The Challenge of Finding Degree Centrality Nodes in Heterogeneous Multilayer Networks

Kiran Mukunda, Abhishek Santra and Sharma Chakravarthy

IT Lab and CSE Department, University of Texas at Arlington, Arlington, Texas, USA

Abstract

Complex data sets with different types of entities and relationships can be elegantly modeled using Heterogeneous Multilayer Networks (HeMLNs), where different sets of nodes are connected within and across layers. To identify highly influential nodes in these networks, it is imperative that we are able to define and compute centrality metrics *directly* on MLNs. Currently, MLNs are converted into a single graph using aggregate and projection alternatives, and centrality and other metrics are computed. However, this approach has been shown to lose information, and structure, and makes result interpretation difficult.

In this paper, we extend the simple graph degree centrality definition to HeMLNs and use the novel decoupling approach. For this, we propose heuristics to develop algorithms for identifying degree centrality nodes in heterogeneous MLNs. The proposed heuristics improve the accuracy with respect to ground truth when additional information from each layer is used to improve accuracy with respect to ground truth. However, identifying that additional minimal information is the challenge. We provide intuition behind the heuristics proposed and provide extensive experimental results using large and diverse synthetic and real-world data sets to demonstrate improved accuracy, precision, and efficiency across graph characteristics. We have also shown that the decoupling approach is significantly more efficient than the computation of ground truth.

Keywords

Heterogeneous Multilayer Networks, Degree Centrality, Efficient Heuristics, Decoupling Approach

1. Motivation

Graphs represent relationships between entities in a system using nodes and edges. This representation allows us to model and perform various types of analysis depending on the relationships in the data. For example, the individuals in a friendship network can be related to their residential cities in a transport network; authors can be related if they publish at the same conferences, etc. As graph data sets are becoming larger and more complicated in the real world, we need to expand not only the analysis methodologies but also representations in appropriate ways.

One way to handle both the size and complexity of relationships in a data set is to use alternative modeling approaches, such as multilayer networks [1, 2, 3, 4]. Instead of a single graph trying to capture all the relationships, a separate layer is used for each relationship making the representation or the model easy to understand. Even the relationships between the

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

kiran.mukunda@mavs.uta.edu (K. Mukunda); abhishek.santra@mavs.uta.edu (A. Santra); sharmac@cse.uta.edu (S. Chakravarthy)

nodes in different layers can be captured in this model. Hence, MLNs are becoming popular for big data analysis. However, the downside is that there are not many algorithms for computing the analysis objectives (e.g., community, centrality, substructure, etc.) on MLNs *directly*. They are typically converted into a single graph representation to use existing algorithms. Research in this direction is becoming important due to the benefits of MLNs for modeling, efficient analysis, and the ability to handle large data sets in a flexible manner. All these are illustrated in this paper for degree centrality computation of heterogeneous MLNs.

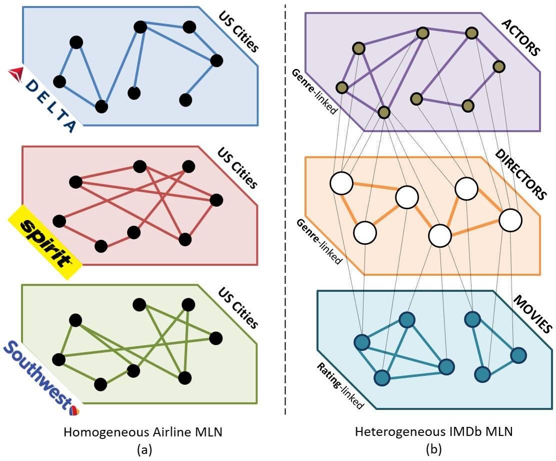


Figure 1: MLN Types

There are two distinct types of multilayer networks: Homogeneous and Heterogeneous. If each layer of a MLN has a common set of entities, they are homogeneous MLN (HoMLNs). For example, the US Airline data set can be modeled using a HoMLN, where nodes in each layer represent the cities and edges correspond to the flights between cities as shown in Figure 1 (a). The other type of multilayer network is the heterogeneous MLN (HeMLN), where the sets of entities are different across layers. IMDb data set requires HeMLNs to model actors, directors, and movies as different layers along with their inter-layer edges as shown in Figure 1 (b) to capture relationships, such as directs-an-actor, directs-a-movie, etc.

Current approaches to centrality detection in MLNs, such as type-independent [5] and projection-based [6], do not support structure and semantics preservation without elaborate mappings as they aggregate (or collapse) layers into a simple graph in different ways. As observed in the literature, *without additional mappings*, currently-used aggregation approaches are likely to result in some information loss [2], distortion of properties [2], or hide the effect of different entity types and/or different intra- or inter-layer relationships as elaborated in [7]. Furthermore, structure and semantics preservation is critical for understanding the layer to which the nodes of interest belong during drill-down analysis of results. From an analysis perspective, lack of structure and semantics makes the drill-down and visualization of results extremely difficult (or even impossible) and hence their understanding. In our approach, analysis results clearly show the structure and ease of drill-down to see patterns in terms of original layers, labels, and relationships.

Centrality nodes are the most influential nodes in a network or graph. Different centrality metrics, such as degree, betweenness (multiple types), closeness, eigenvector, katz centrality, page rank, and percolation centrality are defined for single graphs for various purposes. Some of them are local (e.g., degree) and some are global (e.g., betweenness, closeness) properties of the network. Typically, it takes more effort to compute global measures as compared to local ones. Main memory algorithms exist for computing the above metrics which we leverage in our decoupling-based approach. In this paper, we focus on degree centrality metric. In general, degree centrality defines the relative importance of a node within the given network based on the number of edges incident on it, that is its immediate or 1-hop neighborhood. It can be used to infer the hubs of an airline network. Although there are algorithms for computing it for a graph, to the best of our knowledge, there is no proper definition and algorithm that can be applied directly to MLNs. Using aggregation or projection has a number of disadvantages as discussed earlier. Hence, the focus of this paper is to develop such algorithms using a novel technique termed the decoupling approach.

Contributions of this paper are:

- **Degree Centrality definition** for Heterogeneous MLNs
- **Two heuristics** to improve accuracy, precision, and efficiency of computed results based on **decoupling-based approach**
- **Algorithms** for computing degree centrality nodes *directly* on HeMLNs
- **Experimental analysis** on a number of large (200K vertices and 12 Million edges) synthetic, and real-world graphs with diverse characteristics
- **Accuracy, Precision, and Efficiency comparisons** with ground truth and naive approach

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 introduces the decoupling approach used for MLN analysis and discusses its advantages and challenges. Section 4 gives the degree centrality definition for HeMLNs. Section 4.1 discusses ground truth, naive approach, and the challenges involved in developing techniques for centrality detection. Section 4.2 and 4.3 give the details of the proposed heuristics. Section 5 describes the experimental setup, data sets, and result analysis. Conclusions are in Section 6.

2. Related Work

The concept of centrality of a graph was first proposed by Bavelas in 1948 [8] and it has been researched extensively thereafter. Traditionally, these centrality measures have been implemented as main memory algorithms. With the advent of social media and web 2.0, the amount of data being used for analysis has exploded in volume which has challenged the computation of these on large data sets. Here we summarize the computation of centrality on simple graphs. Our focus, however, is to develop algorithms for centrality on heterogeneous multilayer networks using the decoupling-based framework proposed in [9].

Degree of a node is defined as the total number of edges that are incident on it. It was used as a centrality measure by Shah in 1954. If the network is directed, then the total number of edges that are directed towards the node is called indegree and that are directed outwards from the node is called outdegree [10, 11]. To normalize, the degree of a node is divided by the maximum number of nodes it can have edges with, that is $|V| - 1$ for an undirected graph.

While the above algorithms are focused on a single graph, there is a need to extend them to MLN. There has been some work to detect degree centrality in homogeneous multilayer networks (also called multiplexes), where the *same set* of nodes are connected in multiple layers/networks [12, 13]. In this paper, we focus on computing degree centrality directly on heterogeneous MLNs where each layer has a different set of nodes that are connected by intra- and inter-layer edges. Further, we use the decoupling approach due to its advantages.

3. Decoupling Approach For Multilayer Networks (or MLNs)

Multilayer networks consist of multiple layers of simple graphs where each layer represents a feature of its entities and their relationships in the graph. MLNs are being used to model and analyze large and complex data sets from diverse domains, such as social networks for community detection [1], anomaly detection [14], biological networks to analyze the patterns of human brains [15], finding solutions to oil leakages [16], and so on. However, most algorithms convert the MLN (or a subset of it) into a simple graph using aggregation or projection techniques, leading to a loss of structure, semantics, and information from the final analysis results. Moreover, the existing single graph algorithms cannot be applied directly on the MLNs.

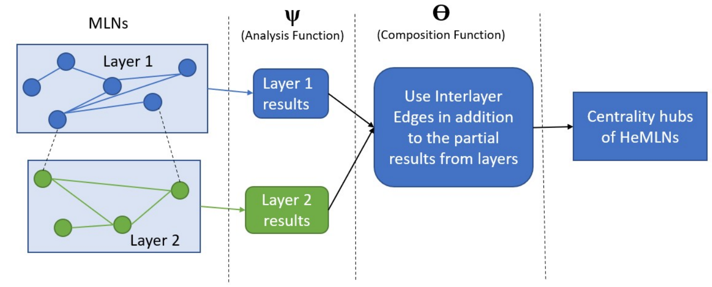

MLNs
Ψ
(Analysis Function)
Θ
(Composition Function)
Layer 1
Layer 2
Layer 1 results
Layer 2 results
Use Interlayer Edges in addition to the partial results from layers
Centrality hubs of HeMLNs

Figure 2: Decoupling Approach for HeMLN Degree Hubs

In this paper, we use a novel decoupling-based framework adopted by a few of the recent works, which preserves the structure and semantics of HeMLNs [17, 18] while performing analysis on complex data sets without losing any information, unlike the traditional approach. The network decoupling approach has been illustrated with respect to the degree centrality computation in Figure 2. It consists of two functions: analysis (Ψ) and composition (Θ). Using the analysis function, each layer in the network is independently analyzed. Then, the partial results from any two layers are combined with the inter-layer edges and processed by the composition function to produce the results for the two layers. This binary composition can be easily extended to n layers by applying it repeatedly on previous results. It is also possible to analyze each layer in parallel to improve the efficiency [9].

Further, due to the layer-wise analysis, each graph is small, which requires less memory for computing layer-wise results. Each layer is analyzed once, and the existing single graph algorithms can be used for analyzing individual layers. The results obtained are then used by the composition function. This approach is also application-independent.

In this approach, the major challenge is to develop the composition algorithm. For accuracy guarantees, it becomes critical to determine the minimal additional information required from each layer during analysis phase to be used for composition, without effecting overall efficiency.

4. Degree Centrality: HeMLN Definition and Heuristics

Degree centrality tells us about the relative importance of a node within the network based on the number of edges it has. While degree centrality is defined for a single graph, there are no definitions/algorithms for HeMLNs that we are aware of[1]. We first define the degree centrality measure for HeMLNs. Note that there are multiple ways to define the degree centrality for a

[1] On the other hand, degree centrality for HoMLNs is defined in [13] as cross-layer degree centrality (or CLDC) which corresponds to degree centrality definition using the Boolean OR operation in [12].

HeMLN. We take the traditional aggregation approach to define them. This definition extends the definition for single graphs to MLNs using union (or Boolean OR) aggregation.

Definition 1. *A heterogeneous multilayer network $HeMLN(G, X)$, is defined by two sets of graphs. The set $G = \{G_1, G_2, \dots, G_n\}$ contains* ***simple*** *graphs, where $G_i(V_i, E_i)$ is defined by a set of vertices V_i and a set of edges E_i. An edge $e(v, u) \in E_i$, connects vertices v and u, where $v, u \in V_i$. The set $X = \{X_{1,2}, X_{1,3}, \dots, X_{n-1,n}\}$ consists of* ***bipartite*** *graphs. Each graph $X_{i,j}(V_i, V_j, L_{i,j})$ is defined by two sets of vertices V_i and V_j and a set of edges (or links) $L_{i,j}$, such that for every link $l(a, b) \in L_{i,j}$, $a \in V_i$ and $b \in V_j$, where V_i (V_j) is the vertex set of graph G_i (G_j). For a HeMLN, the set X is defined only for those layers that have inter-layer edges. For HeMLNs, V_i's are disjoint and some X's may be empty.*

Definition 2. *Degree centrality of a node, z, in a $HeMLN(G, X)$ with n layers where |V| = $\Sigma_{i=1}^{n}|V_i|$ is defined as,*

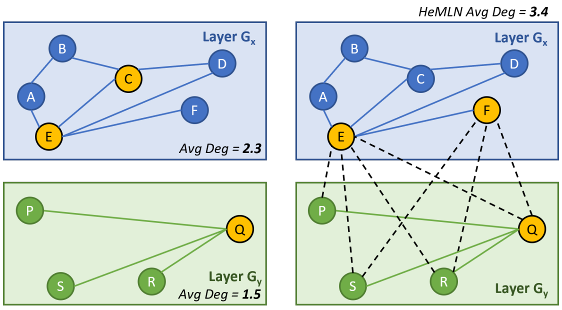
$$HeMLN_{DC}(z) \quad = \frac{degree(z)}{|V| - 1} \tag{1}$$

where, $degree(z)$ denotes the ***number of 1-hop neighbors*** *of node z in the MLN. The high-degree centrality nodes (also called degree hubs) are the ones that have more than or equal to the average value.*

The above definition of degree centrality covers both single graphs and HeMLNs. In this paper, we focus on detecting the degree centrality hubs using the decoupling-based approach for undirected graphs using the above definition.

4.1. Ground Truth, Naive Approach for Baseline Accuracy, and Challenges

The ground truth for the HeMLN is computed on the ground truth graph as per centrality definition given above. This graph is the union (or Boolean OR on the edges of the layers and includes inter-layer edges) of HeMLN layers resulting in a single graph. Existing algorithms are used for computing the degree centrality nodes of the ground truth graph. This can be done for any k layers of the HeMLN. The composition step, being binary, uses two layers. Results of the heuristic-based algorithms using the decoupling approach are validated against this ground truth.



(a)
Degree Hubs in Individual Layers
(inter-layer edges not considered)

(b)
Degree Hubs in the HeMLN
(dotted lines are inter-layer edges)

Figure 3: Degree Hubs (marked in yellow) in the individual layers vs. entire HeMLN

The **naive approach** for computing degree centrality hubs of the HeMLN using the decoupling approach is used as a baseline for comparing and improving the accuracy and precision with proposed heuristic-based approaches. Naive composition takes hubs (computed independently) from each layer, performs union of those hubs (due to OR aggregation), and considers them to be the hubs for the corresponding two-layer HeMLN. This baseline approach *does not use any additional information* from the layers. However, based on observations 1 and 2 below, results from the naive approach are not likely to match the ground truth results due to the presence of *false positives* and *false negatives*, respectively. This is considered a *minimum/baseline accuracy* that we can further improve by using selective additional information from each layer and inter-layer edges during composition.

Observation 1. *A node that is a hub in either layer G_x or G_y, may not be a hub in the HeMLN of G_x, G_y, and $X_{x,y}$*

Example: It can be clearly observed from Figure 3 that node C despite being a hub in layer G_x, does not have enough inter-layer edge connectivity with the nodes in layer G_y, and thus ceases to be a hub in the HeMLN of G_x, G_y, and $X_{x,y}$.

Observation 2. *A node that is not a hub in either layer G_x or G_y, may become be a hub in the HeMLN of G_x, G_y, and $X_{x,y}$*

Example: Again, from Figure 3, it can be observed that node F despite being a node that has low 1-hop connectivity with other nodes in layer G_x (that is a non-hub), has many inter-layer edges with the nodes in layer G_y, and thus becomes a hub in the HeMLN of G_x, G_y, and $X_{x,y}$.

The above observations clearly indicate why false positives and false negatives can be generated by the naive composition depending on layer characteristics and inter-layer edges. Therefore, in order to obtain, in general, accuracy closer to the ground truth (or always matching the ground truth), the challenge is to identify the additional information that needs to be maintained from the layers. Our goal is to develop heuristics for composition functions that *maximize accuracy* (expressed as Jaccard coefficient with respect to ground truth) and *the computation cost is still significantly below the ground truth computation cost.* We test results from heuristic-based algorithms with the ground truth and naive approach. Our heuristics can be applied repeatedly to the results and composed with other layers. This will involve ordering of computation to maximize accuracy. Due to space constraints, in this paper, we validate results on two layers. In general, heuristics can be applied as a binary function for any number of layers.

4.2. Heuristic HeMLN-PG for Precision Guarantee

The question is what additional information from layers can we use that can guarantee reduction or elimination false positives and negatives. What can we gain if we keep the degree value of all the hubs from each layer as well as the information to compute the average degree? It turns out that this can totally eliminate false positives as indicated by the following lemma.

Lemma 1. *By keeping the number of 1-hop neighbors (or only degree information) for **all hubs** along with the number of vertices and edges in layers G_x and G_y, false positives can be completely eliminated in the computation of degree hubs in the HeMLN generated by G_x, G_y, and $X_{x,y}$ using the decoupling approach.*

Proof. Based on observation 2, a non-hub from a layer may become a hub in the HeMLN and is not detected as such. This results in a false negative. With the information kept only for hubs, this can happen. However, a false positive is one where a layer hub becomes a non-hub in the

HeMLN, but is detected as a hub. This cannot happen if degree information for all hubs (from each layer) is kept and their degrees are updated correctly using inter-layer edges. Also, the average degree of the HeMLN can be correctly computed using node, edge information from each layer, and inter-layer edges. Hence, whether a previous layer hub can still be a HeMLN hub or not can be correctly determined without generating any false positives. □

Based on lemma 1, heuristic HeMLN-PG is proposed. Algorithm 1 presents the composition function algorithm.

Algorithm 1 Composition Algorithm Θ for Heuristic HeMLN-PG

```
INPUT:
$DH_i, |V_i|, |E_i|, DH_j, |V_j|, |E_j|$
$deg_i = \{u_{i1} : deg_{i1}, u_{i2} : deg_{i2}, ..., u_{im} : deg_{im}\}$
$deg_j = \{v_{j1} : deg_{j1}, v_{j2} : deg_{j2}, ..., v_{jk} : deg_{jk}\}$
$X_{i,j} = \{(u_{i1}, v_{j1}), (u_{i2}, v_{j2}), ...\}$
ALGORITHM:
$DH_{i,j} \leftarrow \emptyset$
for $(u, v) \in X_{i,j}$ do
    if u $\in DH_i$ || $deg_i[u]$ == 1 then
        $deg_i[u] = deg_i[u] + 1$
    else
        $deg_i[u] = 1$
    end if
    if v $\in DH_j$ || $deg_j[v]$ == 1 then
        $deg_j[v] = deg_j[v] + 1$
    else
        $deg_j[v] = 1$
    end if
end for
$avgDeg_{i,j} = \frac{2*(|E_i|+|E_j|+|X_{i,j}|)}{|V_i|+|V_j|}$
for $deg[u] \in deg_i \cup deg_j$ do
    if $deg[u] \geq avgDeg_{i,j}$ then
        $DH_{i,j} \leftarrow DH_{i,j} \cup u$
    end if
end for
```

During analysis, for each layer G_i, in addition to degree hubs (DH_i), we collect their degree values ($deg_i[]$), total number of nodes ($|V_i|$), and edges ($|E_i|$) as additional information to be used during composition.

In the composition function, for each inter-layer edge ($(u, v) \in X_{i,j}$), the degree value of both the nodes ($deg_i[u]$, $deg_j[v]$) on which the edge is incident upon is incremented by 1. This step will be able to calculate the correct degree of nodes that are hubs in the individual layers and approximate the degree of other nodes as their original layer degree are unavailable. The average degree of HeMLN ($avgDeg_{i,j}$) is calculated as two times the number of total edges in the HeMLN ($|E_i|$ + $|E_j|$ + $|X_{i,j}|$) divided by the total number of nodes in the HeMLN ($|V_i|$ + $|V_j|$). The estimated degree hubs for HeMLN are the ones whose final degree value after scanning all inter-layer edges is greater than or equal to the average degree of the entire network. In order to increase the degree for each inter-layer edge, a hash lookup is carried out on both nodes.

The proposed heuristic has been illustrated in figure 4 (top half). Here, the hubs from layer G_x are nodes C (deg = 3) and E (deg = 4), and the hub from layer G_y is node Q (deg = 3). The degree of these nodes gets increased in the composition function for each inter-layer edge they are a part of. The 6 inter-layer edges are marked (F, Q), (F, R), (F, S), (E, P), (E, S), (E, R), and (E, Q). Therefore, we add to the degrees of nodes E, Q, F, P, R, and S. Since F, P, R, and S are not degree hubs in the layers and we did not record their layer degree information, thus, their degrees are initialized to 1 in the composition function, when the first corresponding inter-layer edge is encountered. Heuristic HeMLN-PG identifies **E**, and **Q** as the degree hubs for the HeMLN

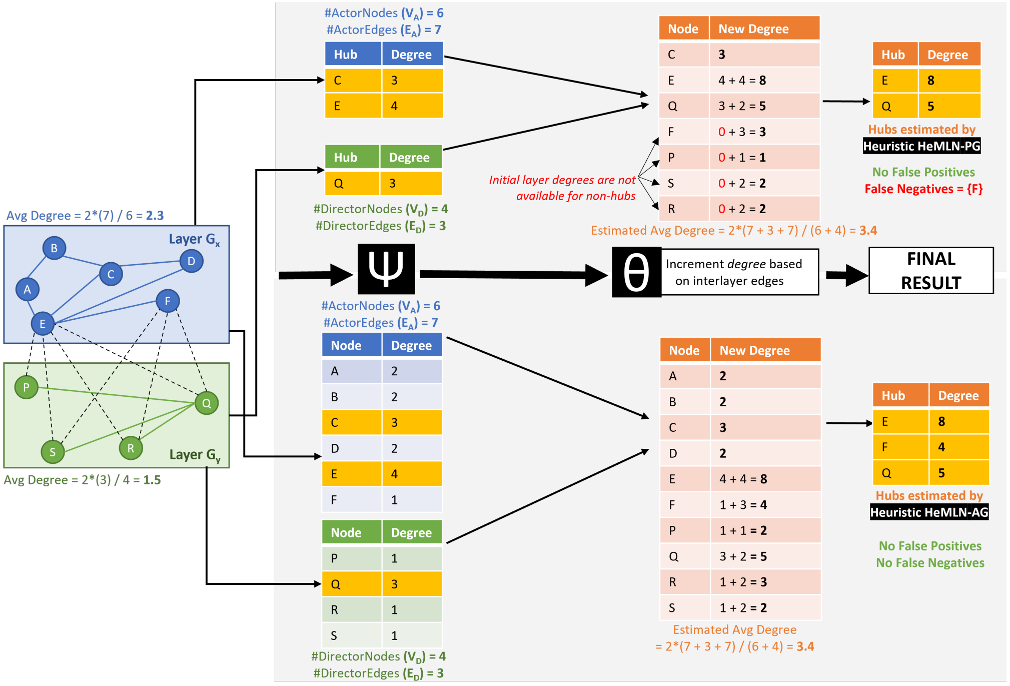


Figure 4: Illustration of Heuristic HeMLN-PG (top half) and Heuristic HeMLN-AG Flow (bottom half) for the HeMLN with ground truth shown in Figure 3

as these nodes have degrees more than or equal to the average degree (≥3.4). In this case, a *false negative* was generated in the form of node **F**, which also came out as a hub as part of the ground truth (**E, <u>F</u>, Q**) in Figure 3.

HeMLN-PG will never generate false positives, as it is able to correctly calculate the degree information of hubs which moreover gets compared against the correct value of the average HeMLN degree. That is, *the precision for this heuristic is always accurate.*

4.3. Heuristic HeMLN-AG for Accuracy Guarantee

The major drawback of HeMLN-PG is that it may generate some false negatives as it does not calculate the correct degree value of the nodes that are not layer-wise hubs, thus leading to inconsistent recall values. However, if we keep degree information for all nodes in each layer, this can be avoided. This leads to the following lemma.

Lemma 2. *It is sufficient to maintain the number of 1-hop neighbors of* ***each node (not just the hubs)*** *along with the number of nodes and edges in layers G_x and G_y to compute degree hubs in the HeMLN generated by G_x, G_y, and $X_{x,y}$ accurately (i.e., to match the ground truth) using the decoupling approach.*

Proof. Using the inter-layer edges, the revised degree of each node in each layer can be **correctly** computed. The average HeMLN degree can also be computed using the number of layer nodes, intra-, and inter-layer edges. Hence, degree hubs for the HeMLN can be correctly computed. □

Note that anything less than the above is not sufficient (although may not be necessary in many cases) to compute degree hubs correctly. Any node can become a hub in the HeMLN depending upon the number of inter-layer edges connected to it which is not available prior to composition step. **In order to understand this better, we conducted experiments using a percentage of high degree nodes (instead of all nodes) and it turns out that a smaller percentage – 25% to 50% – can give very high accuracy (shown in Fig. 9 of Section 5.)**

The two heuristics also clearly indicate the relevance and amount of additional information needed for the decoupling-based approach for improving accuracy. If only precision is required, less additional information can be used. We will also show that even with the additional information used for obtaining ground truth accuracy, the cost incurred by the decoupling-based approach is significantly less than that of ground truth computation cost.

On the basis of lemma 2, the second heuristic has been proposed to eliminate both false negatives and false positives to provide accuracy guarantee. In this case, from the analysis phase *along with the degree hubs information, we also retain the degree values of the remaining nodes* from each layer. We do not add any changes to the composition function, so it remains the same as HeMLN-PG, where the degree of a node is updated if there is an inter-layer edge incident to that node. *Thus, we are able to correctly compute the degree value of every node using this heuristic.* This is also illustrated in the bottom half of Figure 4, where all the hub nodes (**E, F, Q**) are correctly generated.

5. Experimental Analysis

5.1. Data Sets

Both synthetic and real-world data sets have been used for validating accuracy and performance gain. Subgen [19] and Recursive-Matrix (R-MAT) ([20]) are used to create synthetic data sets. A parallel version of R-MAT termed Parallel R-MAT (PaRMAT) has been used to create large data sets.

Synthetic Graphs: Table 2 provides a list of synthetic data sets we used in our experiments along with their characteristics. Graph sizes vary from 25K vertices and 100K edges to 200K vertices and 10 Million edges. Their characteristics vary as well in terms of node distribution in layers, sparsity, number of connected components, and average degree.

Layer	#Nodes	#Intra-Layer Edges	Inter-Layer	#Inter-Layer Edges
Actor	9486	996527	Actor-Director	32033
Director	4511	250845	Actor-Movie	31422
Movie	7952	8777618	Director-Movie	8581

(a) IMDb Data Set

Layer	#Nodes	#Intra-Layer Edges	Inter-Layer	#Inter-Layer Edges
Coauthor	16918	2483	Author-Paper	37142
Papers	10326	12044080	Author-Year	29984
Year	18	18	Paper-Year	10326

(b) DBLP Data Set

Table 1
Graph Characteristics of Real World Data Sets

Real-world Data Sets: In addition to the above, experiments have been performed on the International Movie Database (IMDb), the DBLP Computer Science Bibliography, and data sets from The Laboratory for Web Algorithmics [21, 22]. Only IMDb (Table 1a) and DBLP (Table 1b) details have been provided due to space constraints.

Data Sets	Layers	#Nodes	#Edges	Node distribution (%)	Average degree	Max degree	Min degree	Dangling nodes	#connected components	Largest component	Sparsity (%)
25KV100KE		25000	100000		8	465	0	4385	4422	20541	0.02
	Layer 1	17572	49588	70	5.643979058	241	0	4065	4124	13388	0.02
	Layer 2	7430	8757	30	2.357200538	134	0	3040	3128	4201	0.02
	Interlayer		41655								
25KV200KE		25000	200000		16	926	0	2309	2320	22671	0.03
	Layer 1	17452	97894	70	11.21865689	636	0	2272	2287	15152	0.03
	Layer 2	7550	18174	30	4.814304636	201	0	1962	1987	5536	0.03
	Interlayer		83932								
25KV300KE		25000	300000		24	1307	0	1399	1402	23597	0.05
	Layer 1	17604	150466	70	17.09452397	925	0	1428	1434	16166	0.05
	Layer 2	7398	25380	30	6.861313869	242	0	1459	1469	5921	0.05
	Interlayer		124154								
25KV400KE		25000	400000		32	1716	0	996	999	24000	0.06
	Layer 1	17469	195675	70	22.40254165	935	0	1060	1063	16405	0.06
	Layer 2	7533	36026	30	9.564848002	535	0	1145	1154	6372	0.06
	Interlayer		168299								
35KV400KE		35000	400000		22.85	1262	0	1966	1970	33028	0.03
	Layer 1	24580	194754	70	15.8465419	656	0	2100	2107	22468	0.03
	Layer 2	10422	36535	30	7.011130301	378	0	1985	2003	8401	0.03
	Interlayer		168711								
45KV400KE		45000	400000		17.77	1257	0	4115	4130	40857	0.02
	Layer 1	31442	193573	70	12.3130208	664	0	4062	4082	27342	0.02
	Layer 2	13560	37333	30	5.506342183	392	0	3419	3459	10059	0.02
	Interlayer		169094								
55KV400KE		55000	400000		14.5	1237	0	6130	6153	48824	0.01
	Layer 1	38697	192727	70	9.960823836	832	0	5972	6017	32637	0.01
	Layer 2	16305	37469	30	4.596013493	225	0	4815	4879	11361	0.01
	Interlayer		169804								
100KV1ME		100000	1000000		20	2150	0	8976	9003	90972	0.01
	Layer 1	69984	496482	70	14.18844307	1529	0	8894	8923	61034	0.01
	Layer 2	30018	86976	30	5.794923046	324	0	7567	7658	22268	0.01
	Interlayer		416542								
100KV2ME		100000	2000000		40	4041	0	4260	4265	95732	0.02
	Layer 1	79890	1265733	80	31.69	3227	0	4455	4467	75413	0.02
	Layer 2	20112	83864	20	8.34	630	0	4004	4038	16041	0.02
	Interlayer		650403								
100KV3ME		100000	3000000		60	5686	0	2353	2354	97647	0.03
	Layer 1	90073	2451052	90	54.42	5160	0	2420	2421	87653	0.03
	Layer 2	9929	27996	10	5.64	321	0	2572	2596	7306	0.03
	Interlayer		520952								
100KV4ME		100000	4000000		80	7349	0	1629	1630	98371	0.04
	Layer 1	90108	3241201	90	71.94	6599	0	1722	1723	88386	0.04
	Layer 2	9894	39890	10	8.06	409	0	1997	2013	7866	0.04
	Interlayer		718909								
200KV1ME		200000	1000000		10	1423	0	36881	37097	162684	0.00
	Layer 1	160073	643022	80	8.03	1144	0	34761	35017	124794	0.00
	Layer 2	39929	39062	20	1.96	209	0	19337	19813	19597	0.00
	Interlayer		317916								
200KV5ME		200000	5000000		50	6702	0	7373	7375	192625	0.01
	Layer 1	119908	1792650	60	29.9	2993	0	7923	7938	111957	0.01
	Layer 2	80094	805373	40	20.11	2701	0	8038	8060	72014	0.01
	Interlayer		2401977								
200KV10ME		200000	10000000		100	12486	0	2842	2844	197156	0.03
	Layer 1	80085	1610469	40	40.22	2712	0	3716	3722	76359	0.03
	Layer 2	119917	3584044	60	59.78	7566	0	3438	3440	116477	0.02
	Interlayer		48055487								

Table 2

Graph Characteristics of Large Synthetic Data Sets

5.2. Experimental Results of HeMLN Degree Centrality

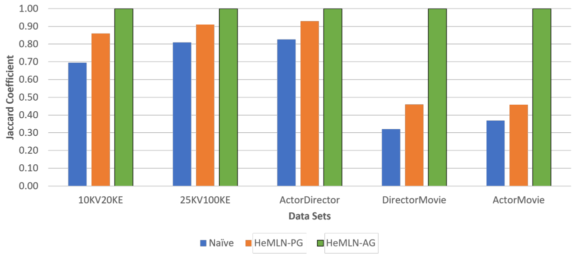


Figure 5: Accuracy: HeMLN-PG and HeMLN-AG consistently better than Naive

In this section, we present our experimental results and analysis of the performance of the two proposed heuristics with respect to accuracy (Jaccard Coefficient) and computational costs.

Figure 5 establishes that both the proposed heuristics give *consistently higher accuracy* as compared to the baseline naive approach with respect to the ground truth.

Figure 6 validates that even though heuristic **HeMLN-PG is able to eliminate false positives and provide a precision guarantee**, it is not able to eliminate false negatives but is still no worse than the naive approach. All our experiments (as per lemma 2) validate that heuristic **HeMLN-AG gives 100% accuracy across all synthetic and real-world data sets of diverse characteristics**. Figure 7 shows selected results for heuristic HeMLN-AG due to space constraints.

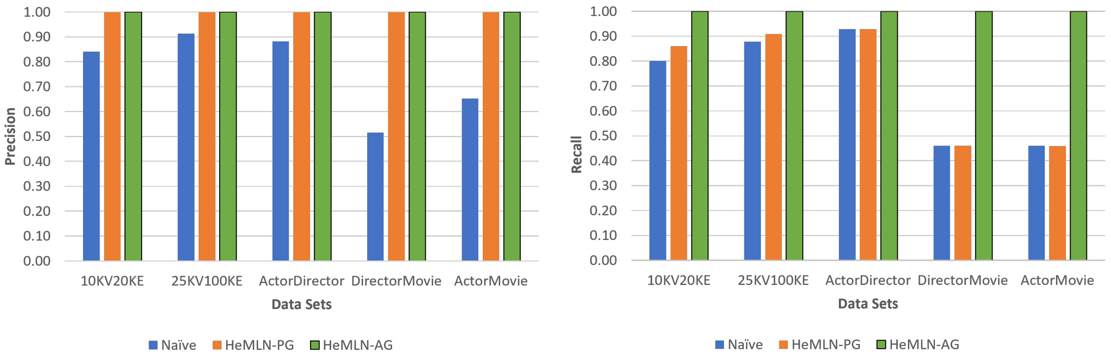


Figure 6: Precision and Recall Comparison of HeMLN-PG and HeMLN-AG

Finally, the efficiency evaluation of heuristic HeMLN-AG has been presented in Figure 8 as compared to the ground truth evaluation. The time taken for HeMLN-AG is calculated as the sum of the maximum time taken for the layers (as layer analysis is done in parallel) and the composition time. The experiments validate that heuristic HeMLN-AG generated 100% accuracy with **savings in computation time** ranging between **15% to 47% for synthetic data sets**, and between **3% to 68% for real-world data sets**.

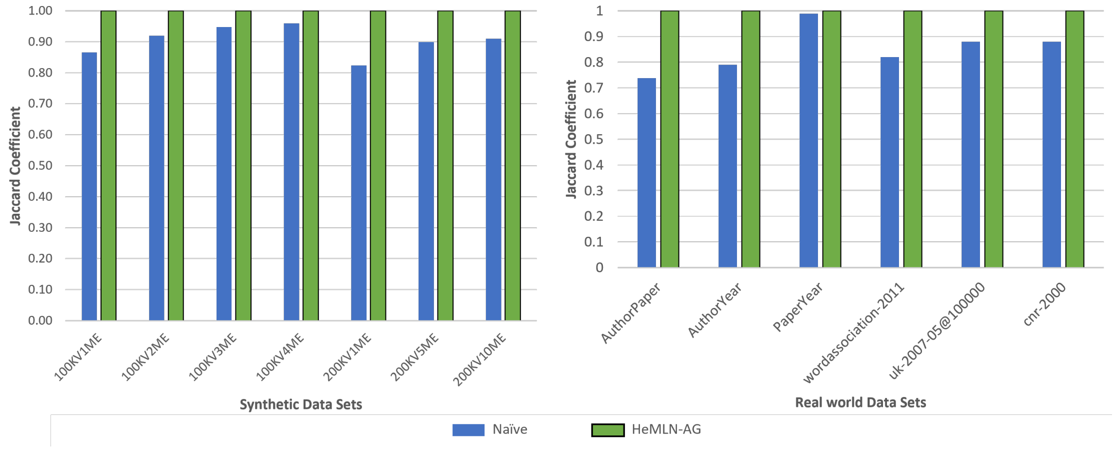


Figure 7: 100% Accuracy of Heuristic 2 across diverse data sets

We have applied HeMLN-AG as a binary function to compute the results of HeMLNs with more than 2 layers. We have analyzed our results of HeMLN-AG up to three layers for real-world data sets IMDb and DBLP and found out the accuracy of data sets remain 100% with significant savings in computational times.

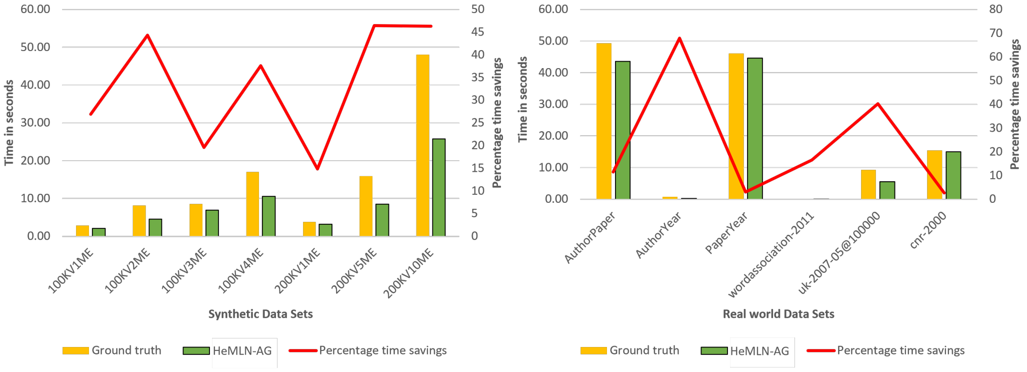


Figure 8: Significant Savings in Computational costs with HeMLN-AG

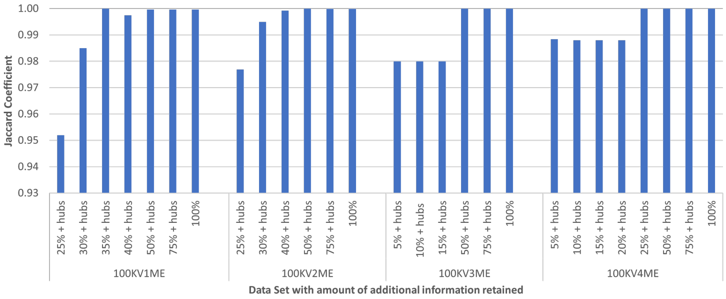


Figure 9: Impact of Additional Information on Accuracy

Effect of Additional Information: We performed experiments to understand the effect of additional layer-wise information. For heuristic HeMLN-PG (*only hub degrees maintained* - one end of the spectrum) and heuristic HeMLN-AG (*all node degrees maintained* - the other end of the spectrum), it can be clearly observed that the increase in accuracy in HeMLN-AG comes at the cost of an increase in the amount of information retained from each layer. Moreover, we modified HeMLN-AG, to include from the layers the degree information of the top x% nodes (sorted in the decreasing order of degree values) + the hub nodes. Figure 9 demonstrates for few of the larger synthetic data sets (100KV1ME to 100KV4ME) how maintaining such additional information gradually increases the accuracy until it reaches a saturation point. However, more layer information increases composition phase cost as well. This validates the **trade-off between the savings in computational cost and the accuracy of the results**.

6. Conclusions

In this paper, we defined degree centrality for heterogeneous multilayer networks. Based on intuition of degree hubs, we proposed two heuristics for the decoupling approach - one that provides **precision guarantee** and the other that provides **accuracy guarantee**. We were able to prove the need for minimal additional information to obtain the guarantees, Each may be useful for different applications. A large number of experiments were performed on diverse synthetic and real-world data sets to validate accuracy, precision, and efficiency of our algorithms. Information retained and accuracy versus efficiency trade-off was also demonstrated.

Acknowledgments: For this work, Drs. Sharma Chakravarthy and Abhishek Santra were partly supported by NSF Grant CCF-1955798. Dr. Sharma Chakravarthy was also partly supported by NSF Grant CNS-2120393.

References

[1] S. Fortunato, C. Castellano, Community structure in graphs, in: Ency. of Complexity and Systems Science, 2009.

[2] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter, Multilayer networks, CoRR abs/1309.7233 (2013).

[3] A. Santra, S. Bhowmick, Holistic analysis of multi-source, multi-feature data: Modeling and computation challenges, in: P. K. Reddy, A. Sureka, S. Chakravarthy, S. Bhalla (Eds.), Big Data Analytics - 5th International Conference, BDA 2017, Hyderabad, India, December 12-15, 2017, Proceedings, volume 10721 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 59–68. URL: https://doi.org/10.1007/978-3-319-72413-3_4. doi:10.1007/978-3-319-72413-3_4.

[4] A. Santra, K. S. Komar, S. Bhowmick, S. Chakravarthy, From base data to knowledge discovery - *A life cycle approach* - using multilayer networks, Data Knowl. Eng. 141 (2022) 102058. URL: https://doi.org/10.1016/j.datak.2022.102058. doi:10.1016/j.datak.2022.102058.

[5] M. D. Domenico, V. Nicosia, A. Arenas, V. Latora, Layer aggregation and reducibility of multilayer interconnected networks, CoRR abs/1405.0425 (2014).

[6] A. Berenstein, M. P. Magarinos, A. Chernomoretz, F. Aguero, A multilayer network approach for guiding drug repositioning in neglected diseases, PLOS (2016).

[7] M. De Domenico, A. Solé-Ribalta, S. Gómez, A. Arenas, Navigability of interconnected networks under random failures, Proceedings of the National Academy of Sciences (2014). doi:10.1073/pnas.1318469111. arXiv:https://www.pnas.org/content/early/2014/05/21/1318469111.full.pdf.

[8] E. Cohen, D. Delling, T. Pajor, R. F. Werneck, Computing classic closeness centrality, at scale, in: Proceedings of the Second ACM Conference on Online Social Networks, COSN '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 37–50. URL: https://doi.org/10.1145/2660460.2660465. doi:10.1145/2660460.2660465.

[9] A. Santra, Analysis of Complex Data Sets Using Multilayer Networks: A Decoupling-based Framework, Ph.D. thesis, The University of Texas at Arlington, 2020. https://itlab.uta.edu/students/alumni/PhD/Abhishek_Santra/ASantra_PhD2020.pdf.

[10] D. Sharma, A. Surolia, Degree Centrality, Springer New York, New York, NY, 2013, pp. 558–558. URL: https://doi.org/10.1007/978-1-4419-9863-7_935. doi:10.1007/978-1-4419-9863-7_935.

[11] L. C. Freeman, Centrality in social networks conceptual clarification, Social networks 1 (1978) 215–239.

[12] A. Santra, S. Bhowmick, S. Chakravarthy, Hubify: Efficient estimation of central entities across multiplex layer compositions, in: R. Gottumukkala, X. Ning, G. Dong, V. Raghavan, S. Aluru, G. Karypis, L. Miele, X. Wu (Eds.), 2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017, IEEE Computer Society, 2017, pp. 142–149. URL: https://doi.org/10.1109/ICDMW.2017.24. doi:10.1109/ICDMW.2017.24.

[13] P. Bródka, K. Skibicki, P. Kazienko, K. Musial, A degree centrality in multi-layered social network, 2011, pp. 237–242. doi:10.1109/CASON.2011.6085951.

[14] B. Oselio, A. Kulesza, A. O. Hero, Multi-layer graph analysis for dynamic social networks, IEEE Journal of Selected Topics in Signal Processing 8 (2014) 514–523.

[15] M. De Domenico, Multilayer modeling and analysis of human brain networks, Giga Science 6 (2017) gix004.

[16] R. Casarin, M. Iacopini, G. Molina, E. Ter Horst, R. Espinasa, C. Sucre, R. Rigobon, Multilayer network analysis of oil linkages, The Econometrics Journal 23 (2020) 269–296.

[17] A. Santra, S. Bhowmick, S. Chakravarthy, Efficient community re-creation in multilayer networks using boolean operations, in: P. Koumoutsakos, M. Lees, V. V. Krzhizhanovskaya, J. J. Dongarra, P. M. A. Sloot (Eds.), International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland, volume 108 of *Procedia Computer Science*, Elsevier, 2017, pp. 58–67. URL: https://doi.org/10.1016/j.procs.2017.05.246. doi:10.1016/j.procs.2017.05.246.

[18] H. R. Pavel, A. Santra, S. Chakravarthy, Degree centrality algorithms for homogeneous multilayer networks, in: F. Coenen, A. L. N. Fred, J. Filipe (Eds.), Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2022, Volume 1: KDIR, Valletta, Malta, October 24-26, 2022, SCITEPRESS, 2022, pp. 51–62. URL: https://doi.org/10.5220/0011528900003335. doi:10.5220/0011528900003335.

[19] AI@WSU, http://ailab.wsu.edu/subdue/download.htm, 2011. [Online].

[20] D. Chakrabarti, Y. Zhan, C. Faloutsos, R-mat: A recursive model for graph mining, in: Proceedings of the 2004 SIAM International Conference on Data Mining, SIAM, 2004, pp. 442–446.

[21] P. Boldi, S. Vigna, The WebGraph framework I: Compression techniques, in: Proc. of the Thirteenth International World Wide Web Conference (WWW 2004), ACM Press, Manhattan, USA, 2004, pp. 595–601.

[22] P. Boldi, M. Rosa, M. Santini, S. Vigna, Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks, in: S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, R. Kumar (Eds.), Proceedings of the 20th international conference on World Wide Web, ACM Press, 2011, pp. 587–596.