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

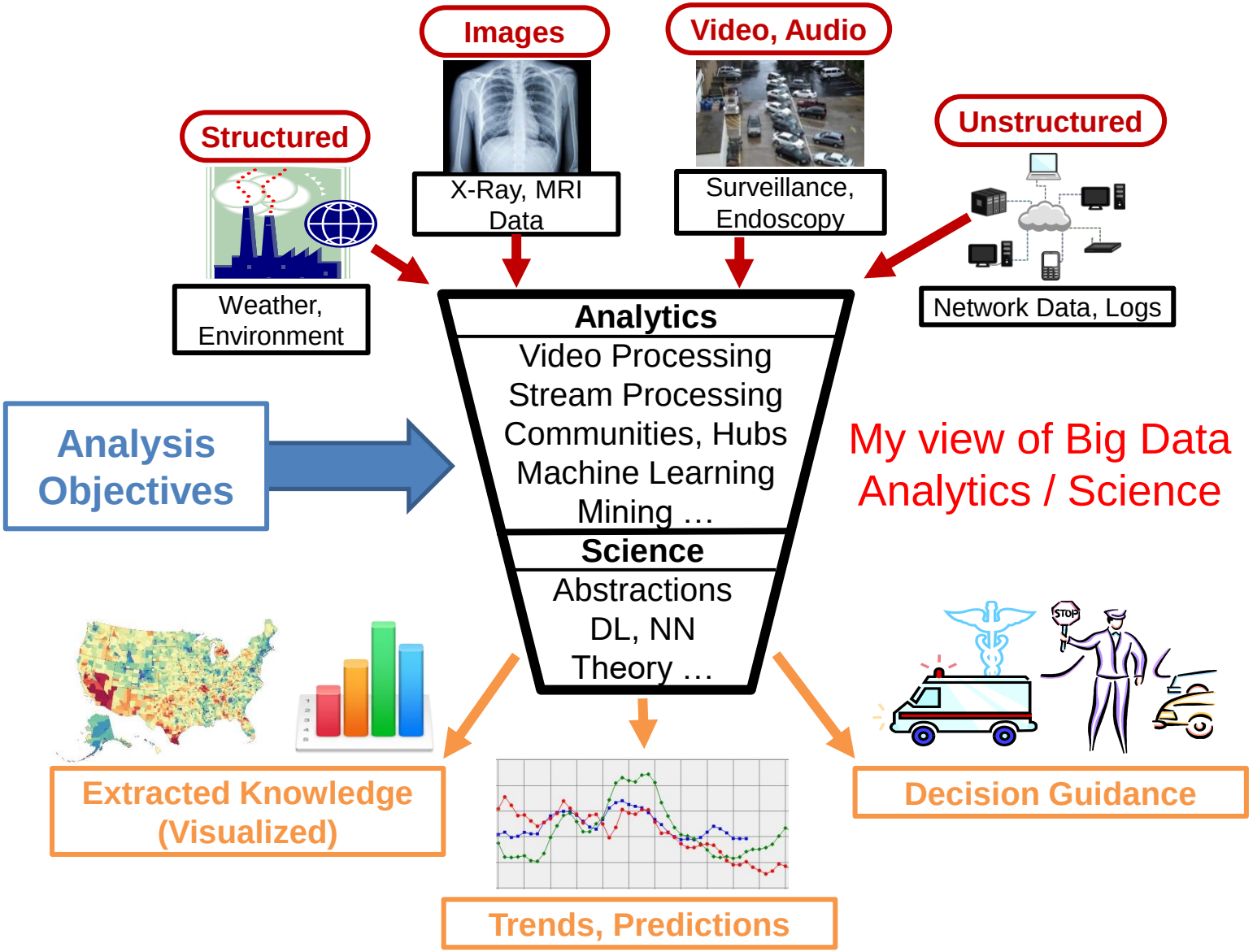
(CODS-COMAD 2021 Tutorial)

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

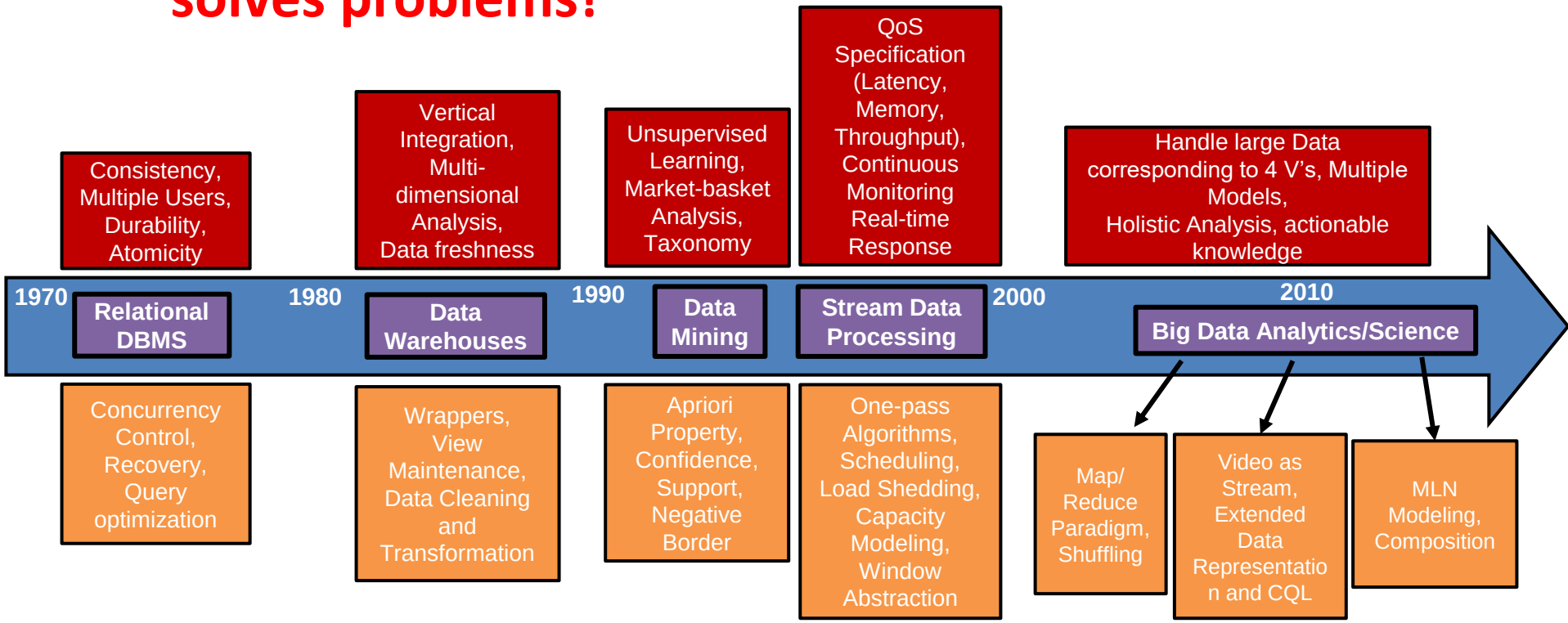
What is Big Data Analysis/Science
Where are we headed?



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**

AADHAAR

Hotels.com

DATA.GOV.UK
Opening up Government

facebook

zomato

LinkedIn

dblp
computer science bibliography

Expedia

NCDC

IMDb
The Internet Movie Database

twitter

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*

Most popular or socially active group of people across platforms?

Most influential set of people?

Airline Connectivity among
same *US cities*

same *Indian Cities*

High central cities (hubs) ?

Next upcoming hub?

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** IMDb entities


Movies


Actors

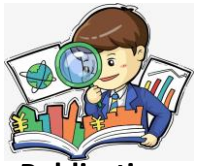

Directors


Genres


Rating

Connectivity among **different** dblp entities


Author Collaborations


Publications


Conferences

WikiCFP
A Wiki for Calls For Papers
Research Domains


Years


Venues

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

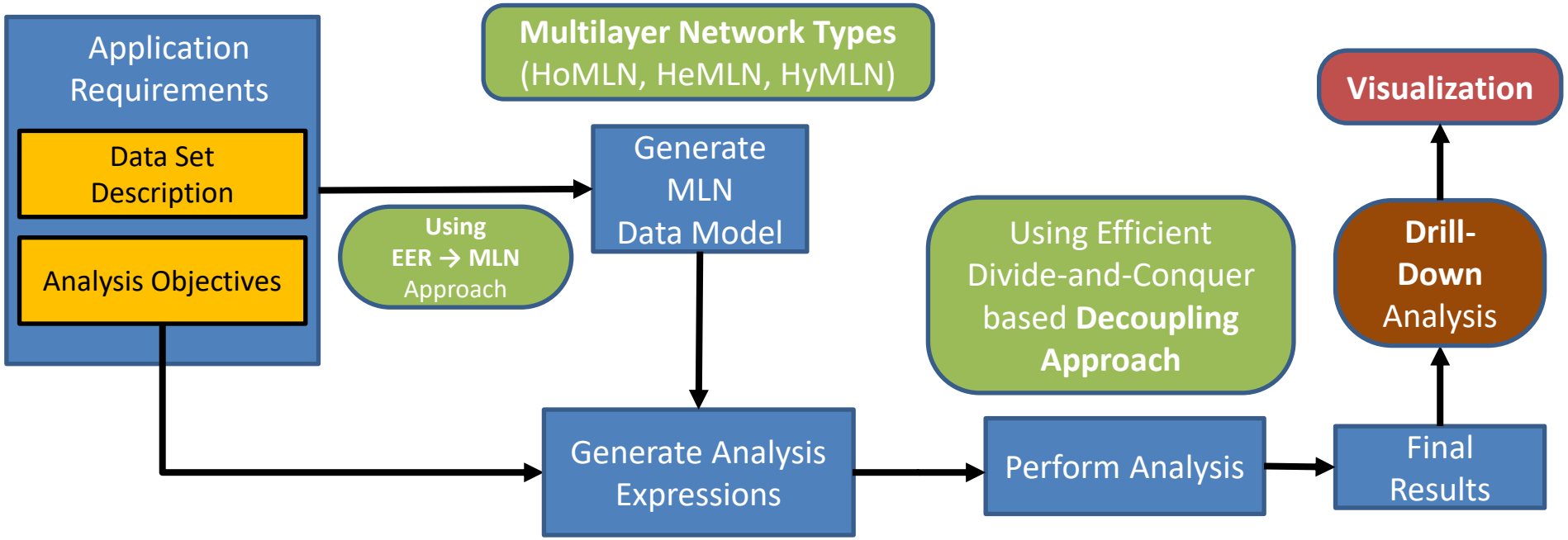
Highly rated actors working in similar genres who have *never* acted together?

Frequently publishing cohesive co-author groups?

Most active periods for popular collaborators?

Heterogeneous Multilayer Networks (HeMLN)

Mining Vs. Big Complex Data Analytics



Motivation

Complex Data Analysis

Traditional Approaches and Their Limitations

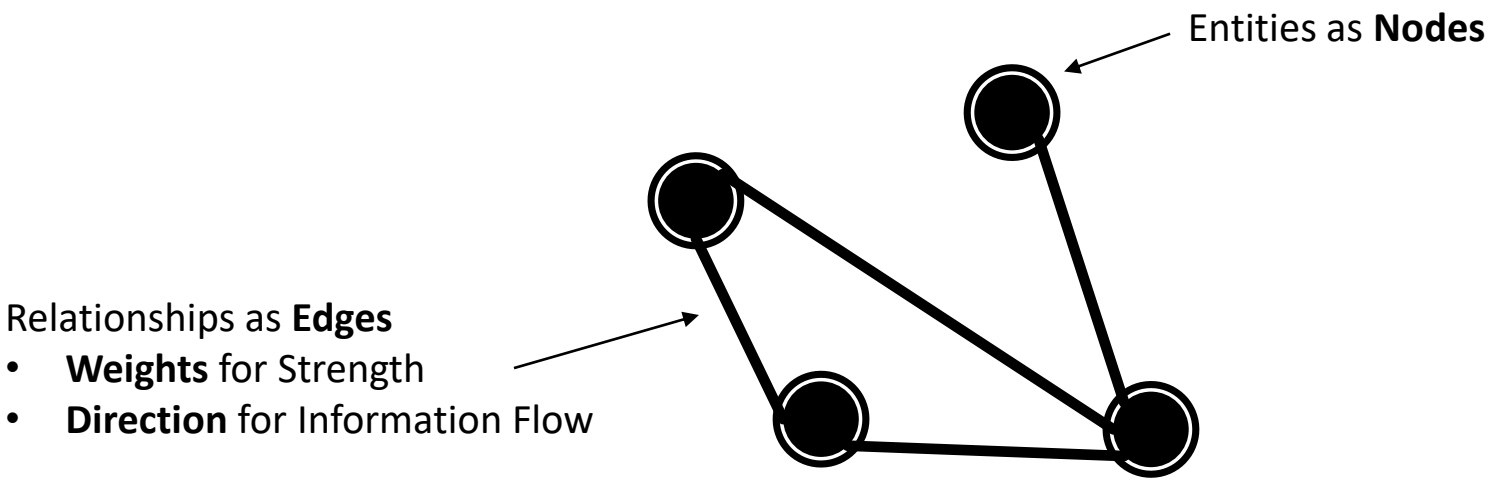
Modeling Using Multilayer Networks

Limitations of Existing Analysis Approaches

Decoupling Approach to Analyze MLNs

Traditional Modeling: Simple Graphs

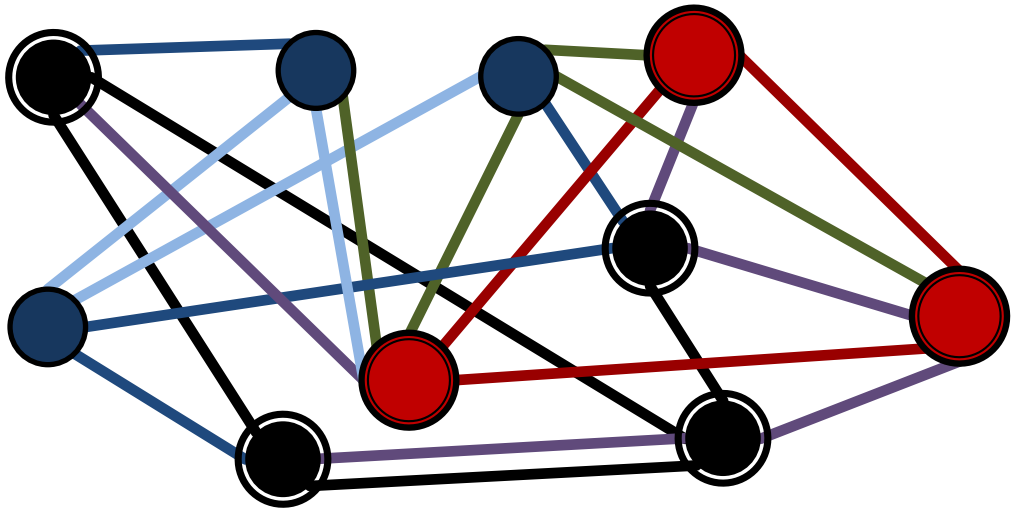
- Nodes: Entities
- Single Edges (weighted or unweighted): Single or Combination of feature-based relationship



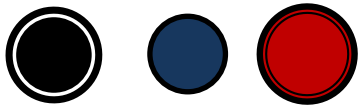
- **Algorithms exist** for communities, hubs, subgraph mining, frequent subgraph counting, etc.

Traditional Modeling: Attributed Graphs

- Nodes: Entities
  - Node Labels: Entity Types
- Multiple Edges: Feature-based relationship (weighted or unweighted)
  - Edge Labels: Feature Types



Entities as **Colored Nodes**



Relationships as **Colored Edges**

- **Algorithms exist** for subgraph mining

Limitations of Traditional Modeling Approaches

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

Previous Work

- **Community Detection in Simple Graphs**
  - Palla, G., Derényi, I., Farkas, I. and Vicsek, T., *Nature*, 2005
  - Rosvall, M. and Bergstrom, *Proceedings of the National Academy of Sciences*, 2008
  - Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, *Journal of statistical mechanics: theory and experiment*, 2008
- **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
  - Kuramochi, M. and Karypis, G., *ICDM*, 2001
  - 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

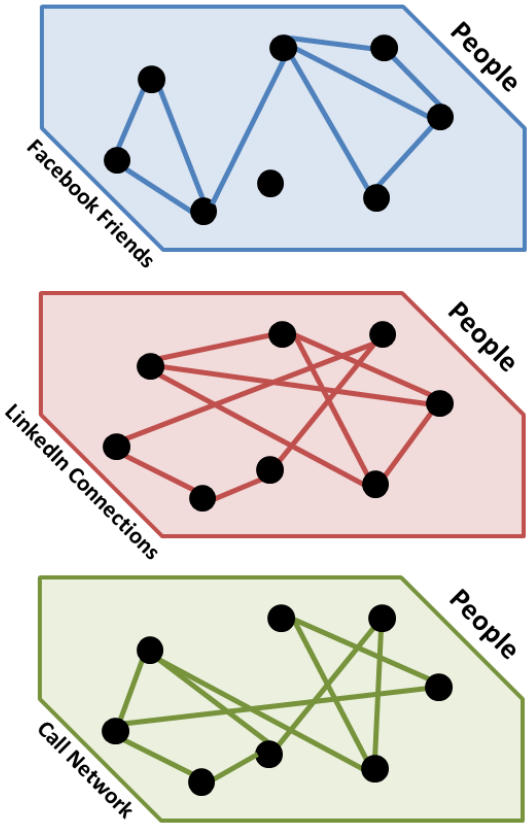
Choice of layer nodes, intra-layer edges depending on the **Semantics of Analysis Objectives**

Interactions among **People**

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

***b. Most influential* set of people?**

Same Entities, Different Relationships



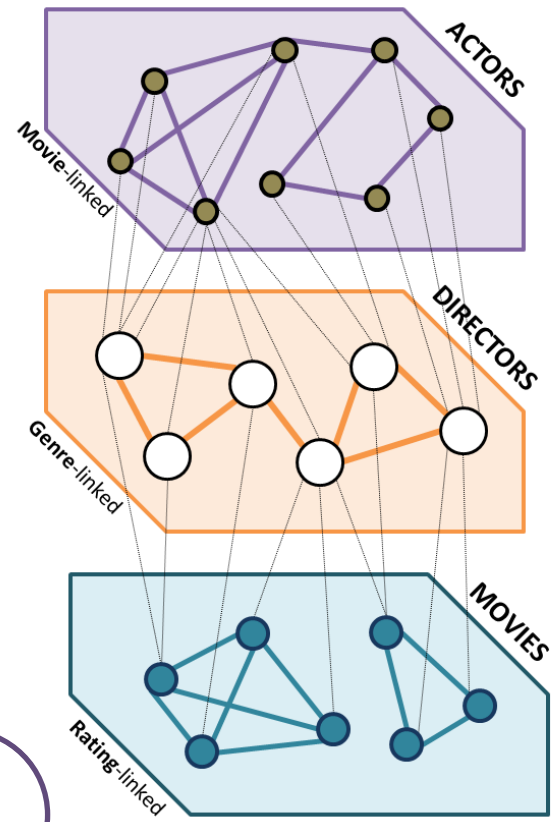
Homogeneous MLN (HoMLN)

Modeling Clarity using MLNs

Choice of layer nodes, intra- <u>and</u> inter-layer edges depending on the **Semantics of Analysis Objectives**

Interactions among IMDb **Entities**

Different Entities, features, and Relationships



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?**

Heterogeneous MLN (HeMLN)

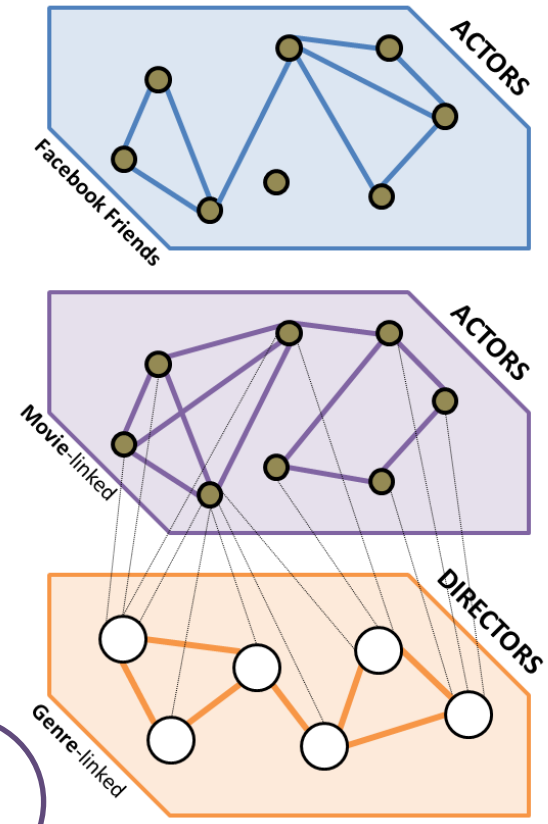
Modeling Clarity using MLNs

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*?

HoMLN
+
HeMLN



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 impilcit**
- **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

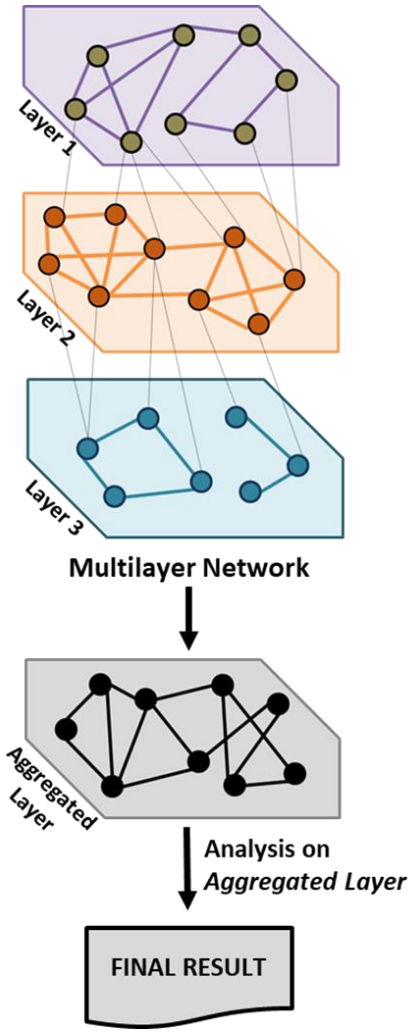
Decoupling Approach to Analyze MLNs

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*

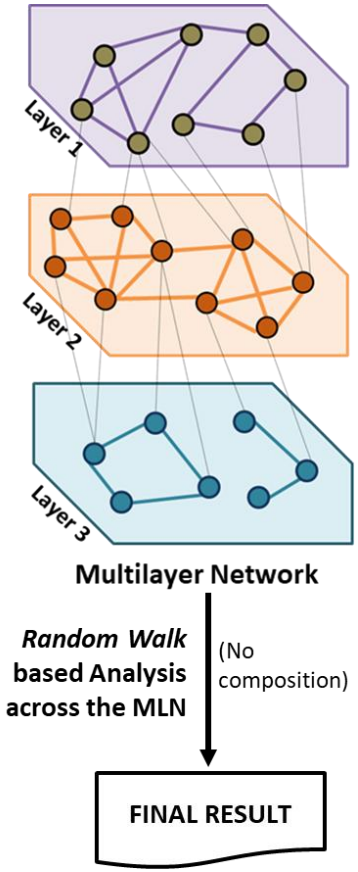
- Loss of information, semantics
- *N* layers ⇒ O (2^N) combined layers !
- Difficult to parallelize and scale



MLNs as a graph

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

- 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**
    - Berenstein, A.J., Magariños, M.P., Chernomoretz, A. and Agüero, F., *PLoS neglected tropical diseases*, **2016**
  - **Type Independent / Aggregation based**
    - Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Del Pozo, F. and Boccaletti, S., *Scientific reports*, **2013**
    - De Domenico, M., Nicosia, V., Arenas, A. and Latora, *CoRR ArXiV*, **2014**
- **MLN as a graph**
  - Sun, Y., Han, J., Yan, X., Yu, P.S. and Wu, T., *Proceedings of the VLDB Endowment*, **2011**
  - Wilson, J.D., Palowitch, J., Bhamidi, S. and Nobel, A.B., *The Journal of Machine Learning Research*, **2017**

Motivation

Complex Data Analysis

Traditional Approaches and Their Limitations

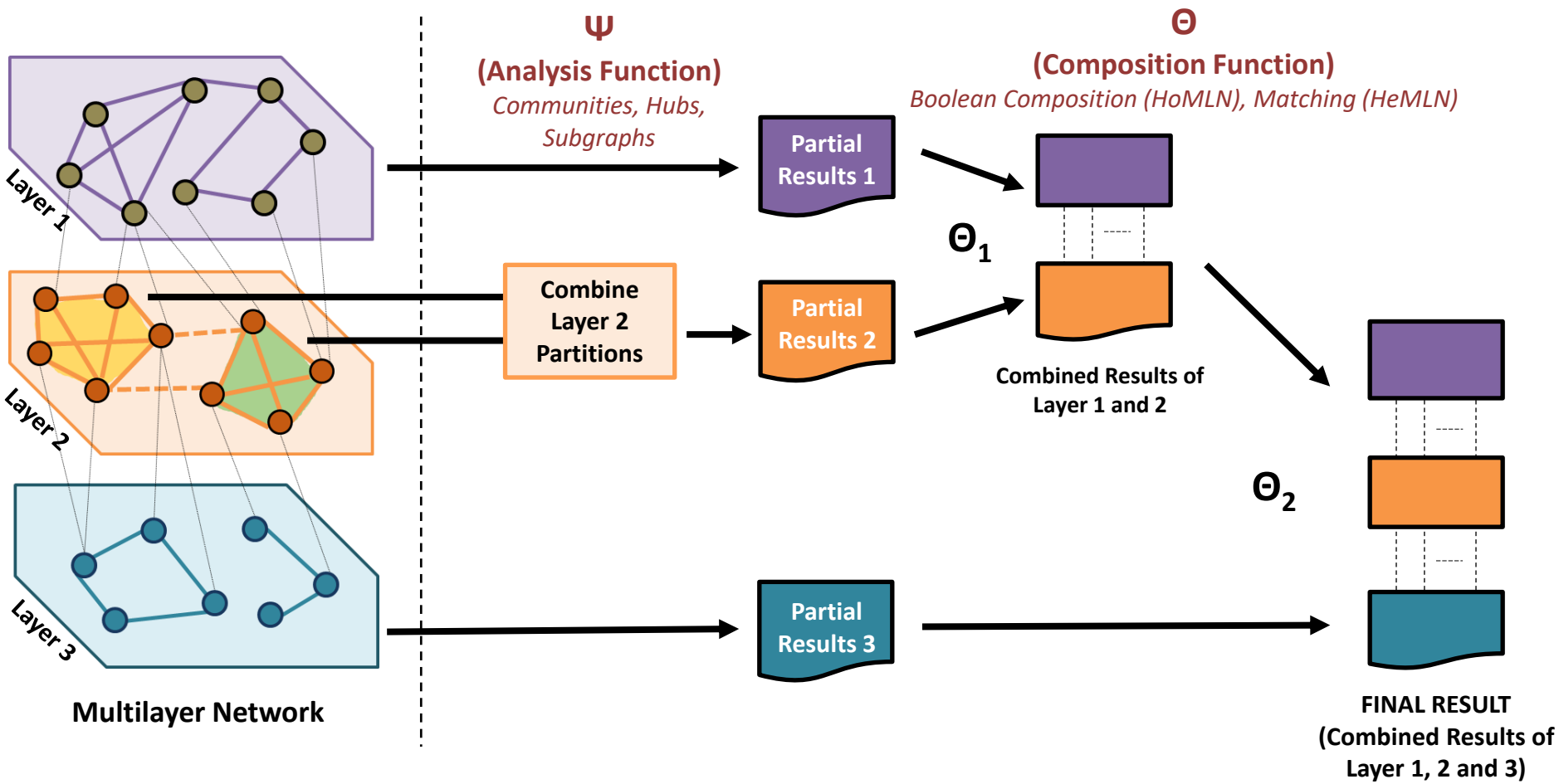
Modeling Using Multilayer Networks

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**



Decoupling Approach Challenges

- Identify **Θ** for various **types of analysis (Ψ)**
  - 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
  1. **[C]** Kanthi Komar, Abhishek Santra, Sanjukta Bhowmick and Sharma Chakravarthy, "Applying ER Approach to Multilayer Network Modeling and Analysis", ER 2020
  2. **[C]** Sharma Chakravarthy, Abhishek Santra, Kanthi Komar, "Why Multilayer Networks Instead of Simple Graphs? Modeling Effectiveness and Analysis Flexibility and Efficiency!", BDA 2019
  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
  5. **[C]** Abhishek Santra and Sanjukta Bhowmick, "Holistic Analysis of Multi-Source, Multi-Feature Data: Modeling and Computation Challenges", BDA 2017
  6. **[C]** Abhishek Santra, Sanjukta Bhowmick and Sharma Chakravarthy, "HUBify: Efficient Estimation of Central Entities across Multiplex Layer Compositions", ICDM Workshops 2017
  7. **[C]** Abhishek Santra, Sanjukta Bhowmick and Sharma Chakravarthy, "Efficient Community Re-creation in Multilayer Networks Using Boolean Operations", ICCS 2017
- Under Review
  8. **[C]** Kunal Samant, Endrit Memeti, Abhishek Santra, Enamul Karim and Sharma Chakravarthy, "CoWiz: Interactive Covid-19 Visualization Based On Multilayer Network Analysis", ICDE 2021
- In preparation
  9. **[J]** Abhishek Santra, Sanjukta Bhowmick, Kamesh Madduri and Sharma Chakravarthy, "Efficient Community Detection in Boolean Composed Multilayer Networks"
  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

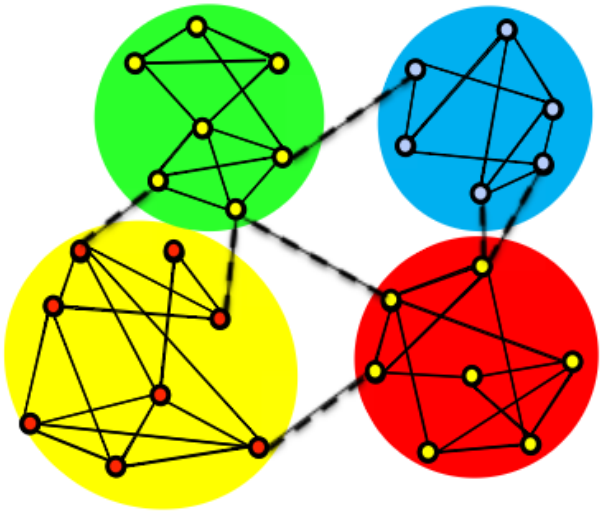
Performance Analysis

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

Case Studies

Communities in Simple Graphs

- **Definition:** **Groups** of related nodes that are **densely inter-connected** 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

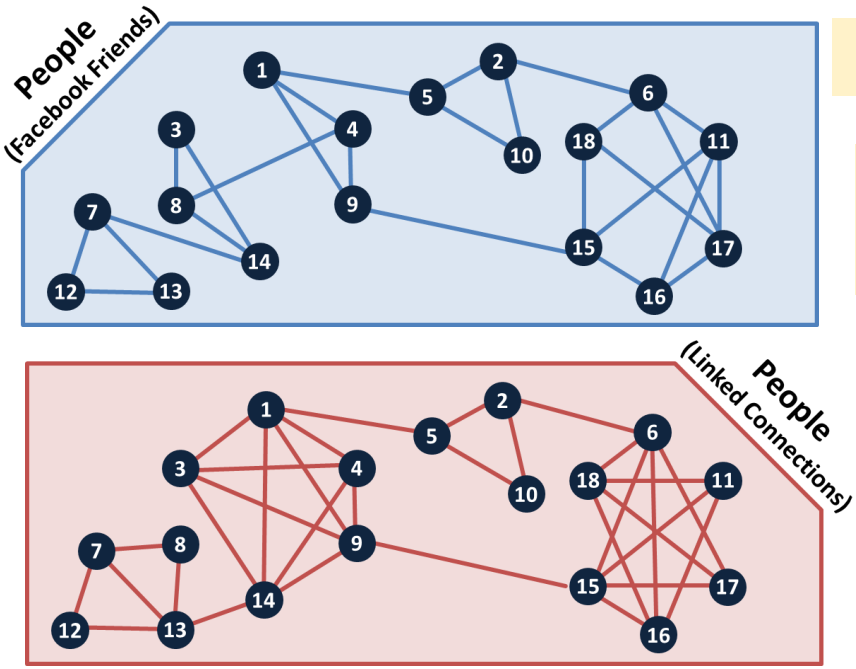
Performance Analysis

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

Case Studies

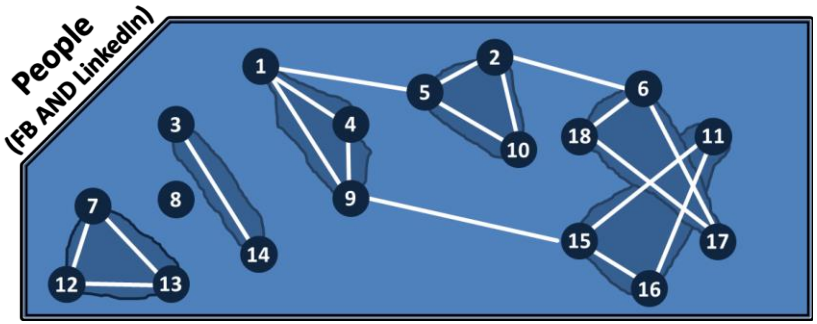
AND Composition using *Single Graph*

- Single Graph Approach for Communities **(C-SG-AND)**
  - **Combine** the required layers using **AND operator**
  - Apply **existing community detection algorithms**
- **Specification: C(G_1 AND G_2 AND … 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?

Combine using **AND**



Communities

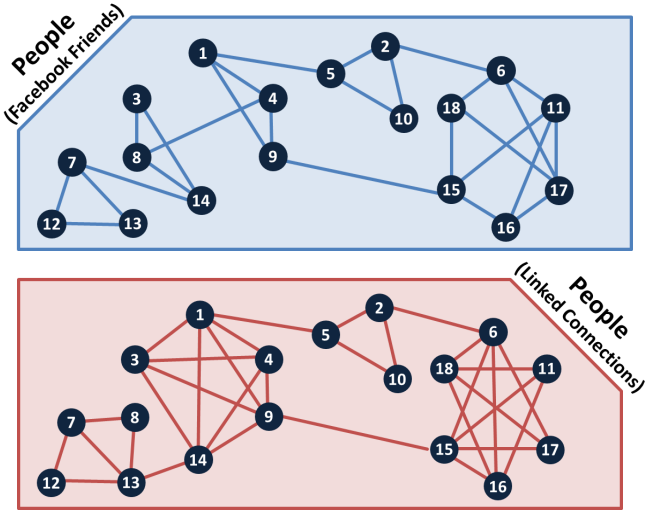
C(FB AND LinkedIn)

AND Composition using *Decoupling Approach*

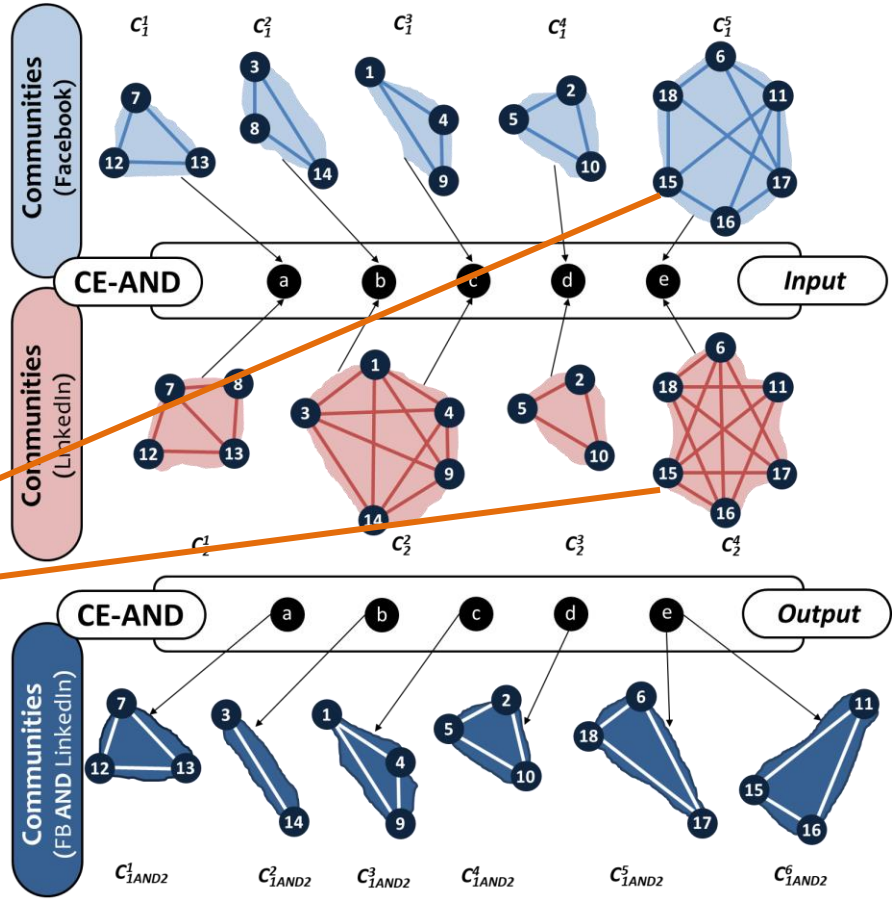
- **Correctness Criterion:** Generate the same communities obtained by ANDing layers into a single graph (**termed C-SG-AND)**
- **Specification: $C(G_1)$ AND $C(G_2)$ AND … AND $C(G_k)$**
  - **$C(G_i)$**: Communities of G_i
- **Approaches for *2 layer Composition***
  - **CV-AND: Node-based intersection** 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*)



Layer-wise
Communities



6 11 15 16 17 18

One Community ✗

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

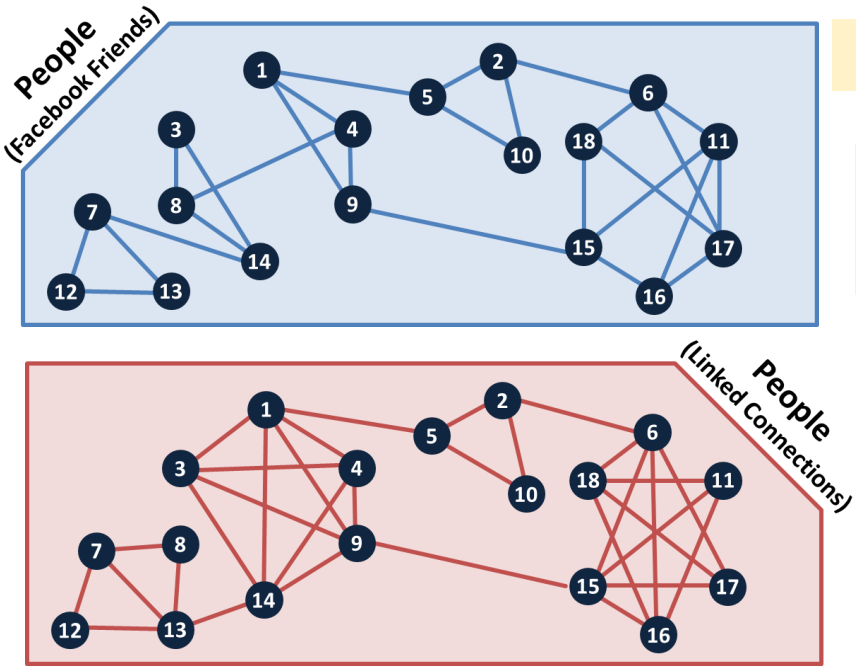
Performance Analysis

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

Case Studies

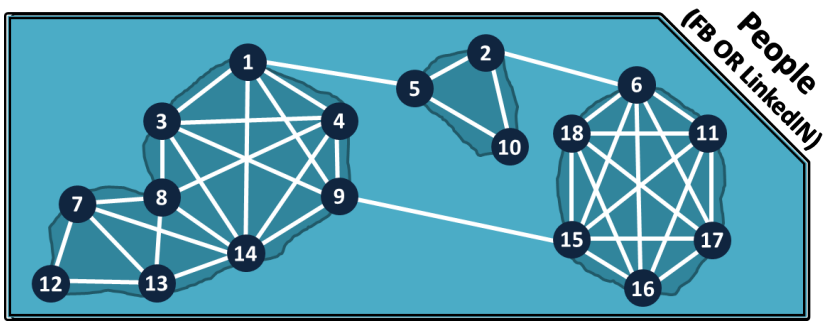
OR Composition using *Single Graph*

- Single Graph Approach for Communities **(C-SG-OR)**
  - **Combine** the required layers using **OR operator**
  - Apply **existing community detection algorithms**
- **Specification: C(G_1 OR G_2 OR … OR G_k)**
  - G_i: Original MLN layer or NOT layer
- Communities used as **Ground Truth** for accuracy calculations



Groups of people connected in at least 1 medium?

Combine using **OR**



Communities

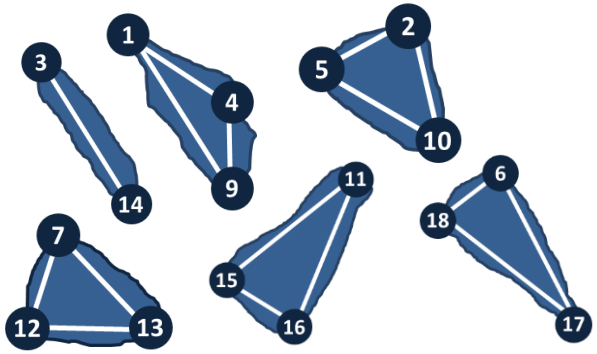
C(FB OR LinkedIn)

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**.

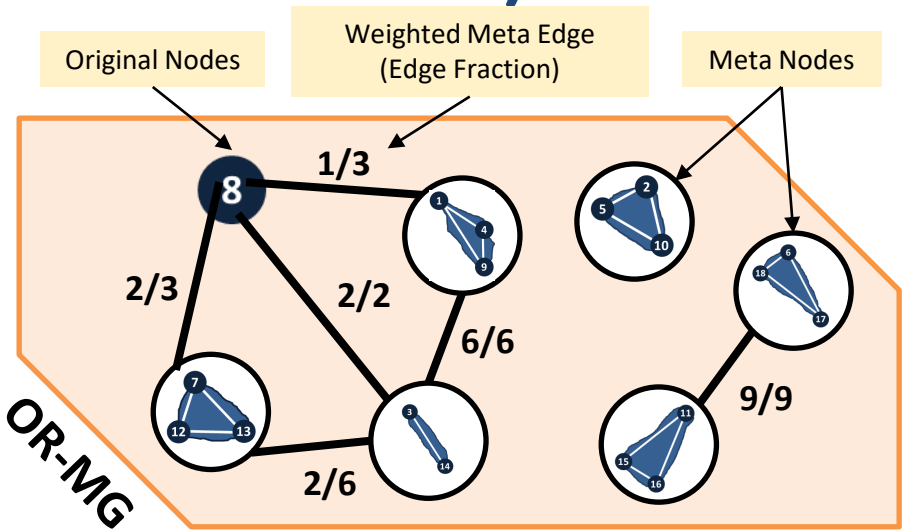
OR Composition using Decoupling Approach

(CE-OR Algorithm - *Illustration*)

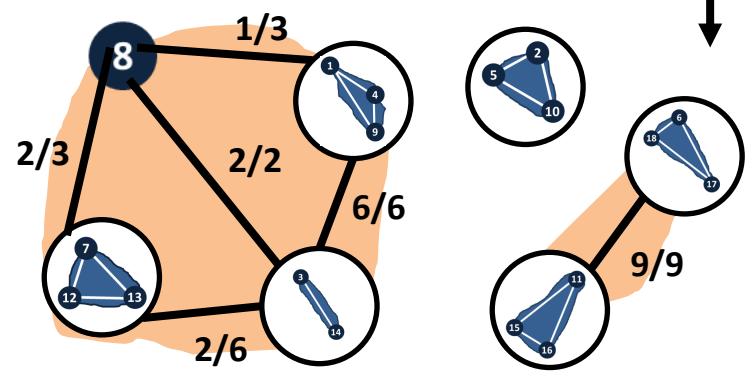


AND Composed Communities (CE-AND)

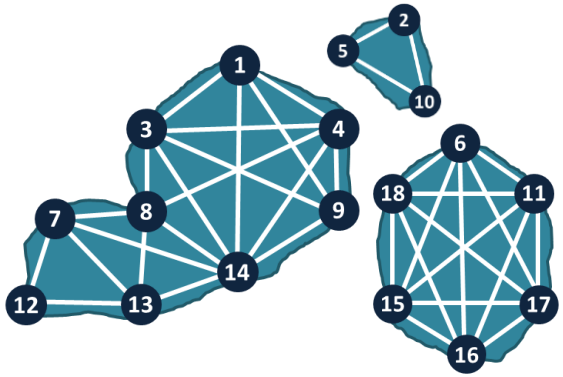
Construct **OR-MG** using union of **intra-community edges**



C(OR-MG)



Expand Communities



OR Composed Communities

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

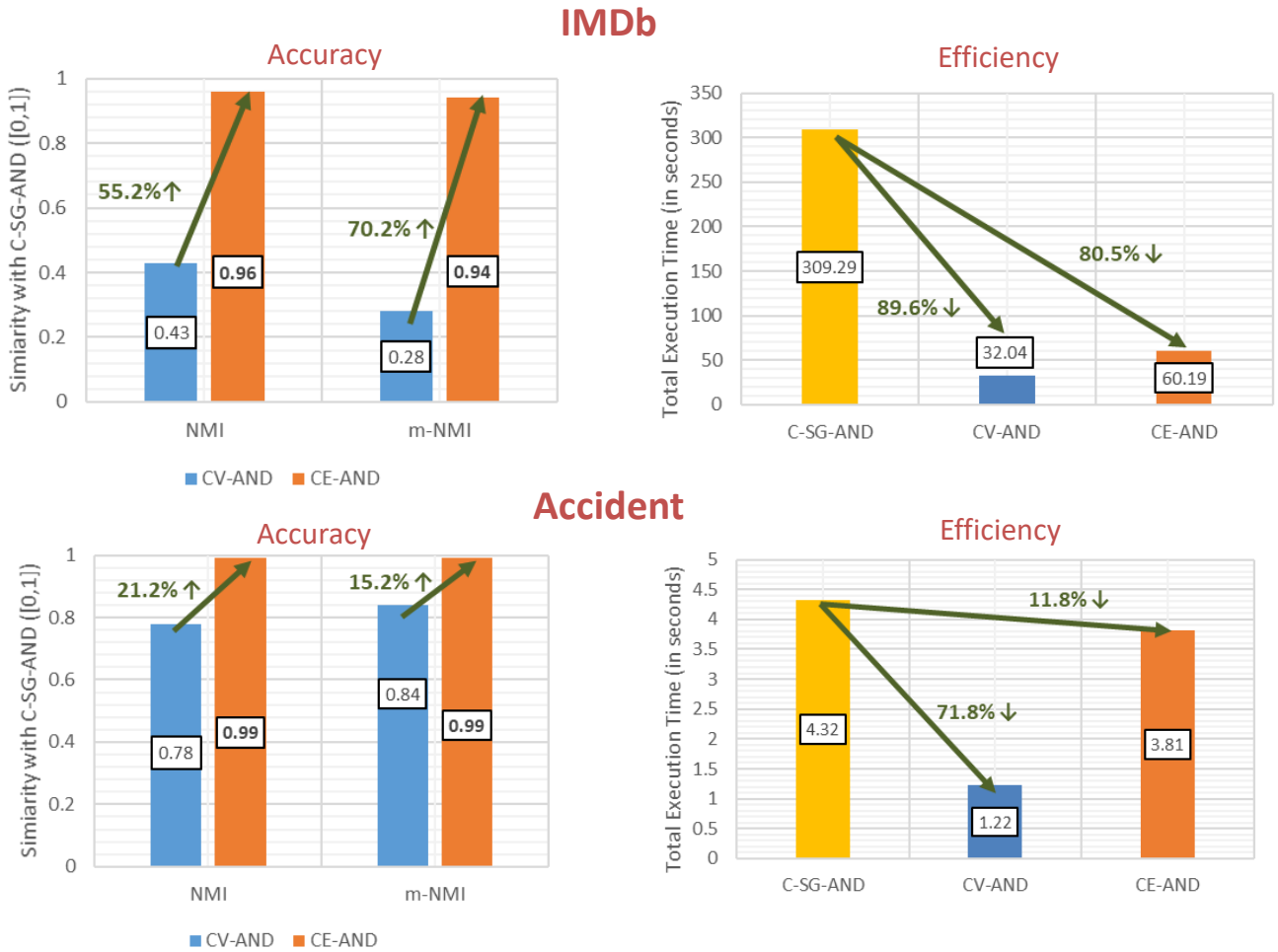
Experimental Set Up

- **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)

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

Accuracy and Efficiency (AND Composition)

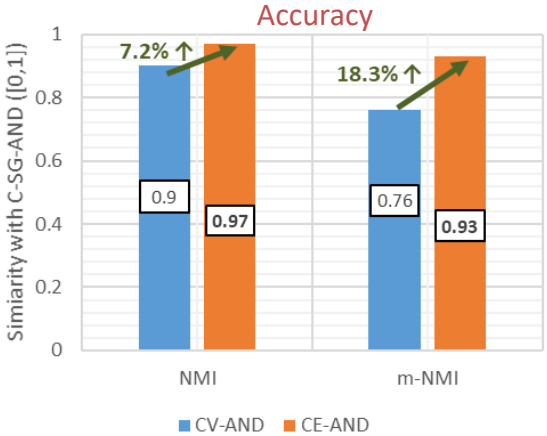


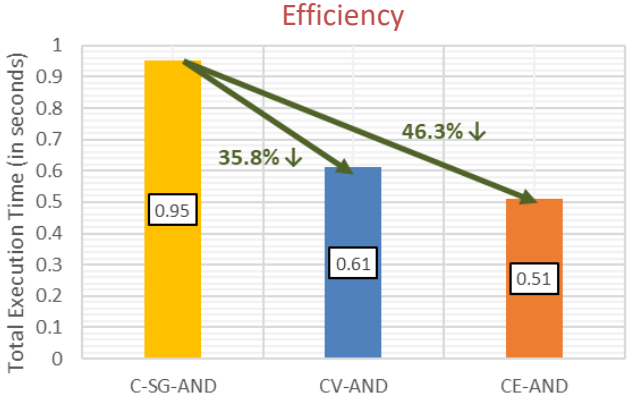
Trade-off: Higher the accuracy, lower is the efficiency

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

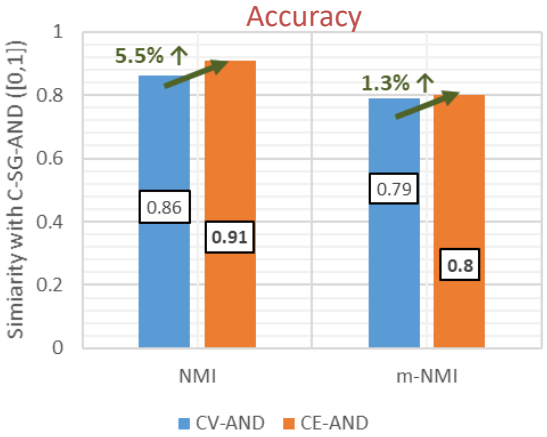
Accuracy and Efficiency (AND Composition)

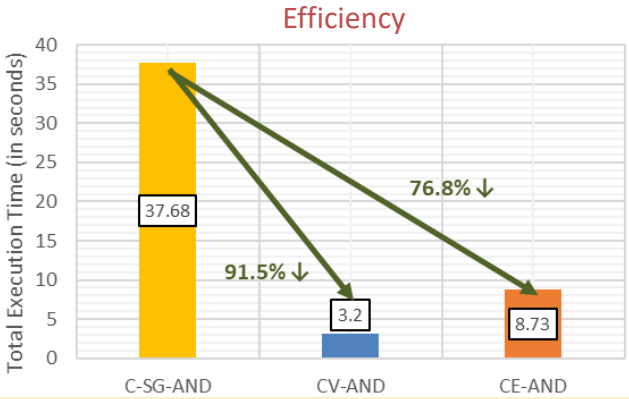
DBLP





R-Mat



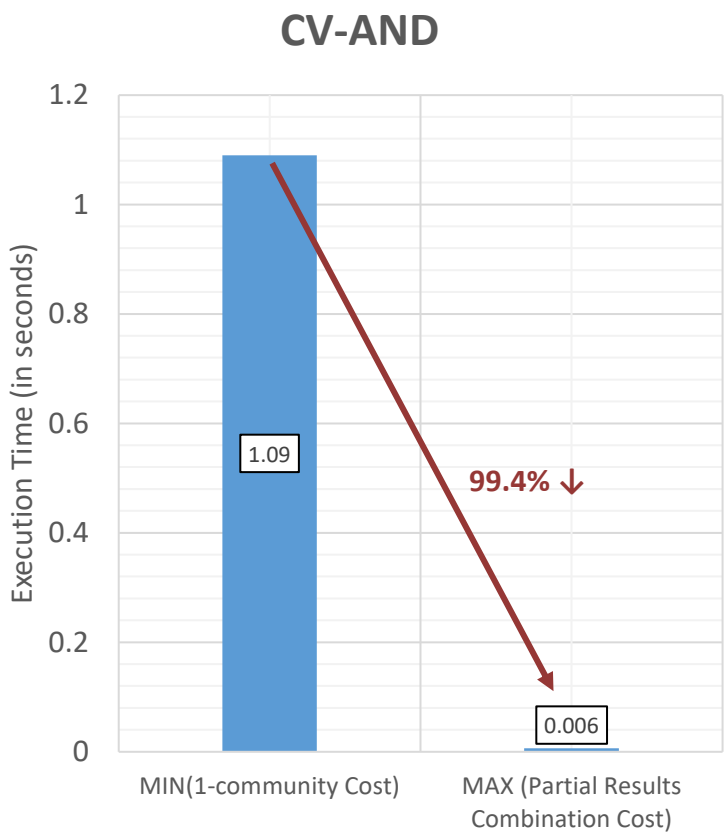


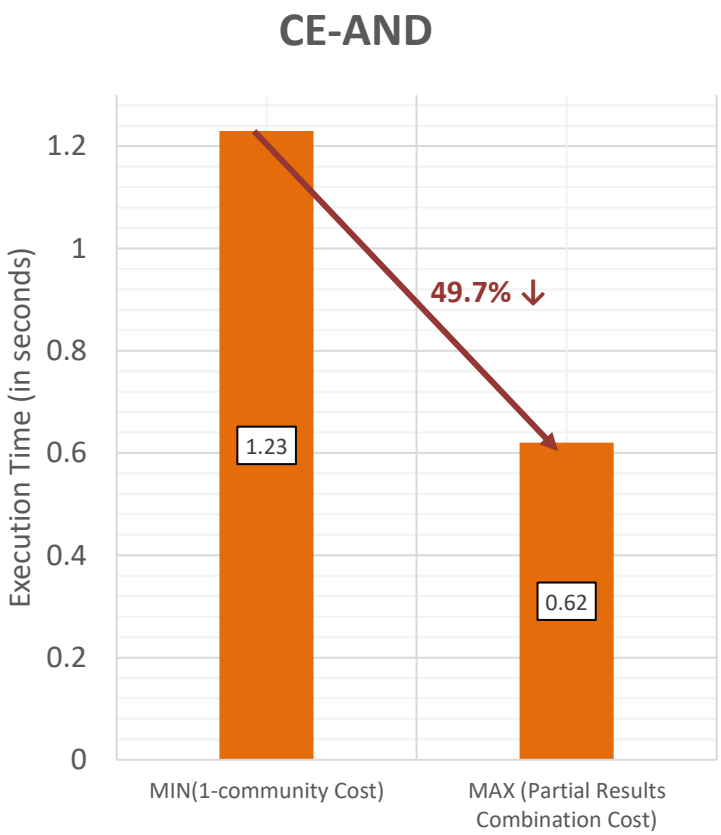
Efficiency improves with more analysis (~$O(2^N)$) for large N

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*)

Component Cost of Decoupling Approaches

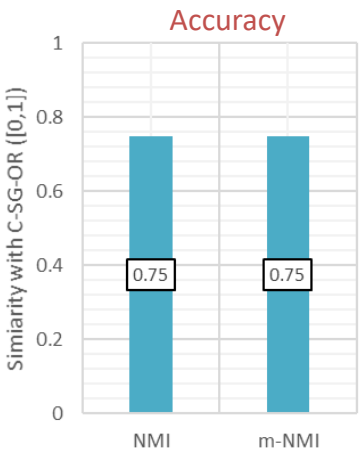


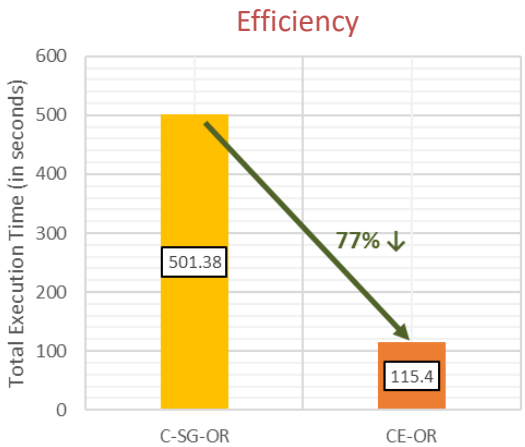


Worst Case Analysis: Maximum cost of combining the partial results is significantly less than the minimum cost to detect 1 layer communities

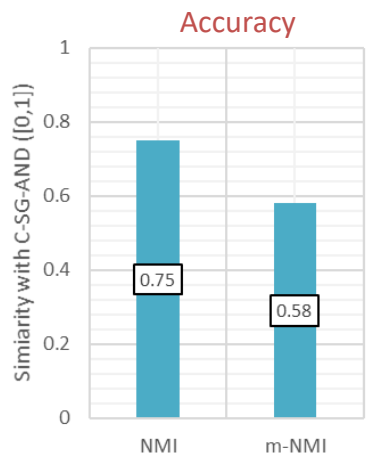
Accuracy and Efficiency (OR Composition)

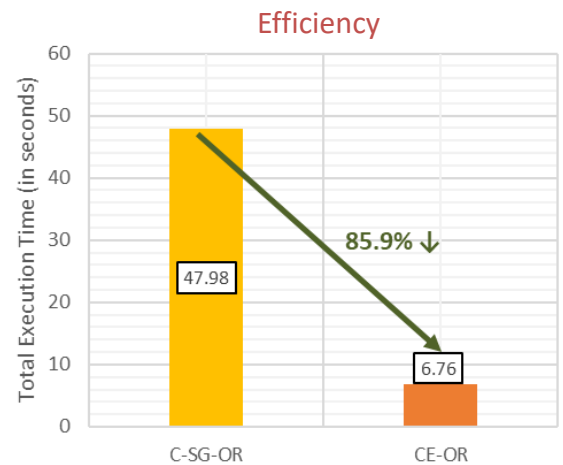
IMDb



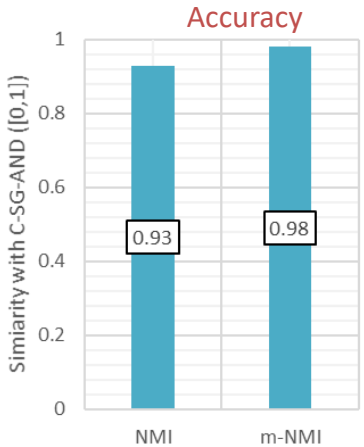


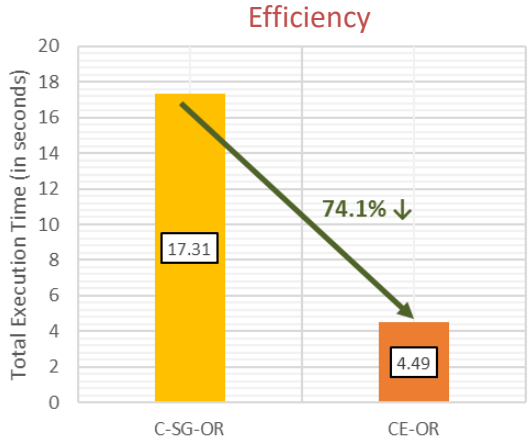
R-Mat

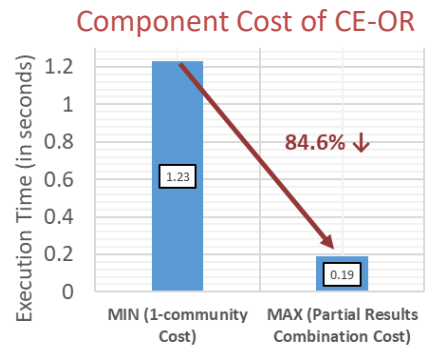




Accident







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

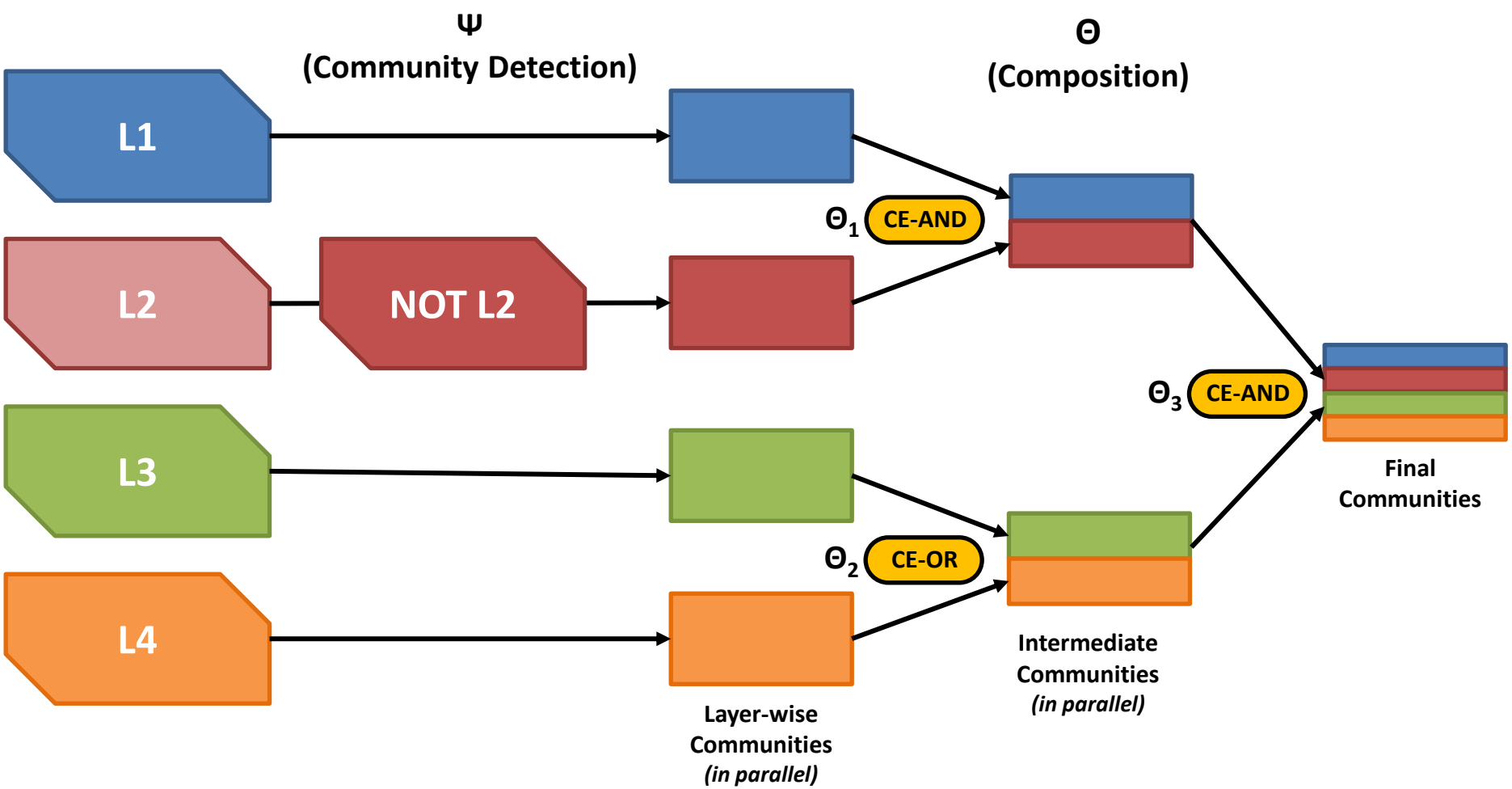
Performance Analysis

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

Case Studies

Applying Decoupling Approach to Boolean Expression

Find Community for: (L1 AND (NOT L2)) AND (L3 OR L4)



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

Number of CE-AND: **2**

Number of CE-AND: **3**

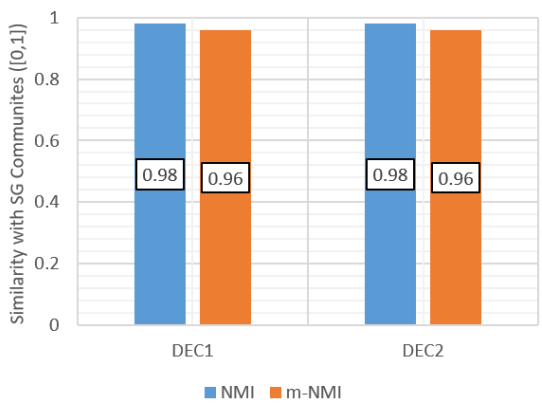
Application of De Morgan's, Distribution Laws

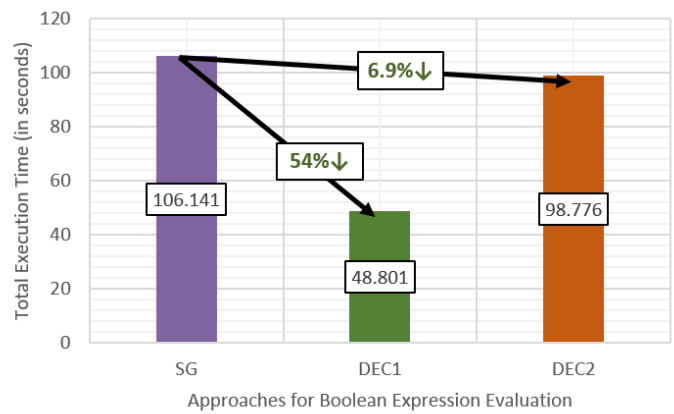
No. of Dense (NOT) layers: **1**

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

Sparse Layers

DEC2 C(VLDB) **CE-AND** C(SIGMOD) **CE-AND** C(NOT DASFAA) **CE-AND** C(NOT DaWaK)





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

	#Nodes	#Edges
Co-Acting	9485	45,581
Genre	9485	996,527
AvgRating	9485	13,945,912

Based on initial set of top 500 actors

DBLP-CoAuthors HoMLN

	#Nodes	#Edges
VLDB	5116	3912
SIGMOD	5116	3303
DASFAA	5116	1519
DaWaK	5116	679

Based on publications from 2003 to 2007

Facebook HoMLN

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

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)

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

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)

NicolasBruno
VivekRNarasayya
ManojSyamala
SurajitChaudhuri
ArunprasadPMarathe
SanjayAgrawal
LuborKollár

TorstenGrust
JensTeubner
PeterABoncz
StefanManegold
JanRittinger
MauriceVanKeulen
SherifSakr

SongtingChen
KSelçukCandan
DivyakantAgrawal
ArsanySawires
Jun'ichiTatemura
OliverPo

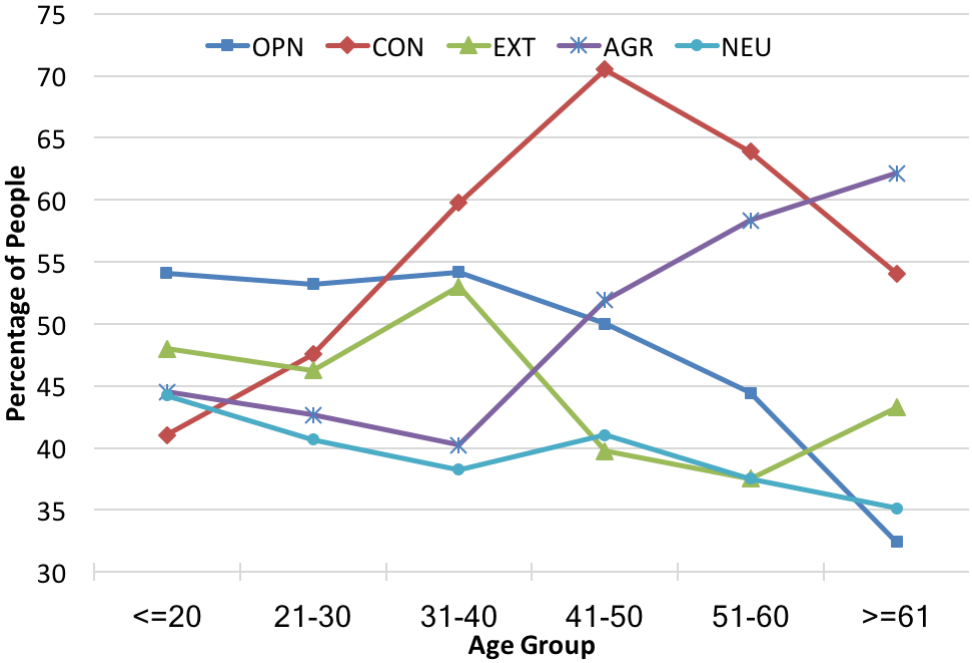
Widely accepted collaboration groups with high quality work

- **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*

Facebook HoMLN

How do the **personality traits evolve with age**?

C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)



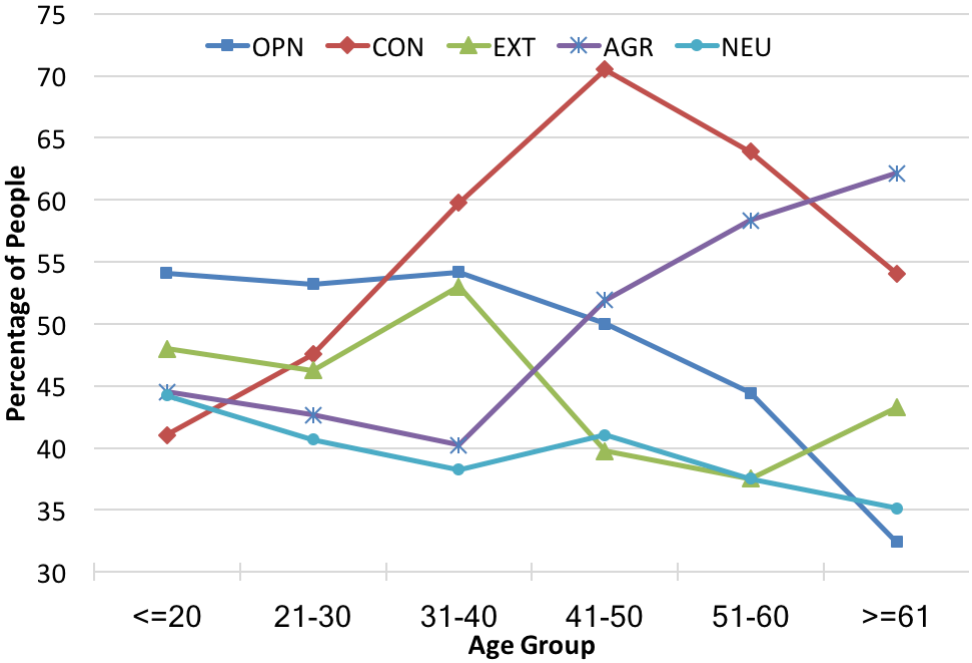
Openness (OPN)

- 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)

Facebook HoMLN

How do the **personality traits evolve with age**?

C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)



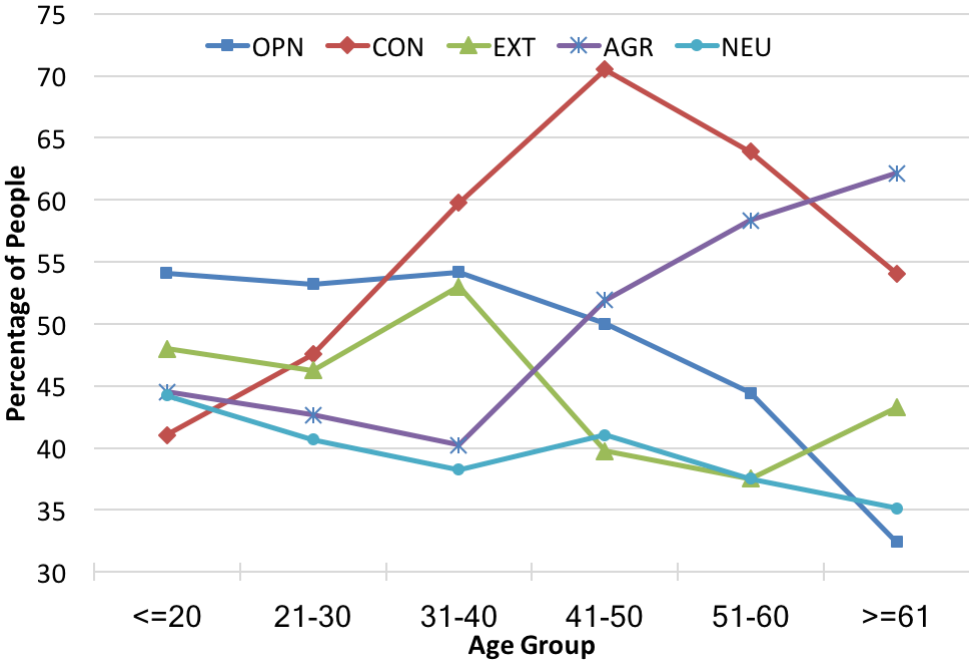
Conscientiousness (CON)

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

Facebook HoMLN

How do the **personality traits evolve with age**?

C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)



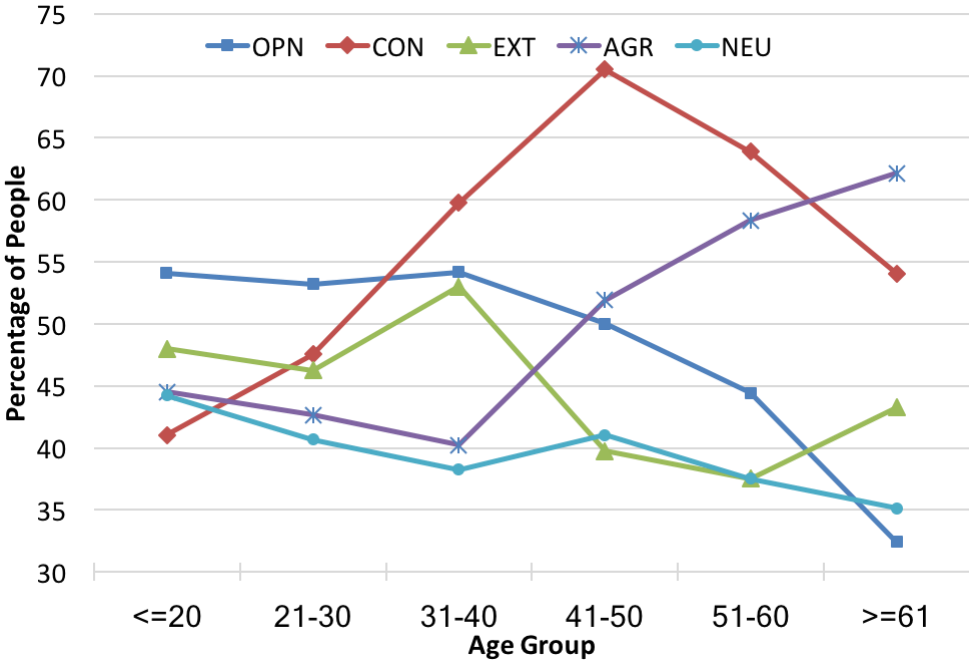
Extraversion (EXT)

- 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

Facebook HoMLN

How do the **personality traits evolve with age**?

C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)



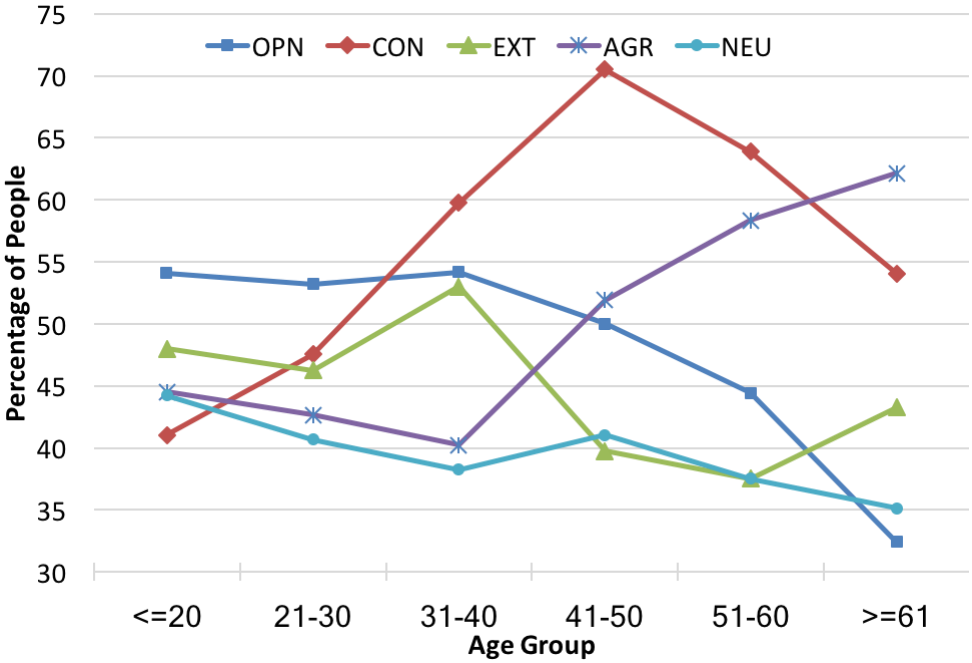
Agreeableness (AGR)

- Reflects a tendency to perceive others in a more positive light
- **Parenthood and grand-parenthood may make the elder generation more empathetic** towards others as compared to the younger lot

Facebook HoMLN

How do the **personality traits evolve with age**?

C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)	C(Age) CE-AND C(OPN)



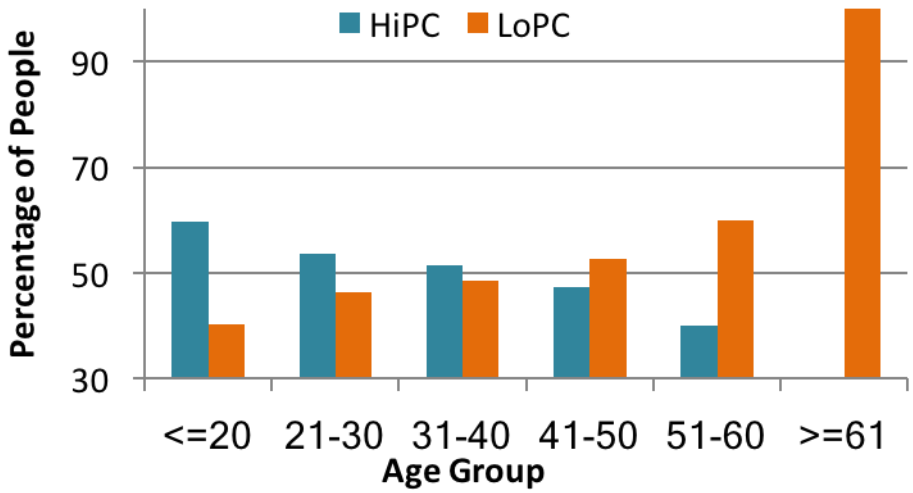
Neuroticism (NEU)

- 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

Facebook HoMLN

How does the **individuals' age correlate** with their **comfort level of sharing personal information on social media**?

C(Age) CE-AND C(Privacy Concern)



- **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

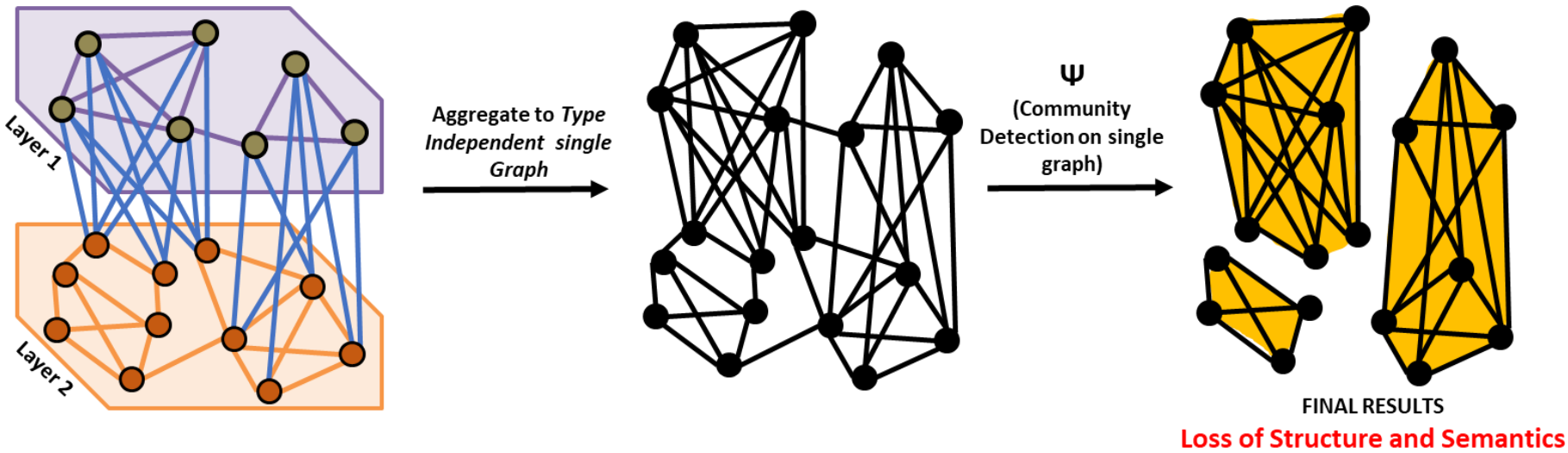
Heterogeneous Community Definition

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**

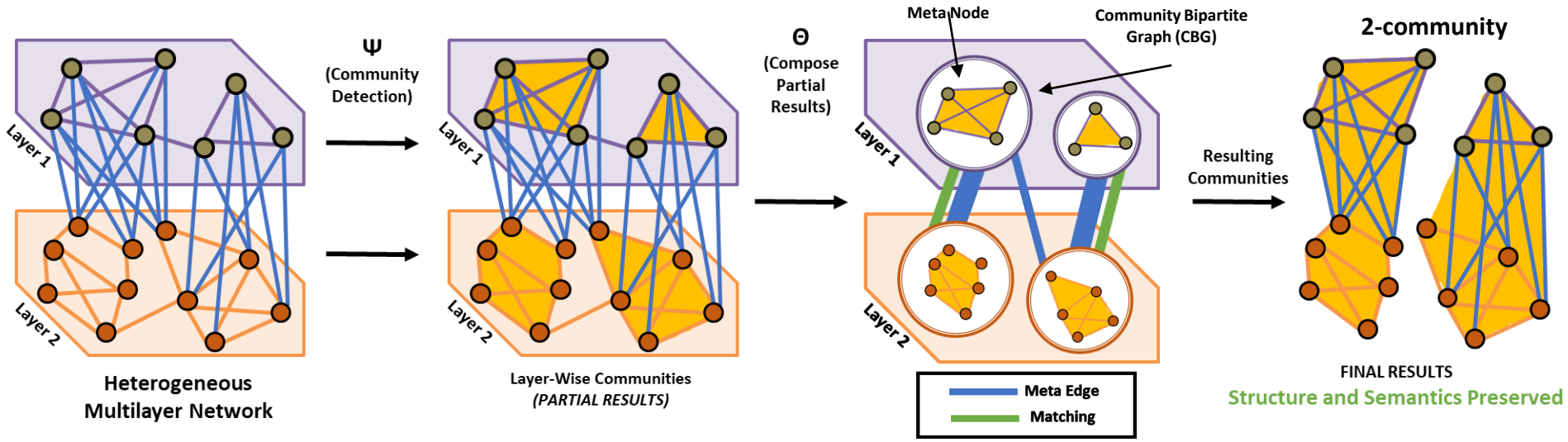


Desirable Properties for HeMLN Communities

- **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**

2-community: The Binary Operator

Community Definition in 2-layer HeMLN



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

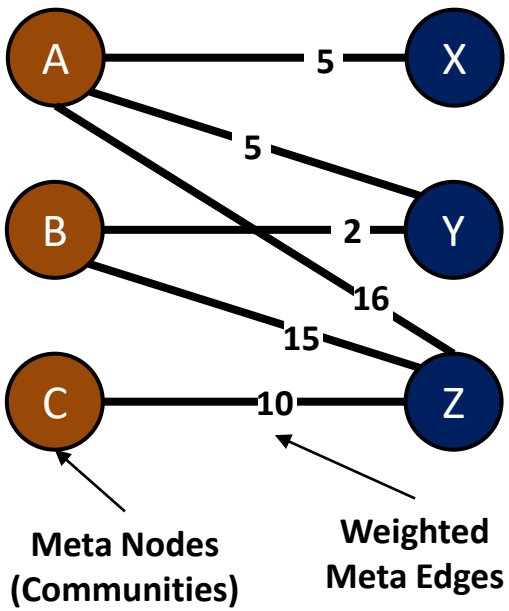
- **Simple nodes** (hiring, dating)
- Weighted Edges supported
- **Unique Matches only**

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



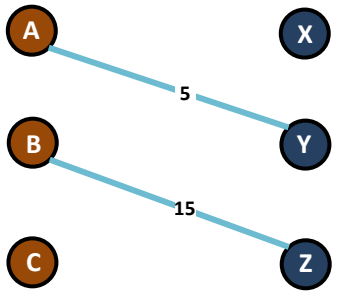
Total Inter-Layer Edge Weight
MWPM <= MWM <= MWRM
MWM <= MWMT

Total Number of Pairs
MWM = MWRM <= MWMT
MWM = MWRM <= MWPM

Traditional Algorithms

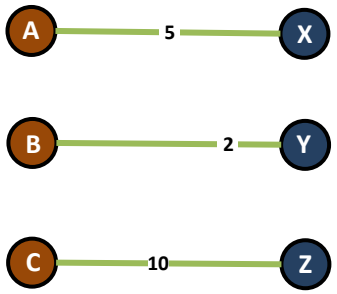
Maximum Weight Match (MWM)

$\sum w$ = **20;** #matches = **2**



Maximum Weight Perfect Match (MWPM)

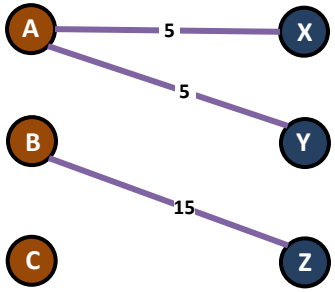
$\sum w$ = **17**; #matches = **3**



Relaxed Algorithms

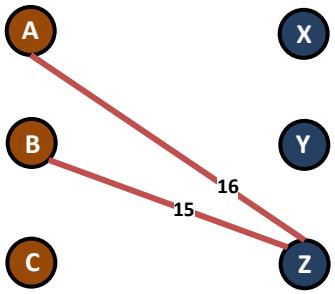
Maximum Weight Match with Ties (MWMT)

$\sum w$ = **25**; #matches = **3**



Maximum Weight with Relaxed Matching (MWRM)

$\sum w$ = **31**; #matches = **2**



k-community Detection Algorithm

1. Specification:

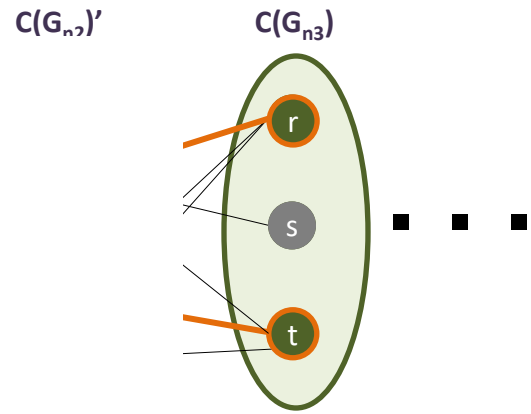
$$C(G_{n1})\ \Theta_{n1,n2}\ C(G_{n2})\ \Theta_{n2,n3} \ldots \Theta_{ni,nk}\ C(G_{nk}),\ \omega,\ MWxx$$

2. ***result*** ← *initialize*(2-community(G_{n1}, G_{n2}), HeMLN, ω, MWxx)

3. left ← *nextLeftSubscript*(Θ), **right** ← *nextRightSubscript*(Θ), k = 2

4. while left ≠ null && right ≠ null

- Construct **CBG for G_{left} and G_{right}**
- **MP** ← *MWxx*(**CBG**)
- For each tuple ∈ ***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** ← *nextLeftSubscript*(Θ) or null
- **right** ← *nextRightSubscript*(Θ) or null



< **a, n, r**; $x_{n1,n2}$, $x_{n2,n3}$ >

<**b, p, t**; $x_{n1,n2}$, , $x_{n2,n3}$ >

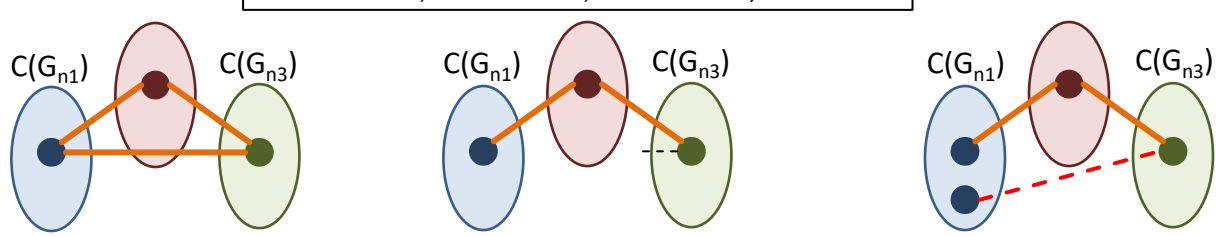
<**c, q, 0**; $x_{n1,n2}$, ϕ >

Cases and Outcomes of MWxx

(in any *recursive step* with CBG for layers G_{left} and G_{right})

(G_{left}, G_{right}) Outcome	Effect on tuple *t*
Case (i): One Processed Layer and One New Layer	
a) Consistent Match	**Extend *t*** with *paired community id* and $x_{i,j}$
b) No Match	**Extend *t*** with 0 *(null id)* and ϕ *(empty set)*
Case (ii): Both are Processed Layers (Cyclic)	
a) Consistent Match	**Update *t*** only with $x_{i,j}$
b) No Match	**Update *t*** with ϕ *(empty set)*
c) Inconsistent Match	**Update *t*** with ϕ *(empty set)*

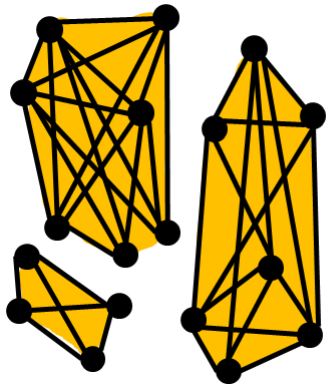
$C(G_{n1})\ \Theta_{n1,n2}\ C(G_{n2})\ \Theta_{n2,n3}\ \mathbf{C(G_{n3})\ \Theta_{n3,n1}\ C(G_{n1})}$



Proposed Weight Metrics for Meta Edge (u, v)

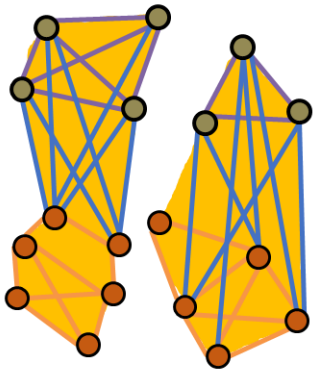
- **Analysis Requirement:** *Maximize number of interactions* between the participating communities
  - **Number of Inter-Community Edges**
    - $\omega_e = number\ of\ interlayer\ edges\ between\ cu\ and\ c_v$
- **Analysis Requirement:** *Strong intra-community and inter-community interaction*
  - **Density and Edge Fraction**
    - $\omega_d = (c_u\ density) * \frac{\omega_e}{|c_u\ nodes| * |c_v\ nodes|} * (c_v\ density)$
- **Analysis Requirement:** *Participation of influential nodes within and between* participating communities
  - **Hub Participation**
    - $\omega_h = (c_u\ ratio\ of\ hubs\ participating) * \frac{\omega_e}{|c_u\ nodes| * |c_v\ nodes|} * (c_v\ ratio\ of\ hubs\ participating)$

Evaluation of Proposed Community Definition (Modularity Comparison)



Type Independent Communities
(Intermixing of Node Types)

v/s



Proposed HeMLN Communities
(Node Types kept Separate)

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

DBLP HeMLN (Author Θ Paper, w_e)

Type-Independent	MWM	MWMT	MWPM	MWRM
0.69	0.69	0.69	0.69	0.69

IMDb HeMLN (Actor Θ Director, w_e)

Type-Independent	MWM	MWMT	MWPM	MWRM
0.78	0.643	0.643	0.698	0.603
	83	220	95	83

Effect of MWxx

Variation in Number of Matches and Overall Weight

Specification	MWM	MWMT	MWPM	MWRM
P-Au, w_d	*Total:* ***6*** ∑w: 0.04988	Total: 9 ∑w: **0.0778**	Total: 6 ∑w: 0.04988	*Total:* ***6*** ∑w: 0.05003
DBLP HeMLN				
P-Au-Y, w_e	*Total:* ***6*** Partial: 0 ∑w: 548	Total: 6 Partial: 0 ∑w: 548	Total: 6 Partial: 0 ∑w: 548	*Total:* ***6*** Partial: 3 ∑w: **986**
A-D, w_e	*Total:* ***50*** ∑w: 9902	Total: 69 ∑w: 9970	Total: 57 ∑w: 5144	*Total:* ***50*** ∑w: **11640**
IMDb HeMLN				
M-A-D-M, w_e	*Total:* ***2*** Partial: 7 $\sum w_{M\text{-}A}$: 6979	Total: 3 Partial: 12 $\sum w_{M\text{-}A}$: 6984	Total: 0 Partial: 9 $\sum w_{M\text{-}A}$: 6979	*Total:* ***2*** Partial: 11 $\sum w_{M\text{-}A}$: **11557**

Total Inter-Layer Edge Weight
MWPM <= MWM <= MWRM
MWM <= MWMT

Total Number of Pairs
MWM = MWRM <= MWMT
MWM = MWRM <= MWPM

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 (**ω_d, ω_h**)
- **Cost of combining intermediate results**
  - Weight Computation Cost + CBG Creation Cost + MWxx Cost
    - **|Meta Edges| << |Inter-layer Edges|**
    - **|Meta Nodes or Communities| << |Layer Nodes|**
- **Cost of combining intermediate results (cost of each iteration) << 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

	#Nodes	#Edges	#Communities	Avg. Comm. Size
Actors (Genre-linked)	9485	996,527	63/190 (Size > 1/All)	148.5
Directors (Genre-linked)	4510	250,845	61/190 (Size > 0/All)	73
Movies (Rating-linked)	7951	8,777,618	9/9 (Size > 0/All)	883.4

Based on initial set of top 500 actors

DBLP HeMLN

	#Nodes	#Edges	#Communities	Avg. Comm. Size
Authors (3 Papers Co-authored)	16,918	2,483	591/15528 (Size > 0/All)	3.3
Papers (Conference-linked)	10,326	12,044,080	6/6 (Size > 0/All)	1721
Years (Range-linked)	18	18	6/6 (Size > 0/All)	3

Based on publications in VLDB, SIGMOD, ICDM, KDD, DASFAA, DaWaK from 2001 to 2018

Analysis Set

DBLP HeMLN

Conference-wise which are the **most cohesive group(s)** of authors who **publish frequently (ties included)**?

P $\Theta_{P,Au}$ Au; w_d; Θ = MWMT

For the **most popular <u>unique</u>** collaborators from each conference, which are the **<u>unique</u> most active** 3-year period(s)?

P $\Theta_{P,Au}$ Au $\Theta_{au,Y}$ Y; w_e; Θ = MWM

IMDb HeMLN

Find the actor and director similar-genre based group pairs such that **overall actor-director collaborations are maximized**?

A $\Theta_{A,D}$ D; w_e; Θ = MWRM

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

A $\Theta_{A,D}$ D; w_h; Θ = MWPM

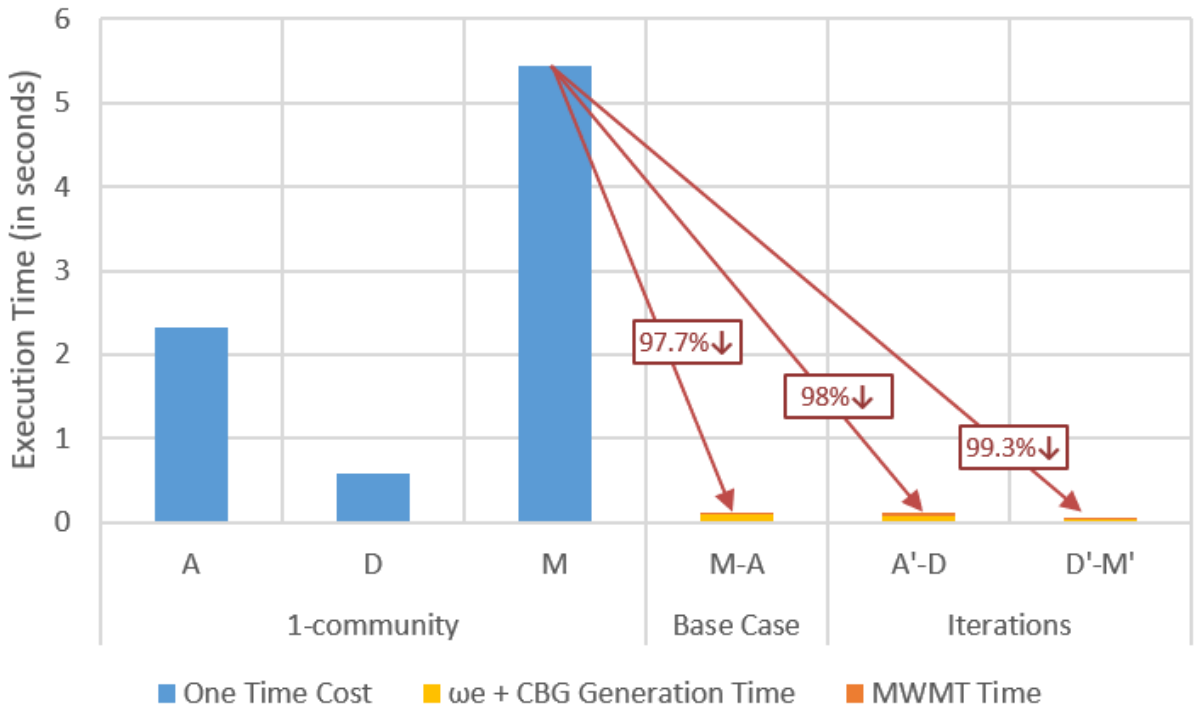
For the **most popular** actor groups, from each movie rating class, find the director groups with **maximum interaction** and who also make movies with similar ratings (**including ties**).

M $\Theta_{M,A}$ A $\Theta_{A,D}$ D $\Theta_{D,M}$ M; w_e; Θ = 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)

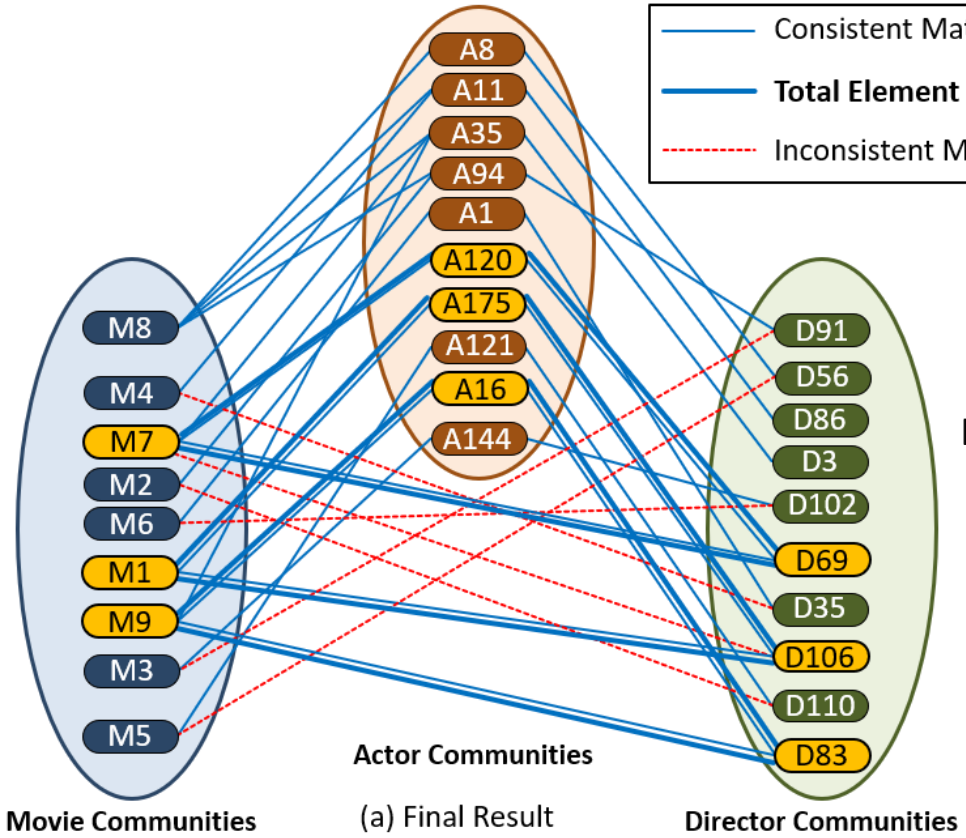


$M\ \Theta_{M,A}\ A\ \Theta_{A,D}\ D\ \Theta_{D,M}\ M$, ω_e, Θ = MWMT

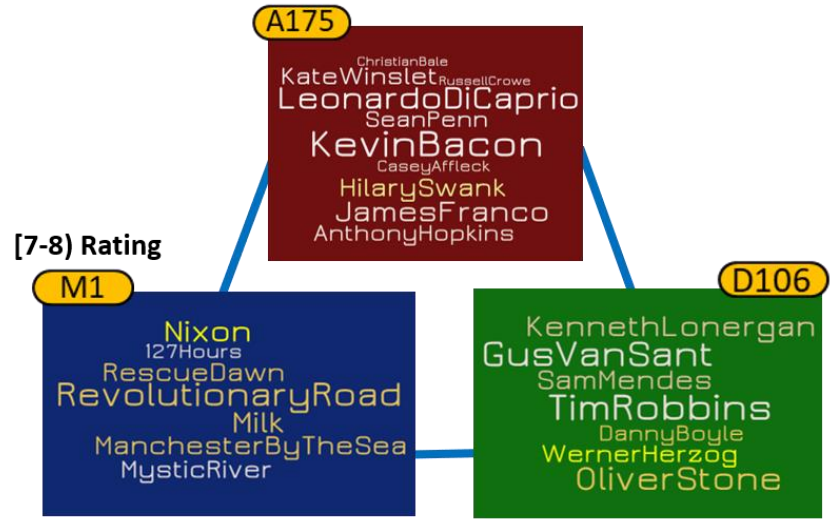
Potential Actor-Director Collaborations (IMDb)

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**)

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



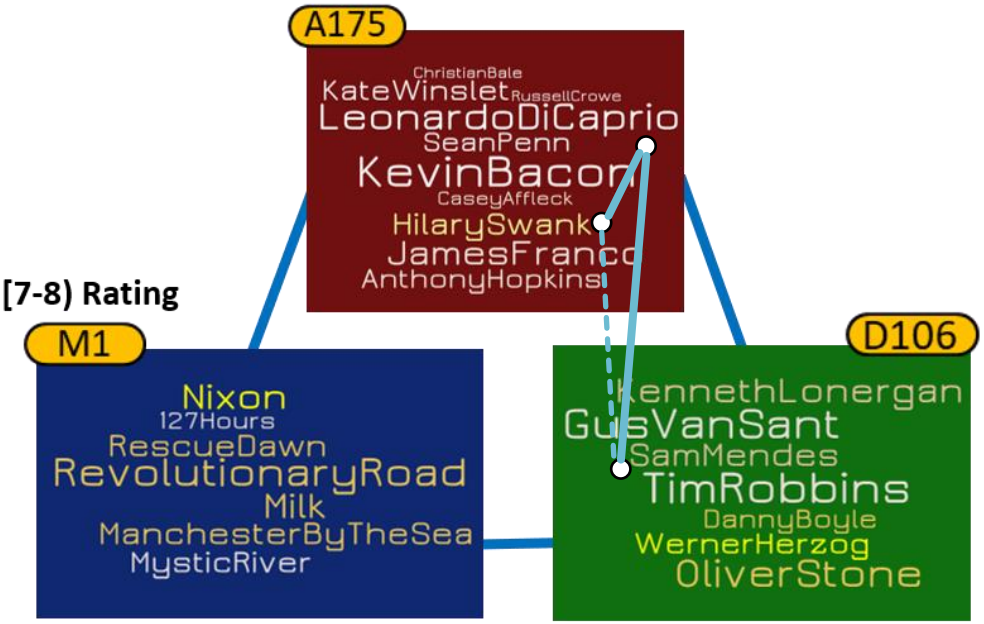
(a) Final Result



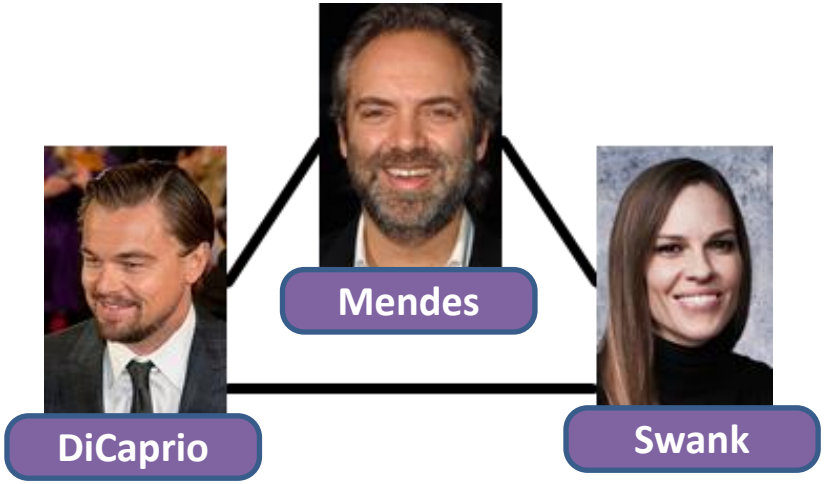
(b) Sample *Movies, Actors and Directors* from a Total Element

Potential Actor-Director Collaborations (IMDb)

IMDb HeMLN Community Results: Similar Expertise, Highly Rated



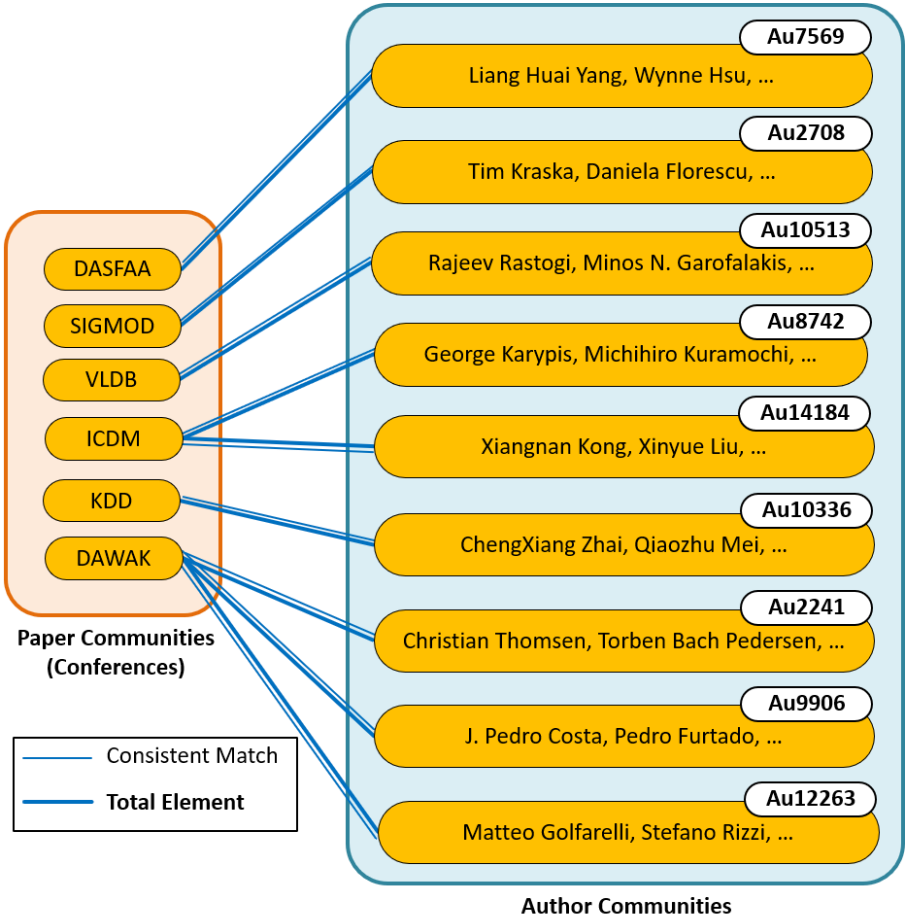
(b) Sample *Movies, Actors and Directors* from a Total Element



Frequently Publishing Groups (DBLP)

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

P $\Theta_{P,Au}$ Au; w_d; Θ = MWMT

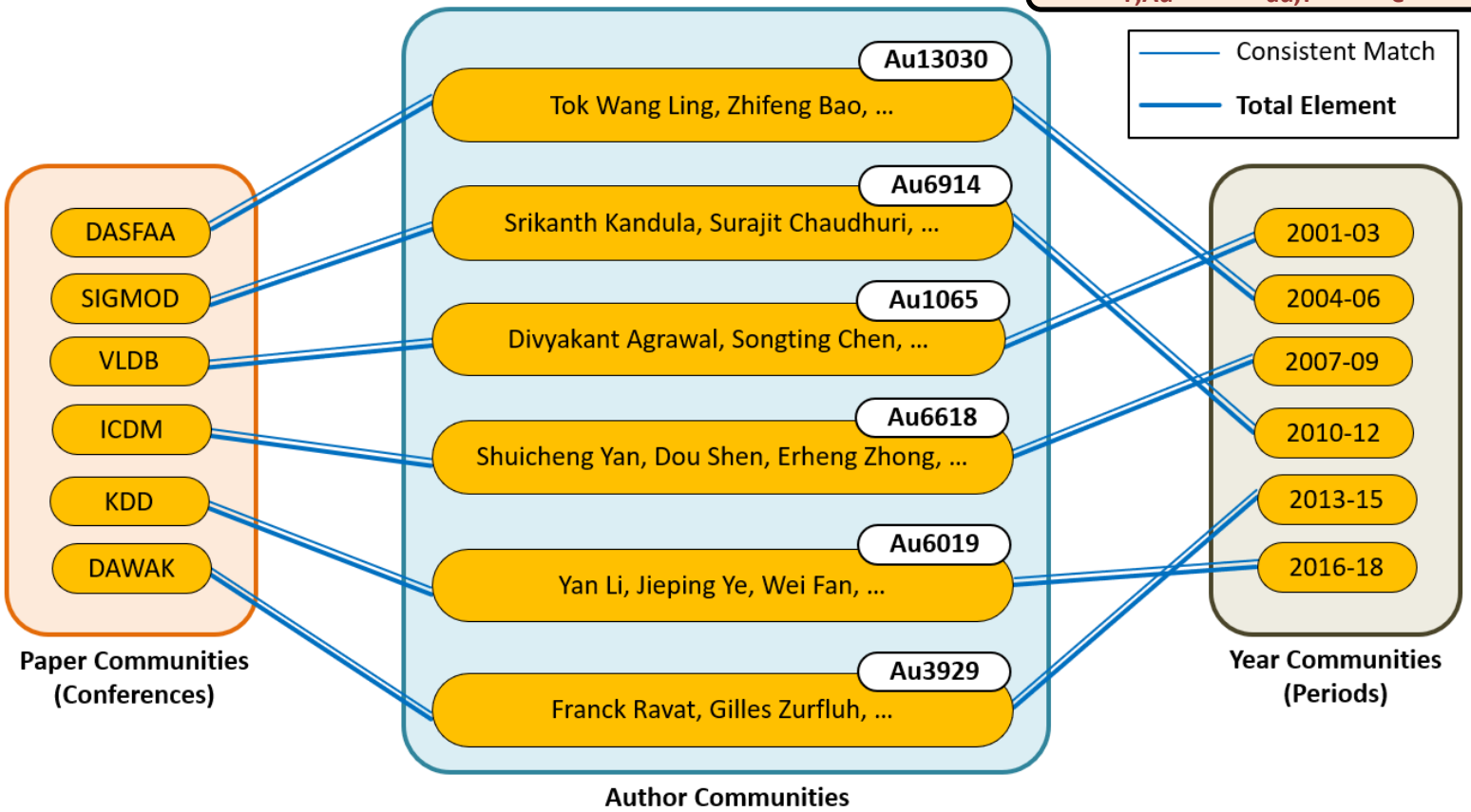


- **ICDM** and **DaWaK** have *multiple author communities* that are equally important
- **Quality Validating Facts:**
  - **SIGMOD: Tim Kraska** has been a recipient of **Best of SIGMOD Award (2008, 2016)**
  - **VLDB: Rajeev Rastogi's** published papers in VLDB (in past 18 years) have received over **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)**

Active Periods for Popular Collaborators (DBLP)

For the **most popular** **unique** author groups from each conference, which are the **unique** **most active** 3-year period(s)?

$P\ \Theta_{P,Au}\ Au\ \Theta_{au,Y}\ Y;\ w_e;\ \Theta = MWM$

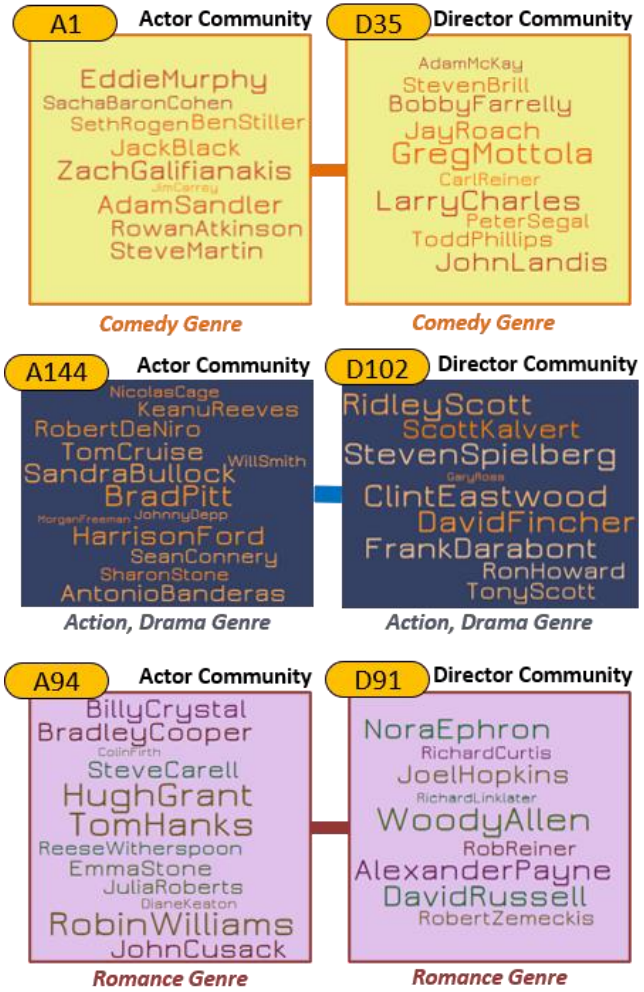


For **SIGMOD, VLDB and ICDM** the **most popular researchers** include **Srikanth Kandula (15188 citations)**, **Divyakant Agrawal (23727 citations)** and **Shuicheng Yan (52294 citations)**, respectively have been active in different periods in the past 18 years

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**

$A\ \Theta_{A,D}\ D;\ w_h;\ \Theta = MWPM$



- 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

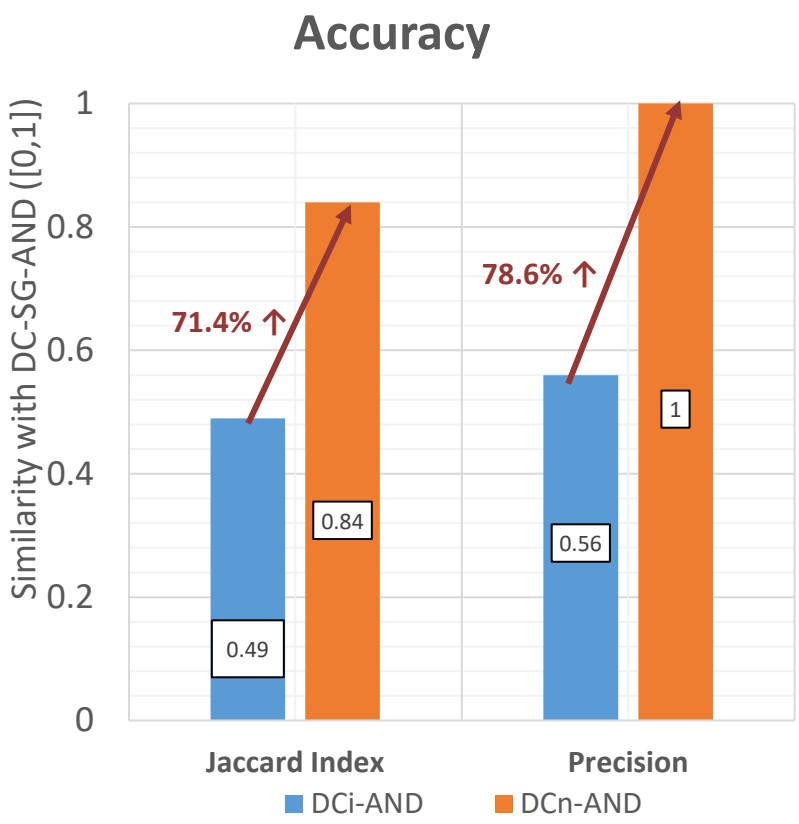
Degree Centrality Heuristics

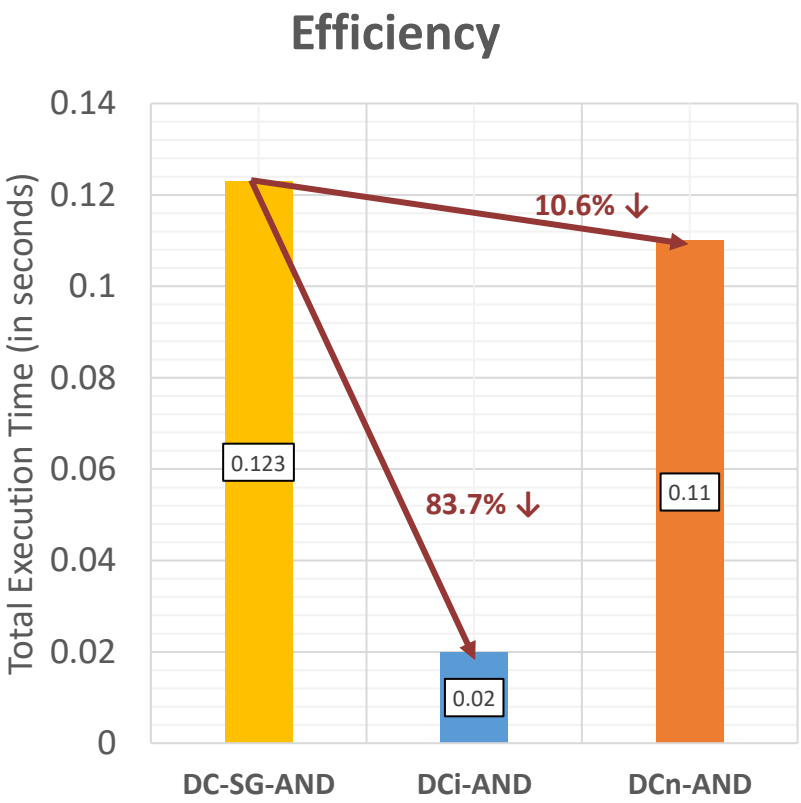
- **DCi-AND: Intersect** the layer-wise hubs
  - **Layers have similar topology:** High Accuracy, Low Overhead
  - **In general, low accuracy** due to presence of *false positives and negatives*
- **DCn-AND:** Check if the **common hubs** have **enough shared neighbors**
  - **Additional Overhead**
    - 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





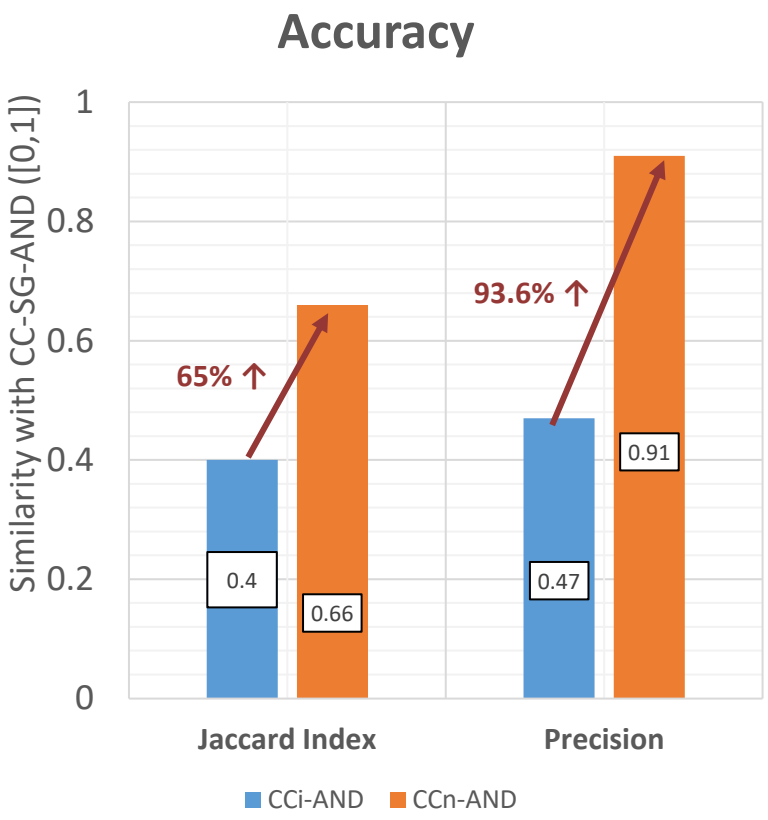
Elimination of False Positive increases the **Precision, Decreases Efficiency**

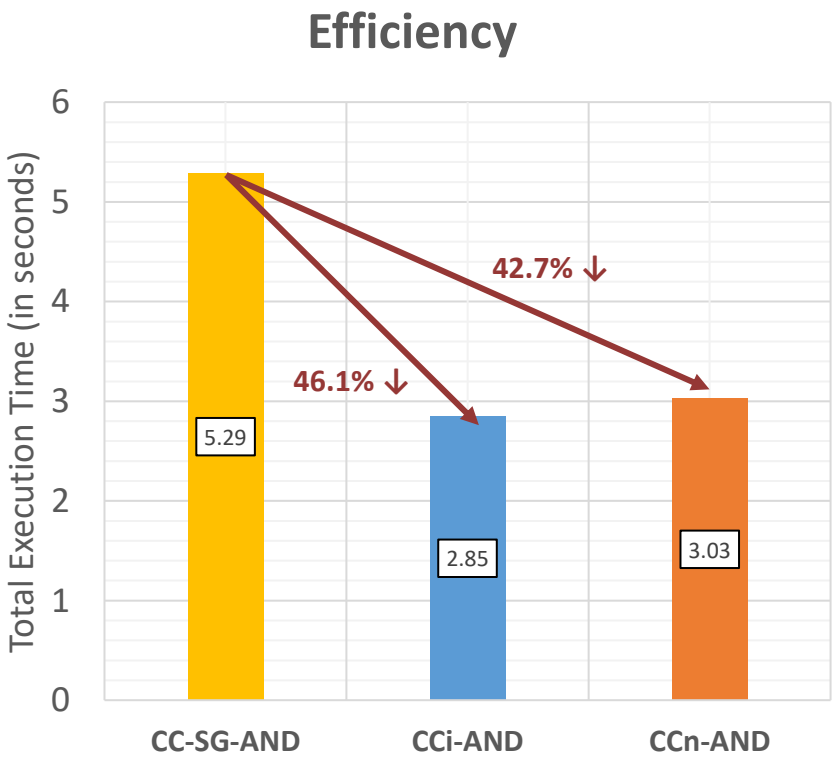
For **large N** (number of MLN layers), **denser layers, more analysis: Efficiency is higher**

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

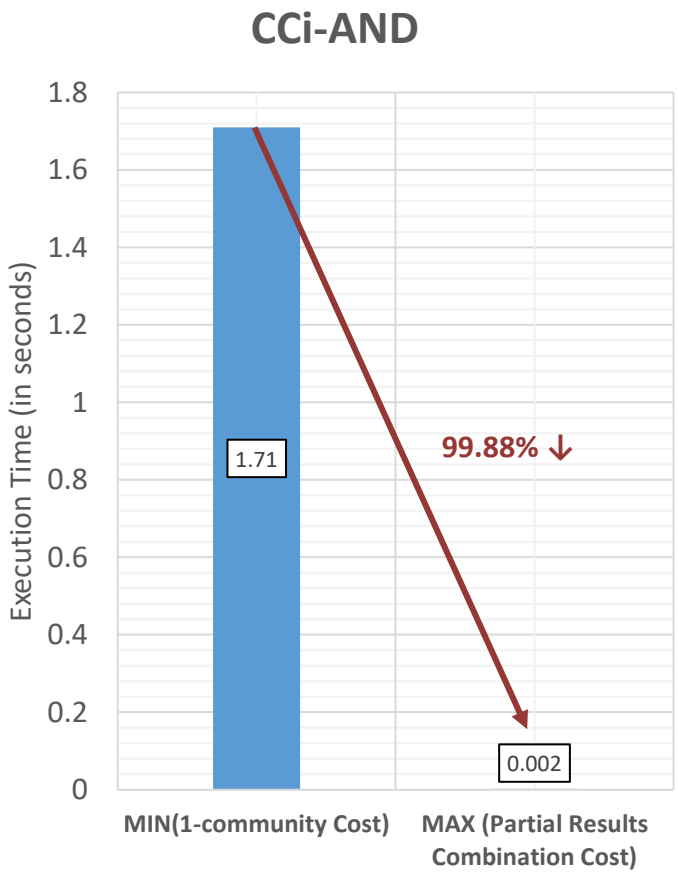


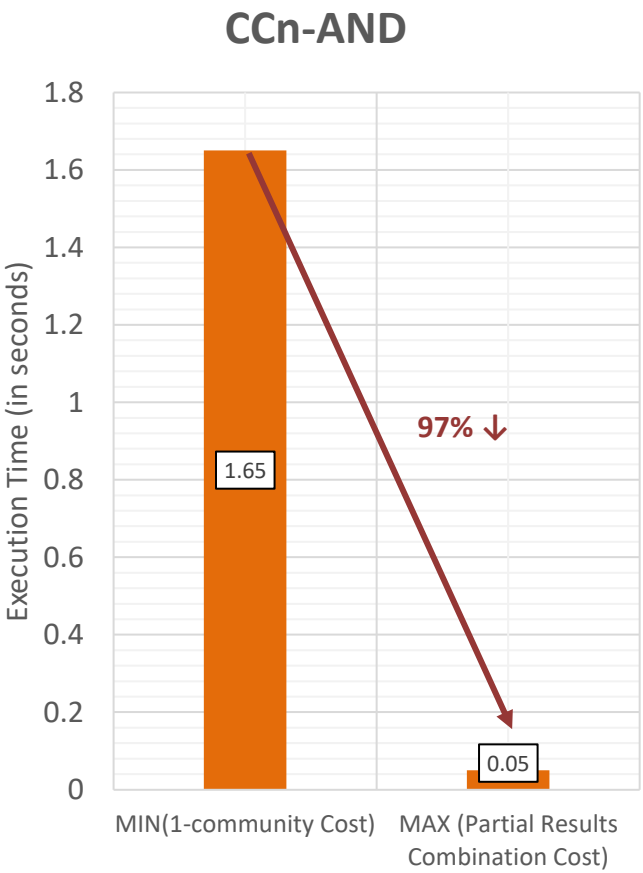


Decrease in False Positives increases the **Precision, Decreases Efficiency**

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

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))

	#Nodes	#Edges
American	290	746
Southwest	290	717
Delta	290	688
Frontier	290	346
Spirit	290	189
Allegiant	290	379

Based on direct flights active in **February 2018**

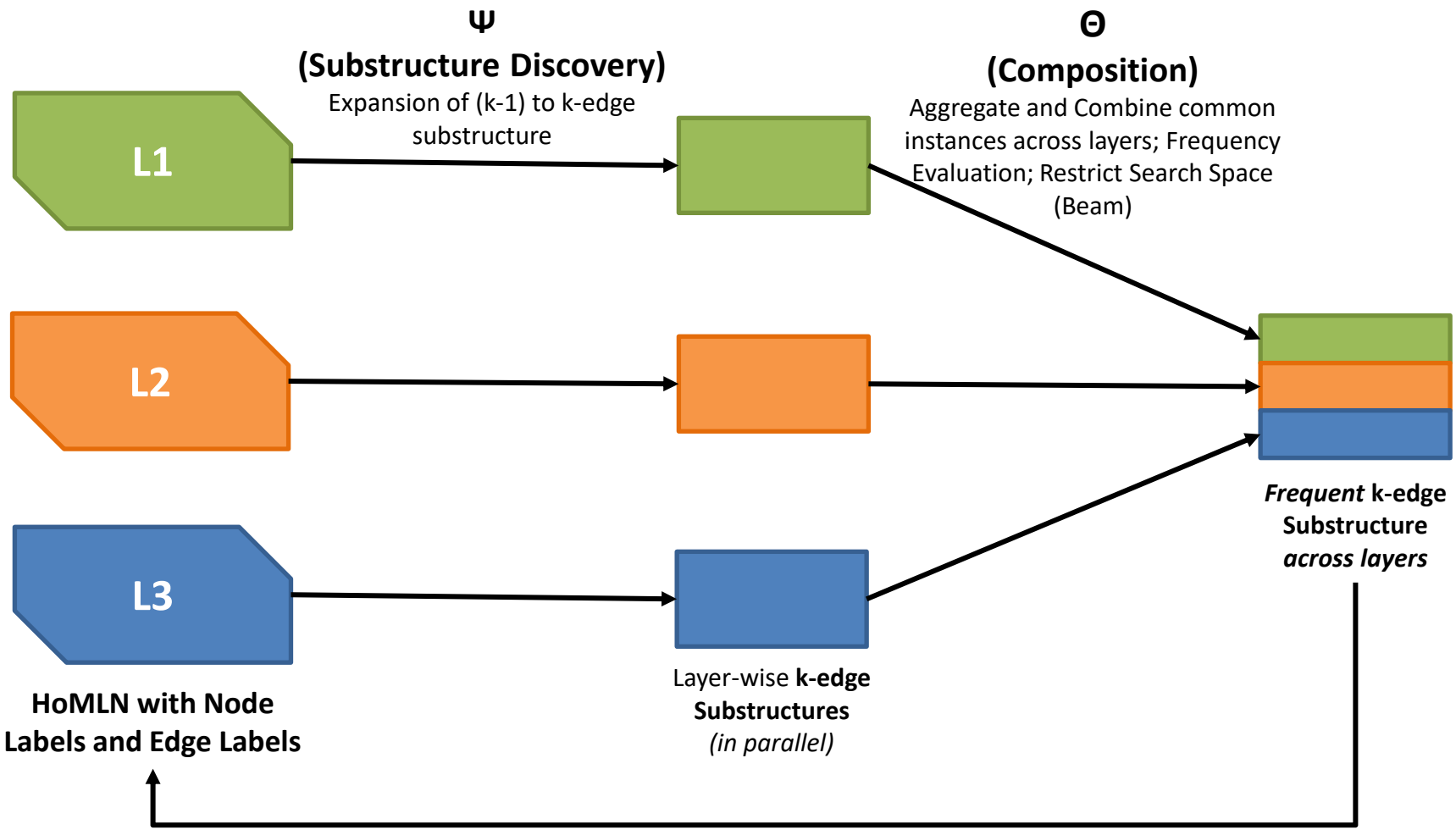
Allegiant v/s All
Grand Rapids
Elko
Montrose

- **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



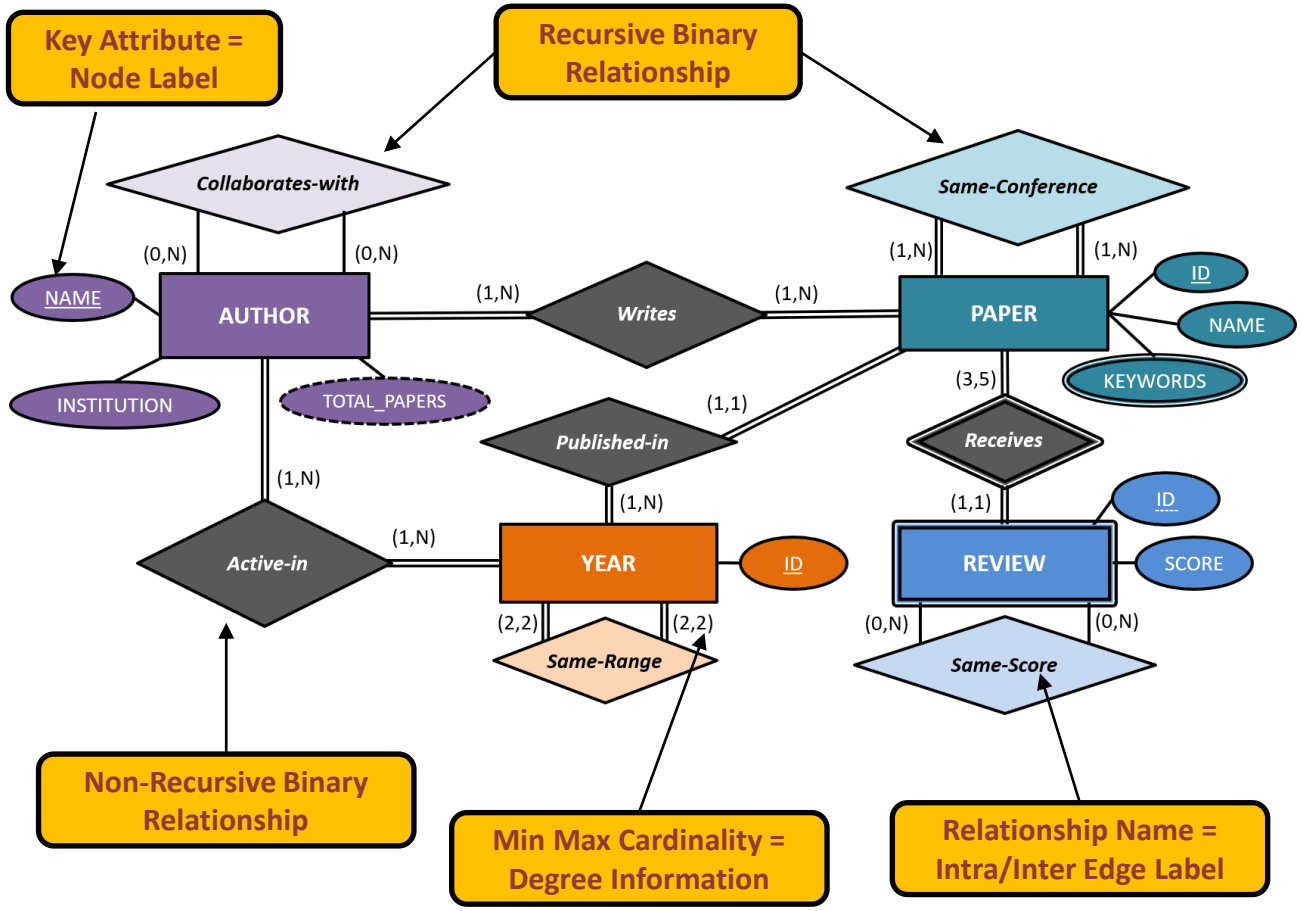
This process continues iteratively until a termination condition is applied
(Limit the maximum length of the frequent substructure)

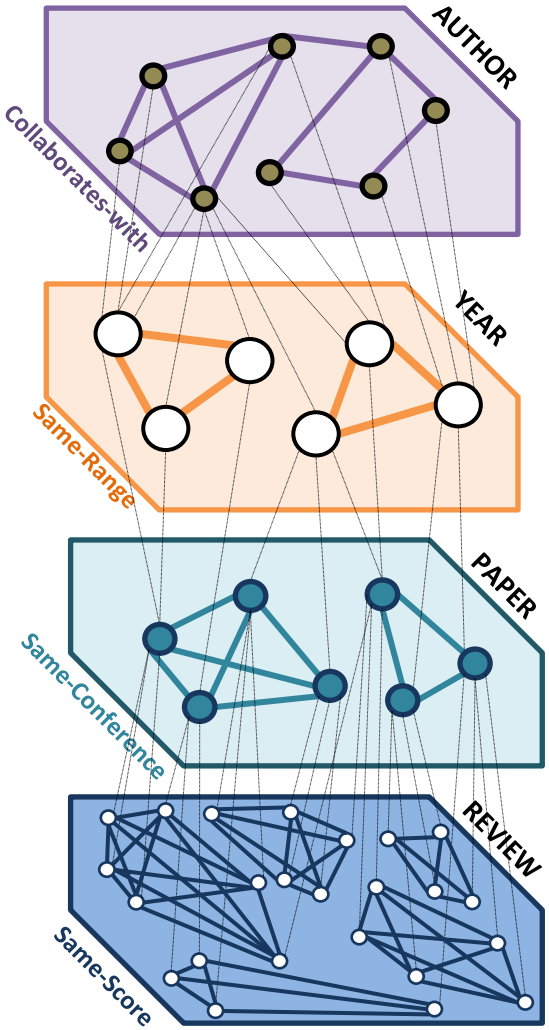
Conversion of an EER Model to MLN Model

8 Step Algorithm

Algorithm for EER Model → MLN Model

Research Paper Publication Dataset Modeling





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?

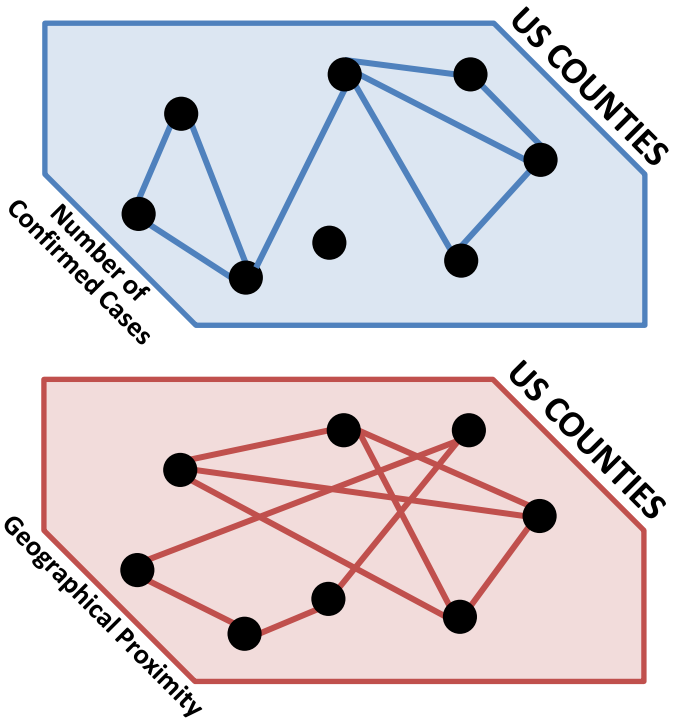
Which dense regions of the country are more susceptible to the virus spread due to the presence of institutions with large work force?

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?

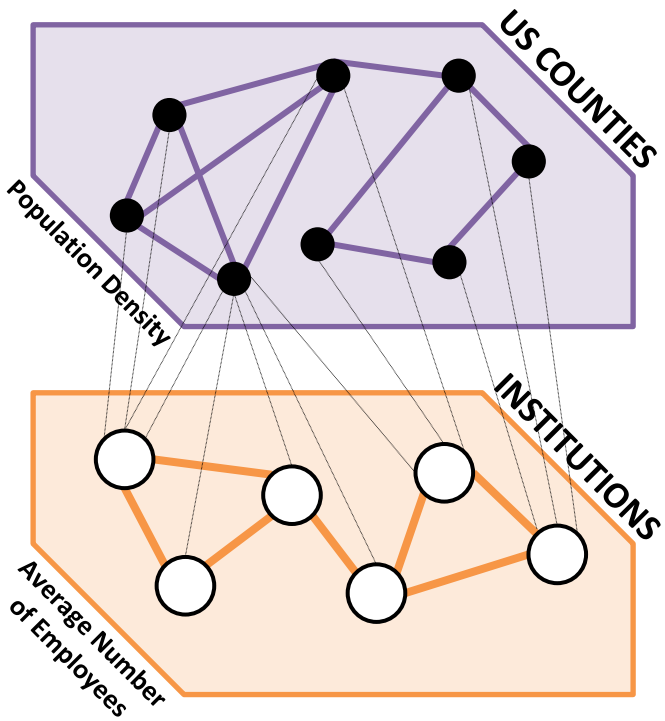


Homogeneous Multilayer Network

Data Analysis Requirements

Multiple Features, Relationships

Multiple Entities



Heterogeneous Multilayer Network

CoWiz

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)

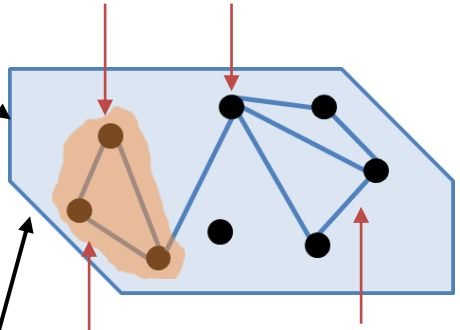
Nov 10, 2020 – Nov 19, 2020

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)

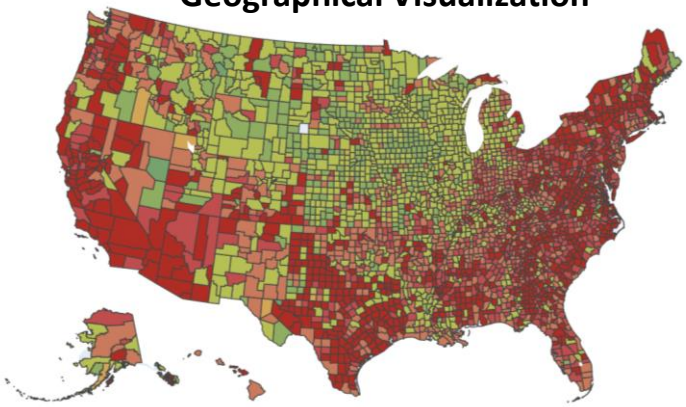
US COUNTIES



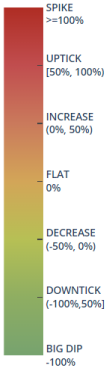
Edge exists if counties have **similar % Change** in the number of **new cases** in the Target Period as compared to the Base Period

County Communities: Geographical Regions with similar % change in new cases; Classifying Regions with **varying effects of event**

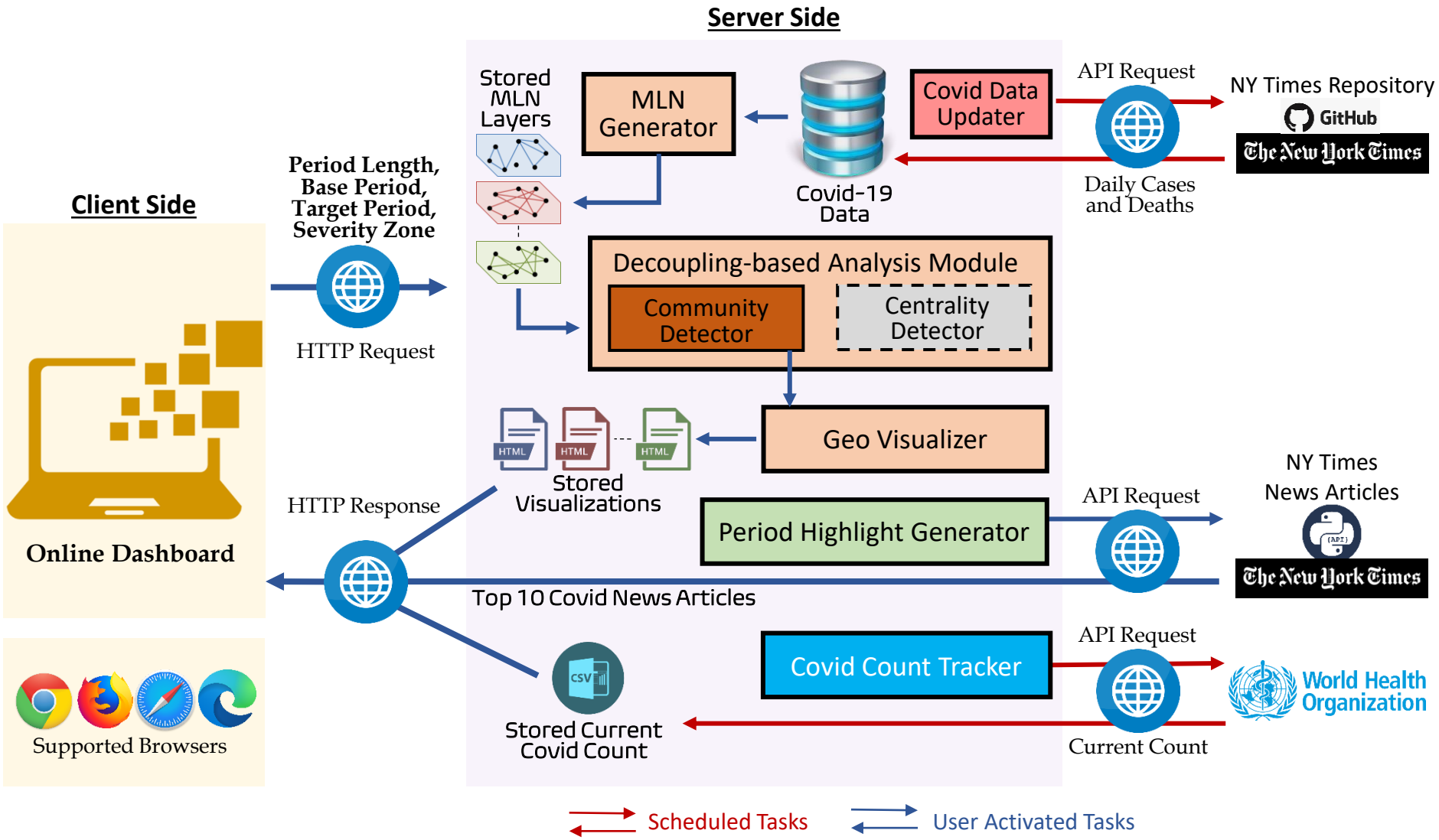
Geographical Visualization



Severity Rate
(Severity of the Spread)



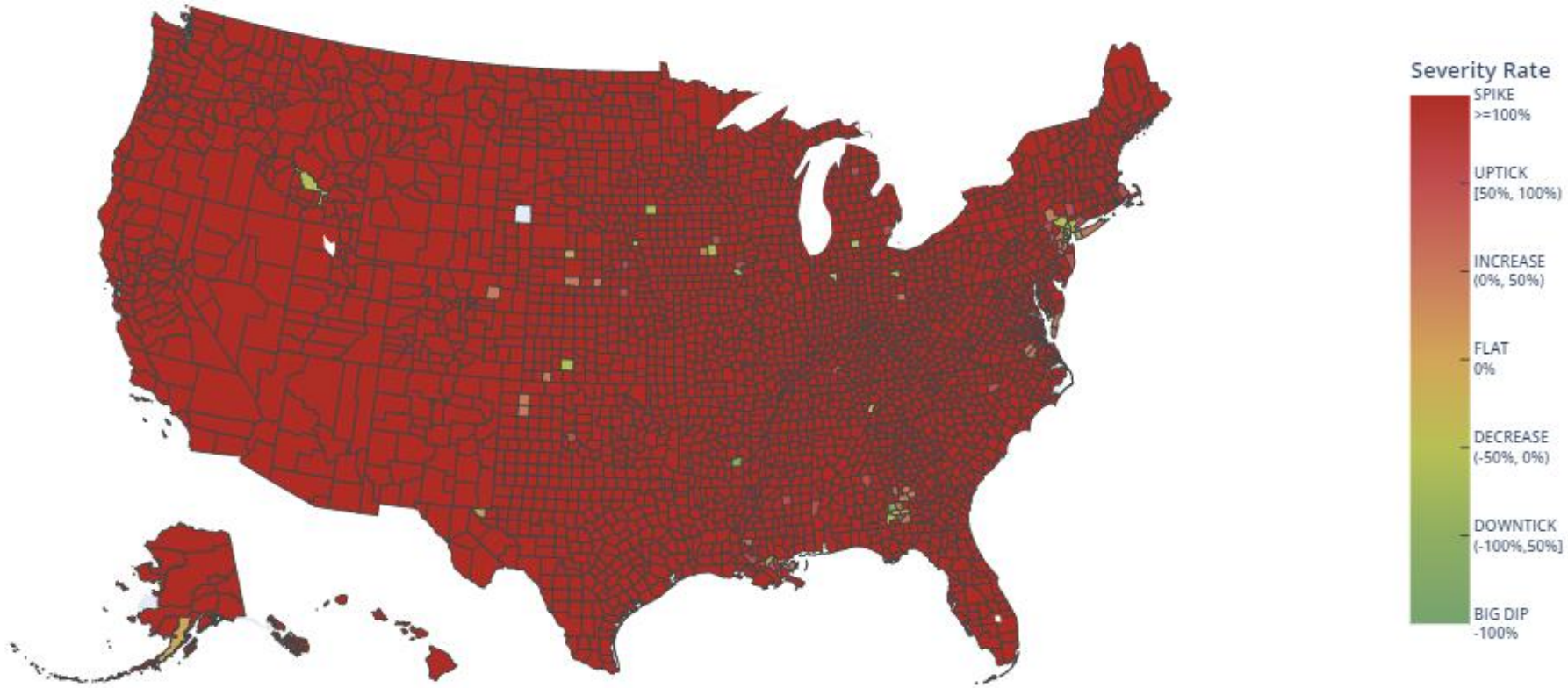
CoWiz: Architecture



CoWiz: Post Lockdown Effect on a Monthly basis

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

Target Period	**Dec 1, 2020 to Dec 30, 2020**



CoWiz

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/
  - India: https://censusindia.gov.in/

  OPEN-SOURCE
- **Traffic Movement:** Miles, County Boundaries, …
  - US: https://www.streetlightdata.com/

  PROPRIETARY
- **Average Mobility in Categories of Places:** Parks, Transit Stations, Workplaces, …

  OPEN-SOURCE
  - Country-wise: https://www.google.com/covid19/mobility/
- **Expenditure Categories:** Grocery, Shopping Malls, Entertainment, …

  PROPRIETARY
  - Anonymized data from credit card companies

Summary

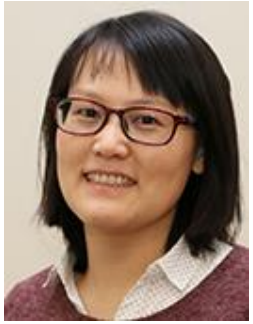
- MLNs v/s Simple/Attributed Graphs
  - Modeling and Computation Challenges
- **Decoupling Approach** for MLN Analysis
- Efficient and lossless composition techniques for various analysis
  - Communities (HoMLN, HeMLN)
  - Hubs, Substructure (HoMLN)
- Community Definition for HeMLN
- Real world applicability 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

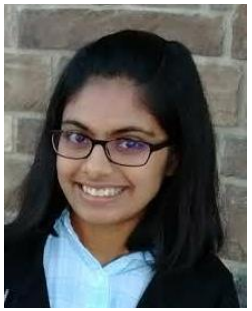
Thank You
Questions?





For more information visit:

https://itlab.uta.edu




References

Samant, K., Memeti E., Santra, A., Karim, E. and Chakravarthy, S., 2021. CoWiz: Interactive Covid-19 Visualization Based On Multilayer Network Analysis. In IEEE International Conference on Data Engineering (ICDE). IEEE.

Santra, A., 2020. Analysis of Complex Data Sets Using Multilayer Networks: A Decoupling-Based Framework (Doctoral Dissertation). The University of Texas at Arlington

Komar, K.S., Santra, A., Bhowmick, S. and Chakravarthy, S., 2020, November. EER → MLN: EER Approach for Modeling, Mapping, and Analyzing Complex Data Using Multilayer Networks (MLNs). In International Conference on Conceptual Modeling (pp. 555-572). Springer, Cham.

Santra, A., Komar, K.S., Bhowmick, S. and Chakravarthy, S., 2020. A new community definition for multilayer networks and a novel approach for its efficient computation. arXiv preprint arXiv:2004.09625.

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

Vu, X.S., Santra, A., Chakravarthy, S. and Jiang, L., 2019. Generic multilayer network data analysis with the fusion of content and structure. In 2019 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).

Santra, A., Bhowmick, S. and Chakravarthy, S., 2017. Efficient community re-creation in multilayer networks using Boolean operations. Procedia Computer Science, 108, pp.58-67.

Santra, A., Bhowmick, S. and Chakravarthy, S., 2017, November. Hubify: Efficient estimation of central entities across multiplex layer compositions. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 142-149). IEEE.

Santra, A. and Bhowmick, S., 2017, December. Holistic analysis of multi-source, multi-feature data: Modeling and computation challenges. In International Conference on Big Data Analytics (pp. 59-68). Springer, Cham.