our case, minimizing the value of α) from which the values of α and δ are obtained. This procedure is independent of the chosen seed of the network.

Following this procedure for the CCSB-Y2H network, the values of α and δ were obtained as 0.083 and 0.583. For the WI-2007 network, the values of α and δ were obtained as 0.034 and 0.720.

These estimates were then used to generate 20 networks out of the two seed networks: Clique-based and Influential-based seeds, explored in detail in subsection 4.2 and subsection 4.3.

	Original	Clique Seed	Influential Seed
Edge Count	1810	102958.25	102883.85
Average Degree	2.828	160.87	160.76

Table 1. Original and seed-based network statistics through Direct Estimation for the CCSB-Y2H dataset.

In Table 1, Clique Seed and Influential Seed refer to the networks generated using Duplicate-Diverge model, grown out of the Max Cliques based seed and the Influential Nodes based seed respectively.

	Original	Clique Seed	Influential Seed
Edge Count	1817	51150.35	51177.3
Average Degree	2.426	68.29	68.33

Table 2. Original and seed-based network statistics through Direct Estimation for the WI-2007 dataset.

Statistics of the seed-based networks in Table 1 and Table 2 were averaged over 20 generated networks. As observed from Table 1 and Table 2, the average number of edges in the generated networks greatly differ from those in the original network, leading to highly inaccurate models. The reason for this is assumed to be that the asymptotic average degree approximation does not hold for the datasets chosen, likely due to their relatively small size. Therefore, the grid search approach was preferred over direct estimation for further analysis.

4.2 Clique based seeding

The largest clique size in the CCSB-Y2H graph was found to be 4. Eight such 4-cliques were present in the network, none of them connected to each other. Therefore, two out of the eight 4-cliques were chosen at random, and the seed graph was constructed accordingly. The growth algorithm was run until the number of nodes reached the original amount (1280).

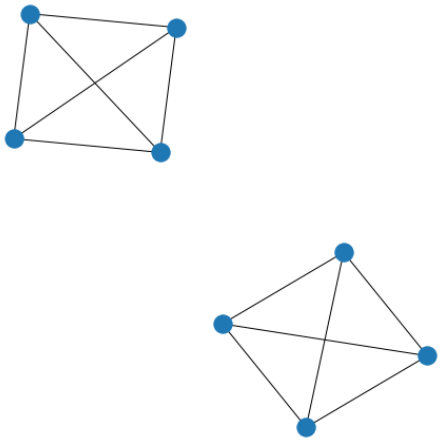

Fig. 1. Maximum-Size Clique-based Seed Graph for CCSB-Y2H

Similarly, The largest clique size in the WI-2007 graph was found to be 4. Five such 4-cliques were present in the network, none of them connected to each other. Therefore, two 4-cliques were chosen at random, and the seed graph was constructed accordingly. The network was grown until the nodes of generated network equalled that of the original network (1489).

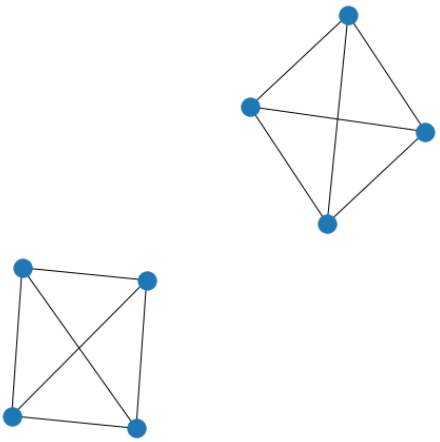

Fig. 2. Maximum-Size Clique-based Seed Graph for WI-2007

4.3 Influential seeding

The proposed influential seeding algorithm that is described in section 3 was implemented on CCSB-Y2H and WI-2007 dataset. The total number of influential nodes was more than 1.5% of total nodes in the network. The seed graph for CCSB-Y2H consisted of 20 nodes as shown in Figure 3.

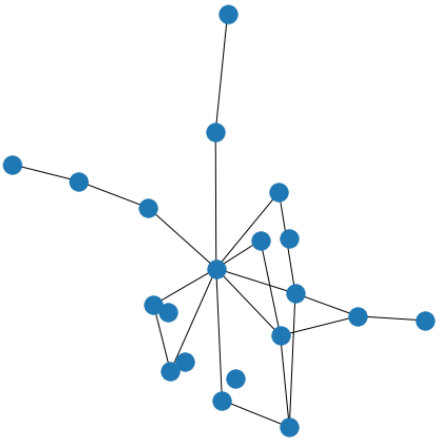

Fig. 3. Most Influential Nodes-based Seed Graph for CCSB-Y2H

WI-2007 dataset also had influential nodes constituting more than 1.5% of total nodes in the graph. Therefore, top 20 influential nodes of the WI-2007 dataset were chosen for the seed graph. The network was grown until the nodes of generated network equalled that of the original network (1489).

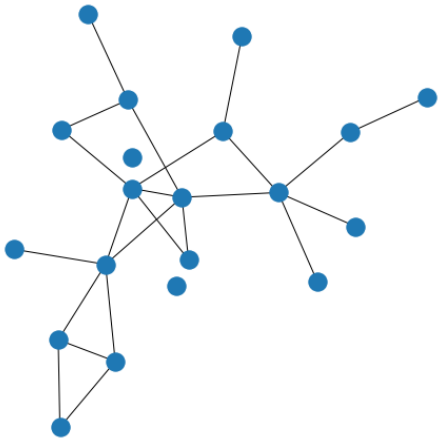

Fig. 4. Most Influential Nodes-based Seed Graph for WI-2007

It may be noted that Figure 4 and Figure 3 contain singleton nodes. In the process of network growth, these nodes were retained and not removed, however any new singletons formed during growth were removed as per the algorithm.

4.4 Results

CCSB-Y2H Dataset: Following the grid-search based parameter estimation technique described in subsection 4.1, the optimal values for α and δ were estimated, for both the clique-based and the influential-based seed, as given in Table 3.

	Clique Seed	Influential Seed
α	0.001	0.001
δ	0.790	0.822

Table 3. Estimated parameters for the CCSB-Y2H network.

Using the values reported in Table 3, 20 networks were generated using each of the two seeds. The descriptive statistics of the network-sets were found and averaged, and subsequently compared with those of the original network.

	Original	Clique Seed	Influential Seed
Edge Count	1810	1848	1769.4
Average Degree	2.828	2.888	2.765

Table 4. Original and seed-based generated statistics for the CCSB-Y2H network.

The three chosen centrality distributions were averaged for each seed type, and the resultant distributions were compared against that of the original network.

In order to collectively compare the two seed-based networks using the three centrality measures, the network-centrality matrix was constructed. The network-centrality matrix for CCSB-Y2H is of dimension 1280×3. In total, we construct three network-centrality matrices, one each for the original, clique-based and influential-based seed generated network. PCA is carried out on the original network centrality matrix and the network-centrality matrix was projected onto the top two principal components, effectively resulting in a reduced centrality matrix of dimension 1280×2. The idea behind this reduction is that the PCs

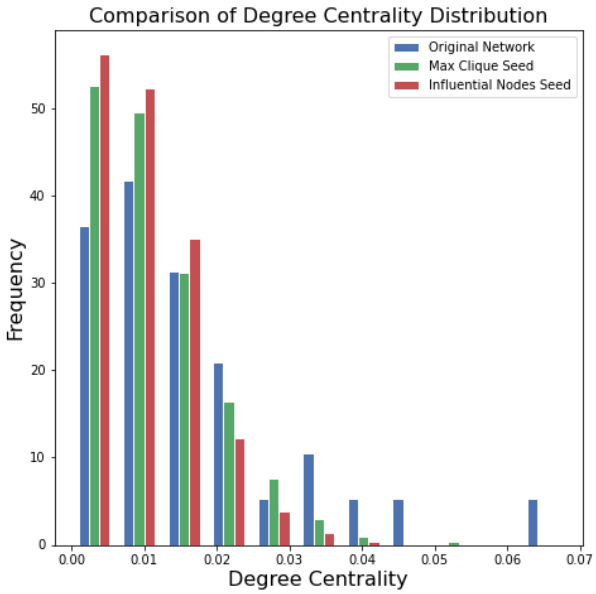

Fig. 5. Degree Centrality Plot for CCSB-Y2H

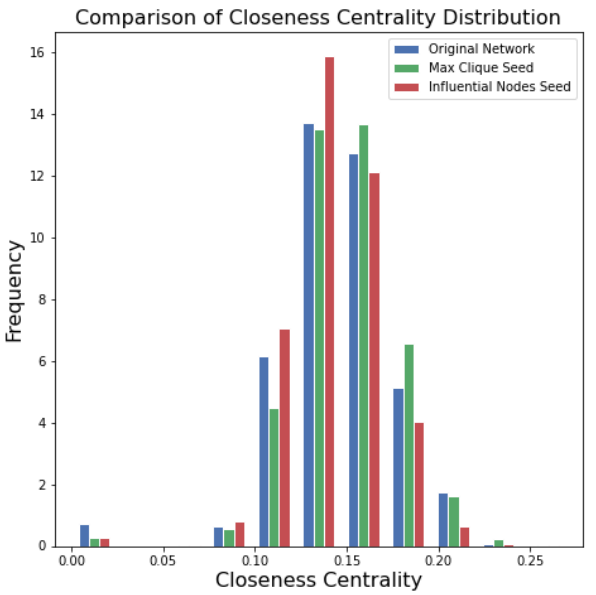

Fig. 6. Closeness Centrality Plot for CCSB-Y2H

represent a weighted combination of the centrality measures, with the weights signifying the relative importance of each measure.

The network-centrality matrices of the two seed-based networks were projected onto the space spanned by the top two PCs of the original network. Thus, the resultant three reduced centrality matrices were compared.

Figure 8 shows two Cloud Plots, each showing the distribution region of the values along the two PCs as loosely-defined clouds, with denser cloud regions indicat-

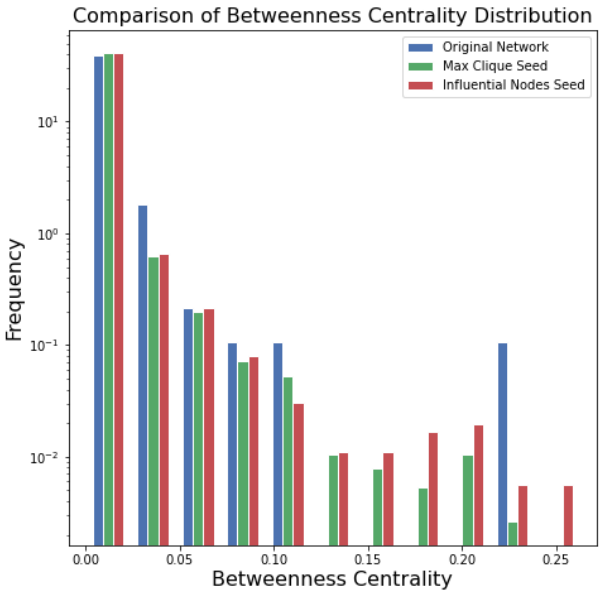

Fig. 7. Betweenness Centrality Plot for CCSB-Y2H

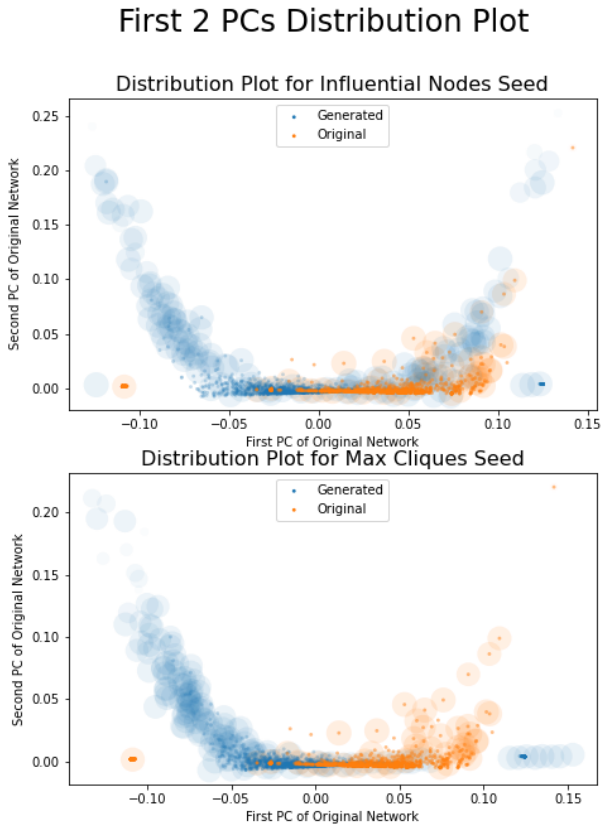

Fig. 8. Original and seed-based generated network-centrality matrices, reduced along the top 2 PCs of the original network-centrality matrix, for the CCSB-Y2H dataset.

ing greater concentration of points. Apart from the clouds, the plots also include a subset of the actual points comprising the distribution. Due to the presence of large number of points, only a subset is displayed, however, the cloudy regions approximately represent the distribution of the entire set of points.

In order to quantify the similarity of the distributions of these plots, the two-sample KS test of similarity was used. The KS test was performed to compare the original network distribution against the clique-based seed distribution, and the original network against the influential nodes-based seed distribution. The Test Statistic values are given in Table 5.

	Clique vs Original	Influential vs Original
KS Test Statistic	0.431	0.378

Table 5. KS Test Statistics of reduced centrality matrix of generated vs original network, for the CCSB-Y2H dataset.

The KS test statistic provides a quantitative measure of the deviation between two distributions. For two sets of values sampled from identical and independent distributions, the test statistic would be close to 0.

WI-2007 Dataset: The same approach as subsubsection 4.4.1 was used to compare the networks generated using the two seed-based approaches on the WI-2007 dataset.

	Clique Seed	Influential Seed
α	0.001	0.001
δ	0.895	0.895

Table 6. Estimated parameters for the WI-2007 network.

	Original	Clique Seed	Influential Seed
Edge Count	1817	1781.4	1821.4
Average Degree	2.426	2.378	2.432

Table 7. Original and seed-based generated statistics for the WI-2007 network.

The three network-centrality matrices were constructed, similar to the manner in which they were constructed for the CCSB-Y2H dataset. These matrices were subsequently reduced to be along the directions of the first two PCs of the network-centrality matrix corresponding to the original network. The Cloud Plots of the distributions obtained are given in Figure 12.

The distributions plotted in Figure 12 are compared using the two-sample KS Test for similarity.

	Clique vs Original	Influential vs Original
KS Test Statistic	0.449	0.446

Table 8. KS Test Statistics of reduced centrality matrix of generated vs original network, for the WI-2007 dataset.

4.5 Inference

In this subsection, we will discuss the results of the comparisons made in subsection 4.4.

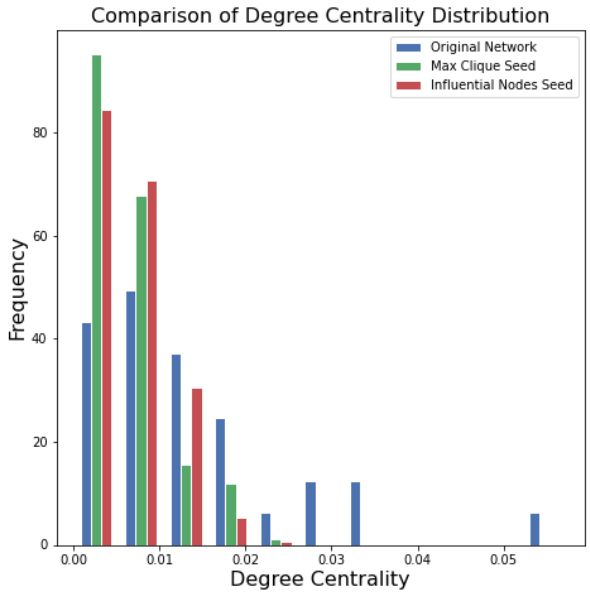

Fig. 9. Degree Centrality Plot for WI-2007

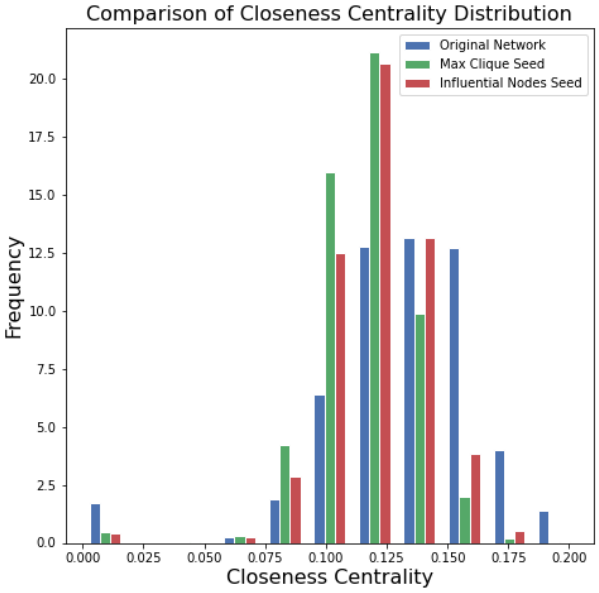

Fig. 10. Closeness Centrality Plot for WI-2007

Efficiency on CCSB-Y2H dataset: From Table 3, we can observe that the estimated parameters using the two approaches have the same α whereas δ of influential seed approach is larger than that of clique seed approach, which resulted in a slightly lower average degree, due to more frequent edge removal. By looking at the plots of degree centrality, closeness centrality and betweenness centrality distributions from Figure 5, Figure 6 and Figure 7 respectively, we can infer that both the influential seed and clique based approach follow the same distributions as that of the original network. The same is suggested by Table 4 where

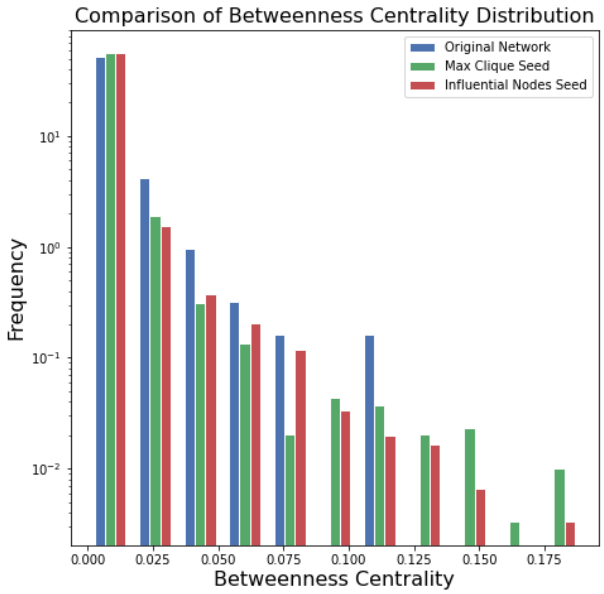

Fig. 11. Betweenness Centrality Plot for WI-2007

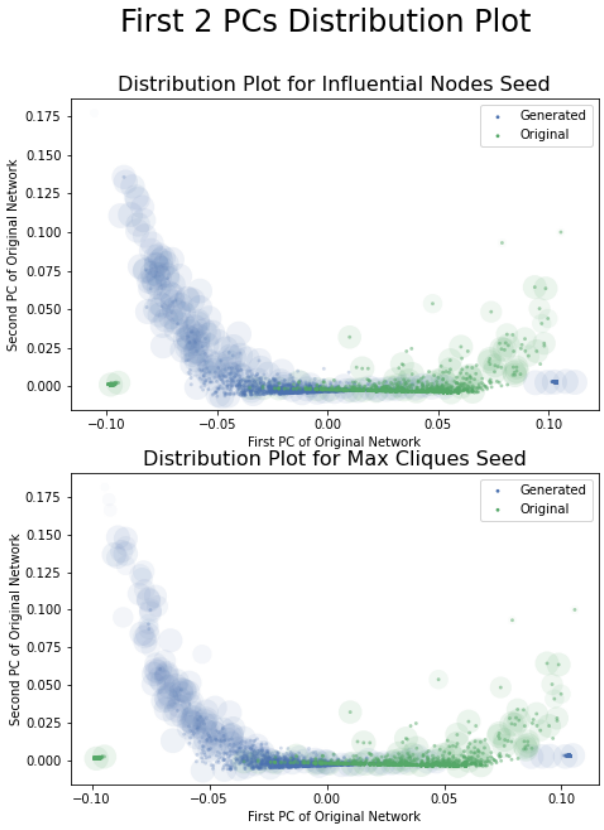

Fig. 12. Original and seed-based generated network-centrality matrices, reduced along the top 2 PCs of the original network-centrality matrix, for the WI-2007 dataset.

the average count of edges and average degree across 20 realisations of the generated networks are close to that of the original network. To further examine the extent of similarity between distributions, we refer to Figure 8 and Table 5. From Figure 8, we can infer that there is a dense overlap between the top two principal components of network generated by influential seed and original network. In Table 5, we observe the KS-test statistic values, and it is seen that the KS-statistic of influential seed approach is lower than that of clique-based approach. The lower the KS-statistic, the better is the match between the two distributions. Therefore, we can conclude that for CCSB-Y2H dataset, influential seed approach is more efficient than clique based approach.

Efficiency on WI-2007 dataset: We perform similar inference for WI-2007 dataset like we did for Y2H-datset. From Table 6 we can observe that the estimated parameters using the two approaches have the same α and δ, indicating similar growth patterns. The plots of degree centrality, closeness centrality and betweenness centrality distributions from Figure 9, Figure 10 and Figure 11 respectively, also indicate that the distributions of the two approaches are quite similar and are capable of capturing the respective centrality distributions of the original network. Finally, the PCA cloud plot and KS-statistics from Figure 12 and Table 8 respectively, confirms that the influential seeding and clique-based approach indeed exhibit similar performance on the WI-2007 dataset.

5. CONCLUSION

In this work, we have analysed the existing seed selection algorithm which was based on finding cliques present in the network. As an alternative, we constructed the seed graph with the most influential nodes of the original network. We compared the two methods by analysing the PCA-based reduced-dimension network-centrality matrix. Results showed us that the proposed method performs better than clique-based method on CCSB-Y2H dataset and equalled the performance on WI-2007 dataset. However, the current work relies on grid-search for parameter estimation, which is not easily scalable in comparison to other methods. Therefore, we believe that a more optimized parameter estimation technique for estimation of α and β could be a potential research area in the future.

REFERENCES

[1] Hormozdiari F, Berenbrink P, Prz˘ulj N, Sahinalp SC (2007) *Not all scale-free networks are born equal: The role of the seed graph in PPI network evolution.* PLOS Computational Biology 3(7): e118. doi:10.1371/journal.pcbi.0030118

[2] Peng Gang Sun, Yi Ning Quan, Qi Guang Miao, Juan Chi, *Identifying influential genes in protein–protein interaction networks*, Information Sciences, Volumes 454–455, 2018, Pages 229-241, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2018.04.078.

[3] Pastor-Satorras R, Smith E, Solé RV. *Evolving protein interaction networks through gene duplication.* J Theor Biol.2003 May 21;222(2):199-210. doi: 10.1016/s0022-5193(03)00028-6. PMID: 12727455.

[4] Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z. et al. *A systematic survey of centrality measures for protein-protein interaction networks. BMC Syst Biol 12, 80 (2018).* https://doi.org/10.1186/s12918-018-0598-2

[5] CCSB-Y2H Dataset, Yeast Interactome Project, CCSB Interactome Database, Harvard Medical School.

[6] WI-2007 Dataset, Worm Interactome Project, CCSB Interactome Database, Harvard Medical School.

[7] Morone, F., Min, B., Bo, L. et al. *Collective Influence Algorithm to find influencers via optimal percolation in massively large social media.* Sci Rep 6, 30062 (2016). https://doi.org/10.1038/srep30062

[8] Etsuji Tomita, Akira Tanaka, Haruhisa Takahashi, *The worst-case time complexity for generating all maximal cliques and computational experiments*, Theoretical Computer Science, Volume 363, Issue 1, 2006, https://doi.org/10.1016/j.tcs.2006.06.015.