DUAL-LAYER NETWORK REPRESENTATION EXPLOITING INFORMATION CHARACTERIZATION

Virginia De Bernardinis[1], Rui Fa[2], Marco Carli[1], Asoke K. Nandi[2,3]

[1]Applied Electronics Department, Università degli Studi Roma TRE, Roma, Italy
[2]Department of Electronic and Computer Engineering, Brunel University, Uxbridge, UB8 3PH, United Kingdom
[3]Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

ABSTRACT

In this paper, a logical dual-layer representation approach is proposed to facilitate the analysis of directed and weighted complex networks. Unlike the single logical layer structure, which was widely used for the directed and weighted flow graph, the proposed approach replaces the single layer with a dual-layer structure, which introduces a provider layer and a requester layer. The new structure provides the characterization of the nodes by the information, which they provide to and they request from the network. Its features are explained and its implementation and visualization are also detailed. We also design two clustering methods with different strategies respectively, which provide the analysis from different points of view. The effectiveness of the proposed approach is demonstrated using a simplified example. By comparing the graph layout with the conventional directed graph, the new dual-layer representation reveals deeper insight into the complex networks and provides more opportunities for versatile clustering analysis.

Index Terms— Information, dual-layer, characterization, clustering

1. INTRODUCTION

Many real systems are often the results of the co-operation of co-existing mechanisms regulated by the constraints that their own nature imposes. The network analysis allows people to study the mechanisms running in the real system. The network describes all that natural concerns, including the physical, the biological, the technological, and the social fields. These architectures need to be analysed according to many perspectives [1][2] [3].

A multi-layer approach was used to investigate the coexisting mechanisms belonging to different points of view and acting simultaneously in the system. In particular, two layers hosting two different topologies were exploited [2] [4], where the lower layer described the positions of the nodes and the upper layer represented the logical connections among the nodes and the flow distribution all over the system. This multiple-layer network analysis is actually constrained on an implicit rule that there is only one layer for each perspective, particularly, one layer for logical layer. The study of robustness is also performed based on this structure. However, for many real world networks, say the file sharing networks, the blogging networks, or gene regulatory networks, such single structure cannot provide more insight of the network beyond the connectivity, since multi-layer structure also exist in the logical layer.

In this paper, we propose a logical dual-layer representation in order to implement a high level network analysis that characterizes the nodes in the system and investigates the topology among them. The dual-layer approach is useful to analyse networks modelling systems whose mechanisms are described by weighted and directed relations. Unlike the multi-layer in[2][4], our model is implemented using a dual-layer structure to present the logical information flow. It replaces the single layer with a dual-layer structure, which introduces a provider layer and a requester layer. By exploiting the new dual-layer network representation, we can carry out the network analysis with deeper insight. For instance, the design of alternative approaches based on its graphical structure can be performed for the clustering and the prediction of links [5][6][7]. The effectiveness of the proposed approach is demonstrated using a simplified example. By comparing the graph layout with the conventional directed graph, the new dual-layer representation reveals deeper insight into the complex networks and provides more opportunities for versatile clustering analysis.

2. PROPOSED APPROACH

In this paper, we propose a dual-layer structure to represent one logical perspective for directed and weighted networks. In particular, we address the network where we assume the existence of a logical connection due to the fulfillment of the information requested from a node by another. In this section, we firstly describe the principle of information flow; then we detail the proposed dual-layer representation for the information layer of the network.

2.1 INFORMATION FLOW LAYER

To infer the networks, we use either the connectivity data directly or the collected attribute information. Suppose that we are given attributes of all nodes, we employ the pair wise Euclidean distance as the information metric between two nodes. Let $\{\boldsymbol{x}|x_i, i = 1, \dots, M\}$ and $\{\boldsymbol{y}|y_i, i = 1, \dots, M\}$ be the attribute vectors representing two nodes respectively, where

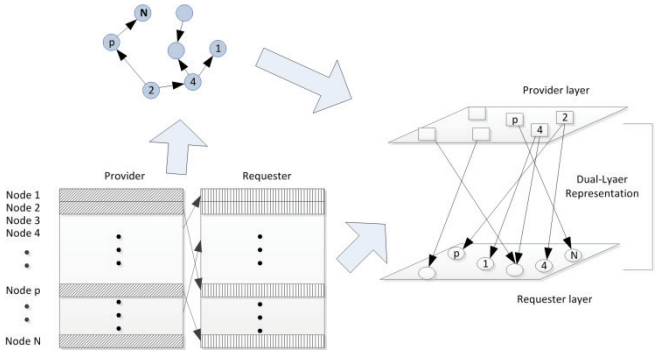

Figure 1 The diagram of the proposed dual-layer representation of the information layer.

M is the dimension. The Euclidean distance between two nodes is given by

$$D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad (1)$$

On the basis of distance between two nodes, the links can be built in the information flow layer.

In a network, a logical edge implies a certain similarity or dissimilarity between the nodes involved, which is known as structural proximity in[5]. For directed networks, the directed link suggests that the attributes characterizing each node conceptually have two disjoint parts. Two parts respectively represent the information sufficiency and the information insufficiency, which are correspondingly called provider part and requester part, respectively. One node may provide a kind of information, but request another kind of information in the meantime. On the other hand, the information always flows from the nodes providing it (sufficiency) to the nodes requesting it (insufficiency). One of the examples is the file sharing networks, where some nodes provide the files (information), some nodes request the files, some nodes simply pass them, and even some nodes do all the providing, requesting and passing. Another example is the blogging networks: some bloggers post the blogs in their own expertise areas; in the meanwhile, they are interested in or subscribe others' blogs, which are very different with their own. These interconnected systems can be well modeled in the information flow point of view. However, there is no structure in the literature reflecting this fact.

2.2 DUAL-LAYER REPRESENTATION

We have presented a brief description of information flow layer, whose links are based on the similarity score (in our case, the Euclidean distance). Here we will detail the implementation of the dual-layer representation based on the idea splitting a node into two parts of the attributes, namely provider and requester. In this case, one node turns to be two nodes, and then the total number of nodes is doubled. Here we assume that each node has both provider part and request part. But it is worth noting that this assumption is not necessarily always the case, e.g., in gene regulatory networks, a large number of genes are targets rather than regulators, so they do not possess provider part. If that is the case, our model is still valid by involving independent providers and requesters. Thus the logical layer turns to be two layers, which are provider layer and requester layer, respectively, as depicted in Figure 1.

Suppose that we have two attribute matrices $\boldsymbol{P}$ and $\boldsymbol{R} \in \mathbb{R}^{N\times M}$, $\boldsymbol{P}$ for providers and $\boldsymbol{R}$ for requesters. The weighted and directed link in the information map represents a logical information flow from the provider node, which is representing the source, to the requester node, which is representing the destination. The new representation does not change the links, but reflects the information flow in a different way. Note that two layers can be superposed to be one graph, in which all nodes, whatever providers or requesters, are positioned according to their information similarities.

The greatest advantage of this representation is that it builds an information landscape, where those logically close nodes are gathered around in the visualization. Comparing with the original directed graph, the links do not reflect the logically similarity or dissimilarity among nodes quantitatively; while in the new dual-layer representation, it clearly illustrates the distributions of both the information source nodes and the information destination nodes, and their relationships. Another advantage is that the new dual-layer representation is beneficial to clustering analysis, which will be discussed in the following sections.

2.3 NETWORK IMPLEMENTATION AND VISUALIZATION

We borrow the idea of the network formation model based on node similarity reported in [8], which only considered the undirected unweighted networks. In our case, the network is built on the similarities between providers and requesters. Let us define a similarity matrix $\boldsymbol{S} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}}$, and a adjacency matrix $\boldsymbol{A} \in \{0,1\}^{N\times N}$. The columns of both $\boldsymbol{S}$ and $\boldsymbol{A}$ represent providers and the rows of them represent requesters. The entries of $\boldsymbol{S}$ are similarities between providers and requesters, which are given by

$$s_{ij} = D(\boldsymbol{p}_i, \boldsymbol{r}_j), \quad (2)$$

where $\boldsymbol{p}_i$ is the *i*-th row vector in $\boldsymbol{P}$ and $\boldsymbol{r}_j$ is the *j*-th row vector in $\boldsymbol{R}$. The adjacency matrix $\boldsymbol{A}$ is a binarised product of $\boldsymbol{S}$ subject to a threshold *T,* which is mathematically written as

$$a_{ij} = \begin{cases} 1 & if\ s_{ij} \le T \\ 0 & if\ s_{ij} > T \end{cases}, \quad (3)$$

where a_{ij} is one of entries in $\boldsymbol{A}$. The adjacency matrix indicates that what nodes are connected. There is another parameter associated with the network, namely sparsity, which is defined as

$$\text{sparsity} = \frac{\sum_{i=1}^{N}\sum_{j=}^{N} a_{ij}}{\dim(\boldsymbol{A})}, \quad (4)$$

where dim($\boldsymbol{A}$) is the total number of potential links, which is *N(N-1)*. The threshold *T* determines the actual number of links in the network, in turn, the sparsity. If given the sparsity, T can be obtained easily by finding the ⌈sparsity ×

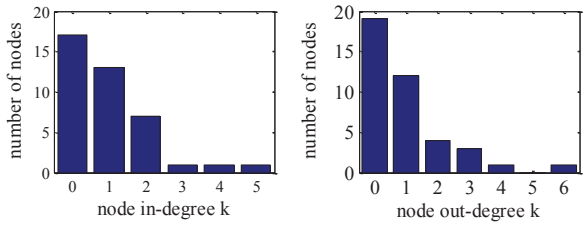

Figure 2 Node in-degree and out-degree distributions in a 40-node network follow the power-law distribution.

dim $(\boldsymbol{A})\rceil$ lowest entry in the similarity matrix *S*, where $\lceil\cdot\rceil$is the ceil operator.

A very important aspect of the proposed dual-layer structure is the visualization. As we mentioned, the position distribution of nodes reflects the information density of a network. However, to illustrate an *M*-dimension data in a two-dimensional space is a challenging job. The classical multi-dimensional scaling (CMDS) [9] is used to exploit the pair wise similarities or dissimilarities of the *M*-attributes of each node to place the provider and requester nodes in the *M*-dimensional space, which depends on the cardinality of distances and the number of the attributes belonging to the subset. In this paper, a non-metric CMDS algorithm according to the Sammon nonlinear mapping criterion [10, 11] is performed, trying to map a high dimensional space to a space of lower dimensionality, but maintaining the inherent structure of the system and the relations among its points. This task is performed by the minimization of an error function:

$$E = \frac{1}{\sum_{1<j} d_{ij}^*} \sum_{1<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \quad , \tag{5}$$

where d_{ij}^* is the distance between the *i*-th node and the *j*-th node in the original space, and d_{ij} is the one in the lower dimensional space.

Therefore, the nodes, which are connected with a directed link, are positioned close to each other. Surely, this map is static snapshot of the network. If considering dynamics, the requesters can move towards their providers time by time because of the information sharing, or in other cases, requesters or providers suddenly modify their information, then the whole network topology will change correspondingly.

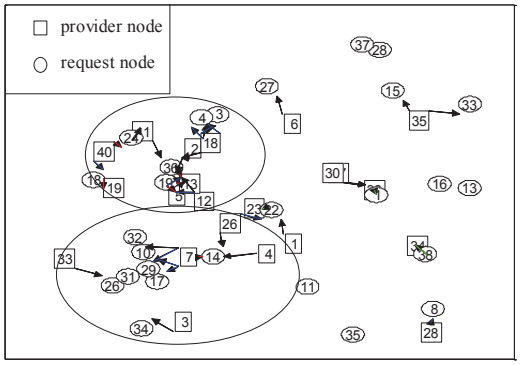

Figure 3 Information map of the 40- node network using the dual-layer representation. It is worth noting that there are at least two areas dense of both providers and requesters, which are marked out.

3. CLUSTERING

The dual-layer structure helps to understand the structure of the network by exploring the information perspective, which, however, was often neglected. Indeed the new structure shows the same logical information in two different modalities that deepen the insight of the analysis. The distribution of the links among nodes can be hence exploited to perform clustering strategies. Depending on the available knowledge of the graph data, the linkage-based clustering and the centroid-based clustering can be performed. We apply different strategies to two clustering respectively. For linkage-based clustering, only the requesters are exploited to produce hierarchical clustering (in this paper, the average linkage is employed). From requester-centric point of view, we believe that in a provider-requester system, the requester is much more important than the provider, since the information is useless if no one requests it. Once the clustering is done, the providers, which have links with the requesters, are injected into the map. Then, we can tell that in the area where only few providers but a lot of requesters locate, those providers are essential to the network.

Instead, in the centroid-based clustering, those nodes that have a higher weighted ratio out-degree/in-degree than the average of the network are considered as sources in the network, since from provider-centric point of view, they provide more information than they require. These providers are assigned to different clusters as centroids. Then, all the nodes that are linked to at least one of the members in the cluster are successively injected into the map.

The new dual-layer representation brings many possibilities to analyze the complex network from different points of view. The versatile clustering analysis is a good example.

4. RESULTS

In this paper, we introduce a simple example to demonstrate the effectiveness of the proposed approach. A network with 40 nodes is simulated randomly. Thus in dual-layer struc-

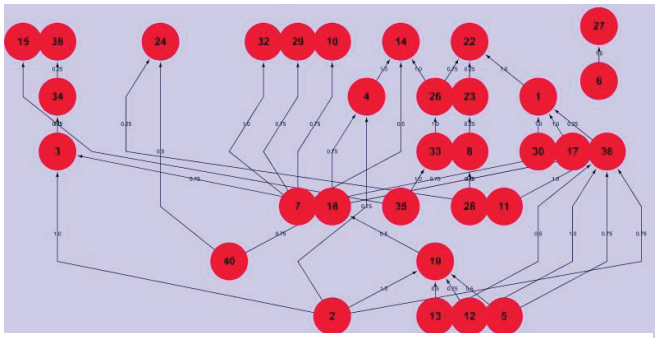

Figure 4 Flow graph of the 40-node network by using the Cytoscape Software, with hierarchical layout.

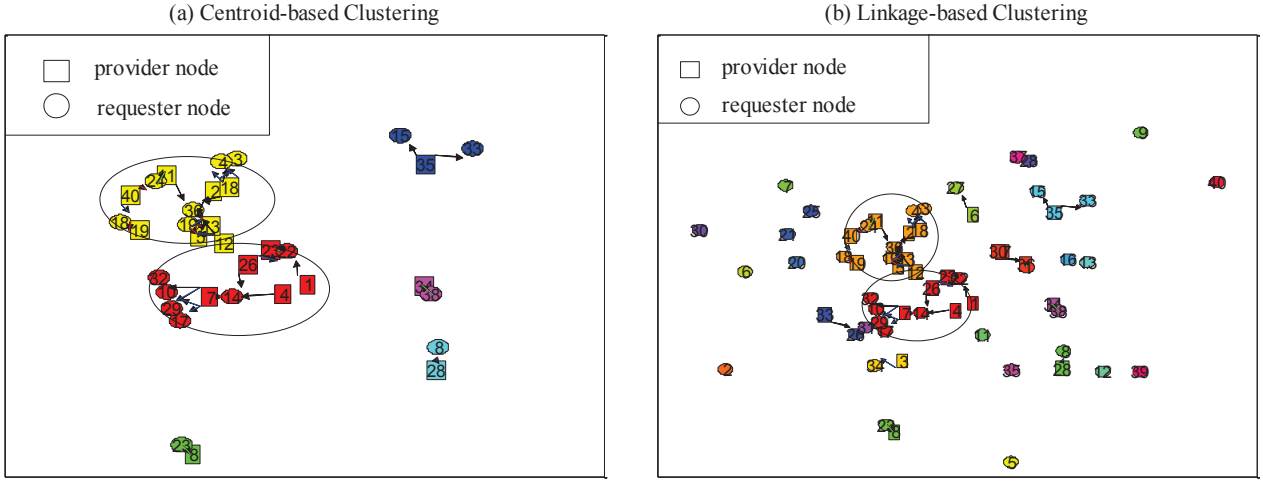

Figure 5 (a) Centroid-based clustering choosing the nodes with an out-degree/in-degree ratio higher than the network average; (b) Linkage-based clustering.

ture, there are 40 providers and 40 requesters. Each provider or requester has four attributes. We set the sparsity in this case equal to 1/40 (for larger networks, the sparsity will be much lower than this). As depicted in Figure 2, both in-degree and out-degree follow the power-law distribution, which illustrates that the network simulation makes some sense.

Then we visualize the dual-layer representation of the network in Figure 3. Note that the graph is the superposition of provider layer and requester layer. In the graph, symbol square represents the provider and symbol circle represents the requester. The directed links indicate the information flowing from providers to requesters. It is worth noting that there are at least two areas dense of both providers and requesters, which are marked out. For comparison, we show the network using Cytoscape [12] in a conventional way in Figure 4. Except the connectivity among the nodes, the conventional directed graph does not provide more information of the network.

Subsequently, we show some clustering results based on the proposed dual-layer structure. The results of centroids-based clustering and linkage-based clustering are shown in Figure 5 (a) and (b), respectively. The clusters are marked in different colors. Although two clustering methods are performed from different points of view, their results indicate some results in common. Two large clusters, as we noticed in Figure 2, are the areas dense of providers and requesters. From the requester-centric point of view, the areas where many requesters locate indicate the main interests of information in the network, thus the providers in those areas are essential to the network. From provider-centric point of view, the providers with high out-degree to in-degree ratio are the important sources. These two points of view provide a consensus analysis to the network rather than a conflict.

Moreover, we can further split the network into small clus-

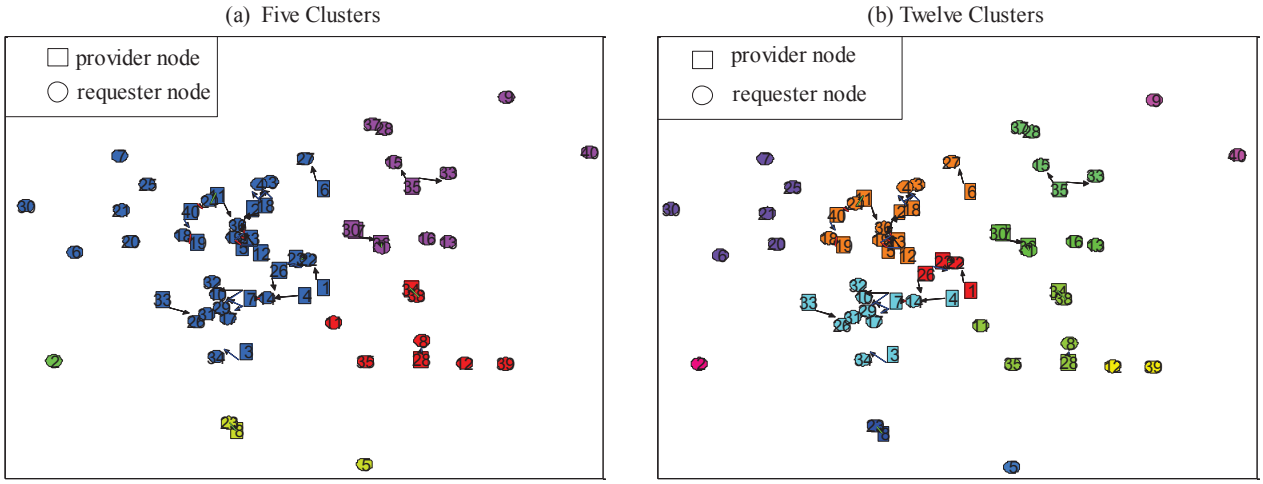

Figure 6 Linkage-based clustering performed on the 40 node network, choosing (a) 5 and (b) 12 as number of clusters.

ters. As shown in Figure 6 (a) and (b), the five-cluster cut and the twelve-cluster cut are provided respectively. This approach can have more power to analyze the larger complex networks.

In summary, the dual-layer representation reveals much deeper insight into the networks and provides more opportunities to perform versatile clustering analysis.

5. CONCLUSION

In this paper, a logical dual-layer representation approach has been proposed for the analysis of directed and weighted complex networks. Unlike the single logical layer structure, which was widely used in the literature for the directed and weighted flow graph, the proposed approach replaces the single layer with a dual-layer structure, which introduces provider layer and requester layer. The new structure provides the characterization of the nodes by the information, which they provide to and they request from the network. Its features have been explained and its implementation and visualization have also been detailed. We also designed two clustering methods with different strategies respectively, which provide the analysis from different points of view. The effectiveness of the proposed approach was demonstrated using a simplified example. By comparing the graph layout with the conventional directed graph, the new dual-layer representation reveals deeper insight into the complex networks and provides more opportunities for versatile clustering analysis.

Most importantly, the new dual-layer structure provides us many opportunities in the future work. One of them is that we may design new link prediction methods, which exploit the dual-layer structure to quantify the probability of a link between two nodes computing the number of their common neighbors, against the existing ones based on the conventional directed graph [5][6][7] [13]. Another possible future work is to study the ergodic behavior of the network entropy exploiting the ergodic theory of dynamical systems on the new graph[14].

7. ACKNOWLEDGMENTS

This article summarises independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0310-1004). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. Professor A. K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship.

8. REFERENCES

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, "Complex networks: Structure and dynamics," *Physics reports,* vol. 424(4), pp. 175-308, 2006.

[2] M. Kurant, P. Thiran and P. Hagmann, "Error and attack tolerance of layered complex networks," *Physical Review E,* vol. 76(2), no. 026103, 2007.

[3] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature,* vol. 464(7291), pp. 1025-1028, 2010.

[4] M. Kurant and P. Thiran, "Layered complex networks," *Physical review letters,* vol. 96(13), no. 138701, 2006.

[5] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications,* vol. 390, no. 6, pp. 1150-1170, 2011.

[6] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology,* vol. 58, no. 7, pp. 1019-1031, 2007.

[7] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (pp. 73-80). IEEE*, 2011, October.

[8] M. Scholz, "Node similarity as a basic principle behind connectivity in complex networks," *arXiv preprint arXiv:1010.0803,* 2010.

[9] J. Z. Wang, "Classical Multidimensional Scaling," in *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, Springer Berlin Heidelberg, , 2011. , pp. 115-129..

[10] K. Hansjörg and J. M. Buhmann, "Data visualization by multidimensional scaling: a deterministic annealing approach," *Pattern Recognition,* vol. 33, no. 4, pp. 651-669, 2000.

[11] J. W. Sammon Jr, "A nonlinear mapping for data structure analysis," *Computers, IEEE Transactions on,* vol. 100, no. 5, pp. 401-409, 1969 .

[12] P. Shannon, a. A., O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research,* vol. 13, no. 11, pp. 2498-2504, 2003.

[13] Y. Dong, Q. Ke, B. Wang and B. Wu, "Link Prediction Based on Local Information," in *In Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on (pp. 382-386). IEEE.*, 2011, July.

[14] L. Demetrius and T. Manke, "Robustness and network evolution—an entropic principle," *Physica A: Statistical Mechanics and its Applications,* vol. 346(3), pp. 682-696, 2005.