Big Data Analysis using Multilayer Networks

Sharma Chakravarthy and Abhishek Santra
Information Technology Laboratory (IT Lab)
Department of Computer Science
University of Texas at Arlington

(CODS-COMAD 2021 Tutorial)
Roadmap

➢ Background
➢ Motivation: The Overall Picture
➢ Community Detection in MLNs
  ▪ Boolean Composition Approaches (HoMLN)
  ▪ Weighted Bipartite Matching (HeMLN)
➢ Hub Detection in HoMLNs
  ▪ Degree and Closeness Centrality Hub Detection Heuristics
➢ Substructure Discovery in MLNs
➢ Case Studies on Real World Datasets
  ▪ Facebook, US Airlines, IMDb, DBLP, ...
➢ Live Dashboard (Demo): https://itlab.uta.edu/cowiz/
  ▪ Analysis and Visualization of COVID-19 using MLNs
  ▪ Youtube: https://www.youtube.com/watch?v=4vJ56FYBSCg
Background
What is Big Data Analysis/Science
Where are we headed?
Transforming Disparate Data into Actionable Knowledge and Decisions
Where are we headed?

➢ Without understanding the past, it is very difficult to appreciate the present and plan for the future!

➢ Technology provides solutions; it does not mean it solves problems!
Motivation

Complex Data Analysis

Traditional Approaches and Their Limitations
Modeling Using Multilayer Networks
Limitations of Existing Analysis Approaches
Decoupling Approach to Analyze MLNs
Big Data Analytics

Influx of data pertaining to the 4Vs, i.e. **Volume, Velocity, Variety** and **Veracity**

Which *class of big data problems* are we looking into?
Problem 1: Analyzing Large Multi Entity, Feature, and Relationship Data Sets

Multiple relationships among same type of entities

Interactions among same set of people

Airline Connectivity among same US cities

Most popular or socially active group of people across platforms?

Most influential set of people?

Homogeneous Multilayer Networks (HoMLN)
Problem 2: Analyzing Large Multi Entity, Feature, and Relationship Data Sets

Multiple relationships among different types of entities

Connectivity among different entities

For the most popular actor groups from each movie rating class, which are the director groups with which they have maximum interaction?

Highly rated actors working in similar genres who have never acted together?

Frequently publishing cohesive co-author groups?

Most active periods for popular collaborators?

Heterogeneous Multilayer Networks (HeMLN)
Mining Vs. Big Complex Data Analytics

Data Set Description

- Clustering, Classification, Association Rule Mining, Anomaly Detection, Regression Subgraph Mining, ...

Final Results

Drill-Down Analysis

Application Requirements

Data Set Description

Analysis Objectives

Multilayer Network Types (HoMLN, HeMLN, HyMLN)

Using EER → MLN Approach

Generate MLN Data Model

Generate Analysis Expressions

Using Efficient Divide-and-Conquer based **Decoupling** Approach

Perform Analysis

Visualization

Drill-Down Analysis

Final Results
Motivation
Complex Data Analysis

Traditional Approaches and Their Limitations
Modeling Using Multilayer Networks
Limitations of Existing Analysis Approaches
Decoupling Approach to Analyze MLNs
Traditional Modeling: Simple Graphs

- Nodes: Entities
- Single Edges (weighted or unweighted): Single or Combination of feature-based relationship

- Algorithms exist for communities, hubs, subgraph mining, frequent subgraph counting, etc.
Traditional Modeling: Attributed Graphs

- Nodes: Entities
  - Node Labels: Entity Types
- Multiple Edges: Feature-based relationship (weighted or unweighted)
  - Edge Labels: Feature Types
- Algorithms exist for subgraph mining

Entities as Colored Nodes

Relationships as Colored Edges
## Limitations of Traditional Modeling Approaches

<table>
<thead>
<tr>
<th></th>
<th>Modeling Clarity</th>
<th>Analysis Flexibility</th>
<th>Computational Efficiency</th>
<th>Drill Down and Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Graph</strong></td>
<td>Not Supported</td>
<td>To some extent</td>
<td>Bad</td>
<td>Difficult</td>
</tr>
<tr>
<td></td>
<td>(Single entity, feature type only supported)</td>
<td>(Communities, Hubs, Subgraph Mining, Frequent Subgraph Counting)</td>
<td>(New graphs re-created for every feature combination; Combination not straightforward)</td>
<td>(Only Single types supported)</td>
</tr>
<tr>
<td><strong>Attributed Graph</strong></td>
<td>To some extent</td>
<td>Not Available</td>
<td>Bad</td>
<td>Difficult</td>
</tr>
<tr>
<td></td>
<td>(Multiple node and edge labels supported)</td>
<td>(Except Subgraph Mining)</td>
<td>(Multiple Traversals required to fetch required combination)</td>
<td></td>
</tr>
</tbody>
</table>

- **Limitations**
  - **Single Graph**
    - Not Supported
      - (Single entity, feature type only supported)
  - **Attributed Graph**
    - To some extent
      - (Multiple node and edge labels supported)
Previous Work

➢ Community Detection in Simple Graphs

➢ Centrality Metric Evaluation in Simple Graphs
  ▪ Freeman, L.C., Social Networks, 1978
  ▪ Page, L., Brin, S., Motwani, R. and Winograd, T., Stanford InfoLab, 1999
  ▪ Dekker, A., Journal of Social Structure, 2005

➢ Subgraph Mining in Simple Graphs
  ▪ Cook, D. J. and Holder L. B., Journal of Artificial Intelligence Research 1, 1994
  ▪ Yan X. and Han, J., ICDM, 2002
Motivation
Complex Data Analysis
Traditional Approaches and Their Limitations
Modeling Using Multilayer Networks
Limitations of Existing Analysis Approaches
Decoupling Approach to Analyze MLNs
Modeling Clarity using MLNs

Choice of layer nodes, intra-layer edges depending on the Semantics of Analysis Objectives

Interactions among People

facebook  LinkedIn  twitter

Same Entities, Different Relationships

Homogeneous MLN (HoMLN)

a. Most popular or socially active group of people across platforms?

b. Most influential set of people?
Modeling Clarity using MLNs

Choice of layer nodes, intra- and inter-layer edges depending on the Semantics of Analysis Objectives

Interactions among IMDb Entities

a. Actor and Director Groups having maximum interaction for each movie rating?
b. Highly rated actors working in similar genres who have never acted together?

Different Entities, features, and Relationships

Heterogeneous MLN (HeMLN)
Modeling Clarity using MLNs

Combination of the two: Choice of layer nodes, intra- and inter-layer edges depending on the Semantics of Analysis Objectives

Interactions among IMDb Entities

For the co-actor groups who are friends on Facebook, which are the director groups with which they have maximum interaction?

Hybrid MLN (HyMLN)
Modeling Clarity (An Overview)

Layers of Networks

- **Layer**: Simple graph capturing the *semantics* of a (or a subset of) feature for the same entity type through intra-layer edges (HoMLN and HeMLN)
- **Inter-layer Edges**: Explicit connection corresponding to relationships between different entity types (HeMLN only)
  - For HoMLN, Inter-layer edges are implicit

Benefits

- **Increased Clarity/Understanding**: Less convoluted, Types (or semantics) preserved
- **Existing single graph algorithms** can be leveraged
- **Layers can be processed in parallel**
- **Elegant handling of dataset updates**
Motivation

Complex Data Analysis

Traditional Approaches and Their Limitations

Modeling Using Multilayer Networks

Limitations of Existing Analysis Approaches

Decoupling Approach to Analyze MLNs
Limitations of Current MLN Analysis Alternatives

Reduce to a Single Graph (SG)

- Aggregate/Combine the desired MLN layers as a single simple graph.
- Process the combined layer using existing algorithms.

MLNs as a graph

- Process the MLN as a whole for analysis.
- Focus on inter-layer edges for HeMLN.

- Loss of information, semantics
- $N$ layers $\Rightarrow O(2^N)$ combined layers!
- Difficult to parallelize and scale

- Need to develop new algorithms
- Difficult to parallelize and scale (Repeated traversals of MLN may be required)
Previous Work

➢ Single Graph (SG) Approaches

▪ Projection Based
  – Sun, Y. and Han, J., *ACM SIGKDD Explorations Newsletter*, 2013

▪ Type Independent / Aggregation based

➢ MLN as a graph


Motivation

Complex Data Analysis
Traditional Approaches and Their Limitations
Modeling Using Multilayer Networks
Limitations of Existing Analysis Approaches

Decoupling Approach to Analyze MLNs
Decoupling Approach to Analyze MLNs

Divide and Conquer Approach: Analysis function-specific partial (or intermediate) results composed systematically to fulfill objective

ψ
(Analysis Function)
Communities, Hubs, Subgraphs

Θ
(Composition Function)
Boolean Composition (HoMLN), Matching (HeMLN)

Partial Results 1

Partial Results 2

Partial Results 3

Combined Results of Layer 1 and 2

FINAL RESULT
(Combined Results of Layer 1, 2 and 3)

Multilayer Network

Combine Layer 2 Partitions

Layer 1

Layer 2

Layer 3
Decoupling Approach Challenges

➢ Identify $\Theta$ for various types of analysis ($\Psi$)
  ▪ Communities
  ▪ Centralities
  ▪ Frequent Substructures
  ▪ Motifs
  ▪ Graph Query

➢ Establish their correctness with respect to the single graph approach

➢ Establish their properties
  ▪ Commutativity, Associativity, Distributivity

➢ Develop efficient algorithms
Benefits of Decoupling Approach

1. Retain the MLN modeling, clarity it brings, and semantics
2. Leverage single graph algorithms
   Infomap, Louvain, Subdue, ...
3. Structure Preservation
   No loss of information, no distortion, clear result semantics
4. Efficiency
   Analysis of $O(2^N)$ combinations of graphs reduced from exponential to linear cost
5. Flexibility
   Arbitrary subsets of features can be analyzed without creating a new graph
6. Amenable to Optimizations
   Possible to Optimize an analysis specification with multiple layers
7. Parallelization Opportunities
8. Ease of dataset updates
   Entails updating affected layers only and their results
9. Application Independent
MLN Publications from IT Lab

Published
3. [C] Xuan-Son Vu, Abhishek Santra, Sharma Chakravarthy, Lili Jiang, “Generic Multilayer Network Data Analysis with the Fusion of Content and Structure”, CiCLing 2019
4. [C] Sharma Chakravarthy, Abhishek Santra, Kanthi Komar, “Humble Data Management to Big Data Analytics/Science: A Retrospective Stroll”, BDA 2018

Under Review

In preparation
10. [J] Abhishek Santra, Kanthi Komar, Sanjukta Bhowmick and Sharma Chakravarthy, “Community Definition For Heterogeneous MLNs And Algorithms For Computing It Efficiently”
11. [C] Abhishek Santra, Kanthi Komar, Sanjukta Bhowmick and Sharma Chakravarthy, “Beyond OLAP: Data-Driven Aggregate Analysis of Multilayer Networks”
12. [C] Anish Rai, Abhishek Santra and Sharma Chakravarthy, “MLN-Subdue: Decoupling Approach-Based Substructure Discovery In Multilayer Networks (MLNs)”
Community Detection in HoMLN

Brief Introduction to Community

Boolean AND Composition
Boolean OR Composition
Performance Analysis
Evaluation of General Boolean Expressions (AND, OR, NOT)
Case Studies
Communities in Simple Graphs

- **Definition:** Groups of related nodes that are densely interconnected and have fewer connections with the rest of the network
  - e.g., community of co-actors, co-authors, FB friends
- Disjoint or overlapping
- Computationally difficult task
- Various detection approaches exist
Widely-used Community Detection Algorithms

- **Hierarchical Clustering:** Hierarchically nodes grouped based on a *similarity measure and threshold*
  - **Louvain method (Maximizing Modularity Function)**
    - Measures *density of links* inside communities compared to links between communities
  - **Infomap method (Reducing Map Function)**
    - Measures *per-step average code length* necessary to describe a random walker's movements on a network partition

- **Minimum-cut method**
  - Equi-sized groups (approx.) where *number of inter-group edges is minimized*

- **Betweenness (Girvan–Newman)**
  - Edges with *high betweenness* value are removed

- **Clique-based Methods**
  - *Maximal cliques* bigger than a minimum size
Community Detection in HoMLN

Brief Introduction to Community

Boolean AND Composition

Boolean OR Composition

Performance Analysis

Evaluation of General Boolean Expressions (AND, OR, NOT)

Case Studies
AND Composition using *Single Graph*

- **Single Graph Approach for Communities (C-SG-AND)**
  - Combine the required layers using **AND** operator
  - Apply existing community detection algorithms

- **Specification:** \( C(G_1 \text{ AND } G_2 \text{ AND } \ldots \text{ AND } G_k) \)
  - \( G_i \): Original MLN layer or NOT layer

- Communities used as **Ground Truth** for accuracy calculations

---

Most Socially Active group of people across all platforms?

Communities
\[ C(\text{FB AND LinkedIn}) \]
AND Composition using Decoupling Approach

- **Correctness Criterion:** Generate the same communities obtained by ANDing layers into a single graph (termed C-SG-AND)

- **Specification:** $C(G_1) \text{ AND } C(G_2) \text{ AND } \ldots \text{ AND } C(G_k)$
  - $C(G_i)$: Communities of $G_i$

- **Approaches for 2 layer Composition**
  - **CV-AND:** Node-based intersection of layer wise communities
    - Accuracy is not very high as community topology not considered. Works well in the presence of cliques.
  - **CE-AND:** Detect connected subgraphs after edge-based intersection of layer wise communities
    - Accuracy improves significantly
    - Additional computation required!
AND Composition using *Decoupling Approach*  
(CE-AND Algorithm - *Illustration*)

CV-AND (Naive Approach)  
Node-based intersection of layer-wise communities
Cost Analysis of Decoupling Approaches

Total Decoupling Approach Cost = One Time Cost + Cost of combining partial results

➢ One Time Cost
  ▪ 1-community: Set of layer-wise communities generated once using existing algorithms
  ▪ When in parallel, time bounded by the densest layer

➢ Cost of combining partial results
  ▪ CV-AND: One scan of community nodes, per required layer
  ▪ CE-AND: One scan of community edges, per required layer

Cost(C-SG-AND) = Cost to generate AND layer + Cost of detecting communities in that

➢ Cost to generate AND layer
  ▪ Requires traversal of all constituent layers

➢ Cost of detecting communities
  ▪ Random walks in a hierarchical fashion until the function is optimized (Infomap/Louvain)

MAX (Partial Result Combination Cost) < MIN(1-community Cost)

Cost(CV-AND) < Cost(C-SG-AND), Cost(CE-AND) < Cost(C-SG-AND)
  ▪ Cost benefit amortized over large analysis space ($2^N$)
Community Detection in HoMLN

Brief Introduction to Community
Boolean AND Composition

Boolean OR Composition

Performance Analysis

Evaluation of General Boolean Expressions (AND, OR, NOT)

Case Studies
OR Composition using *Single Graph*

- **Single Graph Approach for Communities (C-SG-OR)**
  - Combine the required layers using **OR operator**
  - Apply **existing community detection algorithms**

- **Specification:** $C(G_1 \lor G_2 \lor \ldots \lor G_k)$
  - $G_i$: Original MLN layer or NOT layer

- Communities used as **Ground Truth** for accuracy calculations
OR Composition using Decoupling Approach

- **Correctness Criterion:** Generate the same communities obtained by ORing layers into a single graph (**termed** C-SG-OR)

- **Specification:** \( C(G_1) \) OR \( C(G_2) \) OR ... OR \( C(G_k) \)
  - \( C(G_i) \): Communities of \( G_i \)

- **Challenge and Intuition**
  - A group of nodes **tightly knit w.r.t a feature** may break into smaller groups or merge with other groups when **edges** (relationships) **w.r.t to another feature** are included.
  - **AND composed communities** will not break, also part of OR composed communities
    - Uses meta graphs (MG) where AND communities are nodes in the **meta graph**.
OR Composition using Decoupling Approach

(CE-OR Algorithm - Illustration)

Construct OR-MG using union of intra-community edges

Expand Communities

OR Composed Communities

AND Composed Communities (CE-AND)

Original Nodes

Weighted Meta Edge (Edge Fraction)

Meta Nodes

C(OR-MG)
Cost Analysis of Decoupling Approaches

Cost (CE-OR) = One Time Cost + Cost of combining partial results

➢ One Time Cost
   ▪ 1-community generated once in parallel

➢ Cost of combining partial results
   ▪ CV-AND/CE-AND is efficient
   ▪ One scan of community edge files for OR-MG
   ▪ Cost of C(OR-MG) < Cost of C(G_i) < Cost of C(OR layer)
     – Size of OR-MG < Size of OR layer; Number of nodes

Cost(C-SG-OR) = Cost to generate OR layer + Cost of detecting communities in that

➢ Cost to generate OR layer
   ▪ Requires traversal of all constituent layers
   ▪ Note that graph size increases!

➢ Cost of detecting communities
   ▪ Random walks in a hierarchical fashion until the function is optimized (Infomap/Louvain)

MAX (Partial Result Combination Cost) < MIN(1-community Cost)

Cost(CE-OR) < Cost(C-SG-OR)
   ▪ Cost benefit amortized over large analysis space (2^N)
Community Detection in HoMLN

Brief Introduction to Community
Boolean AND Composition
Boolean OR Composition
Performance Analysis
Evaluation of General Boolean Expressions (AND, OR, NOT)
Case Studies
Experimental Set Up

➢ Environment
  ▪ OS: UBUNTU 16.04 | RAM: 8GB
  ▪ Codes implemented in C++

➢ Community Detection Algorithm: Infomap

➢ Accuracy Metrics
  ▪ Normalized Mutual Information (NMI): Measures quality w.r.t. participating entity nodes
  ▪ modified-NMI (m-NMI): Measures quality w.r.t. participating entity nodes and network topology.
    - Misclassification of a strongly connected node should have higher effect as compared to a node on the fringe
## Experimental Set Up (Data Sets)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#Vertices</th>
<th>#Edges in L1</th>
<th>#Edges in L2</th>
<th>#Edges in L3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HoMLN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IMDb</strong> (For top 500 actors then repopulated with co-actors)</td>
<td>9,485 (Actors)</td>
<td>45,581 (Co-Acting)</td>
<td>13,945,912 (Similar Genre)</td>
<td>996,527 (Similar Avg Rating)</td>
</tr>
<tr>
<td><strong>DBLP</strong> (Publications from 2000-2018)</td>
<td>17,204 (Authors)</td>
<td>5,831 (VLDB)</td>
<td>17,737 (SIGMOD)</td>
<td>12,986 (ICDM)</td>
</tr>
<tr>
<td><strong>UK Accident</strong></td>
<td>5,000 (Accidents)</td>
<td>193,860 (Light Cond.)</td>
<td>235,175 (Weather Cond.)</td>
<td>216,397 (Road Cond.)</td>
</tr>
<tr>
<td><strong>R-Mat</strong> (Synthetic Scale-free networks following Power Law degree distributions)</td>
<td>32,768</td>
<td>230,445 (Original Network: $a=0.65$, $b=c=d=0.15$)</td>
<td>230,445 (Cross-perturbation to 1% edges)</td>
<td>230,445 (Cross-perturbation to 5% edges)</td>
</tr>
</tbody>
</table>
Trade-off: Higher the accuracy, lower is the efficiency

CV-AND or CE-AND? Cliques (CV-AND for efficiency), In general (CE-AND for accuracy and efficiency)
Accuracy and Efficiency (AND Composition)

Trade-off: Higher the accuracy, lower is the efficiency

CV-AND or CE-AND? Cliques (CV-AND for efficiency), In general (CE-AND for accuracy and efficiency)

Efficiency improves with more analysis (~O(2^N)) for large N
**Component Cost of Decoupling Approaches**

**Worst Case Analysis:** Maximum cost of combining the partial results is significantly less than the minimum cost to detect 1 layer communities.
Accuracy and Efficiency (OR Composition)

**Accuracy**

- **IMDb**
  - Total Execution Time (in seconds)
  - Similarity with C-SG-OR (0,1)
  - NMI: 0.75, m-NMI: 0.75
  - C-SG-OR: 501.38, CE-OR: 115.4

- **R-Mat**
  - Total Execution Time (in seconds)
  - Similarity with C-SG-AND (0,1)
  - NMI: 0.75, m-NMI: 0.58
  - C-SG-OR: 47.98, CE-OR: 6.76

**Efficiency**

- **IMDb**
  - Total Execution Time (in seconds)
  - Similarity with C-SG-AND (0,1)
  - 77% decrease

- **R-Mat**
  - Total Execution Time (in seconds)
  - Similarity with C-SG-AND (0,1)
  - 85.9% decrease

**Component Cost of CE-OR**

- Significant Savings in Worst Case Component Cost Comparison validates efficiency

**Efficiency improves with more analysis (~O(2^N)) for large N**
Community Detection in HoMLN

Brief Introduction to Community
Boolean AND Composition
Boolean OR Composition
Performance Analysis

Evaluation of General Boolean Expressions (AND, OR, NOT)

Case Studies
Applying Decoupling Approach to Boolean Expression

Find Community for: \((L1 \text{ AND } (\text{NOT } L2)) \text{ AND } (L3 \text{ OR } L4)\)

\(\psi\) (Community Detection)

\(\Theta\) (Composition)

Layer-wise Communities (in parallel)

Intermediate Communities (in parallel)

Final Communities
Decoupling Approach Amenable to Optimization

Which collaboration groups have published in both the highly ranked conferences, but have never published in either of the medium ranked conferences?

**DEC1**
\[\text{C(VLDB) CE-AND C(SIGMOD) CE-AND C(NOT (DASFAA OR DaWaK))}\]

- Number of CE-AND: 2

**DEC2**
\[\text{C(VLDB) CE-AND C(SIGMOD) CE-AND C(NOT DASFAA) CE-AND C(NOT DaWaK)}\]

- Number of CE-AND: 3

**Application of De Morgan’s, Distribution Laws**

**Sparse Layers**

**No. of Dense (NOT) layers:** 1

**No. of Dense (NOT) layers:** 2

---

**Reduce No. of Compositions**

**Start with AND Compositions**

**Reduce NOT Compositions**

Perform cost analysis of all possible specification alternatives for a Boolean expression.
Community Detection in HoMLN

Brief Introduction to Community
Boolean AND Composition
Boolean OR Composition
Performance Analysis
Evaluation of General Boolean Expressions (AND, OR, NOT)

Case Studies: Results, Drill-down and Visualizations
Real Life HoMLNs

**IMDb-Actors HoMLN**

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-Acting</td>
<td>9485</td>
<td>45,581</td>
</tr>
<tr>
<td>Genre</td>
<td>9485</td>
<td>996,527</td>
</tr>
<tr>
<td>AvgRating</td>
<td>9485</td>
<td>13,945,912</td>
</tr>
</tbody>
</table>

Based on initial set of top 500 actors

**DBLP-CoAuthors HoMLN**

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLDB</td>
<td>5116</td>
<td>3912</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>5116</td>
<td>3303</td>
</tr>
<tr>
<td>DASFAA</td>
<td>5116</td>
<td>1519</td>
</tr>
<tr>
<td>DaWaK</td>
<td>5116</td>
<td>679</td>
</tr>
</tbody>
</table>

Based on publications from 2003 to 2007

**Facebook HoMLN**

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>2695</td>
<td>1,228,223</td>
</tr>
<tr>
<td>Gender</td>
<td>2695</td>
<td>1,813,638</td>
</tr>
<tr>
<td>Relationship Status</td>
<td>2695</td>
<td>1,119,592</td>
</tr>
<tr>
<td>Political Views</td>
<td>2695</td>
<td>494,974</td>
</tr>
<tr>
<td>Locale</td>
<td>2695</td>
<td>2,799,160</td>
</tr>
<tr>
<td>Trait: OPN</td>
<td>2695</td>
<td>1,020,306</td>
</tr>
<tr>
<td>Trait: CON</td>
<td>2695</td>
<td>840,456</td>
</tr>
<tr>
<td>Trait: EXT</td>
<td>2695</td>
<td>795,691</td>
</tr>
<tr>
<td>Trait: AGR</td>
<td>2695</td>
<td>718,201</td>
</tr>
<tr>
<td>Trait: NEU</td>
<td>2695</td>
<td>627,760</td>
</tr>
<tr>
<td>Privacy Concern</td>
<td>2695</td>
<td>2,191,659</td>
</tr>
</tbody>
</table>

Based on psychometric tests and FB profile in period (2007-2012)
## Potential Actor Collaborations (IMDb)

Which **highly rated** actors work in **similar genres** but have **not co-acted together** in any movie?

C(NOT CoActing) CE-AND C(Genre) CE-AND C(AvgRating)

<table>
<thead>
<tr>
<th>Actor/Actresses</th>
<th>Prominent Genres</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willem Dafoe, Russell Crowe</td>
<td>Action, Crime</td>
</tr>
<tr>
<td>Hilary Swank, Kate Winslet</td>
<td>Drama</td>
</tr>
<tr>
<td>Tom Hanks, Reese Witherspoon, Cameron Diaz</td>
<td>Comedy, Romance</td>
</tr>
<tr>
<td><strong>Johnny Depp, Tom Cruise</strong></td>
<td><strong>Adventure, Action</strong></td>
</tr>
<tr>
<td><strong>Leonardo DiCaprio, Ryan Gosling</strong></td>
<td><strong>Crime, Romance</strong></td>
</tr>
<tr>
<td>Nicolas Cage, Antonio Banderas</td>
<td>Action, Thriller</td>
</tr>
<tr>
<td>Hugh Grant, Kate Hudson, Emma Stone</td>
<td>Comedy, Romance</td>
</tr>
</tbody>
</table>

**Reports:** In 2017, talks of casting **Johnny Depp** and **Tom Cruise** in pivotal roles in Universal Studios' cinematic universe titled Dark Universe
Widely Accepted Collaborators (DBLP-CoAuthors)

Which collaboration groups have published in both the highly ranked conferences, but have never published in either of the medium ranked conferences?

C(VLDB) CE-AND C(SIGMOD) CE-AND C(NOT DASFAA) CE-AND C(NOT DaWaK)

Widely accepted collaboration groups with high quality work

- **Surajit Chaudhari** won the VLDB 10-Year Best Paper Award (2007) with Vivek Narasayya and VLDB Best Paper Award (2008) with Nicolas Bruno, apart from winning ACM SIGMOD Contributions Award (2004)

- **Divyakant Agrawal** has 24000+ citations (Google scholar)

- **Peter A. Boncz** and **Stefan Manegold** published a highly cited paper (350+ citations for MonetDB/XQuery) in SIGMOD 2006 and won the VLDB 10-year award
How do the personality traits evolve with age?

Openness (OPN)

- Reflects one’s preference for new experiences and to engage in self-examination
- Increases with age and peaks around the 30s (54.2% in age group of 31-40)
- Older people prefer to go with the tried-and-tested approach (67.6% of the people above 60 years old resist new experiences)
Conscientiousness (CON)

- Associated with achievement and working systematically, methodically and purposefully
- Analysis shows that the age group with most conscientiousness is 41-50 years old
- Recent survey (2018): Average age of founders and entrepreneurs is 45 years old
How do the personality traits evolve with age?

Extraversion (EXT)

➢ Describes one's sociability and enjoy to be the center of attention

➢ Seems to peak at two age groups (i.e., [31,40] and >=61) in the dataset
Agreeableness (AGR)

➢ Reflects a tendency to perceive others in a more positive light

➢ Parenthood and grandparenthood may make the elder generation more empathetic towards others as compared to the younger lot
Neuroticism (NEU)

- Reflects one's ability to deal with emotion states, such as stress and anxiety
- Younger lot does not deal very well with stress
  - Study (2009): Around 80-90% adolescent suicides are linked to common psychiatric disorders, such as depression and anxiety
- Trait (NEU) seems to be most stable over age compared to other traits
How does the individuals' age correlate with their comfort level of sharing personal information on social media?

- **People (<= 40 years old)** prefer higher level of privacy
  - More aware of the cons of sharing sensitive personal information on the web such as identity theft

- **Status updates of people (>= 41 years old)** contain more personal information and this trend increases with age
  - Reflects a lower level of privacy-concern probably due to unawareness of the potential harm from disseminating personal information on social media
Community Detection in HeMLN

Heterogeneous Community Definition
Weighted Maximum Bipartite Matching Composition
Performance Analysis and Case Studies
Limitations of existing HeMLN communities

➢ No structure-preserving definition
  ▪ Type-independent, projection-based definitions

➢ Posses undesirable properties: Loss of information/semantics, Distortion of Data

![Diagram showing aggregation and type-independent single graph](image)
Desirable Properties for HeMLN Communities

➢ Structure Preservation Required for Semantics
  ▪ Preserve layer community structure (including types)
  ▪ Preserve inter-layer edges (including relationships)

➢ Combined communities should have high modularity

➢ Should be able to use Decoupling Approach

➢ Detection must be Computationally Efficient

➢ Support for Drill-down analysis
**k-community**: Community Definition for k HeMLN layers

**Analysis Specification**: \( C(G_{n1}) \Theta_{n1,n2} C(G_{n2}) \Theta_{n2,n3} \ldots \Theta_{ni,nk} C(G_{nk}) \)

A set, where each element represents

- **One community** from each of k layers obtained through
  - Ensures strong connectivity **within a layer**

- **progressively strong coupling**, using the binary operator
  - Ensures strong connectivity **across layers**

**Each Output element**: \(< c_{n1}^{m1}, c_{n2}^{m2}, \ldots, c_{nk}^{mk} ; x_{n1,n2}, x_{n2,n3}, \ldots, x_{ni,nk} >\)
k-community (Formalism)

- **CBG}_{i,j}(U_i, U_j, L'_{i,j}):** Community Bipartite Graph
  - **U_i, U_j (Meta Node sets):** Representing communities from layer \( G_i, G_j \)
  - **L'_{i,j} (Weighted Meta Edge set):** One edge for a pair of meta nodes, if an inter-layer edge exists between any node pair from corresponding communities

- **2-community:** CBG Meta node pairs that maximize total inter-layer edge weights (along with the overall modularity) between the two CBG meta node sets

- **k-community:** Recursive application of 2-community to compose k-community for k layers
Community Detection in HeMLN

Heterogeneous Community Definition

Weighted Maximum Bipartite Matching Composition

Performance Analysis and Case Studies
Need For Additional Bipartite Matching Algorithms

Traditional Weighted Matchings

➢ Simple nodes (hiring, dating)
➢ Weighted Edges supported
➢ Unique Matches only

HeMLN Community REQUIREMENT

➢ Nodes are Communities
➢ Meta edge weights need to reflect participating community characteristics
➢ Need for non-unique matches
  ▪ To increase total edge weight for the same number of pairings and,
  ▪ Deal with (or include) ties of edge weights incident on the same nodes of unique pairings (instead of choosing one randomly!)
Weighted Maximum Bipartite Matching Algorithms

Original Community Bipartite Graph

Meta Nodes (Communities)

Weighted Meta Edges

Traditional Algorithms

Maximum Weight Match (MWM)

\[ \sum w = 20; \text{ #matches} = 2 \]

Maximum Weight Perfect Match (MWPM)

\[ \sum w = 17; \text{ #matches} = 3 \]

Relaxed Algorithms

Maximum Weight Match with Ties (MWMT)

\[ \sum w = 25; \text{ #matches} = 3 \]

Maximum Weight with Relaxed Matching (MWRM)

\[ \sum w = 31; \text{ #matches} = 2 \]
k-community Detection Algorithm

1. Specification:
   \[ C(G_{n1}) \Theta_{n1,n2} C(G_{n2}) \Theta_{n2,n3} \ldots \Theta_{ni,nk} C(G_{nk}), \omega, MWxx \]

2. result \( \leftarrow \) initialize(2-community\((G_{n1}, G_{n2}), \text{HeMLN}, \omega, MWxx\))

3. left \( \leftarrow \) nextLeftSubscript(\(\Theta\)), right \( \leftarrow \) nextRightSubscript(\(\Theta\)), \(k = 2\)

4. while left \(\neq\) null \&\& right \(\neq\) null
   - Construct CBG for \(G_{\text{left}}\) and \(G_{\text{right}}\)
   - MP \(\leftarrow\) MWxx(CBG)
   - For each tuple \(\in\) result
     - Update the tuple
       (if case (ii): both processed layers)
     - Extend the tuple and increment k
       (if case (i): one processed layer and one new layer,
         left \(\leftarrow\) nextLeftSubscript(\(\Theta\)) or null
         right \(\leftarrow\) nextRightSubscript(\(\Theta\)) or null
Cases and Outcomes of MWxx
(in any recursive step with CBG for layers $G_{\text{left}}$ and $G_{\text{right}}$)

<table>
<thead>
<tr>
<th>($G_{\text{left}}, G_{\text{right}}$) Outcome</th>
<th>Effect on tuple $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case (i): One Processed Layer and One New Layer</strong></td>
<td></td>
</tr>
<tr>
<td>a) Consistent Match</td>
<td>Extend $t$ with <em>paired community id</em> and $x_{i,j}$</td>
</tr>
<tr>
<td>b) No Match</td>
<td>Extend $t$ with 0 (<em>null id</em>) and $\phi$ (<em>empty set</em>)</td>
</tr>
<tr>
<td><strong>Case (ii): Both are Processed Layers (Cyclic)</strong></td>
<td></td>
</tr>
<tr>
<td>a) Consistent Match</td>
<td>Update $t$ only with $x_{i,j}$</td>
</tr>
<tr>
<td>b) No Match</td>
<td>Update $t$ with $\phi$ (<em>empty set</em>)</td>
</tr>
<tr>
<td>c) Inconsistent Match</td>
<td>Update $t$ with $\phi$ (<em>empty set</em>)</td>
</tr>
</tbody>
</table>

\[ C(G_{n1}) \Theta_{n1,n2} C(G_{n2}) \Theta_{n2,n3} C(G_{n3}) \Theta_{n3,n1} C(G_{n1}) \]
Proposed Weight Metrics for Meta Edge \((u, v)\)

➢ **Analysis Requirement:** *Maximize number of interactions* between the participating communities

▪ **Number of Inter-Community Edges**
  
  \[ \omega_e = \text{number of interlayer edges between } cu \text{ and } cv \]

➢ **Analysis Requirement:** *Strong intra-community and inter-community interaction*

▪ **Density and Edge Fraction**
  
  \[ \omega_d = (c_u \text{ density}) \times \frac{c_u \text{ nodes}}{|c_u \text{ nodes}|} \times \frac{c_v \text{ nodes}}{|c_v \text{ nodes}|} \times (c_v \text{ density}) \]

➢ **Analysis Requirement:** *Participation of influential nodes within and between* participating communities

▪ **Hub Participation**
  
  \[ \omega_h = (c_u \text{ ratio of hubs participating}) \times \frac{c_u \text{ nodes}}{|c_u \text{ nodes}|} \times \frac{c_v \text{ nodes}}{|c_v \text{ nodes}|} \times (c_v \text{ ratio of hubs participating}) \]
Evaluation of Proposed Community Definition (Modularity Comparison)

**Good Community Structure**
Modularity > 0.5

**Type Independent**: Node Types mix up to produce smaller denser communities => Higher Modularity

**k-community**: Community pairings vary based on the choice of MWxx, leading to change in modularity values

<table>
<thead>
<tr>
<th></th>
<th>DBLP HeMLN (Author (\Theta) Paper, (w_e))</th>
<th>IMDb HeMLN (Actor (\Theta) Director, (w_e))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type-Independent</strong></td>
<td><strong>MWM</strong></td>
<td><strong>MWMT</strong></td>
</tr>
<tr>
<td><strong>0.69</strong></td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>0.78</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Effect of MWxx

Variation in Number of Matches and Overall Weight

<table>
<thead>
<tr>
<th>Specification</th>
<th>MWM</th>
<th>MWMT</th>
<th>MWPM</th>
<th>MWRM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P-Au, ( w_d )</strong></td>
<td>Total: 6</td>
<td>Total: 9</td>
<td>Total: 6</td>
<td>Total: 6</td>
</tr>
<tr>
<td></td>
<td>( \sum w: 0.04988 )</td>
<td>( \sum w: 0.0778 )</td>
<td>( \sum w: 0.04988 )</td>
<td>( \sum w: 0.05003 )</td>
</tr>
<tr>
<td><strong>DBLP HeMLN</strong></td>
<td><strong>Total: 6</strong></td>
<td><strong>Total: 6</strong></td>
<td><strong>Total: 6</strong></td>
<td><strong>Total: 6</strong></td>
</tr>
<tr>
<td></td>
<td>Partial: 0</td>
<td>Partial: 0</td>
<td>Partial: 0</td>
<td>Partial: 3</td>
</tr>
<tr>
<td></td>
<td>( \sum w: 548 )</td>
<td>( \sum w: 548 )</td>
<td>( \sum w: 548 )</td>
<td>( \sum w: 986 )</td>
</tr>
<tr>
<td><strong>P-Au-Y, ( w_e )</strong></td>
<td>Total: 6</td>
<td>Total: 6</td>
<td>Total: 6</td>
<td>Total: 6</td>
</tr>
<tr>
<td></td>
<td>Partial: 0</td>
<td>Partial: 0</td>
<td>Partial: 0</td>
<td>Partial: 3</td>
</tr>
<tr>
<td></td>
<td>( \sum w: 548 )</td>
<td>( \sum w: 548 )</td>
<td>( \sum w: 548 )</td>
<td>( \sum w: 986 )</td>
</tr>
<tr>
<td><strong>IMDb HeMLN</strong></td>
<td>Total: 50</td>
<td>Total: 69</td>
<td>Total: 57</td>
<td>Total: 50</td>
</tr>
<tr>
<td></td>
<td>( \sum w: 9902 )</td>
<td>( \sum w: 9970 )</td>
<td>( \sum w: 5144 )</td>
<td>( \sum w: 11640 )</td>
</tr>
<tr>
<td><strong>A-D, ( w_e )</strong></td>
<td>Total: 50</td>
<td>Total: 69</td>
<td>Total: 57</td>
<td>Total: 50</td>
</tr>
<tr>
<td></td>
<td>( \sum w: 9902 )</td>
<td>( \sum w: 9970 )</td>
<td>( \sum w: 5144 )</td>
<td>( \sum w: 11640 )</td>
</tr>
<tr>
<td><strong>M-A-D-M, ( w_e )</strong></td>
<td><strong>Total: 2</strong></td>
<td><strong>Total: 3</strong></td>
<td><strong>Total: 0</strong></td>
<td><strong>Total: 2</strong></td>
</tr>
<tr>
<td></td>
<td>Partial: 7</td>
<td>Partial: 12</td>
<td>Partial: 9</td>
<td>Partial: 11</td>
</tr>
<tr>
<td></td>
<td>( \sum w_{M-A}: 6979 )</td>
<td>( \sum w_{M-A}: 6984 )</td>
<td>( \sum w_{M-A}: 6979 )</td>
<td>( \sum w_{M-A}: 11557 )</td>
</tr>
</tbody>
</table>

**Total Inter-Layer Edge Weight**

- MWPM \( \leq \) MWM \( \leq \) MWRM
- MWM \( \leq \) MWMT

**Total Number of Pairs**

- MWM = MWRM \( \leq \) MWMT
- MWM = MWPM \( \leq \) MWMT

**MWMT and MWRM improve the Overall Sum of Weights**

**MWM and MWRM produce least number of pairings**
k-community Characteristics

➢ Family of Composition Algorithms (4 MWxx, 3 w)
➢ Commutative, Not Associative: Community pairs depend on order in which layers are combined
➢ Space of Analysis Alternative: If HeMLN (G,X) is assumed to be a simple graph of G as nodes and X as edges, then number of analysis is a function of,
   ▪ Number of connected subgraphs
   ▪ Number of possible orderings
   ▪ Number of metrics for the edge weights
   ▪ Number of bipartite pairing choices (Θ)
Cost Analysis of Decoupling Approach

➢ **One Time Cost**

  ▪ **Generating 1-community**
  ▪ Can be done in **parallel**. Bounded by the **densest layer**
  ▪ Parts of meta edge weights \((\omega_d, \omega_h)\)

➢ **Cost of combining intermediate results**

  ▪ Weight Computation Cost + CBG Creation Cost + MWxx Cost
    - \(|\text{Meta Edges}| \ll |\text{Inter-layer Edges}|\)
    - \(|\text{Meta Nodes or Communities}| \ll |\text{Layer Nodes}|\)

➢ **Cost of combining intermediate results (cost of each iteration) \ll One Time Cost**
Community Detection in HeMLN

Heterogeneous Community Definition
Weighted Maximum Bipartite Matching Composition

Performance Analysis and Case Studies: Results, Drill-Down and Visualizations
Experimental Set Up

➢ **Environment**
  - OS: Windows 10 | RAM: 8GB
  - Codes implemented in Python

➢ **1-community Detection Algorithm: Louvain**
### Experimental Set up (Data Sets)

#### IMDb HeMLN

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
<th>#Communities</th>
<th>Avg. Comm. Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actors</strong> (Genre-linked)</td>
<td>9485</td>
<td>996,527</td>
<td>63/190 (Size &gt; 1/All)</td>
<td>148.5</td>
</tr>
<tr>
<td><strong>Directors</strong> (Genre-linked)</td>
<td>4510</td>
<td>250,845</td>
<td>61/190 (Size &gt; 0/All)</td>
<td>73</td>
</tr>
<tr>
<td><strong>Movies</strong> (Rating-linked)</td>
<td>7951</td>
<td>8,777,618</td>
<td>9/9 (Size &gt; 0/All)</td>
<td>883.4</td>
</tr>
</tbody>
</table>

Based on initial set of top 500 actors

#### DBLP HeMLN

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
<th>#Communities</th>
<th>Avg. Comm. Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authors</strong> (3 Papers Co-authored)</td>
<td>16,918</td>
<td>2,483</td>
<td>591/15528 (Size &gt; 0/All)</td>
<td>3.3</td>
</tr>
<tr>
<td><strong>Papers</strong> (Conference-linked)</td>
<td>10,326</td>
<td>12,044,080</td>
<td>6/6 (Size &gt; 0/All)</td>
<td>1721</td>
</tr>
<tr>
<td><strong>Years</strong> (Range-linked)</td>
<td>18</td>
<td>18</td>
<td>6/6 (Size &gt; 0/All)</td>
<td>3</td>
</tr>
</tbody>
</table>

Based on publications in VLDB, SIGMOD, ICDM, KDD, DASFAA, DaWaK from 2001 to 2018
<table>
<thead>
<tr>
<th>DBLP HeMLN</th>
<th>IMDb HeMLN</th>
</tr>
</thead>
</table>
| Conference-wise which are the **most cohesive group(s)** of authors who **publish frequently** (ties included)?  
\[ P \Theta_{\text{P, Au}} \text{ Au}; w_d; \Theta = \text{MWMT} \] | Find the actor and director similar-genre based group pairs such that **overall actor-director collaborations are maximized**?  
\[ A \Theta_{A, D}; w_e; \Theta = \text{MWRM} \] |
| For the **most popular unique** collaborators from each conference, which are the **unique most active** 3-year period(s)?  
\[ P \Theta_{\text{P, Au}} \text{ Au }\Theta_{\text{au,Y}} \ Y; w_e; \Theta = \text{MWM} \] | Based on genres, list the **maximum number** of **unique** actor and director groups whose majority of the **most versatile members interact**?  
\[ A \Theta_{A, D}; w_h; \Theta = \text{MWPM} \] |
| For the **most popular** actor groups, from each movie rating class, find the director groups with **maximum interaction** and who also make movies with similar ratings (**including ties**).  
\[ M \Theta_{\text{M,A}} A \Theta_{\text{A,D}} D \Theta_{\text{D,M}} M; w_e; \Theta = \text{MWMT} \] |  |
**Efficiency**

The additional incremental cost for computing a k-community is extremely small validating the efficiency of decoupled approach.

The cost of all iterations together (0.27 sec) is more than an order of magnitude less than the largest one-time cost (5.43 sec for Movie layer).

\[ M \Theta_{M,A} A \Theta_{A,D} D \Theta_{D,M} M, \omega, \Theta = \text{MWMT} \]
Potential Actor-Director Collaborations (IMDb)

For the most popular actor groups, from each movie rating class, find the director groups with maximum interaction and who also make movies with similar ratings (including ties).

$M \Theta_{M,A} A \Theta_{A,D} D \Theta_{D,M} M; \ w_e; \ \Theta = MWMT$

(a) Final Result

(b) Sample Movies, Actors and Directors from a Total Element
Potential Actor-Director Collaborations (IMDb)

IMDb HeMLN Community Results: Similar Expertise, Highly Rated

(b) Sample Movies, Actors and Directors from a Total Element

DiCaprio
Mendes
Swank

[7-8] Rating

M1
D106
A175

Nixon
127Hours
RescueDawn
RevolutionaryRoad
Milk
ManchesterByTheSea
MysticRiver

KennethLonergan
GusVanSant
SamMendes
TimRobbins
DannyBoyle
WernerHerzog
OliverStone

ChristianBale
KateWinslet
RussellCrowe
LeonardoDiCaprio
SeanPenn
KevinBacon
CaseyAffleck
HilarySwank
James Franco
AnthonyHopkins
Frequently Publishing Groups (DBLP)

Conference-wise which are the most cohesive/collaborative author group(s) who publish frequently (ties included)?

- ICDM and DaWaK have *multiple author communities* that are equally important
- Quality Validating Facts:
  - SIGMOD: Tim Kraska has been a recipient of Best of SIGMOD Award (2008, 2016)
  - VLDB: Rajeev Rastogi’s published papers in VLDB (in past 18 years) have received over 900 citations
  - ICDM: George Karypis has been a recipient of IEEE ICDM 10-Year Highest-Impact Paper Award (2010) and IEEE ICDM Research Contributions Award (2017)
For the most popular unique author groups from each conference, which are the unique most active 3-year period(s)?

For SIGMOD, VLDB and ICDM the most popular researchers include Srikanth Kandula (15188 citations), Divyakant Agrawal (23727 citations) and Shuicheng Yan (52294 citations), respectively have been active in different periods in the past 18 years.
Similar Actor-Director Groups (IMDb)

Based on genres, list the **maximum number of unique** actor and director groups whose majority of the **most versatile members** interact

- Directors who make movies **prominently in some genre** **pair up** with actors who **primarily act** in similar kind of movies

- Action/Drama: Clint Eastwood, Ridley Scott and Steven Spielberg **pair up** with Brad Pitt, Tom Cruise and Will Smith

- Comedy: Bobby Farrelly, Todd Phillips, John Landis **pair up** with Jim Carrey, Zach Galifianakis, Eddie Murphy

- Romance: Woody Allen, Tim Burton **pair up** with Diane Keaton, Emma Stone and Hugh Grant
Hub Detection in HoMLN

Introduction to Centrality Metrics

Boolean AND Composition for Centrality Hubs (Overview)

Case Study
Hubs in Simple Graphs

- **Definition:** Nodes having the centrality metric value higher/lower than the average
  - e.g., popular person on Facebook/Twitter, airport hubs, popular co-actors etc.

- Centrality Metrics used
  - **Degree Centrality**
    - *Number of links/edges* incident on a vertex
    - Higher the degree, greater the influence on immediate neighborhood
  - **Closeness Centrality**
    - *Average shortest path* between a node and all other nodes in the graph
    - Information spreads quickly across a network through these hubs

- **Other Metrics:** Betweenness, Eigenvector
Hub Detection in HoMLN

Introduction to Centrality Metrics

Boolean AND Composition for Centrality Hubs (Overview)

Case Study
Degree Centrality Heuristics

➢ **DCi-AND:** Intersect the layer-wise hubs
  - Layers have similar topology: High Accuracy, Low Overhead
  - In general, low accuracy due to presence of *false positives and negatives*

➢ **DCn-AND:** Check if the common hubs have enough shared neighbors
  - Additional Overhead
    - AND layer average degree needs to be estimated
    - One hop neighbors needs to be stored
  - False positives eliminated, Higher Precision
Experimental Results

➢ Setup (IMDb HoMLN)

▪ Nodes: 5000 Actors

▪ Layers: 2 nodes connected if the actors have acted in a Comedy movie (Layer C) or a Drama movie (Layer D) or an Action movie (Layer A)

▪ 4 AND Composition Analysis
  – C AND A, A AND D, C AND D, (C AND A) AND D

➢ Accuracy Metrics

▪ Precision to check “how relevant are the resulting hubs”

▪ Jaccard Index used to compare the hub sets
Trade-off between Accuracy and Efficiency

Elimination of False Positive increases the Precision, Decreases Efficiency

For large N (number of MLN layers), denser layers, more analysis: Efficiency is higher
Closeness Centrality Heuristics

- **CCi-AND** - Intersect the layer-wise hubs
  - **Layers have similar topology**: High Accuracy, Low Overhead.
  - **In general, low accuracy** due to presence of *false positives and negatives*

- **CCn-AND** – High degree neighborhood within 1 hop distance used
  - **Higher Precision**: False positives decreased
Trade-off between Accuracy and Efficiency

Decrease in False Positives increases the Precision, Decreases Efficiency

For large $N$ (number of MLN layers), denser layers, more analysis: Efficiency is higher
Component Cost of Decoupling Approaches

Worst Case Analysis: Maximum cost of combining the partial results is significantly less than the minimum cost to detect 1 layer hubs

CCi-AND

MIN(1-community Cost)  MAX (Partial Results Combination Cost)

1.71  0.002

CCn-AND

MIN(1-community Cost)  MAX (Partial Results Combination Cost)

1.65  0.05

99.88% ↓  97% ↓
Hub Detection in HoMLN

Introduction to Centrality Metrics

Boolean AND Composition for Centrality Hubs (Overview)

Case Study
Upcoming Airline Hubs (US Airlines)

Identify preferred cities for an airline to expand its operations taking all its competitors into consideration

CC(Allegiant) – ActualHubs(Allegiant) – ( CC(American) CCI-AND CC(Southwest) CCI-AND CCI-AND CC(Delta) CCI-AND CC(Spirit) CCI-AND CC(Frontier) )

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>290</td>
<td>746</td>
</tr>
<tr>
<td>Southwest</td>
<td>290</td>
<td>717</td>
</tr>
<tr>
<td>Delta</td>
<td>290</td>
<td>688</td>
</tr>
<tr>
<td>Frontier</td>
<td>290</td>
<td>346</td>
</tr>
<tr>
<td>Spirit</td>
<td>290</td>
<td>189</td>
</tr>
<tr>
<td>Allegiant</td>
<td>290</td>
<td>379</td>
</tr>
</tbody>
</table>

Based on direct flights active in February 2018

➢ Intuition: Cities for expansion?
  ▪ Reduce Cost of Expansion: Fair amount of coverage (high centrality nodes)
  ▪ Minimize Competition against Competitors: Competitor airlines have less coverage (low centrality nodes)

➢ Validating Fact: Grand Rapids is one of the cities converted to a hub by Allegiant from July 6, 2019
Substructure Discovery in HoMLN
Iterative Decoupling Approach (Overview)
Iterative Decoupling Approach for Substructure Discovery in HoMLN

### Substructure Discovery
Expansion of (k-1) to k-edge substructure

### Composition
Aggregate and Combine common instances across layers; Frequency Evaluation; Restrict Search Space (Beam)

This process continues iteratively until a termination condition is applied (Limit the maximum length of the frequent substructure)

**HoMLN with Node Labels and Edge Labels**

**Layer-wise k-edge Substructures (in parallel)**

**Frequent k-edge Substructure across layers**
Conversion of an EER Model to MLN Model

8 Step Algorithm
Algorithm for EER Model $\rightarrow$ MLN Model

Research Paper Publication Dataset Modeling

Key Attribute = Node Label

Recursive Binary Relationship

Min Max Cardinality = Degree Information

Non-Recursive Binary Relationship

Relationship Name = Intra/Inter Edge Label

<table>
<thead>
<tr>
<th>Relationship Name</th>
<th>Key Attribute</th>
<th>Min Max Cardinality</th>
<th>Degree Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborates-with</td>
<td>NAME</td>
<td>(0,N)</td>
<td></td>
</tr>
<tr>
<td>Writes</td>
<td>NAME</td>
<td>(1,N)</td>
<td></td>
</tr>
<tr>
<td>Published-in</td>
<td>TOTAL_PAPERS</td>
<td>(1,N)</td>
<td></td>
</tr>
<tr>
<td>Active-in</td>
<td>INSTITUTION</td>
<td>(1,N)</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td>(1,N)</td>
<td></td>
</tr>
<tr>
<td>Same-Range</td>
<td></td>
<td>(2,2)</td>
<td></td>
</tr>
<tr>
<td>Same-Score</td>
<td></td>
<td>(0,N)</td>
<td></td>
</tr>
<tr>
<td>Same-Range</td>
<td></td>
<td>(2,2)</td>
<td></td>
</tr>
<tr>
<td>Same-Range</td>
<td></td>
<td>(2,2)</td>
<td></td>
</tr>
<tr>
<td>Same-Score</td>
<td></td>
<td>(0,N)</td>
<td></td>
</tr>
</tbody>
</table>

Coding: CODS - COMAD 2021 Tutorial
Analysis and Visualization of COVID-19 using Multilayer Networks

Live Dashboard
Some Critical Covid-19 Analysis Questions

- In which regions, was the lockdown most effective? Did geographical proximity of regions play a role?
- Were the regions with an uptick caused by the election rallies with mass gatherings?
- What is the effect of air and land traffic movement on new cases? How to choose a county (centrality) for lockdown/vaccine administration for maximum impact?
- Which dense regions of the country are more susceptible to the virus spread due to the presence of institutions with large work force?
- Which regions got significantly affected due to major events like Halloween, Thanksgiving, Christmas and New Year? What precautions need to be taken for future events – State Elections, Festivals?
Data Modeling: Multilayer Networks

In which regions, was the lockdown most effective? Did geographical proximity of regions play a role?

Which dense regions are more susceptible to virus spread due to presence of institutions with large work force?

Data Analysis Requirements

Multiple Features, Relationships

Multiple Entities

Homogeneous Multilayer Network

Heterogeneous Multilayer Network

In which regions, was the lockdown most effective? Did geographical proximity of regions play a role?

Which dense regions are more susceptible to virus spread due to presence of institutions with large work force?

Data Analysis Requirements

Multiple Features, Relationships

Multiple Entities

Homogeneous Multilayer Network

Heterogeneous Multilayer Network
Interactive COVID Data Visualization using MLNs

https://itlab.uta.edu/CoWiz/
CoWiz: Analyzing the Impact of Events (Flow)

Event for Analysis

Thanksgiving Weekend in the US (Nov 26, 2020 – Nov 30, 2020)

**Base Period**
(Period Prior to the Event
Number of New Cases form the basis of Comparison)

- Nov 10, 2020 – Nov 19, 2020

**Equal Period Length:** 10 Days

- Dec 3, 2020 – Dec 12, 2020

**Target Period**
(Period After the Event
Number of New Cases compared against the Target Period)

- **US COUNTIES**
- **County Communities:** Geographical Regions with similar % change in new cases; Classifying Regions with varying effects of event
- **Severity Rate** (Severity of the Spread)

**Severity Rate**
(Severity of the Spread)

- **Edge exists if counties have similar % Change in the number of new cases in the Target Period as compared to the Base Period**
CoWiz: Post Lockdown Effect on a Monthly basis

Lockdown in most counties: April 1, 2020 to April 30, 2020 (Base Period)

Target Period

- May 1, 2020 to May 30, 2020
- June 1, 2020 to June 30, 2020
- July 1, 2020 to July 30, 2020
- August 1, 2020 to August 30, 2020
- September 1, 2020 to September 30, 2020
- October 1, 2020 to October 30, 2020
- November 1, 2020 to November 30, 2020
- December 1, 2020 to December 30, 2020
Interactive COVID Data Visualization using MLNs

https://itlab.uta.edu/CoWiz/

DEMO
Additional Layers for the Complete Picture

➢ **Census Data:** Population, Land Area, Income, ...
  ▪ US: [https://www.census.gov/](https://www.census.gov/)  
  ▪ India: [https://censusindia.gov.in/](https://censusindia.gov.in/)

➢ **Traffic Movement:** Miles, County Boundaries, ...
  ▪ US: [https://www.streetlightdata.com/](https://www.streetlightdata.com/)

➢ **Average Mobility in Categories of Places:** Parks, Transit Stations, Workplaces, ...
  ▪ Country-wise: [https://www.google.com/covid19/mobility/](https://www.google.com/covid19/mobility/)

➢ **Expenditure Categories:** Grocery, Shopping Malls, Entertainment, ...
  ▪ Anonymized data from credit card companies
Summary

➢ MLNs v/s Simple/Attributed Graphs
  ▪ Modeling and Computation Challenges

➢ Decoupling Approach for MLN Analysis

➢ Efficient and lossless composition techniques for various analysis
  ▪ Communities (HoMLN, HeMLN)
  ▪ Hubs, Substructure (HoMLN)

➢ Community Definition for HeMLN

➢ Real world applicability of MLN Analysis
Food for Thought

➢ Subgraph Mining in HoMLN and HeMLN
  ▪ MDL/Frequency Definition, Composition Techniques

➢ Querying in MLNs

➢ Hub Detection in HeMLN
  ▪ Definition, Composition Techniques

➢ Composition techniques for *weighted and directed* MLN layers

➢ Processing approaches for distributed MLN
Thank You Questions?

For more information visit: https://itlab.uta.edu
References


Komar, K.S., 2019. Data-Driven Modeling of Heterogeneous Multilayer Networks for Computing Communities Using Bipartite Graphs (Master’s Thesis). The University of Texas at Arlington


