Iranian Journal of Mathematical Chemistry, Vol. 1, No. 2, September 2010, pp. 81 – 96

IJMC

A fast approach to the detection of all-purpose hubs in complex networks with chemical applications

SARAH MICHELE RAJTMAJER[1] AND DAMIR VUKIČEVIĆ[2,•]

[1]*IMC, University of Dubrovnik, Cira Carica 4, HR-20000 Dubrovnik, Croatia*
[2]*Department of Mathematics, University of Split, Teslina 12, HR-21000 Split, Croatia*

(*Received September 7, 2010*)

ABSTRACT

A novel algorithm for the fast detection of hubs in chemical networks is presented. The algorithm identifies a set of nodes in the network as most significant, aimed to be the most effective points of distribution for fast, widespread coverage throughout the system. We show that our hubs have in general greater closeness centrality and betweenness centrality than vertices with maximal degree, while having comparable or higher degree than vertices with greatest closeness centrality and betweenness centrality. As such, they serve as all-purpose network hubs. Several theoretical and real world chemical and biological networks are tested and results are analyzed.

Key words: Chemical networks; complex networks; network hubs; vertex centrality.

1 INTRODUCTION

Complex networks have proven effective models of many social, biological, and technological systems as diverse as neural networks [1], food webs [2], the Internet [3], the World Wide Web [4], systems of social interaction [5,6,7,8], acquaintance networks [9], scientific collaborations [10], problem solving networks [11], and linguistic networks [12]. In biology and chemistry in particular, complex networks are used to represent protein structures [13,14]. Whereas cell biology has traditionally identified proteins based on their individual rules as catalysts, signaling molecules, or building blocks of cells and microorganisms, recently a post-genomic view has expanded the protein's role, regarding it as an element in a network of protein-protein interactions as well. The local and global topological properties of these protein interaction networks can reveal protein structure and

• Author for correspondence. E-mail address: vukicevi@pmfst.hr

function. Also, complex networks play an important role in the analyses of interaction of drug fragments [15].

Furthermore, it has been shown that many real-world networks are „scale-free“, following a power law degree distribution [16,17]. The overwhelming majority of nodes in the network have relatively low degree, just a few connections, while a small percentage of “hubs” have very high degree. It has furthermore been shown that identifying and targeting hubs is an effective way to influence network behavior [18,19,20]. Hubs are the most influential nodes in the network, key to the spread of network processes and effects, and crucial with respect to network resilience to intentional attacks [21]. Indeed, this has been demonstrated to be true in protein interaction networks as well. Consider for example the protein interaction network of the yeast, *S. Cerevisiae*, which is described by 1879 proteins as nodes, connected by 2240 identified direct physical interactions [22,23]. It has been found that random mutations in the genome of *S. Cerevisiae*, modeled by the removal of randomly selected yeast proteins, do not affect the overall topology of the network. In contrast, when the most connected proteins are computationally eliminated, the network diameter increases rapidly [24].

Ultimately, however, there has not been established a precise definition of a network hub. Generally, three measures of vertex centrality are used – degree centrality, closeness centrality and betweenness centrality [25], and network hubs are usually taken as nodes with maximal degree centrality. In this paper, we calculate a novel connectivity value for each vertex in the network taking into consideration not only the degree of the vertex, but also its distances to other vertices in the system. We choose vertices with highest connectivity, so-called *distribution points*, as hubs. Our distribution points have greater closeness centrality and betweenness centrality than vertices with maximal degree, and on the whole, comparable or greater degree than vertices with maximal closeness centrality or betweenness centrality. We propose that this type of all-purpose centrality measure provides a more powerful classification of hubs takes into consideration not only the degree of each node.

2 HUB DETECTION

We consider the general case, for a network of n nodes. In the first step of the algorithm, each vertex is assigned a value according to the proximity of its neighbors. For each vertex i , calculate the *connectivity value of* i :

$$V(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{\propto^{d(i,j)-1}} \tag{1}$$

where d (i, j) is the shortest path distance between i and j .Choose i with maximal initial connectivity value as the Initial Distribution Point, DP_0 . If there are several vertices with equivalent values, they are all chosen as initial distribution points, the set { DP_0}.Proceed to choose the vertex or set of vertices with the next highest connectivity as {DP_1} , and continue to choose vertex hubs in order of their connectivity until all vertices have been chosen.

3 MODIFICATION OF THE ALGORITHM FOR VARYING DIFFUSION

Additionally, we can modify the algorithm to accommodate different kinds of spread in different types of networks, and enable adjustments to the hub detection algorithm for modeling varied effects. In social networks, for example, influence may aggregate, while modeling a power grid or an epidemic, we may wish to disallow compounded influence on one node. Accordingly, we include an additional weight in our connectivity value calculation, so-called *cumulative handicap*, assigned to vertices based on their proximity to a nearest prior-chosen distribution point. Higher weights indicate a preference to target a more widespread section of the network, at the expense of total cumulative impact.

Specifically, the modified algorithm proceeds as follows for each vertex i calculate the *initial connectivity value of i*:

$$V_0(i) = \sum_{\substack{i=1 \\ j \neq i}}^{n} \frac{1}{\propto^{d(i,j)-1}} \tag{2}$$

where d(i, j) is the shortest path distance between i and j. Choose i with maximal initial connectivity value as the Initial Distribution Point, DP_0. Again, if there are several vertices with equivalent values, they are all chosen as initial distribution points, the set DP_0. Next, assign a weight to each vertex in the network $\notin\{DP_0\}$ based on its distance to the nearest member of $\{DP_0\}$, given by $\{d(i, DP_0)\}$. That is:

$$W_0(i) = \frac{1}{\beta^{d(i,\{DP_0\})-1}} \tag{3}$$

The First Distribution Point(s) {DP_1}are chosen as the vertex or vertices which maximize the value of

$$V_1(i) = \max\left\{ \sum_{\substack{j=1 \\ j \neq i, j \notin \{DP_0\}}}^{n} \left(\frac{1}{\alpha^{d(i,j)-1}} - W_0(j) \right), 0 \right\}. \tag{4}$$

Subsequently, new weights are assigned to every vertex in the network (not an initial or first distribution point) :

$$W_1(i) = \max\left\{\frac{1}{\beta^{d(i,\{DP_0\})-1}}, \frac{1}{\beta^{d(i,\{DP_1\})-1}}\right\}. \quad (5)$$

The algorithm repeats; at the *th x* step, a vertex is assigned a weight based on its distance from the nearest distribution point, from the set of all distribution points chosen in the first $x-1$ steps. Vertices $\{DP_x\}$ are chosen as those which maximize

$$V_x(i) = \max\left\{\sum_{\substack{j=1 \\ j\neq i, j\notin\{DP\}}}^{n} \left(\frac{1}{\propto^{d(i,j)-1}} - W_{x-1}(j)\right), 0\right\}, \quad (6)$$

where $\{DP\}=\{\{DP_i\}_{i=0}^{x-1}\}$ is the set of all previous distribution points, and

$$W_{x-1}(j) = \max\left\{\frac{1}{\beta^{d(i,\{DP_0\})-1}}, \frac{1}{\beta^{d(i,\{DP_1\})-1}}, \cdots, \frac{1}{\beta^{d(i,\{DP_{x-1}\})-1}}\right\}. \quad (7)$$

The algorithm concludes when the connectivity values of all remaining vertices is 0.

4 MODIFICATION OF THE ALGORITHM FOR DIRECTED AND WEIGHTED GRAPHS

Many real-world networks are directed or weighted. The neural network of C. Elegans [1], an important model organisim, is both directed and weighted.

A natural adaptation of our notion of vertex connectivity can tackle these cases. In particular, for directed networks, connectivity of vertex *i* is again given by

$$V_x(i) = \max\left\{\sum_{\substack{j=1 \\ j\neq i, j\notin\{DP\}}}^{n} \left(\frac{1}{\propto^{d(i,j)-1}} - W_{x-1}(j)\right), 0\right\}, \quad (8)$$

where $d(i, j)$ is the distance from *i* to *j* given by length of the shortest directed path, and

$$W_{x-1}(j) = \max\left\{\frac{1}{\beta^{d(i,\{DP_0\})-1}}, \frac{1}{\beta^{d(i,\{DP_1\})-1}}, \cdots, \frac{1}{\beta^{d(i,\{DP_{x-1}\})-1}}\right\}. \quad (9)$$

where (d_i, DP_j) is the distance from *i* to DP_j given by length of the shortest directed path. Weighted networks follow similarly; we use shortest weighted path length as the distance measurement.

5 HUB TESTING

We test the validity of our choice of hubs by distribution points, considering degree centrality, closeness centrality and betweenness centrality [25]. We show on several

benchmark network graphs, that distribution points tend to have greater closeness centrality and betweenness centrality than vertices with maximal degree, while they have comparable or greater degree than vertices with maximal closeness centrality or maximal betweenness centrality. As such, we propose that distribution points serve as all-purpose, holistic hubs, crucial to various, diverse measures of connectivity within their networks.

For testing, we consider networks which accommodate aggregated influence on individual nodes. That is, we set the cumulative handicap to zero, $W_i(j) = 0, \forall(i,j)$, the original method presented in section 2.

We use α = 2 in order to balance the contributions of the first neighbors with the contributions of the vertices at a distance greater than 1.

Recall that in scale-free networks, there are a few nodes with very high degree while the vast majority of nodes have much smaller degree. Since we choose hubs one by one in order of significance, we need only consider the first small percentage of nodes chosen. We will examine up to 10 percent of graph vertices, though in reality the number of hubs in a network is even fewer. A powerful new set of benchmark graphs recently proposed by Santo Fortunato [26] make possible the generation of testing networks with demonstrated real-world properties, and with choice of various input parameters, including extent of mixing (μ), average degree (k), max degree (maxk), and sizes (minc and maxc) as well as overlap of network communities. We examine our results on several of these networks.

Figure 1 and Figure 2 illustrate a series of comparisons between distribution points and hubs chosen via the other three centrality measures on two of the Fortunato benchmarks with given parameters.

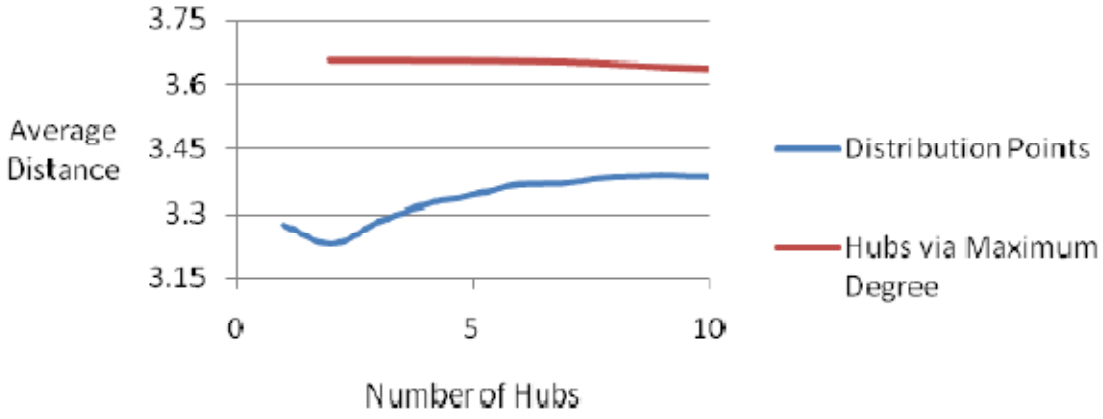

Figure 1a. Closeness centrality. Average distance from each chosen hub to all other vertices in the network. Fortunato benchmark graph (N = 100, k = 5, maxk = 10, μ = 0.1, minc = 5, maxc = 30).

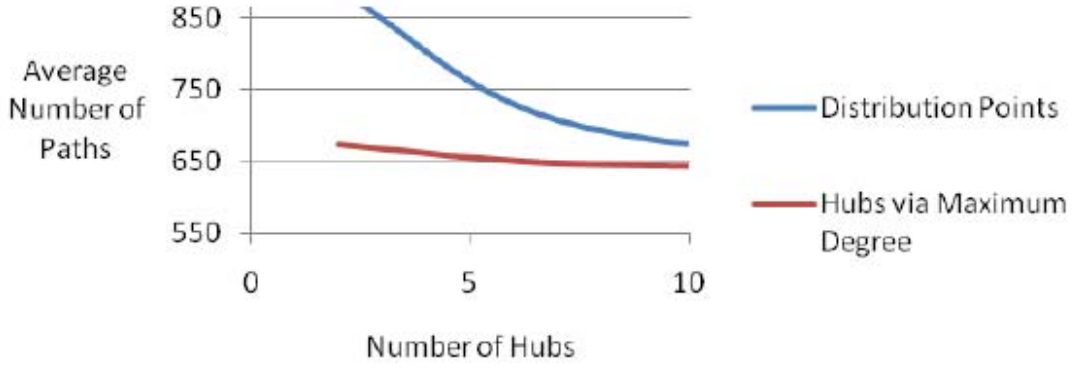

Figure 1b. Betweenness centrality. Average number of shortest paths running through each chosen hub.Fortunato benchmark graph (N = 100, k = 5, maxk = 10, μ=0.1, minc = 5, maxc = 30).

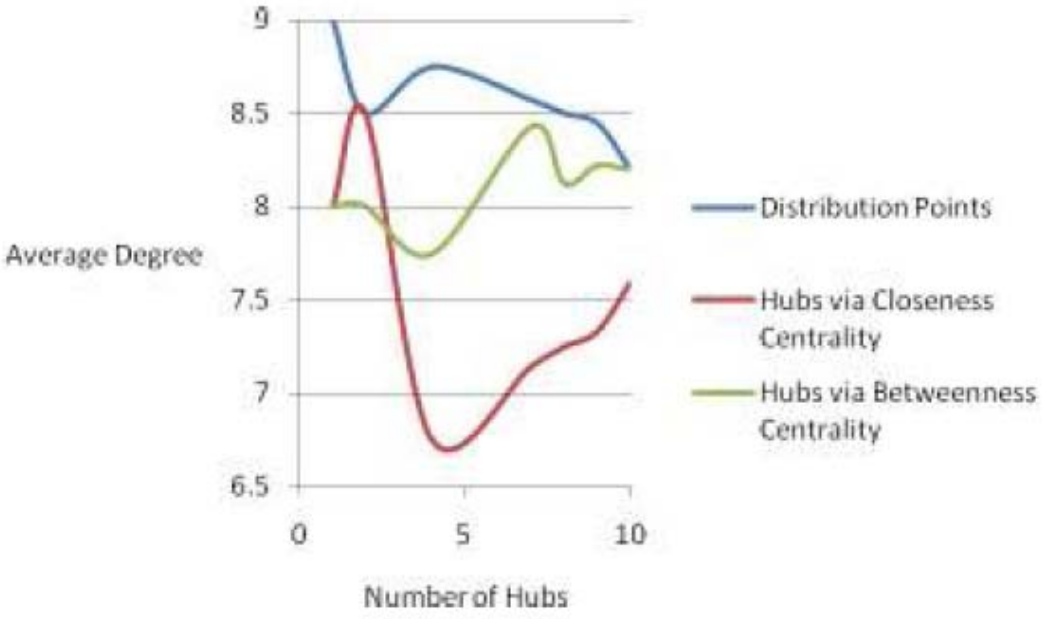

Figure 1c. Degree centrality. Average degree of each chosen hub. Fortunato benchmark graph (N = 100,k = 5, maxk = 10,μ = 0.1, minc = 5, maxc = 30).

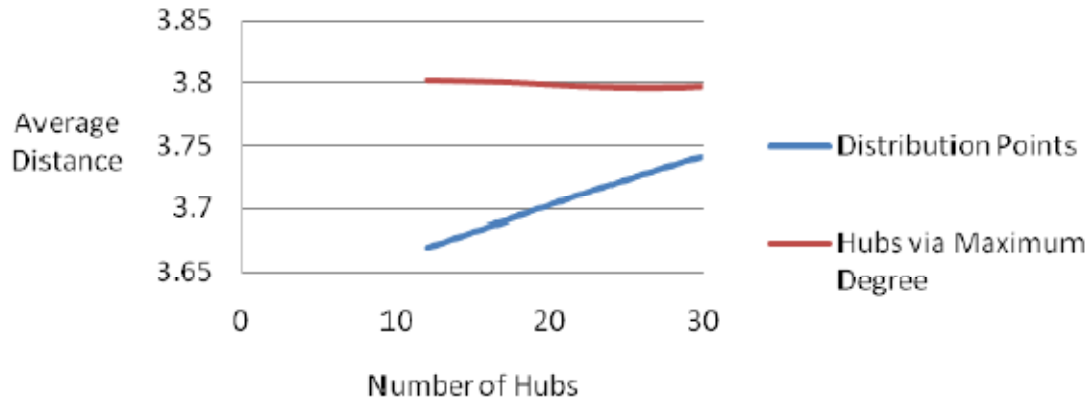

Figure 2a. Closeness centrality. Average distance from each chosen hub to all other vertices in the network. Fortunato benchmark graph (N = 300, k = 5, maxk = 10, μ= 0.1, minc = 5 , maxc = 100).

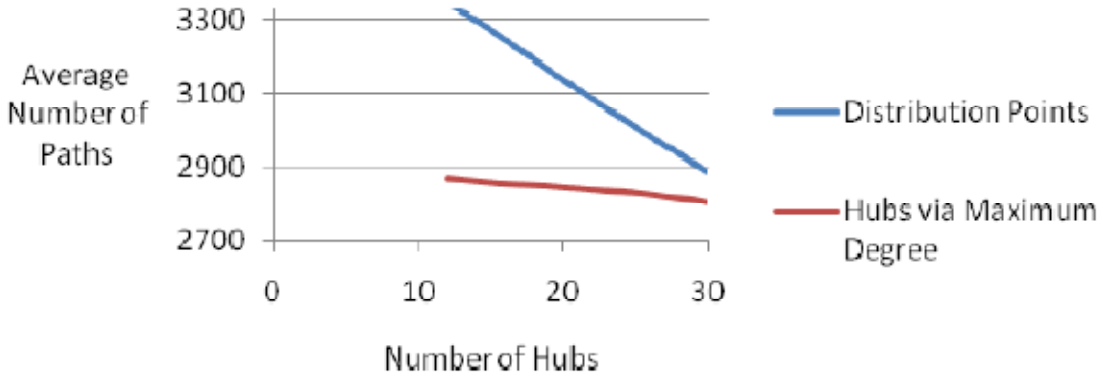

Figure 2b. Betweenness centrality. Average number of shortest paths running through each chosen hub.Fortunato benchmark graph (N = 300, k = 5, maxk = 10,μ= 0.1, minc = 5,
maxc=100).

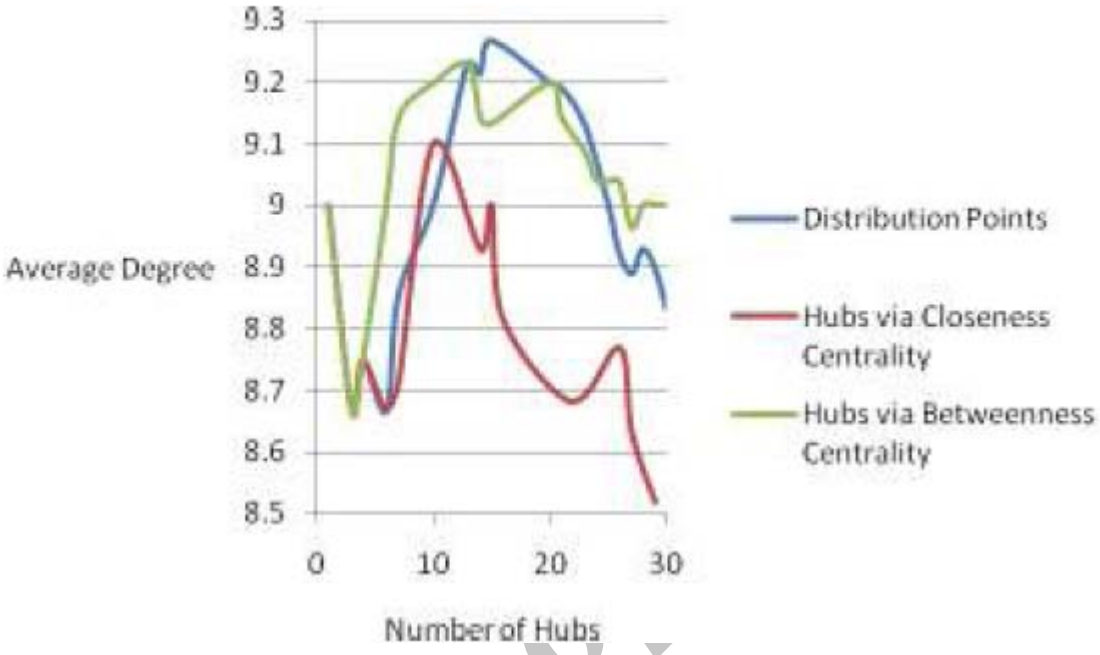

Figure 2c. Degree centrality. Average degree of each chosen hub. Fortunato benchmark graph (N = 300,k = 5, maxk = 10,μ= 0.1, minc = 5, maxc = 100).

We also consider four chemical real-world networks. First, the protein interaction network of *S.cerevisiae*, consisting of 2114 proteins and 4480 interactions between them [24]. Second, the protein interaction network of *A. fulgidus* is considered. The *A. fulgidus* network consists of 32 proteins and 36 validated interactions between them [27]. Third, the network of direct transcriptional regulation between 328 operons in *E. coli* is analyzed [27]. And lastly, we consider the drug network on 616 nodes, analyzed by Ernesto Estrada and his team [15].

Figures 3, 4, 5 and 6 show the comparison of hubs' centrality measures on chemical and biological networks, where hubs are again chosen either as distribution points or as vertices with maximal degree, closeness centrality, or betweenness centrality.

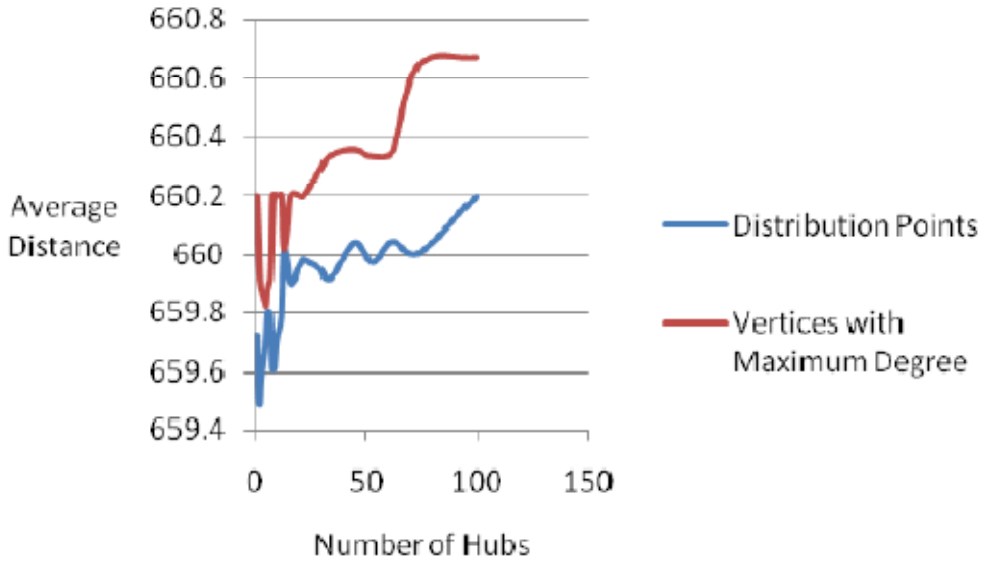

Figure 3a. Closeness centrality on the protein interaction network of S.cerevisiae.

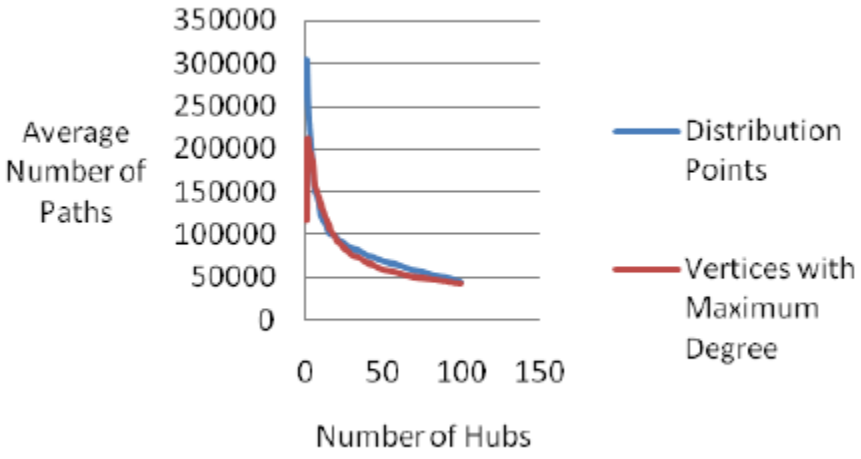

Figure 3b. Betweenness centrality on the protein interaction network of S.cerevisiae.

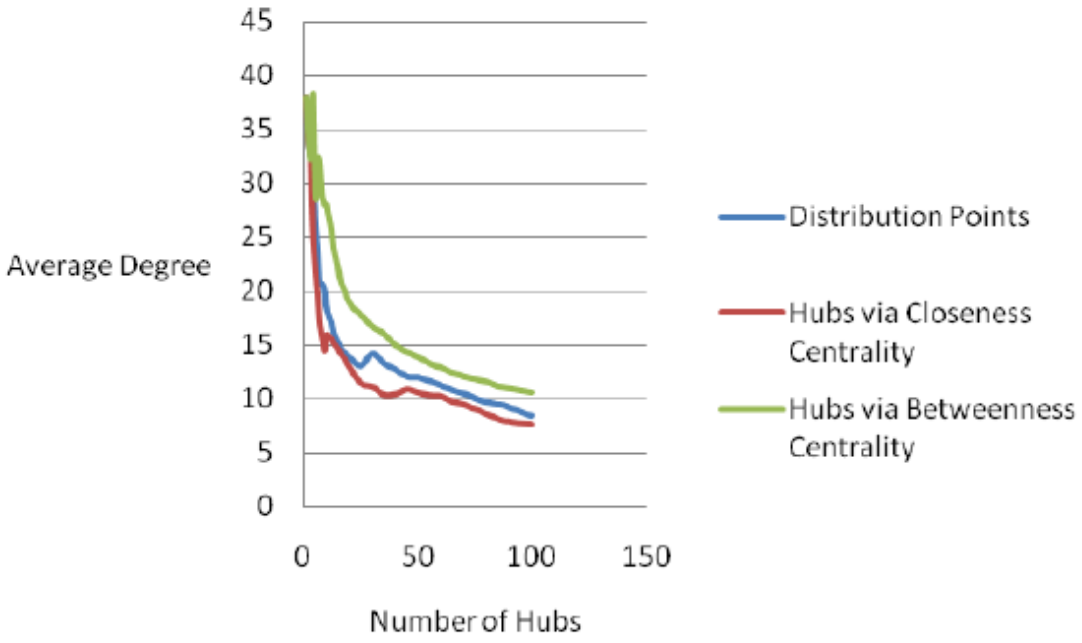

Figure 3c. Degree centrality on the protein interaction network of S.cerevisiae.

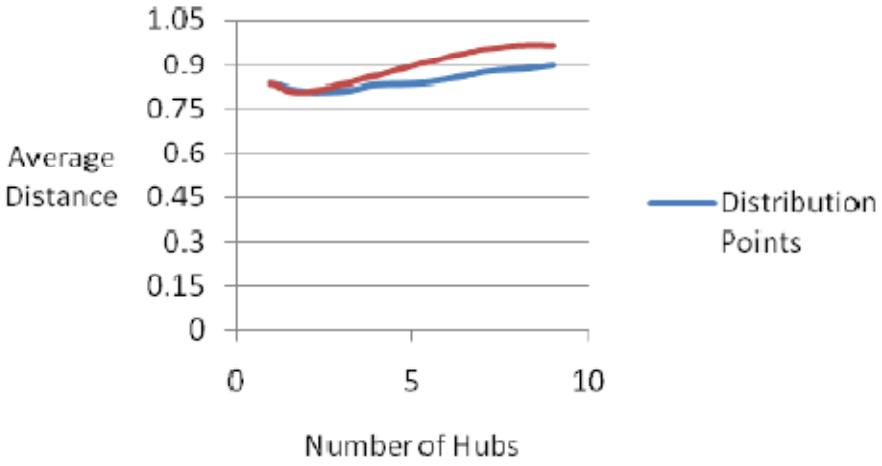

Figure 4a. Closeness centrality on the protein interaction network of A. fulgidus.

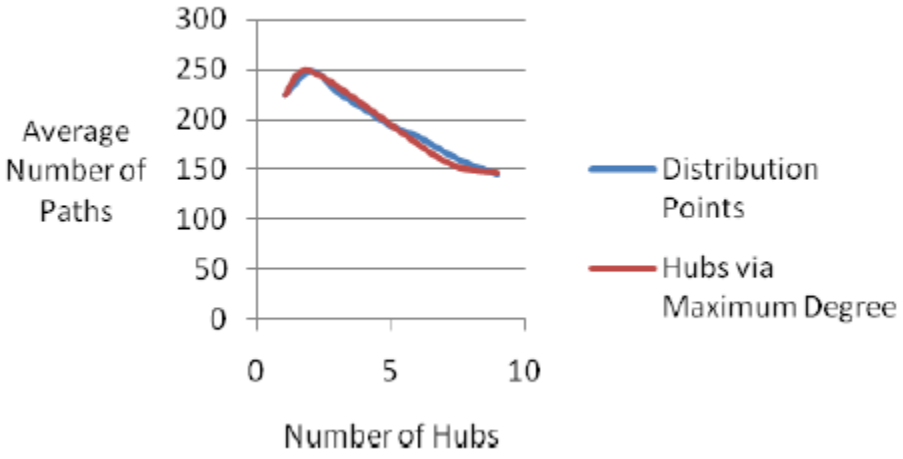

Figure 4b. Betweenness centrality on the protein interaction network of A. fulgidus.

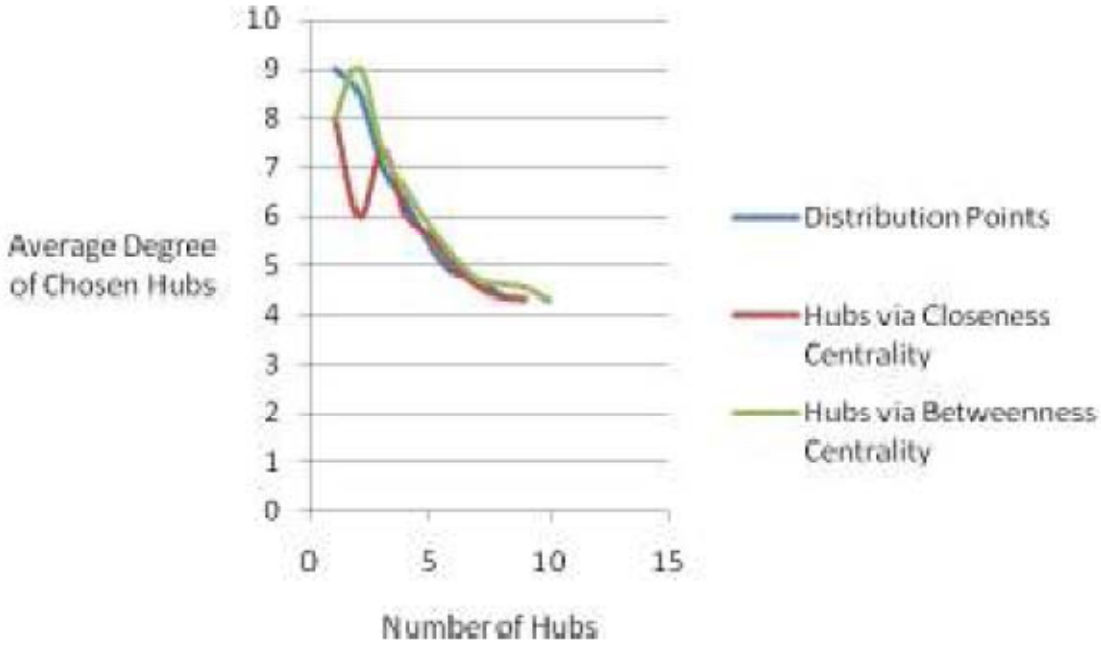

Figure 4c. Degree centrality on the protein interaction network of A. fulgidus.

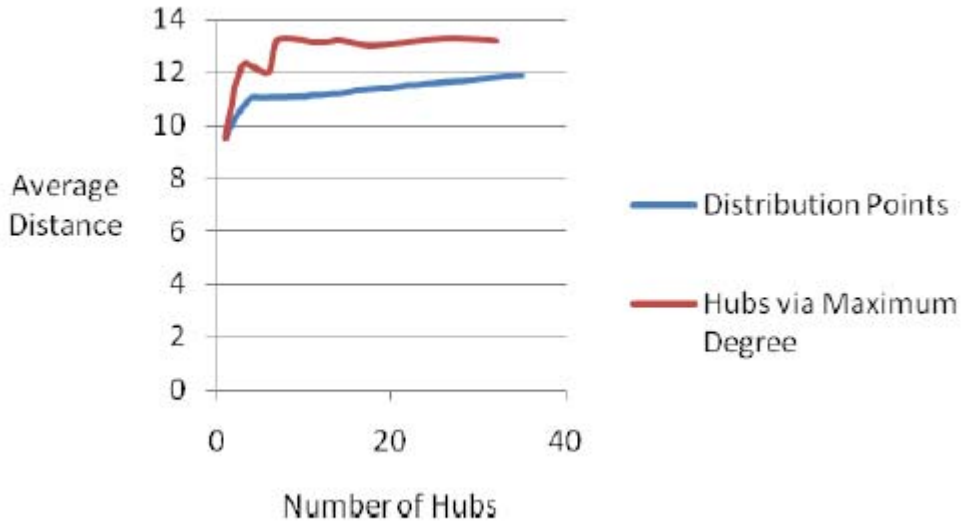

Figure 5a. Closeness centrality on the transcriptional regulation network of E. coli.

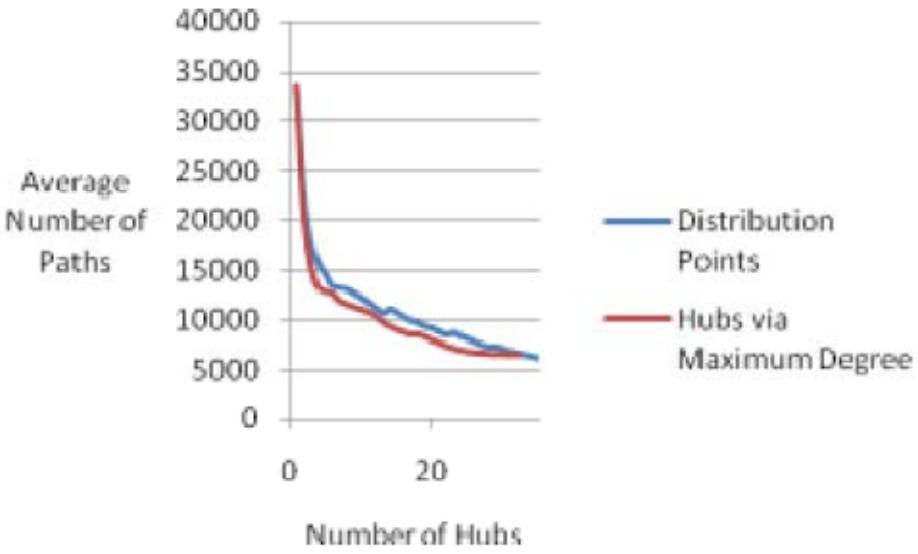

Figure 5b. Betweenness centrality on the transcriptional regulation network of E. coli.

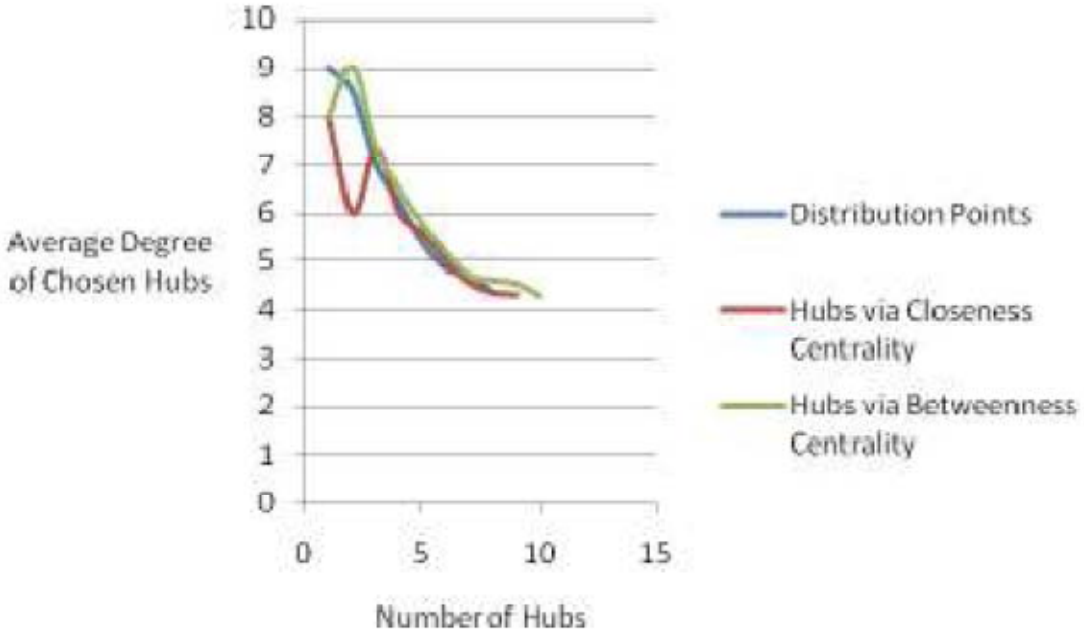

Figure 5c. Degree centrality on the transcriptional regulation network of E. coli.

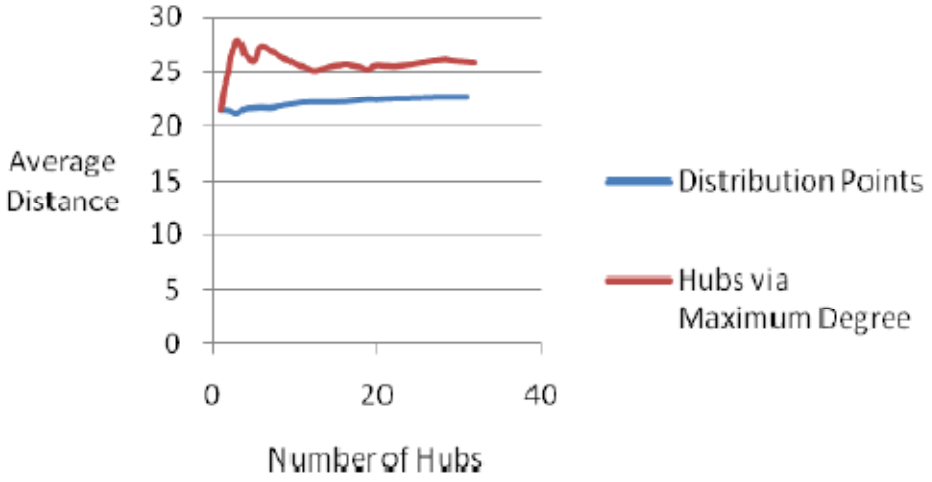

Figure 6a. Closeness centrality on Drugs network.

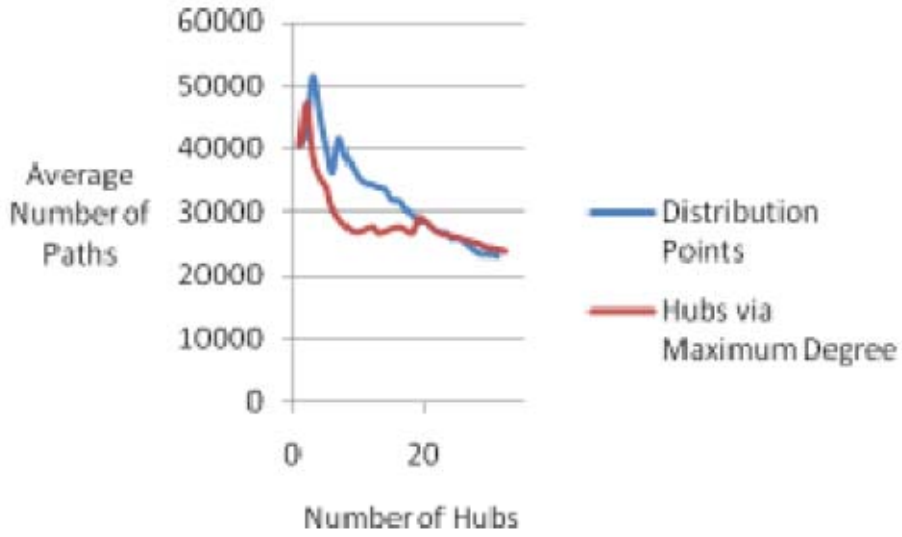

Figure 6b. Betweenness centrality on Drugs network.

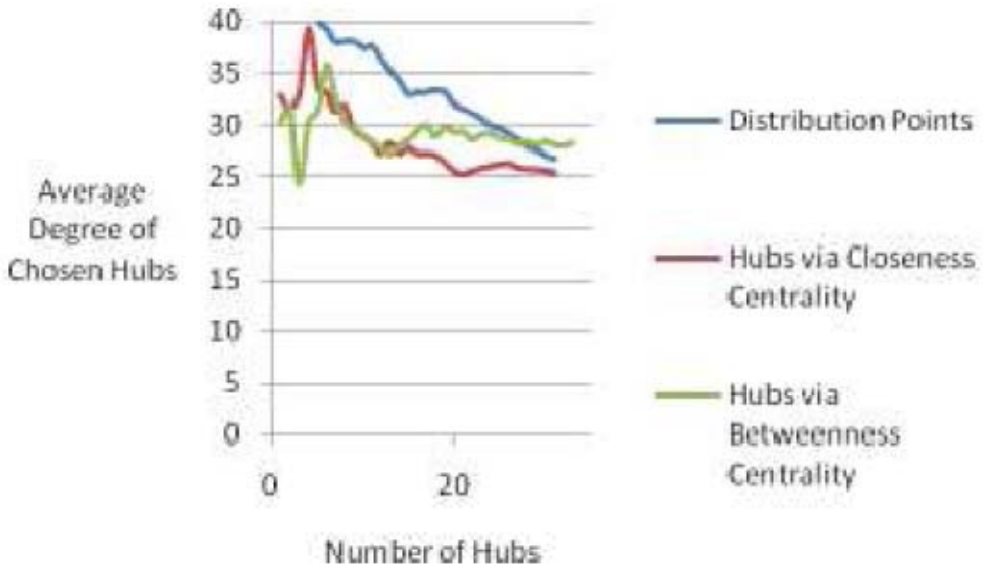

Figure 6c. Degree centrality on Drugs network.

6 CONCLUSIONS

Many methods of hub detection exist. In practice, they all have advantages and drawbacks. Various types and number of hubs are important in different types of networks, where some centrality measures are more representative of network behavior than others. Our distribution points are a compromise between the three types of vertex centrality. As such, our all-purpose hub detection method is aimed to be a fast, useful, universal estimate of node importance, accounting for the inherent difficulties of using snapshots of data to represent the structure of dynamic networks. Several examples illustrate possible chemical applications.

REFERENCES

1. R. J. Williams and N. D. Martinez, Simple rules yield complex food webs, *Nature* **404** (6774) (2000) 180–183.
2. M. Faloutsos, P. Faloutsos and C. Faloutsos, On power-law relationships of the internet Topology, *SIGCOMM Computer Communication Review* **29** (4) (1999) 251–262.
3. R. Albert, H. Jeong and A.-L. Barabasi, Internet: Diameter of the world-wide web, Nature **401** (6749) (1999) 130–131.
4. N. Rashevsky, *Mathematical Theory of Human Relations: An Approach to Mathematical Biology of Social Phenomena*, Principia Press, 1947.
5. D. J. Watts and S.H. Strogatz, Collective dynamics of small-world networks, *Nature* **393** (6684) (1998) 440–442.
6. S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (*Structural Analysis in Social Sciences*), 1st Edition, Cambridge University Press, 1994.
7. S. Redner, How popular is your paper? An empirical study of the citation distribution, *European Physical Journal B* **4** (1998) 131–134.
8. J. P. Scott, *Social Network Analysis: A Handbook*, SAGE Publications, 2000.
9. L. A. N. Amaral, A. Scala, M. Barthelemy and H. E. Stanley, Classes of small-world networks, Proceedings of the National Academy of Sciences of the Unites States of America **97** (21) (2000) 11149–11152.
10. M.E.J. Newman, The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences of the United States of America **98** (2001) 404–409.
11. D. Braha and Y. Bar-Yam, Topology of large-scale engineering problem-solving networks, *Physical Review E* **69** (2004) 016113.

12. R. Ferrer, R. Sole and R. Kohler, Patterns in syntactic dependency networks, *Physical Review E* **69** (2004) 051915.
13. L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, From molecular to modular cell biology, *Nature* **402** (1999) C47–52.
14. D. Eisenberg, E. M. Marcotte, I. Xenarios and T. O. Yeates, Protein function in the post-genomic era, *Nature* **405** (2000) 823–826.
15. E. Estrada, personal communication, 2010.
16. A.–L. Barabasi and R. Albert, Emergence of scaling in random networks, *Science* **286** (5439) (1999) 509–512.
17. R. Albert and A.–L. Barabasi, Statistical mechanics of complex networks, *Reviews of Modern Physics* **74** (1) (2002) 47–97.
18. R. Pastor–Satorras and A. Vespignani, Immunization of complex networks. Physical Review E 65 (2002) 036104.
19. M. Barthelemy, A. Barrat, R. Pastor-Satorras and A. Vespignani, Dynamical patterns of epidemic outbreaks in complex heterogeneous networks, *Journal of Theoretical Biology* **235** (2005) 275–288.
20. Z. Dezso and A.-L. Barabasi, Halting viruses in scale-free networks, *Physical Review E* **65** (2002) 055103.
21. R. Albert, H. Jeong and A.-L. Barabasi, Error and attack tolerance of complex networks, *Nature* **406** (2000) 378–482.
22. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. J. Yang, M. Johnston, S. Fields, J. M. Rothberg, A comprehensive analysis of protein-protein interactions in Saccharomyces cervisiae, *Nature* **403** (2000) 623–627.
23. I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, DIP: the database of interacting proteins, *Nucleic Acids Res*. **28** (2000) 289–291.
24. H. Jeong, S. Mason, A.-L. Barabasi and Z. N. Oltvai, Lethality and centrality in protein networks, *Nature* **407** (2001) 651–654.
25. L. C. Freeman, A set of measures of centrality based on betweenness, *Sociometry* **40** (1) (1977) 35–41.
26. A. Lancichinetti, S. Fortunato, F. Raddichi, Benchmark graphs for testing community structure detection algorithms, *Physical Review E* **78** (2008) 046110.
27. E. Estrada, D. J. Hingham and N. Hatano, Communicability betweenness in complex networks, *Physica A: Statistical Mechanics and its Applications* **388** (2009) 764–774.