HUBify: Efficient Estimation of Central Entities across Multiplex Layer Compositions

Abhishek Santra[1], Sanjukta Bhowmick[2] and Sharma Chakravarthy[3]

[1,3]Information Technology Laboratory, CSE Department,
University of Texas at Arlington, Arlington, Texas, USA
[2]CSE Department, University of Nebraska at Omaha, Omaha, Nebraska, USA
Email: [1]abhishek.santra@mavs.uta.edu, [2]sbhowmick@unomaha.edu, [3]sharma@cse.uta.edu

Talk Outline

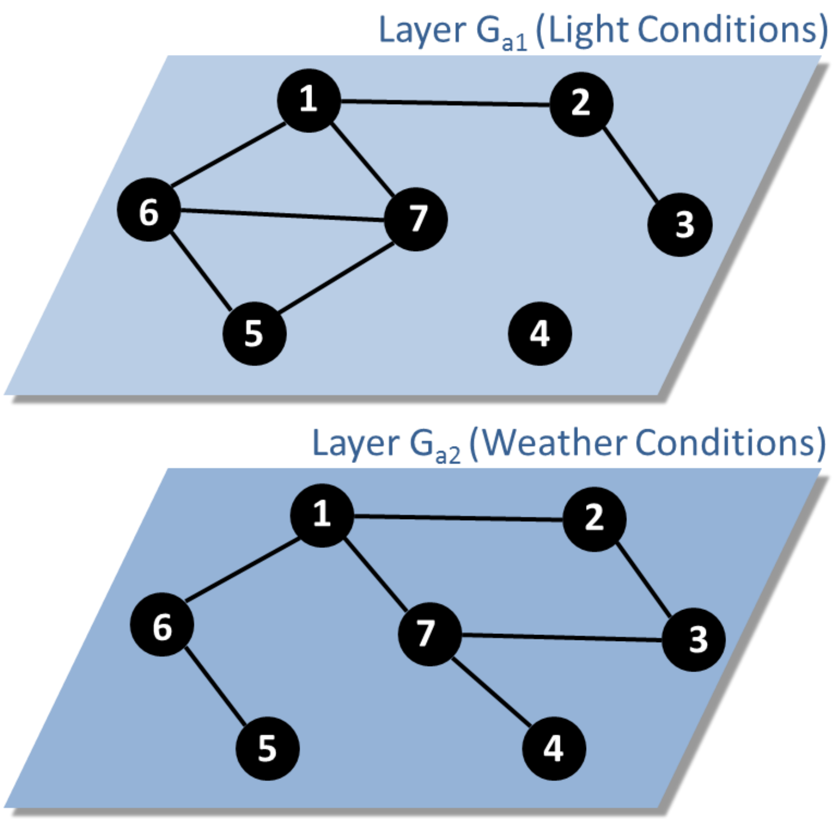
- Motivation
- Problem Formulation
- Proposed Heuristics
- Experimental Analysis
- Conclusions and Future Work

Motivation

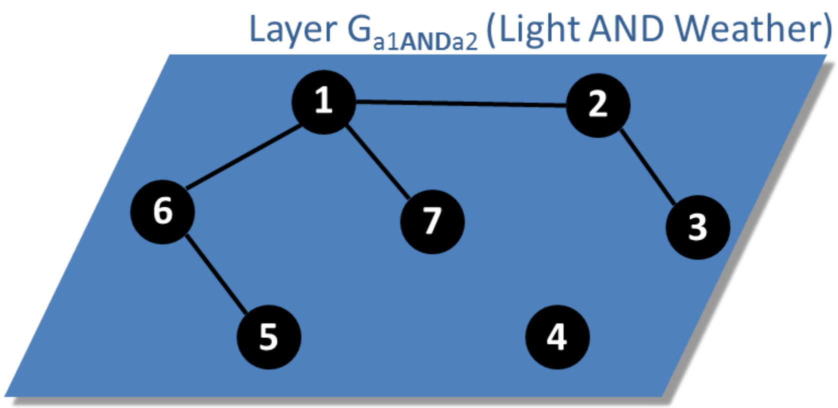
Multiplex-based Modeling

- Existence of **diverse relationships** among **a set of entities**
  - **Interactions among people** varies with the mode of communication like *email, social networking and phone calls*
  - **Similarity among traffic accidents** varies based on factors like *light, weather and road surface conditions*
  - **Co-actor relationship** can vary based on the movie genre like *action, comedy and drama*
- *Same set of entities* are *inter-connected* in *multiple ways* based on features
  - Use **network of graphs (or *layers*) - Multiplexes**

Multiplex-based Modeling
(Same Nodes Linked Differently into Layers)



Accident Multiplex

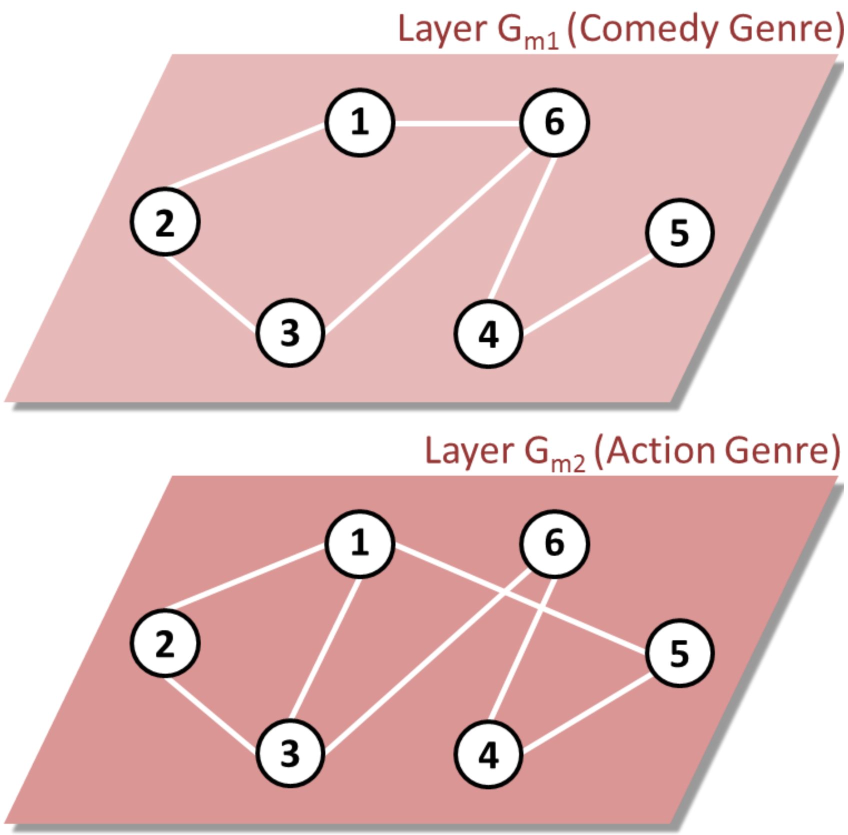


Combined layer with entity-entity relationships present in *all constituent individual layers*

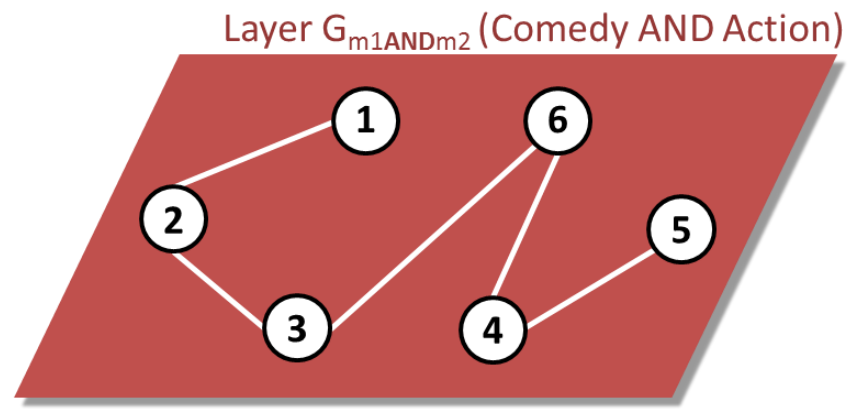
(AND-Composed Layer)

Multiplex-based Modeling

(Same Nodes Linked Differently into Layers)



IMDb Multiplex



Combined layer with entity-entity relationships present in *all constituent individual layers*

(AND-Composed Layer)

Holistic view of Multiplex System

- Insightful Analytical Questions
  - Most **influential people** across different communication platforms (Advertisement Agencies)
  - Most **dominating accident locations** w.r.t poor lighting conditions and bad roads (Accident Prevention Measures)
  - Highly **popular/preferred co-actors** for various genre combinations (Casting and Production Houses)
- **Solution:** Generate **highly central vertices (i.e. hubs)** in the required individual or AND-composed multiplex layer

Challenges in Computing Hubs

- Existing algorithms detect hubs for single-layer networks (or monoplexes)
- For a holistic view, need to **generate, store and re-compute hubs for each of the 2^N layer combinations**
  - **Computationally Expensive**
- **No principled approach exists for *flexibly analyzing any layer combination*, without having to construct that combined layer**

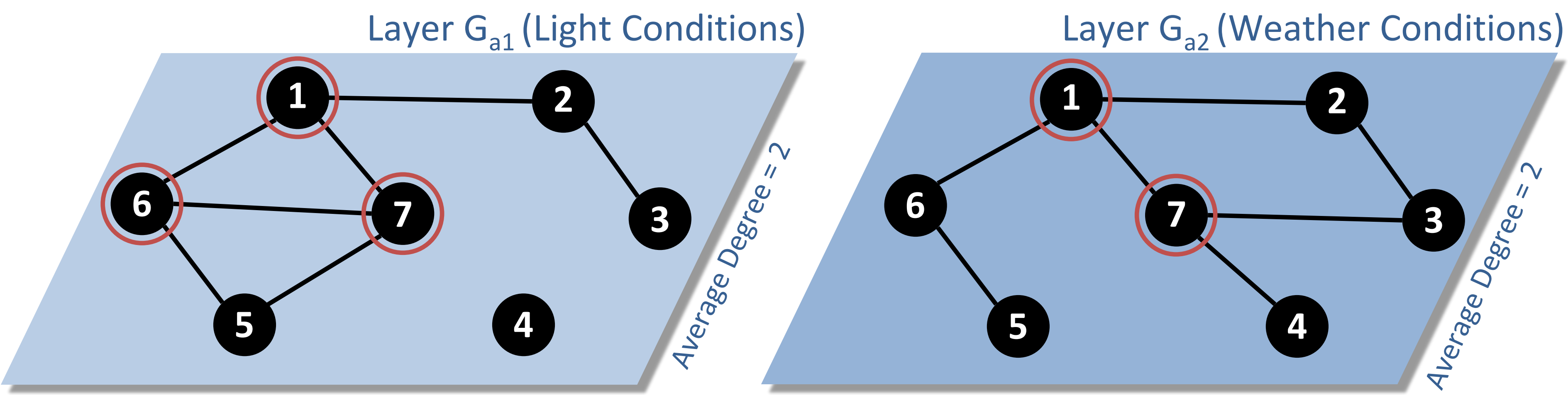
Problem Formulation

Problem Statement

"Identify the hub sets in any AND-composed layer by using information about the hubs from the participating individual layers"

Quantifying Hubs (Degree Centrality Hubs)

- Higher the degree, greater the influence on immediate neighborhood
- Degree Centrality Hub: A node having the degree above the average degree

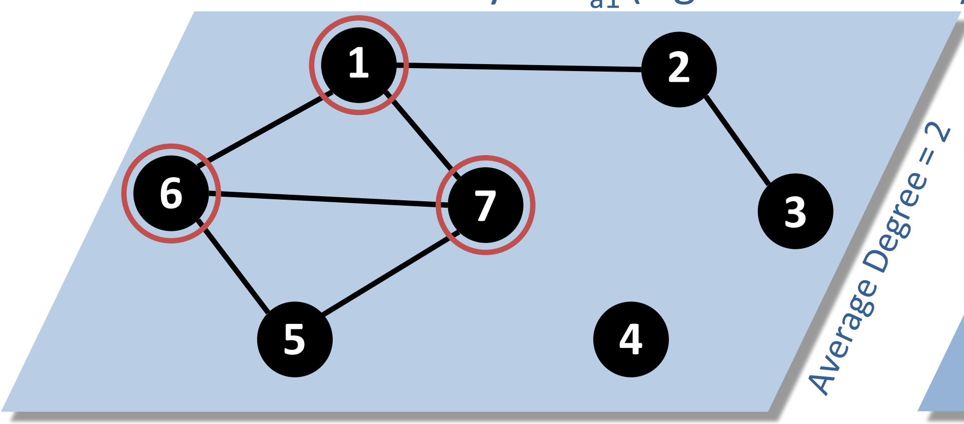


Proposed Solutions

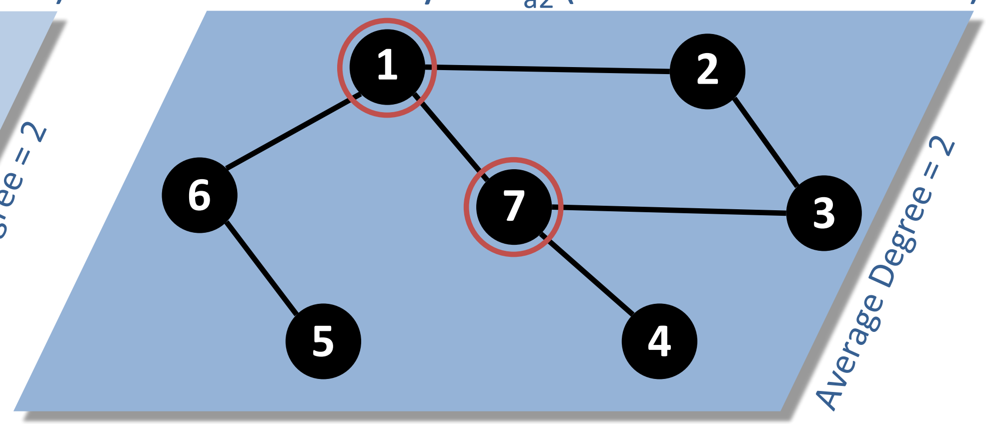
The Naïve Approach

Intersect the Layer-wise Hub Sets

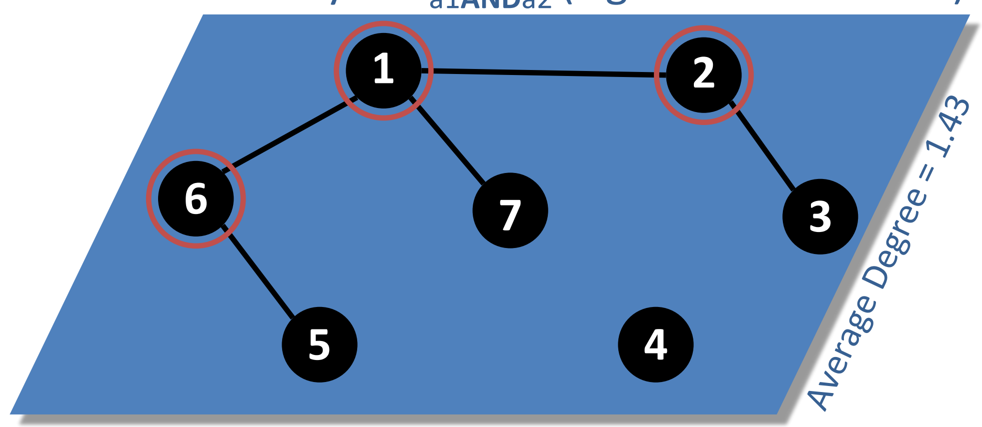
Layer G_{a1} (Light Conditions)



Layer G_{a2} (Weather Conditions)



Layer $G_{a1ANDa2}$ (Light AND Weather)



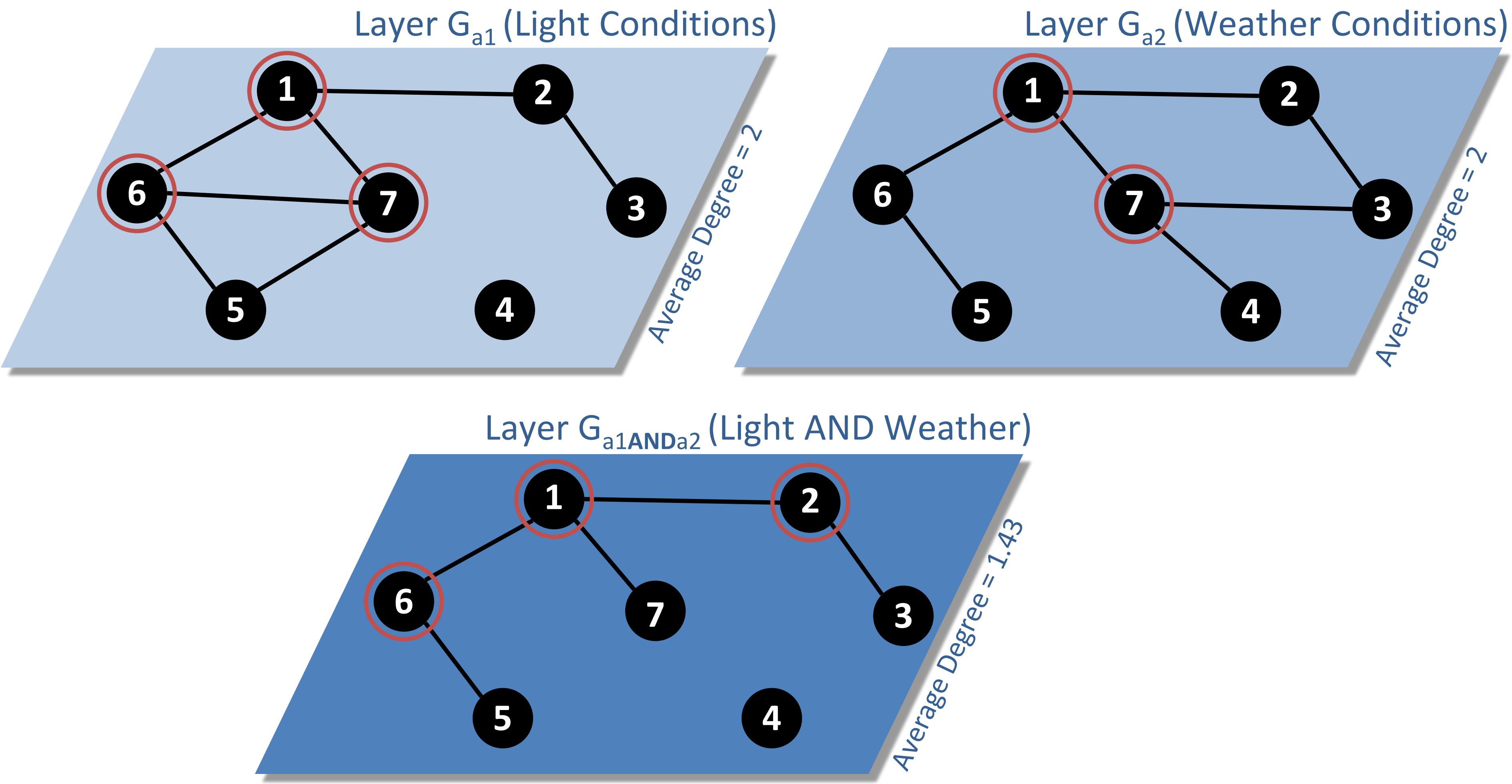
Estimated Hub Set 1 7

False Positive 7

False Negatives 2 6

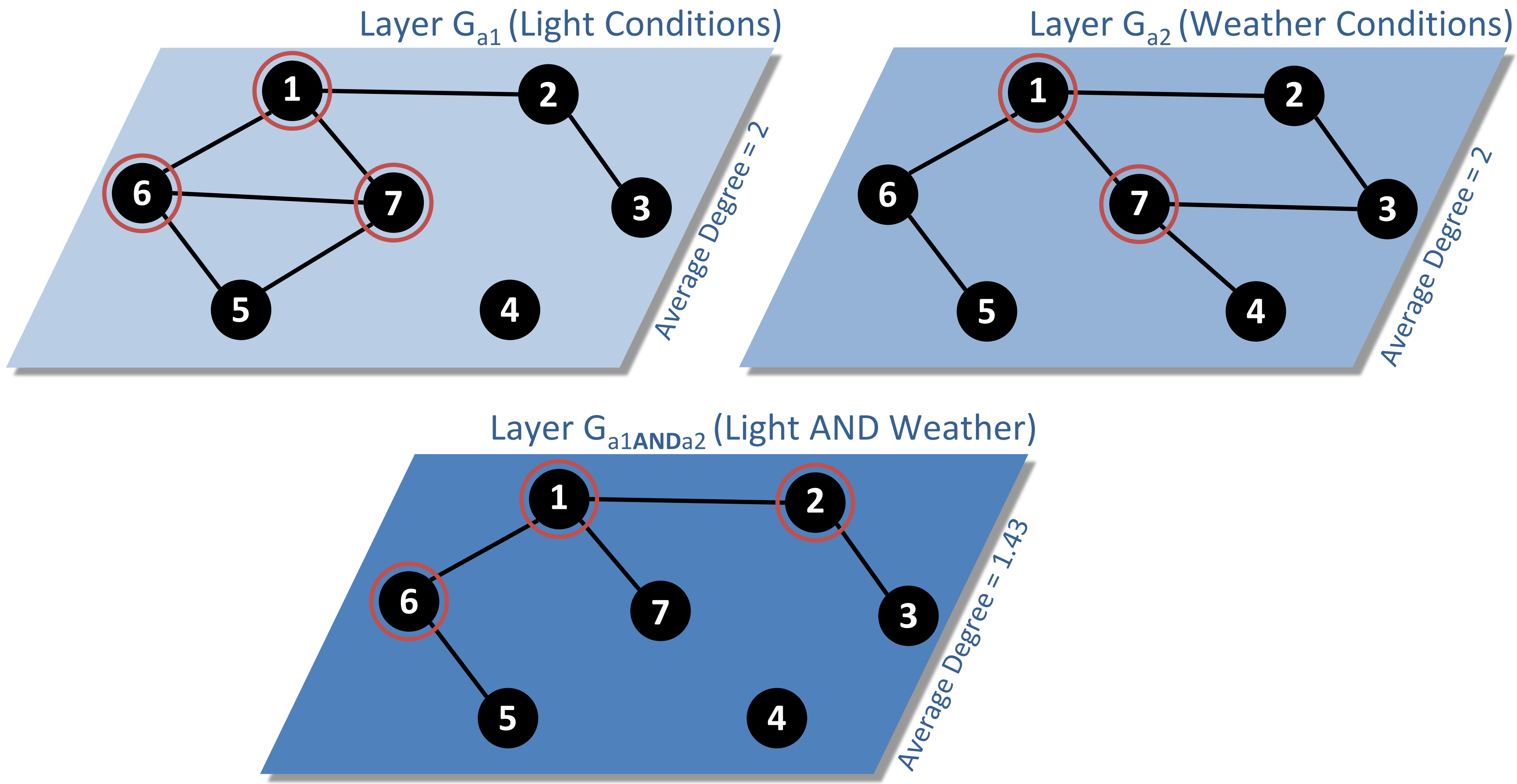
Non-Triviality of the Task (Case 1)

Hubs in individual layers may not be hubs in the AND-composed layer



Non-Triviality of the Task (Case 2)

Non-hubs in individual layers may be **hubs** in the AND-composed layer



Estimating Hub Set for 2-layer AND-Composition

- **Aim:** To reduce naïve approach false positives and negatives
- **Efficient heuristics** are proposed that use the **layer-wise hubs** and their **distance-1 neighbors**, for estimating the degree centrality hub sets

Layer G_{a1} Hub Set

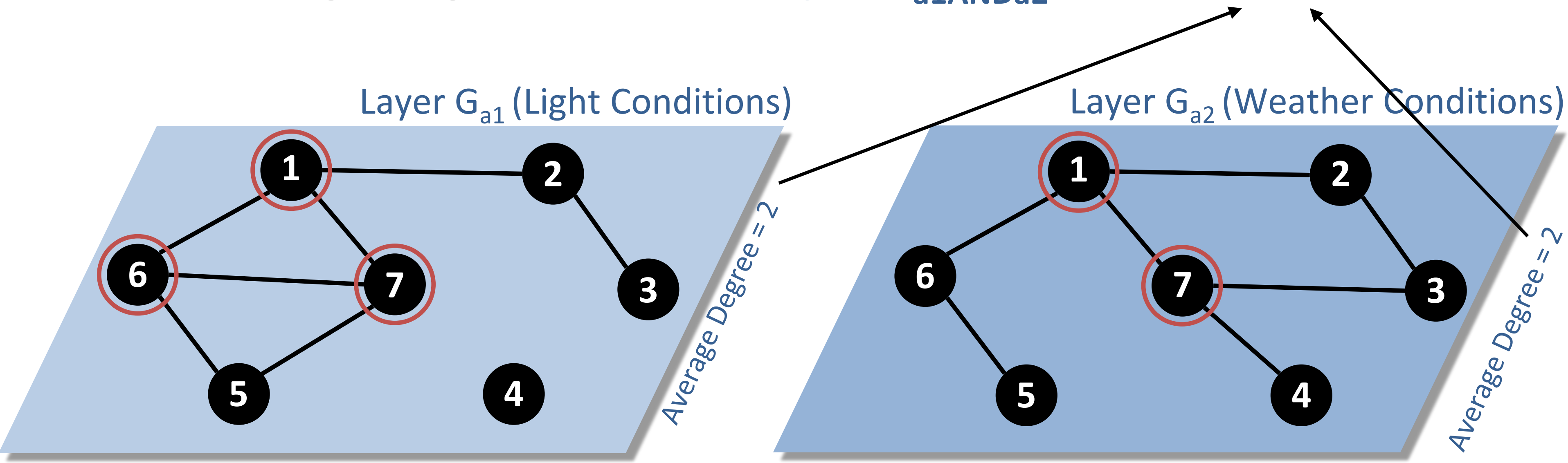
V_i	Neighborhood(V_i)
1	2 6 7
6	1 5 7
7	1 5 6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)
1	2 6 7
7	1 3 4

Degree Centrality Heuristic DC1

Average Degree Estimate Layer $G_{a1ANDa2}$ = MIN (2, 2)



Degree Centrality Heuristic DC1

Average Degree Estimate Layer $G_{a1ANDa2}$ = 2

Layer G_{a1} Hub Set

V_i	Neighborhood(V_i)		
→ 1	2	6	7
6	1	5	7
7	1	5	6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)		
→ 1	2	6	7
7	1	3	4

2 6 7 ∩ 2 6 7

$|\,2\ 6\ 7\,| = 3$

3 > 2 √

Estimated Hub Set 1

Degree Centrality Heuristic DC1

Average Degree Estimate Layer $G_{a1ANDa2}$ = 2

Layer G_{a1} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
6	1	5	7
→ 7	1	5	6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
→ 7	1	3	4

1 5 6

$\cap$

4 3 1

$|1| = 1$

1 > 2 X

Estimated Hub Set 1

No False Positives. May **increase false negatives**

Degree Centrality Heuristic DC2

- Improved average degree estimate using vertex degrees. Closer to the actual average degree

$$\frac{\text{minDeg } 1 + \text{minDeg } 2 + \text{minDeg } 3 + \text{minDeg } 4 + \text{minDeg } 5 + \text{minDeg } 6 + \text{minDeg } 7}{7}$$

Improved Average Degree Estimate Layer $G_{a1ANDa2}$ = 1.71

Actual Average Degree Layer $G_{a1ANDa2}$ = 1.43

- **No False Positives**
- **False Negatives (DC2) ≤ False Negatives (DC1)**
- **More Overhead Information**

Degree Centrality Heuristic DC3 (ϵ)

Considers Selected Layer-wise Non-Hubs based on parameter ϵ

Improved Average Degree Estimate Layer $G_{a1ANDa2}$ = 1.71

Layer G_{a1} Hub Set

V_i	Neighborhood(V_i)		
→ 1	2	6	7
2	1	3	
6	1	5	7
7	1	5	6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)		
→ 1	2	6	7
2	1	3	
3	2	7	
6	1	5	
7	1	3	4

2 6 7 ∩ 2 6 7

| 2 6 7 | = 3

3 > 1.71 √

Estimated Hub Set 1

Degree Centrality Heuristic DC3 (ϵ)

Selected Layer-wise Non-Hubs based on parameter ϵ included

Improved Average Degree Estimate Layer $G_{a1ANDa2}$ = 1.71

Layer G_{a1} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
2	1	3	
6	1	5	7
7	1	5	6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
2	1	3	
3	2	7	
6	1	5	
7	1	3	4

1 3 $\cap$ 1 3

$|\{1, 3\}| = 2$

2 > 1.71 √

Estimated Hub Set 1 2

Degree Centrality Heuristic DC3 (ϵ)

Selected Layer-wise Non-Hubs based on parameter ϵ included

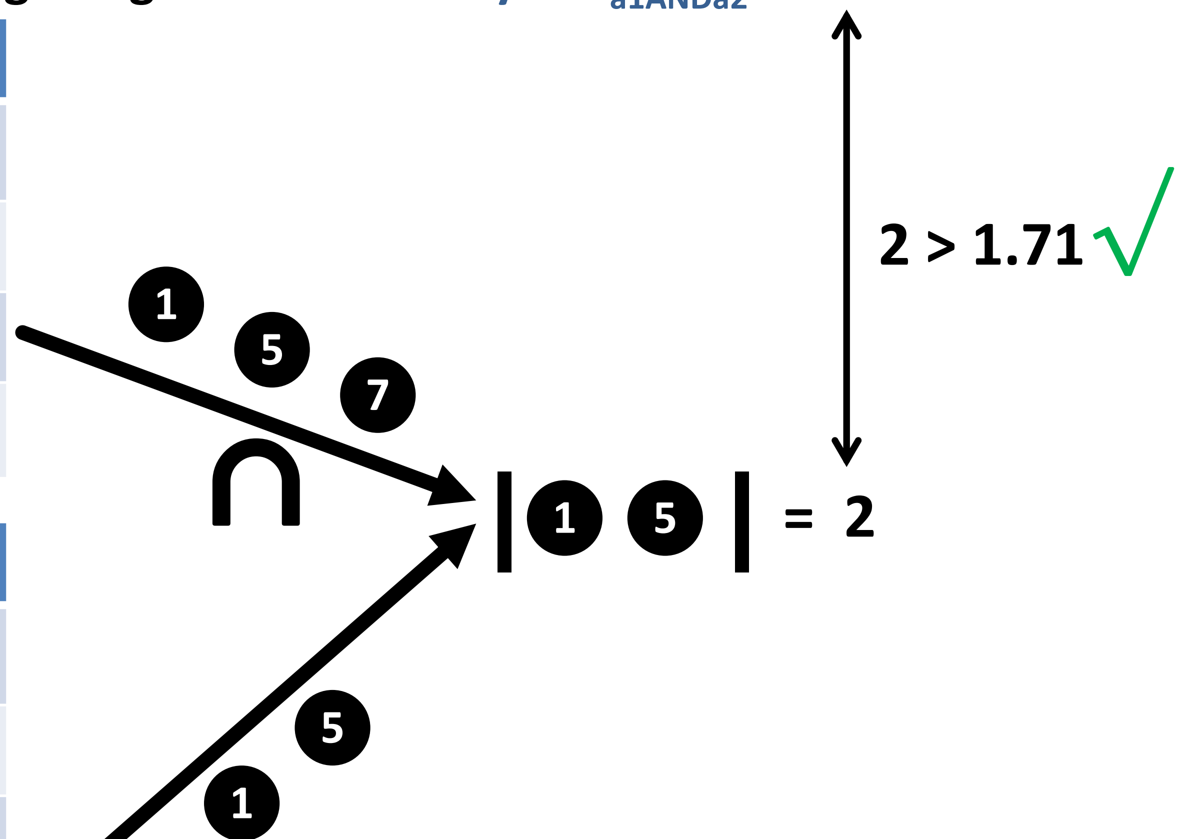
Improved Average Degree Estimate Layer $G_{a1ANDa2}$ = 1.71

Layer G_{a1} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
2	1	3	
6	1	5	7
7	1	5	6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
2	1	3	
3	2	7	
6	1	5	
7	1	3	4



Estimated Hub Set 1 2 6

Degree Centrality Heuristic DC3 (ϵ)

Selected Layer-wise Non-Hubs based on parameter ϵ included

Improved Average Degree Estimate Layer $G_{a1ANDa2}$ = 1.71

Layer G_{a1} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
2	1	3	
6	1	5	7
→ 7	1	5	6

Layer G_{a2} Hub Set

V_i	Neighborhood(V_i)		
1	2	6	7
2	1	3	
3	2	7	
6	1	5	
→ 7	1	3	4

1 5 6 ∩ 1 3 4

$|\ 1\ | = 1$

1 > 1.71 X

Estimated Hub Set 1 2 6

No False Positives. Higher the ϵ, better the estimated hub set, more is the overhead Information.

Advantages of Proposed Heuristics

- Closeness Centrality based heuristic has also been proposed *(details in the paper)*
- Heuristics are **commutative and associative**
- **Flexible Composition**
  - **Any k-layer AND-composition hub set** can be estimated by using the **2-layer heuristic as a subroutine**, **in parallel**
- **Reduce computational complexity**
  - **Eliminate the need to generate, store and re-compute** degrees or shortest paths for $2^N - N$ layer compositions (N: number of layers)
  - **$2^N - N$ AND-composition hub sets** estimated by only **using N layer-wise hub sets and minimal neighborhood information**

Experimental Analysis

Experimental Setup

- **Datasets**
  - *Accident Multiplex:* 1000 random UK traffic accidents from 2014, 3 conditions-based layers (Light, Weather, Road Surface Conditions)
  - *IMDb Multiplex:* 5000 random actors, 3 genre-based layers (Comedy, Action, Drama)
- **Environment:** UBUNTU 13.10, 4GB RAM, C++ codes
- **Comparison Metrics**
  - **Accuracy**: **Jaccard Index** used to compare estimated (X) and actual (Y) hub sets, $J(X,Y) = |X \cap Y|/|X \cup Y|$
  - **Generation Time**
    - **Actual Hub Set**: Time to generate the AND-composed layer + Time to compute the hub set
    - **Estimated Hub Set**: Time to apply the proposed heuristic

Inefficiency of the Naïve Approach

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%**

Low accuracies due to **presence of False Positives and Negatives** (IMDb Multiplex)

Performance of Heuristic DC1

High Accuracies due to absence of false positives, **Low hub generation times**

IMDb 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%↓)

Accident Multiplex

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%↓)

Performance of Heuristic DC2

Better average degree estimate lead to **improved accuracies** as compared to DC1, **but at increased overhead costs**

Accident Multiplex

AND-Composed Layer (Actual Average Degree)	Average Degree $DC1_{est}$	$DC2_{est}$	% Change in Accuracy
$G_{a1AND a2}$ (11.2)	14.92	12.988	5.2%↑
$G_{a1AND a3}$ (10.18)	14.92	12.847	4.4%↑
$G_{a2AND a3}$ (14.35)	16.44	15.257	1.6%↑
$G_{a1AND a2AND a3}$ (9.28)	14.92	12.045	2.7%↑
Overall	–	–	3.5%↑

Improved Accuracy: Accident – 79.5%(DC1), 83.04%(DC2)
IMDb – 82.8%(DC1), 83.9% (DC2)

Fall in Overall Computation Time Savings: Accident – 70.8%(DC1), 58.4%(DC2)
IMDb – 41.9%(DC1), 12.2% (DC2)

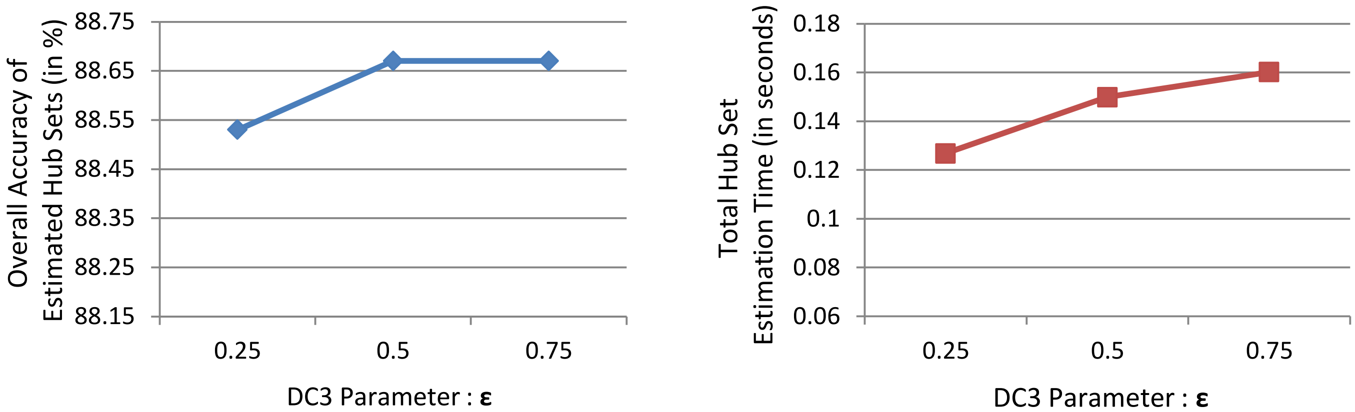
Performance of Heuristic DC3 with Parameter ϵ

Increasing ϵ:

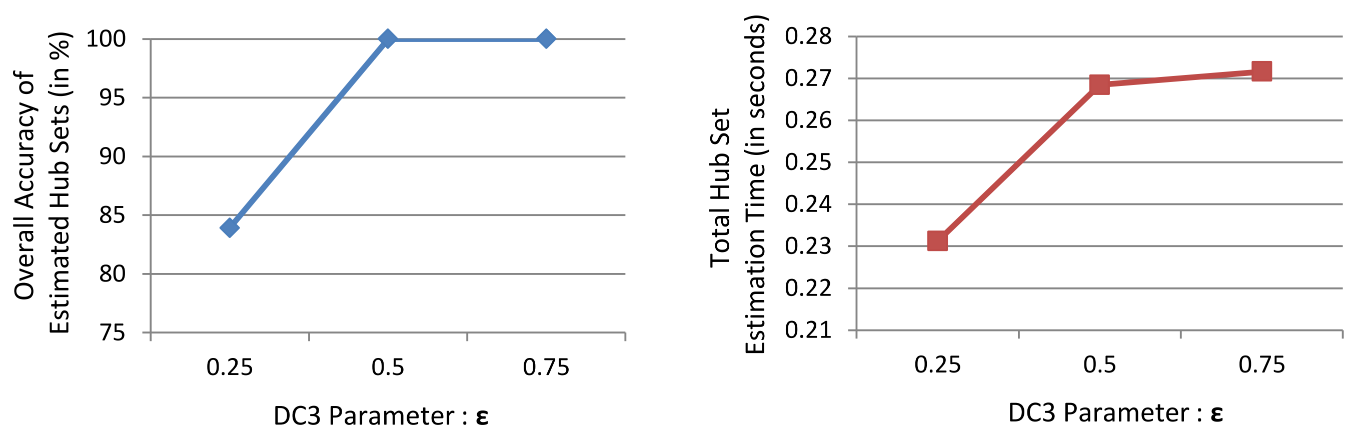
Overall accuracy increases as the number of false negatives are reduced.

Increases Hub Estimation Times as more layer-wise non-hubs are carried forward

Accident Multiplex

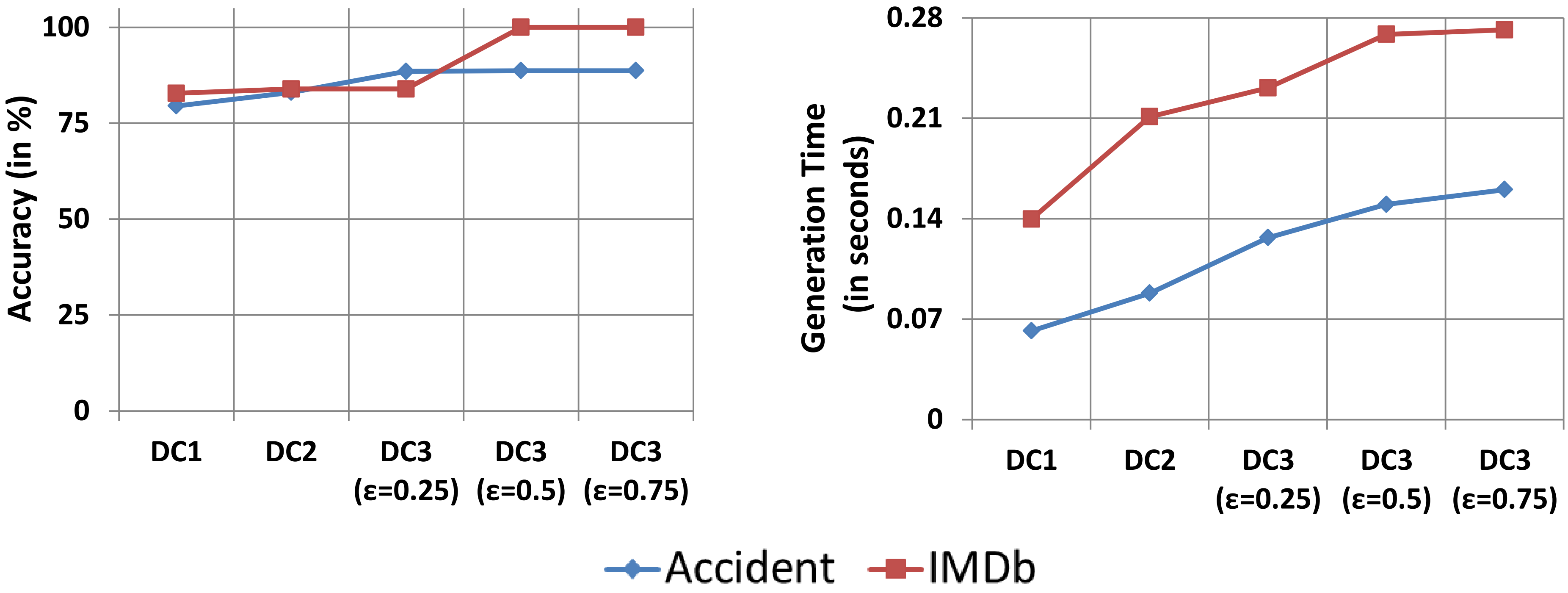


IMDb Multiplex



Comparison between DC1, DC2 and DC3

Trade-off between Accuracy and Savings in Computational Costs



Performance of Heuristic CC1

Good Accuracies due to absence of false positives, **Low hub generation times**

IMDb 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%↓)**

Accident Multiplex

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%↓)**

Conclusions and Future Work

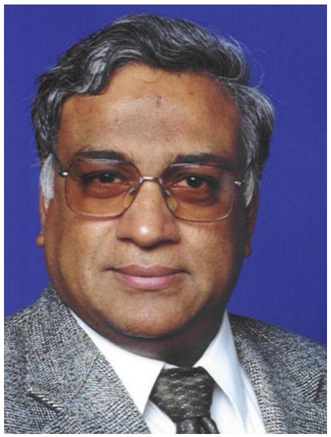
Conclusions

- Efficient heuristics proposed that only **use layer-wise degree or closeness centrality results** to identify hubs in conjunctively composed multiplex layers
- Real-world, multi-feature datasets (IMDb and Accident) empirically 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%**
- **Eliminate the need to separately generate, store and analyze** any AND-composed layer

Future Work

- **Generalize the heuristics** based on layer characteristics and obtain a confidence interval on the accuracy
- Extend to other centrality measures – **eigenvector, betweenness**
- Handle layers with **weighted and directed edges**
- Extend to other composition schemes like **disjunction and negation**

Questions?



Sharma Chakravarthy
Professor
sharma@cse.uta.edu

Abhishek Santra
PhD Student
abhishek.santra@mavs.uta.edu

Sanjukta Bhowmick
Associate Professor
sbhowmick@unomaha.edu

For more information visit:

http://itlab.uta.edu