HUBify: Efficient Estimation of Central Entities across Multiplex Layer Compositions

Abhishek Santra*, Sanjukta Bhowmick† and Sharma Chakravarthy‡
* ‡Information Technology Laboratory, CSE Department, University of Texas at Arlington, Arlington, Texas, USA
†CSE Department, University of Nebraska at Omaha, Omaha, Nebraska, USA
Email: *abhishek.santra@mavs.uta.edu, †sbhowmick@unomaha.edu, ‡sharma@cse.uta.edu

***Abstract*—Graphs or networks are a natural way to analyze inter-related set of entities. When these entities are associated with a diverse number of features, each denoting a specific perspective, then the representation can be simplified by forming a network of layers (one for each feature) or multiplexes. Vertices with high centrality values in the multiplexes represent the most influential vertices. However, detecting central entities in multiplexes for different combinations of features becomes computationally expensive, as the number of layers increases.**

In this paper, we address the task of efficiently identifying high centrality vertices for any conjunctive (AND) combination of features (as represented by multiplex layers.) We propose efficient heuristics that only use results from individual layers to identify high degree and high closeness centrality vertices. Our approaches, when applied to real-world, multi-featured datasets such as IMDb and traffic accidents, show that we can identify the high centrality vertices with an average accuracy of more than *70-80%* while reducing the overall computational time by at least *30%*.

***Index Terms*—Multiplexes; Graph Analysis; Degree Centrality; Closeness Centrality; Lossless Composability;**

I. Introduction

Networks (or graphs) are used to represent pair-wise relationship between entities in a system. In many cases, entities may be connected by not one but multiple relations. For example, a pair of traffic accidents may be related if they occurred in the same location, or under the same light condition, weather condition etc. Similarly, two actors may be related if they acted in the same genre, such as action, comedy, etc. When multiple features are present, then the relationship pertaining to each feature can be represented as a network. Multiplexes are thus a network of networks, where each individual network (termed layer) denotes a distinct relationship (through edges) based on a feature among the *same set of entities* (or nodes.)

Each individual layer of a multiplex, represents the relationship corresponding to a single feature. While there exist several algorithms for analyzing individual networks, the challenge in analyzing a multiplex is that the analysis has to be recomputed for each combination of layers.

In this paper, we concentrate on finding high degree and closeness centrality vertices, also called hubs, in AND-composed layers of multiplex networks. AND-composed layers denoting the conjunction of perspectives can be obtained by combining the individual layers such that *only edges that are present in every individual layer are retained.* High centrality vertices in the accident dataset can help us in identifying the most dominating traffic accident locations with respect to poor lighting conditions and bad roads and this information can be used to devise appropriate accident prevention techniques. However, in order to obtain a holistic view of the multiplex system with n layers, we have to generate, store and analyze a total of $2^n - 1$ networks, leading to extremely expensive operations for multiplexes with large number of layers (for example the network in [4] has 300 layers.)

Problem Formulation and Contributions: Given this challenge of efficiently finding hubs in multiplexes, the main problem we aim to solve is as follows. Given a dataset with multiple entities that are related via a number of distinct features, how can we efficiently find the most influential entities based on any conjunctive (AND) combination of features.

To solve this problem, we use multiplexes for representing such multi-featured datasets and *present elegant techniques for estimating the hubs for any conjunctively composed multiplex layer, without actually constructing that composed layer.*

Our main contributions are two-fold. **First** we show that finding high centrality vertices in the AND composed multiplexes, based on only analyzing the individual layers is a non-trivial problem, and the naive approach of simply taking the intersection of the hubs from each layer does not produce accurate results. **Second**, we present four heuristics (3 for degree centrality and 1 for closeness centrality) to identify hubs in the AND-composed using only the hubs detected in individual layers and their distance-1 neighbors. Our results show that we can identify the vertices with $70-80\%$ accuracy while reducing the computation time by at least 30%.

Our proposed methods can be extended to any number of layers. This approach significantly reduces the complexity of analyzing the AND-composed network and also the storage as only n individual layers are constructed and analyzed.

The remainder of this paper is organized as follows: In Section III we give an overview of how a multiplex is formed and how to conjunctively combine networks to produce new AND-composed layers. In Section IV, we detect high degree and closeness centrality vertices in each layer. We show how these hub sets vary across different individual and AND-composed layers. In Section V, we present four heuristics to improve the accuracy of computing the degree or closeness centrality based hubs of any conjunctive combination of layers by using the required layer-wise hubs.

In Section VI, we empirically validate the quality of the hub

sets generated by executing our algorithms on two diverse data sets: traffic accidents and IMDb. We use the Jaccard Index to compare the set of hubs obtained through our heuristics with the actual set of hubs. We show that our approach can significantly reduce the computational costs of finding hubs in the composed networks.

II. Related Work

Recently, significant amount of work has been done in the area of multilayer networks [3], [11] to handle varying interactions among the same set of entities such as co-authorship relationship in different conferences [4], citation relationship across different topics [13], interaction relationships based on calls/bluetooth scans [9], connection relationships across different social media platforms [12] and multilayer protein-protein interactions [8]. Most of this work focuses on *overall multiplex diagnostics* by considering the multiplex layers *individually*. However, in order to understand the effect of multiple features using composition of individual multiplex layers, we need a principled approach to arbitrarily combine features without having to construct combined layers and analyze them.

Using multiplex representation schemes such as adjacency tensors [5] are also not efficient as computations based on any subset of layers will require the loading of the entire multiplex tensor, thus increasing the computational complexity.

Santra et. al. [15] proposed an approach for efficiently re-creating communities of any combination of layers by performing Boolean operations on the communities obtained from the individual layers. In this paper, we take inspiration from their work and propose novel cost-effective heuristics that are able to estimate highly accurate hub sets for any conjunctive combination of layers.

Degree centrality [10] and closeness centrality [7], [14] have been used in monoplex (single layer network) to detect high centrality nodes. There has been work in determining centrality measures by aggregating all the layers of a multiplex [6] or performing walks across layers [16]. However, to the best of our knowledge, the problem of *inferring the degree centrality or closeness centrality hubs of any arbitrary conjunctively combined network from hubs of individual layers, in a cost-effective manner, has not been addressed earlier*.

III. Multiplexes: A Brief Overview

In this section, we give an overview of how multi-featured datasets like Internet Movie Database (IMDb) and traffic accident dataset can be modeled as multiplexes. We also show how conjunctive composition of layers presents a new perspective and discuss the benefits of multiplex-based modeling.

Multi-Source or Multi-featured Datasets: In multi-featured datasets, the relationship between any two entities can be defined in multiple ways. For example, the interaction among people can be through various media such as email, phone conversations, social networking, etc., the similarity among the accidents can be based on different factors such as light, weather, road conditions, etc., two actors can be related based on the different movie genres in which they have acted together, such as comedy, action, etc. In a multiplex, the relation due to each feature is represented through a network. Two vertices are connected if they exhibit a relation based on feature represented in the network. The networks for each feature together form a multiplex. The set of entities, represented by nodes, remains the same in each layer. For example, Figure 1 (a) shows an accident multiplex depicting the similarity among 7 accident occurrences based on light (G_{a1}) and weather (G_{a2}) conditions. Similarly, in Figure 1 (b), the IMDb multiplex depicts the co-actor relationship among 6 actors based on the movie genres, comedy (G_{m1}) and action (G_{m2}). The notations mentioned in Table I have been used to formalize the various concepts discussed in this paper.

TABLE I
List of notations used for defining the concepts.

I	Set of entities
f	Set of features/perspectives
$G(V_k, E_k)/G_k$	The k^{th} layer
u_i^k	Representative node for i^{th} entity in the k^{th} layer
$NBD_k(u_i^k)$	Set of nodes adjacent to the i^{th} node in the k^{th} layer
deg_i^k	Degree of the i^{th} node in the k^{th} layer
$avgDeg^k$	Average degree of the k^{th} layer
clo_i^k	Closeness centrality of the i^{th} node in the k^{th} layer
$avgClo^k$	Average closeness centrality of the k^{th} layer
V_k	Set of nodes in the k^{th} layer
(u_i^k, u_j^k)	An edge in the k^{th} layer
E_k	Set of edges in the k^{th} layer
DH_k	Set of degree centrality based hubs in k^{th} layer
CH_k	Set of closeness centrality based hubs in k^{th} layer

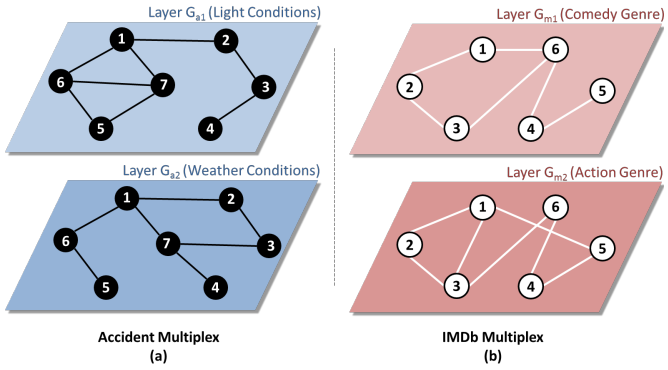

Fig. 1. Snapshots of accident and IMDb multiplexes

The distinct co-actor (or accident-accident) connectivity in each layer shows that every genre (or factor) presents a unique way of analyzing the same set of actors (or accidents). For instance - accident 4 and accident 7 were not caused by the same lighting conditions, but the weather conditions at the time of occurrence were similar. Similarly, actor 3 is one of the most paired actors in the action genre, whereas in the comedy genre actor 6 has worked with most of the other actors.

Composition of Multiplex Layers: In addition to analyzing individual layers, it is also important to study the effect of different combinations of features on the given set of entities. In this paper, we compose any two individual layers in a conjunctive (AND-based) manner, i.e. link will exist in the composed layer if it exists in *both* the individual layers.

Formally, if G_x and G_y are two individual layers of a multiplex, then the AND-composed layer, G_{xANDy}, will be constructed by including the edges that are part of both G_x and G_y. For example, Figure 2 (a) shows the AND-composed layer, $G_{a1ANDa2}$ generated by linking those accidents that have similar lighting and weather conditions at the time of occurrences. Similarly, in Figure 2 (b) the AND-composed layer $G_{m1ANDm2}$ denotes the co-actors present in both the comedy layer, G_{m1} and the action layer, G_{m2}. Any AND-composed layer will have same set of nodes as its constituent layers. However, the upper bound on the number of edges, $|E_{iANDj}|$, will be $min(|E_i|, |E_j|)$. The AND-composition can be extended to multiple layers of the network.

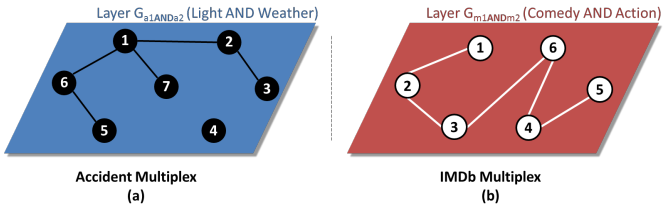

Fig. 2. AND Compositions using the individual layers from Figure 1

Benefits of Multiplex-based Modeling: Modeling of multi-featured data as multiplexes allows *ease of handling the dataset incrementally* through the addition of nodes (when a new accident or actor is encountered), edges (to represent the new entity's relationships with the earlier entities) or layers (to account for fresh perspectives). Moreover, a latest snapshot of multiplex can be easily maintained through the deletion of obsolete entities (nodes), relationships (edges) or perspectives (layers). Further, this modeling facilitates the study of relationships among the entities with respect to individual as well as combination of different features.

IV. Hubs (High Centrality Vertices) across Multiplex Layers

Entities vary in their influencing capability with respect to the occurrence of events, interaction networks and so on. For example, a particular person might be considered highly influential if he/she is connected to a large majority of people on Facebook. Thus, an advertisement agency will prefer this person in order to enhance their information transfer. However, he/she may not be equally influential on LinkedIn. Thus, in case of multi-featured data, the influencing capability for a particular entity may vary substantially with features. With respect to multiplexes, this translates to generating the hubs across different individual or AND-composed layers.

Degree Centrality (deg_i^k): The number of nodes adjacent to the i^{th} vertex in the k^{th} multiplex layer defines a vertex's layer specific degree. The higher is the degree of a node, greater is its influence on the immediate neighborhood. We define high centrality nodes or hubs in the k^{th} layer (or feature) as the ones that have a degree greater than the average degree of the layer, $avgDeg^k$, which is computed by $\frac{2|E_k|}{|V_k|}$. Figure 3 (a) encircles the accident nodes in red that have been detected as hubs due to their greater than average degree.

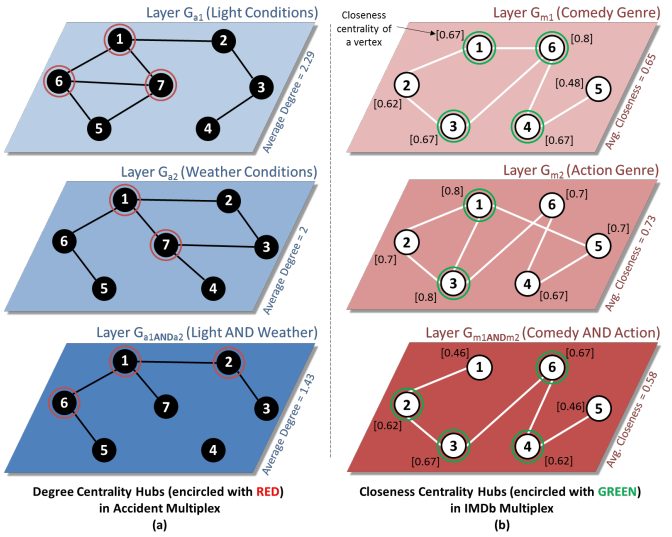

Fig. 3. Variation in the Degree and Closeness Centrality based Hubs across Different Individual and Composed Multiplex Layers

Closeness Centrality (clo_i^k): The closeness centrality of a node measures how close are the other nodes in the network from it. Therefore, closeness centrality of the i^{th} vertex in the k^{th} multiplex layer is defined by the average of the summation of reciprocal of shortest paths between the i^{th} node and every other node in the layer. We use the valued closeness centrality variant proposed in [7], [14] as any multiplex layer need not be comprised of a single connected component. Therefore, $clo_i^k = \frac{1}{|V_k|-1} \sum_{j=1, j\neq i}^{|V_k|} \frac{1}{d(u_i^k, u_j^k)}$, where $d(u_i^k, u_j^k)$ is the shortest path between the i^{th} and the j^{th} vertex in the k^{th} layer. The higher is the closeness centrality of a node, closer it is from all other nodes in the layer and greater will be its influence on the network. We define the high centrality nodes or hubs in the k^{th} layer (or feature) as the ones that have their closeness centrality metric value greater than the average closeness centrality of the layer, $avgClo^k$, which is computed by $\frac{\sum_{i=1}^{|V_k|} clo_i^k}{|V_k|}$. Figure 3 (b) encircles the actor nodes in green that have been detected as hubs based on closeness centrality.

Characteristics of Hubs in the Composed Layers: In Figure 3, we show using simple examples that finding hubs of the composed layer from the individual hubs is a non-trivial problem. In some cases, such as for actor 4 (or accident 6) *a vertex may be a hub in the composed layer even if it is not a hub in both the layers.* Further, the actor 1 and accident 7 illustrate that *a node that is a hub in both individual layers may not be a hub in the AND-composed layer*. Moreover, there

can be some entities like actor 2 and accident 2 that are *hubs in the AND-composed layer in spite of not being a hub in either of the individual layers.* This is due to the fact that edge connectivity varies across individual and composed layers, thus effecting the values of degree centrality and closeness centrality. Our goal is to develop heuristics that can take into account these connectivity patterns and identify the hubs in the AND-composed layer using the hubs of the individual layers.

V. Identifying Hubs in AND-composed Multiplexes

In this section, we introduce four heuristics to identify the degree or closeness centrality hub sets in the AND-composed layer using information about the hubs in the individual layers. Our techniques eliminate the need to generate, store and compute degrees and shortest paths for the AND-composed layers, thus reducing the computational complexity.

For the following discussion, let us assume the two individual layers to be G_x and G_y, with degree centrality based hub sets, DH_x and DH_y, respectively, and closeness centrality based hub sets, CH_x and CH_y, respectively. Further, let us suppose that DH_{xANDy} and CH_{xANDy} are the *actual* degree and closeness centrality based hub sets, respectively, for the AND-composed layer, G_{xANDy}.

A. Estimating Hubs based on Degree Centrality

As shown in Figure 3 (a), a) a node that is not high degree in the individual layers may share enough neighbors across layers to become a hub in the AND-composed layer, whereas b) the node that is a hub across layers may lose its hub property after AND-composition due to the absence of common neighbors. Therefore, the naive way of taking the intersection of layer-wise hubs to find the hubs in the AND-composed layer will generate a large number of false positives and false negatives. Here we propose and discuss three heuristics to estimate degree centrality based hub set of the AND-composed layer.

Heuristic DC1: To reduce the false positives, we estimate the average degree of the AND-composed layer, $avgDeg_{est}^{xANDy}$. Note that the upper bound on the average degree in the AND-composed networks will be the minimum average degree from the individual layers. Therefore, $avgDeg^{xANDy} \leq min(avgDeg^x, avgDeg^x)$. We set the estimated average degree of the AND-composed network to this upper bound: $avgDeg_{est}^{xANDy} = min(avgDeg^x, avgDeg^x)$.

We first obtain the vertices from the intersection of the hubs in the individual layers, i.e. all nodes $u \in DH_x \cap DH_y$. We then check whether these nodes have a common set of one hop neighbors in their individual layers. The larger the set of common neighbors, the greater the degree in the AND-composed network. Formally we only retain the vertex u as a hub if $|NBD_x(u) \cap NBD_y(u)| > avgDeg_{est}^{xANDy}$, where $NBD_x(u)$ and $NBD_y(u)$ denote the sets of one hop neighbors of vertex u in G_x and G_y, respectively.

Heuristic DC2: In the above heuristic, if $avgDeg_{est}^{xANDy}$ is much larger than $avgDeg^{xANDy}$, then a common hub in spite of sharing enough neighbors across the individual layers will not be generated as a hub in the composed layer. A better estimate for the AND-composed layer's average degree is obtained by *maintaining the degree of each vertex in every individual layer*. In the AND-composed layer, the number of neighbors for any vertex will be at most that vertex's least degree among all individual layers. That is, $deg_i^{xANDy} \leq min(deg_i^x, deg_i^y)$. This implies, $avgDeg^{xANDy} \leq \frac{1}{|v_x|}\sum_{i=1}^{V_x} min(deg_i^x, deg_i^y)$. We set the estimated average degree of the AND-composed network to this upper bound, $avgDeg_{est}^{xANDy} = \frac{1}{|v_x|}\sum_{i=1}^{V_x} min(deg_i^x, deg_i^y)$. We execute the steps in heuristic DC1 with this improved estimate. This method provides a better accuracy as compared to DC1, but the computational cost increases.

Algorithm 1 Procedure for Heuristic DC1

Require: DH_x, $avgDeg^x$, DH_y, $avgDeg^y$, $DH'_{xANDy} = \emptyset$

1: $avgDeg_{est}^{xANDy} = min(avgDeg^x, avgDeg^x)$.
2: **for all** $u \in DH_x$ **do**
3: $NBD_x(u) \leftarrow$ one hop neighbors of u in G_x
4: **end for**
5: **for all** $u \in DH_y$ **do**
6: $NBD_y(u) \leftarrow$ one hop neighbors of u in G_y
7: **end for**
8: **for all** $u \in DH_x \cap DH_y$ **do**
9: **if** $|NBD_x(u) \cap NBD_y(u)| > avgDeg_{est}^{xANDy}$ **then**
10: $DH'_{xANDy} \leftarrow DH'_{xANDy} \cup u$
11: **end if**
12: **end for**

Heuristic DC3: Heuristics DC1 and DC2 reduce false positives but cannot handle false negatives. Specifically they miss out vertices that are hubs in the AND-composed layer but are not hubs in at least one of the individual layers. For handling this case, we maintain few low degree nodes from each individual layer that have a degree close to the average degree. That is, *if* $deg_i^x > (1-\epsilon)avgDeg^x$*, then insert the vertex in* DH_x*, where* $0 \leq \epsilon \leq 1$, and we similarly update DH_y. Therefore, executing heuristic DC2 with these updated layer-wise hub sets, will also generate those nodes that are non-hubs in at least one of the individual layers, but share enough neighbors across layers to become hubs in the AND-composed layer. The higher is the value of ϵ, more accurate will be the estimated hub set. This increased accuracy comes at a cost of maintaining more overhead information. Thus, from DC2 and DC3 it is evident that there is a trade-off between accuracy and savings in computational costs.

Discussion: If the topology of the individual layers, G_x and G_y is similar, then most of the layer-wise hubs will also be hubs in the AND-composed networks and the naive approach can give a good estimation. Also note that if the average degree estimate for the AND-composed layer is not close enough to the actual average degree then even an ϵ value of 1 may not give 100% accuracy due to the exclusion of common hubs and non-hubs that share more than actual but less than estimated average degree number of neighbors across layers. Therefore, the effectiveness of our heuristics depends on the fraction of

Algorithm 2 Procedure for Heuristic DC3

Require: DH_x, $deg_i^x \ \forall \ u_i^x$, $avgDeg^x$, DH_y, $deg_i^y \ \forall \ u_i^y$, $avgDeg^x$, ϵ, $DH'_{xANDy} = \emptyset$

for all $u_i^x \in V_x$ **do**
  if $deg_i^x > (1-\epsilon)avgDeg^x$ **then**
    $DH_x \leftarrow DH_x \cup u_i^x$
  end if
end for
for all $u_i^y \in V_y$ **do**
  if $deg_i^y > (1-\epsilon)avgDeg^y$ **then**
    $DH_y \leftarrow DH_y \cup u_i^y$
  end if
end for
execute Heuristic DC2 with updated DH_x and DH_y.

AND-composition hubs that are common to the layers, average degree estimate and the value of ϵ.

B. Estimating Hubs based on Closeness Centrality

Closeness centrality depends on the shortest paths between any two nodes. As shown in Figure 3 (b) that even if a certain node is closest to all the remaining nodes in the individual layers, it may not be a hub in the AND-composed layer due to the absence of common paths between this node and every other node, that are short enough. Therefore, the naive way of intersecting the layer-wise closeness centrality based hubs will generate false positives. We propose and analyze a heuristic that maintains minimal neighborhood information to estimate the closeness centrality hubs for the AND-composed layer.

Heuristic CC1: From a high closeness centrality node we can traverse the entire network in *minimum number of hops*. Therefore, if high degree nodes are close to a node, the chances of this node becoming a high closeness centrality node increase. Therefore, one way of eliminating the false positives is to check whether the common closeness centrality hubs share high degree neighbors across layers.

Based on this observation, we propose the following heuristic. Initially, for every node, $u \in CH_x$ (or, $u \in CH_y$), we obtain the set of degree based hubs present in its one hop neighborhood, $degNBD_x(u)$ (or $degNBD_y(u)$). We estimate the degree based hub set for AND-composed layer, DH'_{xANDy}, using one of the heuristics discussed above. We then obtain the set of common closeness centrality hubs from CH_x and CH_y. For each of these vertices, we obtain the set of those common degree based hubs in the one hop neighborhood that are also estimated to be hubs in the AND-composed layer. The larger the size of this set, greater are the chances of a node to remain a high closeness centrality node even in the AND-composed layer. Formally, we only retain a vertex u as a closeness centrality based hub if $|degNBD_x(u) \cap degNBD_y(u) \cap DH'_{xANDy}| \geq 1$.

Discussion: If the topology of layer G_x is similar to G_y, then the shortest paths between most of the node pairs will be common. In such a case, the naive approach is capable of generating good hub set estimates of the layer G_{xANDy}.

Algorithm 3 Procedure for Heuristic CC1

Require: CH_x, DH_x, CH_y, DH_y, DH'_{xANDy}, $CH'_{xANDy} = \emptyset$

for all $u \in CH_x$ **do**
  $degNBD_x(u) = \emptyset$
  for all $v \in NBD_x(u)$ **do**
    if $v \in DH_x$ **then**
      $degNBD_x(u) \leftarrow degNBD_x(u) \cup v$
    end if
  end for
end for
for all $u \in CH_y$ **do**
  $degNBD_y(u) = \emptyset$
  for all $v \in NBD_y(u)$ **do**
    if $v \in DH_y$ **then**
      $degNBD_y(u) \leftarrow degNBD_y(u) \cup v$
    end if
  end for
end for
for all $u \in CH_x \cap CH_y$ **do**
  if $|degNBD_x(u) \cap degNBD_y(u) \cap DH'_{xANDy}| \geq 1$ **then**
    $CH'_{xANDy} \leftarrow CH'_{xANDy} \cup u$
  end if
end for

Maintaining information about the alternate paths to every degree based hub beyond 2-3 hops from the closeness centrality hubs and similar path information about some layer-wise non-closeness centrality based hubs will improve the accuracy of the heuristic. However, due to the large overhead costs the computational time will significantly increase.

C. Estimation of Hubs in k-layer AND Compositions

The input to any of the above heuristics is two hub sets that may either be the actual hub sets of individual layers or the estimated hub sets of AND-composed layers. For any 3 layers, G_x, G_y and G_z, the average degree estimation and neighborhood intersection are both commutative and associative. Therefore, the four proposed heuristics are also commutative ($DH'_{xANDy} = DH'_{yANDx}$, $CH'_{xANDy} = CH'_{yANDx}$) and associative ($DH'_{(xANDy)ANDz} = DH'_{xAND(yANDz)}$, $CH'_{(xANDy)ANDz} = CH'_{xAND(yANDz)}$). Therefore, to estimate the hub sets of a k-layer AND-composed network, any heuristic is applied on the k/2 pairs of hub sets, in parallel, generating k/2 AND-composed hub sets, and so on until the final estimated set of hubs, corresponding to the k-layer AND-composed network, is obtained. Thus, in this way for a multiplex with n layers, the $2^n - n$ AND-composition hub sets can be estimated by only using n layer-wise hub sets and minimal overhead information.

VI. Experimental Analysis

In this section we present our experimental results on the performance of the four proposed heuristics to estimate the hub

sets of the AND-composed multiplex layers with respect to accuracy and computational costs. Specifically, we i) construct multiplexes for datasets from diverse domains, ii) generate the AND-composed layers and the actual sets of high centrality nodes, iii) obtain the estimated hub set based on our heuristics and iv) compute accuracy of the estimated hubs based on the actual hub set.

Experimental Setup and Datasets: Our codes are implemented in C++ and executed on a Linux machine with 4 GB RAM and installed with UBUNTU 13.10.

Our experiments are performed on two different multiplexes built from real-life datasets collected from diverse domains - UK Traffic Accidents [2], Internet Movie Database - IMDb [1]. Detailed structure of these multiplexes is as follows:

Accident Multiplex: We use 1000 random road accidents that occurred in the United Kingdom in the year 2014. This multiplex has 3 basic layers with respect to Light Conditions (Domain = {daylight, darkness: lights lit, darkness: lights unlit, darkness: no lighting, darkness: lighting unknown}), Weather Conditions (Domain = {fine + no high winds, raining + no high winds, snowing + no high winds, fine + high winds, raining + high winds, snowing + high winds, fog or mist, other}) and Road Surface Conditions (Domain = {dry, wet or damp, snow, frost or ice, flood, oil or diesel, mud}). An edge in any layer represents that the corresponding accidents occurred within 10 miles of each other and are similar based on light conditions (layer G_{a1}), weather conditions (layer G_{a2}) or road surface conditions (layer G_{a3}).

IMDb Multiplex: This 3-layer multiplex is built with 5000 random actors. An edge in any basic layer signifies that the corresponding actors have worked together in at least one movie that belongs to the Comedy genre (layer G_{m1}), Action genre (layer G_{m2}) or Drama genre (layer G_{m3}).

Actual Hub Sets in the Individual and AND-composed Layers: Apart from the individual multiplex layers, four AND-composed layers each, for the accident multiplex - $G_{a1ANDa2}$, $G_{a1ANDa3}$, $G_{a2ANDa3}$ and $G_{a1ANDa2ANDa3}$, and IMDb multiplex - $G_{m1ANDm2}$, $G_{m1ANDm3}$, $G_{m2ANDm3}$ and $G_{m1ANDm2ANDm3}$, are generated. Every cell in Table II lists percentage of hubs followed by the average degree or closeness centrality for the individual and AND-composed multiplex layers. Variation in this information across layers shows that any combination of layers (or features) presents a unique perspective of analyzing the same set of entities.

Comparison Metrics: We compare the similarity of the estimated hub sets with the actual hub sets using the jaccard index. For any two sets, X and Y, jaccard index, $J_{X,Y} = \frac{|X \cap Y|}{|X \cup Y|}$. If two sets completely overlap, then jaccard index is 1, denoting highest accuracy of 100%. We compute overall accuracy of a heuristic as the mean of the accuracies obtained by estimating hub sets of every AND-Composed layer.

The computational time to generate the actual hub set for any AND-composition includes the time to generate the AND-composed layer followed by the time it takes to compute degree based hubs or shortest paths for closeness centrality based hubs. On the other hand, the time to estimate the hub set for the same AND-composed layer includes time it takes to apply the proposed heuristics using the layer-wise hub sets.

AND-Composed Layer	Accident *(x = a)*		IMDb *(x = m)*	
	$\lvert DH_k \rvert$ avgDeg	$\lvert CH_k \rvert$ avgClo	$\lvert DH_k \rvert$ avgDeg	$\lvert CH_k \rvert$ avgClo
G_{x1}	23.4% 14.92	30.6% 0.0324	34.9% 1.4404	29.4% 0.0181
G_{x2}	20.5% 17.99	36.3% 0.0462	29.4% 0.8564	19% 0.0071
G_{x3}	21.3% 16.44	28.5% 0.0347	47.1% 1.92	39.4% 0.031
$G_{x1ANDx2}$	21% 11.2	28% 0.0251	9.6% 0.1948	9.6% 0.00009
$G_{x1ANDx3}$	20.4% 10.18	25.2% 0.0202	22.7% 0.5176	10.5% 0.0016
$G_{x2ANDx3}$	18.2% 14.35	26.2% 0.0302	11.8% 0.24	9.3% 0.0002
$G_{x1ANDx2ANDx3}$	18.2% 9.28	24.1% 0.0186	1.6% 0.0228	1.6% 0.000005

TABLE II
VARYING HUB INFORMATION DENOTING THE DIVERSE PERSPECTIVES OBTAINED THROUGH MULTIPLEX LAYERS

The Naive Approach: Table III shows that the naive approach of intersecting the layer-wise degree or closeness centrality based hub sets will not guarantee a highly accurate estimated hub set for the AND-composed layers, due to the presence of a large number of false positives. Absence of common immediate neighboring nodes and common shortest paths between nodes across the layers may lead to such low accuracies with the naive approach. However, we observed that the Accident multiplex layers have similar topology due to which the naive approach gives relatively better accuracies as most of the layer-wise hubs are also hubs in the composed layers (Table IV).

AND-Composed Layers	Degree Centrality	Closeness Centrality
$G_{m1ANDm2}$	59%	**43.3%**
$G_{m1ANDm3}$	67.9%	55.4%
$G_{m2ANDm3}$	54.4%	**48.1%**
$G_{m1ANDm2ANDm3}$	**14.1%**	**13.5%**
Overall	**48.9%**	**40.1%**

TABLE III
LOW ACCURACIES OF THE NAIVE APPROACH TO ESTIMATE AND-COMPOSITION HUB SETS (IMDB MULTIPLEX)

AND-Composed Layers	Degree Centrality	Closeness Centrality
$G_{a1ANDa2}$	84.8%	93%
$G_{a1ANDa3}$	82.6%	82.1%
$G_{a2ANDa3}$	85.4%	93.3%
$G_{a1ANDa2ANDa3}$	79.2%	87.4%
Overall	83%	88.9%

TABLE IV
SIMILAR TOPOLOGY ACROSS LAYERS LEADING TO GOOD ACCURACIES OF THE NAIVE APPROACH TO ESTIMATE AND-COMPOSITION HUB SETS (ACCIDENT MULTIPLEX)

Estimating Degree Centrality based Hubs: Here we empirically evaluate the performance of the three degree-based hub estimation heuristics.

Performance of Heuristic DC1: In DC1, the average degree estimate for an AND-composed layer is obtained by taking the minimum of the two layer-wise average degrees. This heuristic generates only those common layer-wise hubs that share more than this estimated number of neighbors across layers, thus striking out the possibility of any false positive's presence from the estimated hub sets. Table V and VI show that the overall accuracy of the estimated hub sets is **79.5%** and **82.8%** for the accident and IMDb multiplexes, respectively. Moreover, there is an overall saving of **70.8%** and **41.9%** in computation time for generating the hub sets of accident and IMDb multiplexes, respectively.

AND-Composed Layer	Accuracy	Hub Set Generation Time (secs)	
		Actual	Estimated by DC1
$G_{a1ANDa2}$	78.6%	0.0523	0.0166
$G_{a1ANDa3}$	77.5%	0.0423	0.0152
$G_{a2ANDa3}$	85.7%	0.0711	0.0152
$G_{a1ANDa2ANDa3}$	76.4%	0.0458	0.0147
Overall	**79.5%**	0.2115	0.0618 (**70.8%**↓)

TABLE V
EFFECTIVE PERFORMANCE OF DC1: HIGH ACCURACIES AND LOWER HUB SET GENERATION TIMES (ACCIDENT MULTIPLEX)

AND-Composed Layer	Accuracy	Hub Set Generation Time (secs)	
		Actual	Estimated by DC1
$G_{m1ANDm2}$	88.2%	0.0597	0.0302
$G_{m1ANDm3}$	74.6%	0.0681	0.0483
$G_{m2ANDm3}$	82.4%	0.0634	0.0385
$G_{m1ANDm2ANDm3}$	85.9%	0.0492	0.0226
Overall	**82.8%**	0.2403	0.1396 (**41.9%**↓)

TABLE VI
EFFECTIVE PERFORMANCE OF DC1: HIGH ACCURACIES AND LOWER HUB SET GENERATION TIMES (IMDB MULTIPLEX)

Note that for IMDB the overall accuracy improved from **48.9%** in the naive scheme to **82.8%**. However, the accuracy for the Accident multiplex decreased. This is because the estimated average degree was far larger than the actual average degree of the AND-composed networks. To solve this issue we apply heuristic DC2.

Performance of Heuristic DC2: Table VII shows that the improved average degree estimate for the AND-composed layers can also improve the accuracy. Using heuristic DC2, increases the overall accuracy from **79.5% to 83.04%** for the Accident Multiplex. Similarly, the accuracy of estimated hub set for IMDb Multiplex increases from from **82.8% to 83.9%**. The proximity of this estimate to the actual average degree allows the generation of some common layer-wise hubs that were excluded by DC1, however the computational costs increase. Therefore, for instance, in case of the Accident multiplex hub set estimation process the overall savings in computational time falls from 70.8% to 58.4%.

Performance of Heuristic DC3: To consider the case where non-hub layer-wise nodes become hubs in the AND-composed layer, few low degree nodes from each layer are maintained such that their degree is at least $(1-\epsilon)$ times the individual layer's average degree, where $0 \leq \epsilon \leq 1$. Figure 4 (a) and (c)

AND-Composed Layer (Actual Average Degree)	Average Degree		% Change in Accuracy
	$DC1_{est}$	$DC2_{est}$	
$G_{a1ANDa2}$ (11.2)	14.92	12.988	5.2%↑
$G_{a1ANDa3}$ (10.18)	14.92	12.847	4.4%↑
$G_{a2ANDa3}$ (14.35)	16.44	15.257	1.6%↑
$G_{a1ANDa2ANDa3}$ (9.28)	14.92	12.045	2.7%↑
Overall	–	–	**3.5%**↑

TABLE VII
IMPROVED ACCURACIES OF DC2 OVER DC1 (ACCIDENT MULTIPLEX)

show that by increasing the value of ϵ the overall accuracy increases as the number of false negatives are reduced. However, higher the value of ϵ, more is the number of layer-wise non-hubs carried forward to the estimation process. Therefore, this increased overhead cost increases the time to estimate hub sets (Figure 4 (b) and (d)).

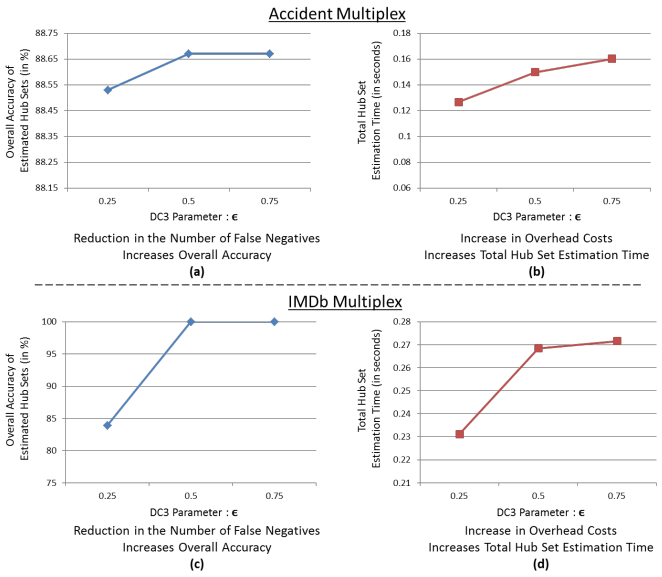

Fig. 4. Performance of DC3 with respect to the parameter ϵ

Figure 4 (c) shows that the average degree estimate for the IMDb multiplex is good enough to give a perfectly accurate estimate for an $\epsilon = 0.5$. However, the average degree estimate becomes a bottleneck in the case of Accident multiplex due to which even with increasing ϵ, the rate of increase in the overall accuracy is low (Figure 4 (a)). A better average degree estimate in these cases will prove to be helpful.

The overall accuracy and total hub set estimation times shown in each cell for the three proposed heuristics in the Summary Table VIII justify that there is an evident trade-off between accuracy and savings in the computational costs.

Estimating Closeness Centrality based Hubs using *Heuristic CC1*: In every layer, high degree neighbors for each high closeness centrality node are maintained. The intuition is that if a common high closeness centrality node shares

DC1 Accuracy Time (secs)	DC2 Accuracy Time (secs)	DC3 $\epsilon = 0.25$ Accuracy Time (secs)	DC3 $\epsilon = 0.5$ Accuracy Time (secs)	DC3 $\epsilon = 0.75$ Accuracy Time (secs)
Accident Multiplex				
79.5% 0.0618	83.04% 0.088	88.5% 0.1268	88.7% 0.1499	88.7% 0.1602
IMDb Multiplex				
82.8% 0.1396	83.9% 0.211	83.9% 0.2312	100% 0.2685	100% 0.2716

TABLE VIII
SUMMARIZING THE PERFORMANCES OF THE THREE DEGREE BASED HUB ESTIMATION HEURISTICS

high degree neighbors across layers that are also part of the hub set estimated by heuristic DC2, then its chances of being accessible via less number of hops from every other node in AND-composed layer increase. Table IX and X show that for both accident and IMDb multiplexes, this heuristic estimates hub sets that have an overall accuracy of **73.8%** and **66.5%**, respectively. Moreover, this process leads to a saving of at least **30%** in computation time.

AND-Composed Layer	Accuracy	Hub Set Generation Time (secs)	
		Actual	Estimated by CC1
$G_{a1ANDa2}$	73.1%	0.3086	0.2028
$G_{a1ANDa3}$	68.9%	0.2834	0.2004
$G_{a2ANDa3}$	78.2%	0.345	0.2017
$G_{a1ANDa2ANDa3}$	75.1%	0.237	0.2051
Overall	**73.8%**	1.174	0.81 (**31%**↓)

TABLE IX
EFFECTIVE PERFORMANCE OF CC1: HIGH ACCURACIES AND LOWER HUB SET GENERATION TIMES (ACCIDENT MULTIPLEX)

AND-Composed Layer	Accuracy	Hub Set Generation Time (secs)	
		Actual	Estimated by CC1
$G_{m1ANDm2}$	60.4%	2.0534	1.5153
$G_{m1ANDm3}$	71.3%	2.6168	1.5255
$G_{m2ANDm3}$	70.1%	2.0432	1.5159
$G_{m1ANDm2ANDm3}$	64.1%	2.029	1.5071
Overall	**66.5%**	8.7424	6.0637 (**30.64%**↓)

TABLE X
EFFECTIVE PERFORMANCE OF CC1: HIGH ACCURACIES AND LOWER HUB SET GENERATION TIMES (IMDB MULTIPLEX)

The similar topology among the Accident Multiplex layers means that most of the shortest paths among the node pairs across layers are common leading to the naive approach giving a higher accuracy as compared the proposed heuristic that excludes some common layer-wise hubs as it only considers shared one hop high degree neighbors. Even though this heuristic gives good accuracies for the estimated hub sets, but it can be improved by maintaining the path information to high degree nodes beyond 2-3 hops from the high closeness centrality hubs in each layer. However, as stated earlier, maintaining such longer path information will significantly increase the computational costs.

VII. CONCLUSION AND FUTURE WORK

In this paper, various heuristics have been presented and validated to efficiently estimate hubs in any conjunctively composed layer of a multiplex. Using real-life datasets from diverse backgrounds, we have empirically shown that by maintaining minimal neighborhood information along with the layer-wise hubs, it is possible to estimate good quality degree or closeness centrality based hub sets of any AND-composed layer with an overall accuracy exceeding 80% or 70%, respectively, while reducing the computation time by at least 30%. Further, such techniques eliminate the need to generate and store any composed layers, thus saving storage space too.

We plan to extend hub estimation to other centrality measures like betweenness and eigenvector, and handle weighted and/or directed edges. In addition to conjunction, we plan on extending this composition to disjunction and negation.

REFERENCES

[1] The internet movie database. ftp://ftp.fu-berlin.de/pub/misc/movies/database/

[2] Road safety - accidents 2014. https://data.gov.uk/dataset/road-accidents-safety-data/resource/1ae84544-6b06-425d-ad62-c85716a80022

[3] Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gmez-Gardees, J., Romance, M., Sendia-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks. Physics Reports 544(1), 1 – 122 (2014), http://www.sciencedirect.com/science/article/pii/S0370157314002105, the structure and dynamics of multilayer networks

[4] Boden, B., Gnnemann, S., Hoffmann, H., Seidl, T.: Mining coherent subgraphs in multi-layer graphs with edge labels. In: Proc. of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2012), Beijing, China. pp. 1258–1266 (2012)

[5] De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M.A., Gómez, S., Arenas, A.: Mathematical formulation of multilayer networks. Physical Review X 3(4), 041022 (2013)

[6] De Domenico, M., Solé-Ribalta, A., Omodei, E., Gómez, S., Arenas, A.: Centrality in interconnected multilayer networks. arXiv preprint arXiv:1311.2906 (2013)

[7] Dekker, A.: Conceptual distance in social network analysis. Journal of Social Structure (JOSS) 6 (2005)

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

[9] Dong, X., Frossard, P., Vandergheynst, P., Nefedov, N.: Clustering with multi-layer graphs: A spectral perspective. CoRR abs/1106.2233 (2011), http://dblp.uni-trier.de/db/journals/corr/corr1106.html#abs-1106-2233

[10] Freeman, L.C.: Centrality in social networks conceptual clarification. Social networks 1(3), 215–239 (1978)

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

[12] Magnani, M., Rossi, L.: Formation of multiple networks. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. pp. 257–264. Springer (2013)

[13] Ng, M.K.P., Li, X., Ye, Y.: Multirank: co-ranking for objects and relations in multi-relational data. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1217–1225. ACM (2011)

[14] Rochat, Y.: Closeness centrality extended to unconnected graphs: The harmonic centrality index. In: ASNA. No. EPFL-CONF-200525 (2009)

[15] Santra, A., Bhowmick, S., Chakravarthy, S.: Efficient community re-creation in multilayer networks using boolean operations. In: International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. pp. 58–67 (2017), https://doi.org/10.1016/j.procs.2017.05.246

[16] Solé-Ribalta, A., De Domenico, M., Gómez, S., Arenas, A.: Centrality rankings in multiplex networks. In: Proceedings of the 2014 ACM conference on Web science. pp. 149–155. ACM (2014)