Big Data Analysis using Multilayer Networks

Sharma Chakravarthy, Abhishek Santra

Information Technology Laboratory (ITLab)

Department of Computer Science

University of Texas at Arlington

(BDA 2019 Tutorial)

Roadmap

- **Motivation**
  - *The Overall Picture*
- **Community Detection in MLNs**
  - *Boolean Composition Approaches in **HoMLNs***
  - *Maximum Weighted Bipartite Matching Approaches in **HeMLNs***
- **Hub Detection in HoMLNs**
  - *Degree and Closeness Centrality Hub Detection Heuristics*
- **Case Studies on Real World Datasets**
  - *Facebook, US Airlines, IMDb, DBLP, …*
- **Publications**

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: 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*

same *Indian Cities*

Problem: 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

Problem: Handling Analysis Flexibility

Ability to analyze the dataset using combinations of features (or perspectives)

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?

Problem: Handling Analysis Flexibility

Ability to analyze the dataset using combinations of features (or perspectives)

Connectivity among **different** IMDb entities

Movies

Actors

Directors

Genres

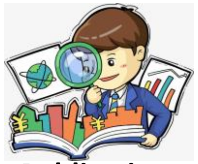
Rating

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?

Connectivity among **different** dblp entities

Author Collaborations



Publications

Conferences

WikiCFP
A Wiki for Calls For Papers
Research Domains

Years

Venues

Frequently publishing cohesive co-author groups?

Most active periods for popular collaborators?

Motivation

Complex Data Analysis

Traditional Approaches and Their Limitations

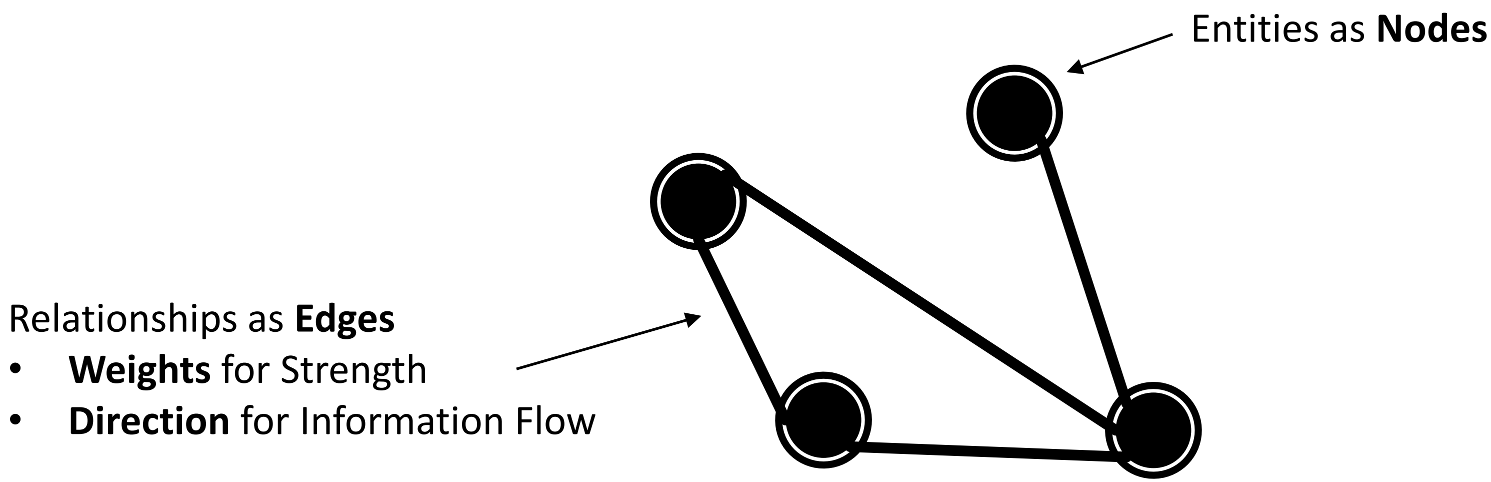
Modeling Using Multilayer Networks

Limitations of Existing Analysis Approaches

Decoupling Approach to Analyze MLNs

Traditional Modeling: Simple Graphs

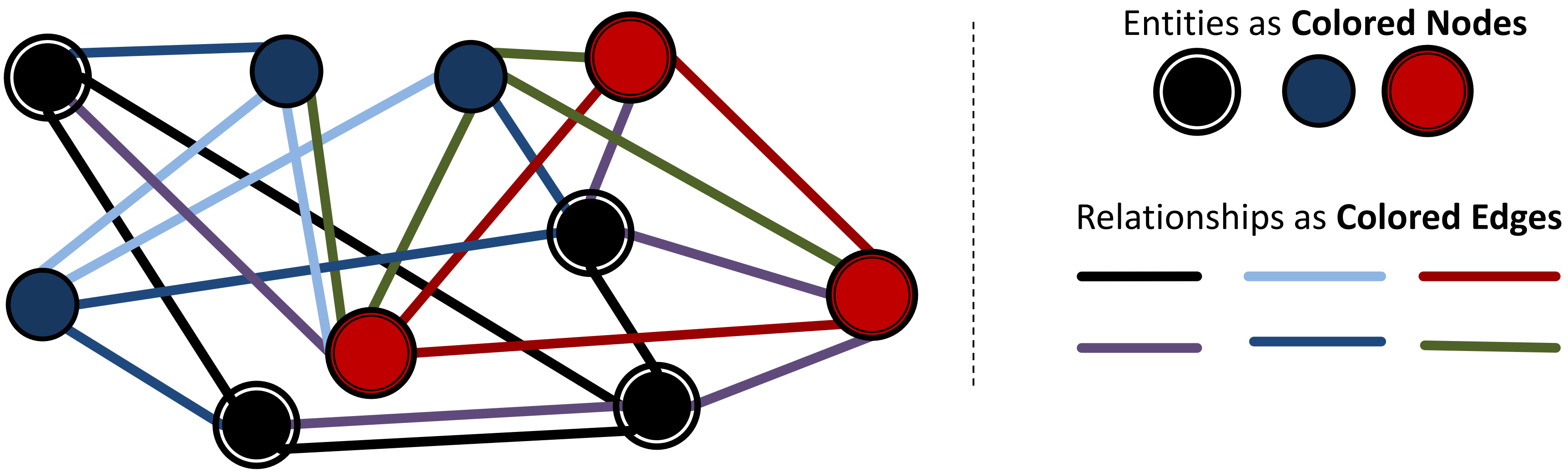
- Nodes: Entities
- Single Edges (weighted or unweighted): Single or Combination of feature-based relationship



- **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

Complex *(Multi-Entity, Multi-Feature)* Data Analysis

	Modeling Clarity	Analysis Flexibility	Computational Efficiency
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)
Attributed Graph	**To some extent** (Multiple node and edge labels supported)	**Not Available** (Except Subgraph Mining)	**Bad** (Multiple Traversals required to fetch required combination)
Multilayer Networks	**Good**	**Good**	**Good (for cases shown)**

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

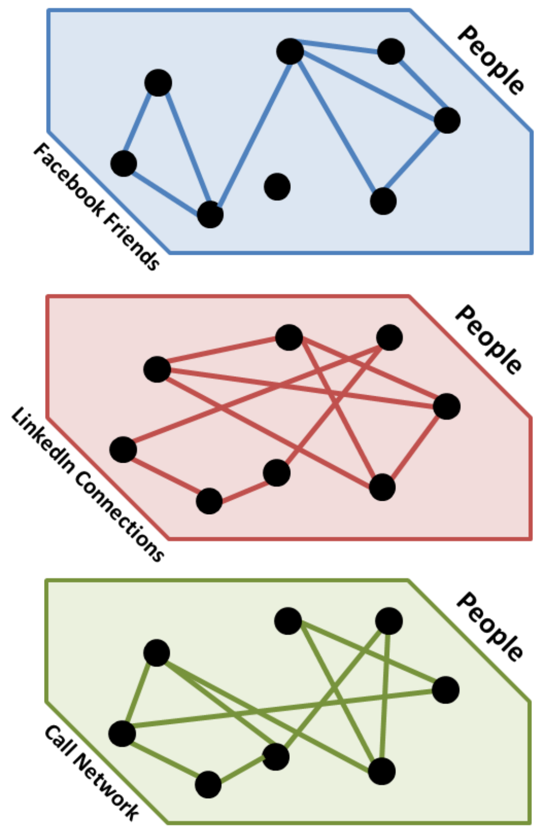
Decoupling Approach to Analyze MLNs

Modeling Clarity using MLNs

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

Interactions among **People**

Same Entities, Different Relationships



Homogeneous MLN (HoMLN)

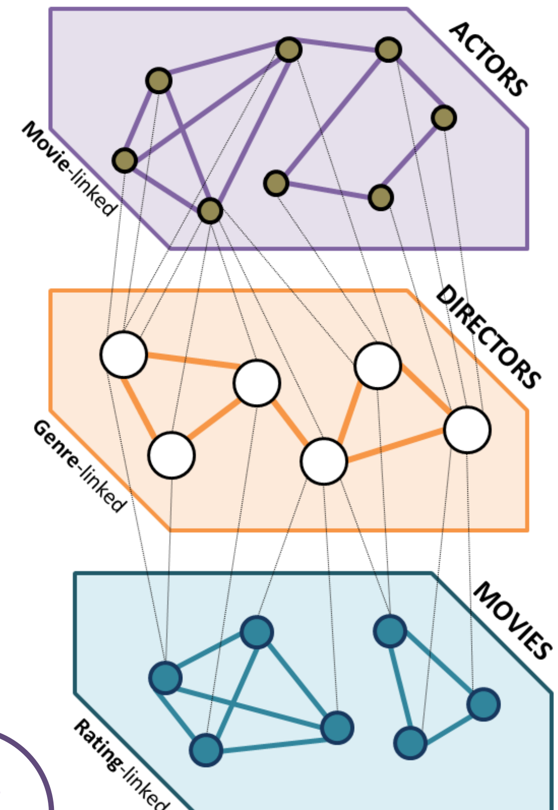
a. **Most popular or socially active group of people across platforms?**
b. ***Most influential* set of people?**

Modeling Clarity using MLNs

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

Interactions among IMDb Entities

Different Entities, features, and Relationships



a. For the *most popular actor groups* from each *movie rating class*, which are the *director groups* with which they have *maximum interaction*?
b. Highly rated actors working in similar genres who have never acted together?

Heterogeneous MLN (HeMLN)

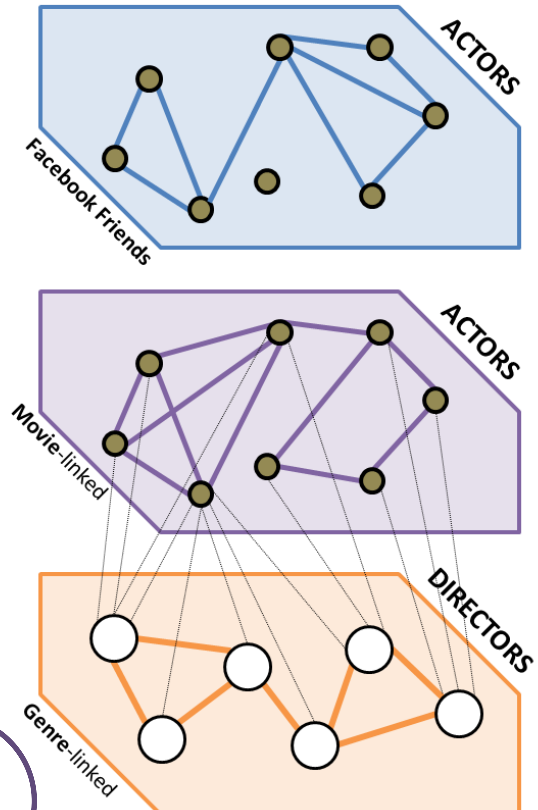
Modeling Clarity using MLNs

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

Interactions among **IMDb** **Entities**

a. For the *co-actor groups* who are friends on *Facebook,* which are the *director groups* with which they have *maximum interaction*?
b. Co-actor groups who are not 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

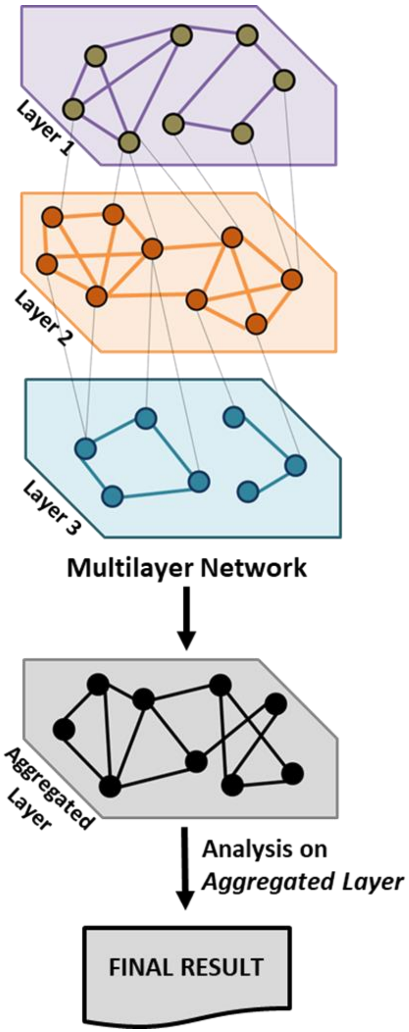
Limitations of Existing Analysis Approaches

Decoupling Approach to Analyze MLNs

Current MLN Analysis Alternatives

Reduce to a Single Graph (SG)

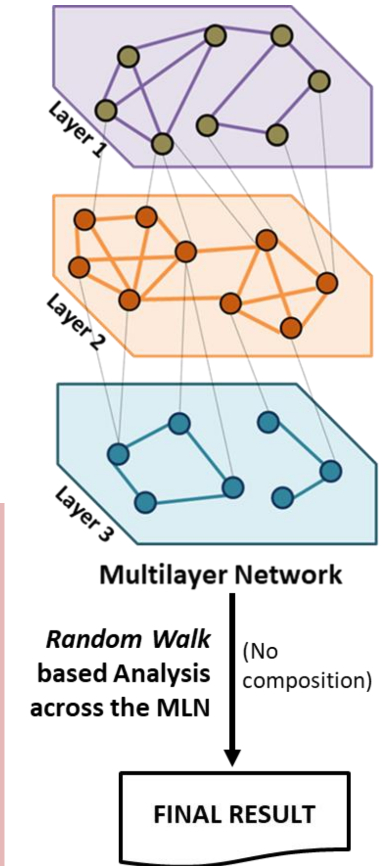
- Aggregate/Combine the desired MLN layers as a single simple graph.
- Process the *combined layer* using *existing algorithms*



- **Loss of information, semantics**
- ***N*** layers $\Rightarrow$ **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

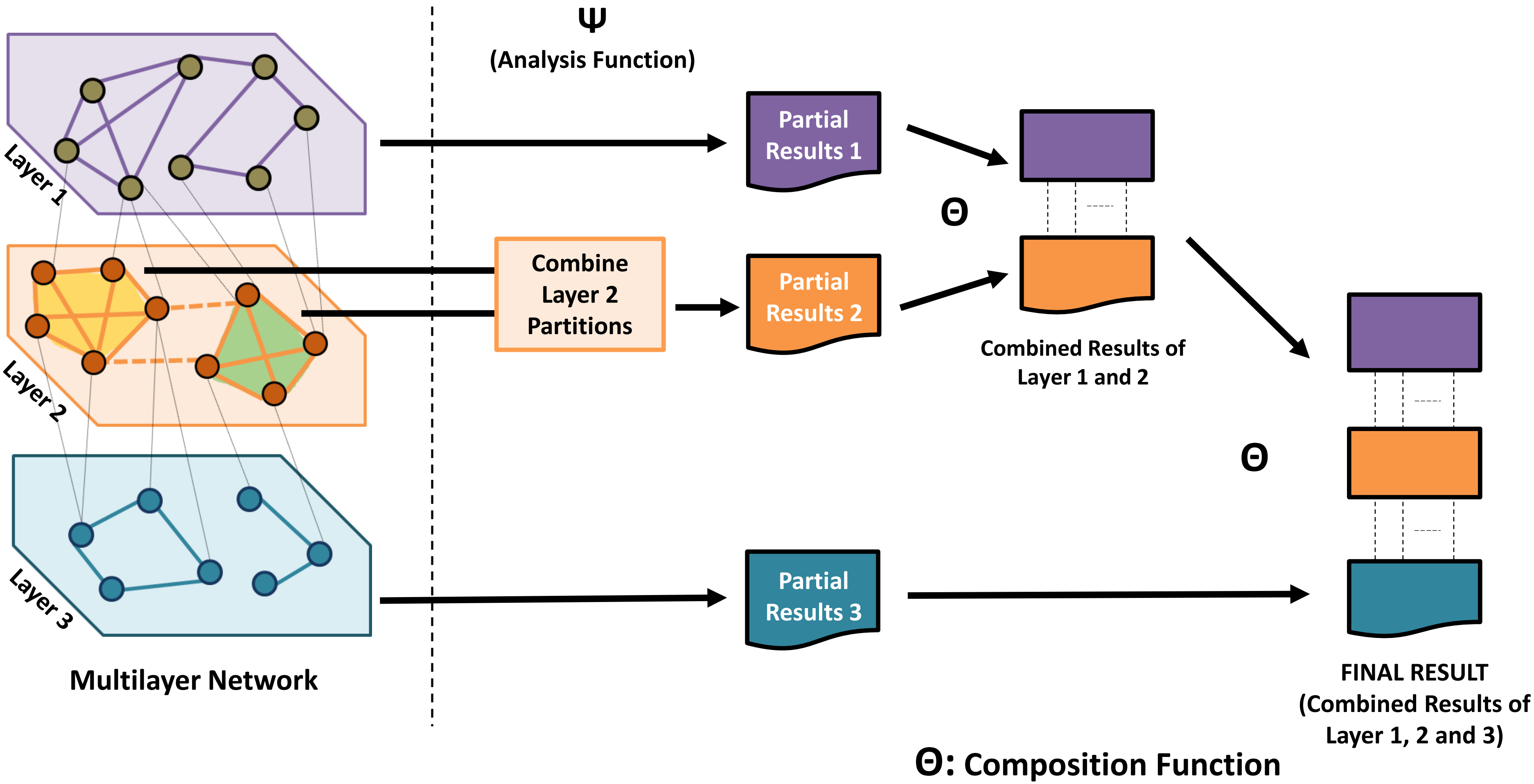
Modeling Using Multilayer Networks

Limitations of Existing Analysis Approaches

Decoupling Approach to Analyze MLNs

Overview of Decoupling Approach

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



Decoupling Approach

- A "**Divide-and-Conquer**" approach
  - Use an **Analysis function** (Ψ) to generate layer-wise results (termed *partial results*) based on **analysis objectives**
    - *E.g., communities, centrality, subgraphs, …*
  - Use a **Composition function** (Θ) to **correctly (loss-less, no distortion)** combine generated layer-wise partial results
    - *E.g., maximal weighted bipartite matching, …*
- Challenge:
  - Identify Ψ and Θ for various types of analysis and establish their correctness
  - 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. **Parallelization Opportunities**
7. **Ease of** dataset updates

   Entails *updating affected layers only and their results*
8. **Application Independent**

Community Detection in HoMLN

Brief Introduction to Community

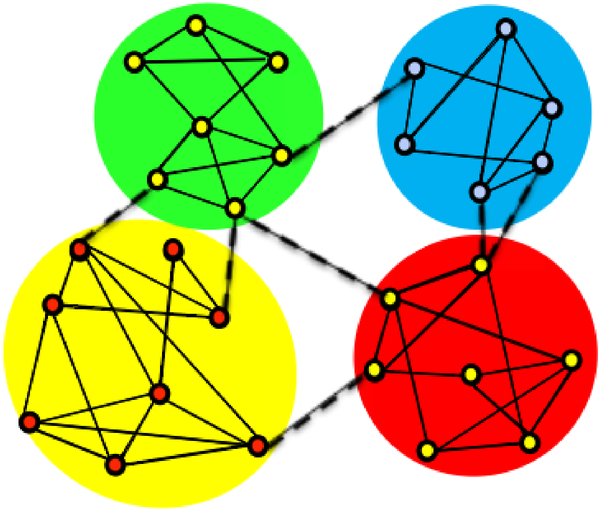
Boolean AND Composition

Boolean OR Composition

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

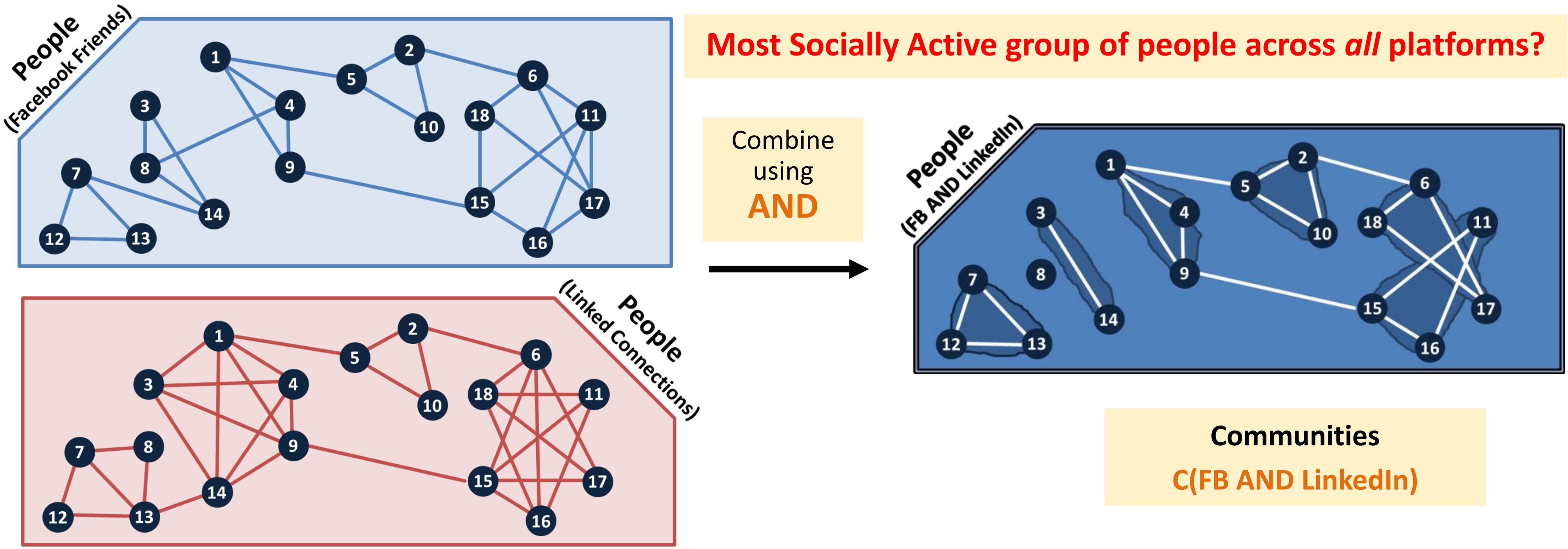
Boolean OR Composition

Case Studies

AND Composition using *Single Graph*

- ➢ **S**ingle **G**raph Approach for **C**ommunities **(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?

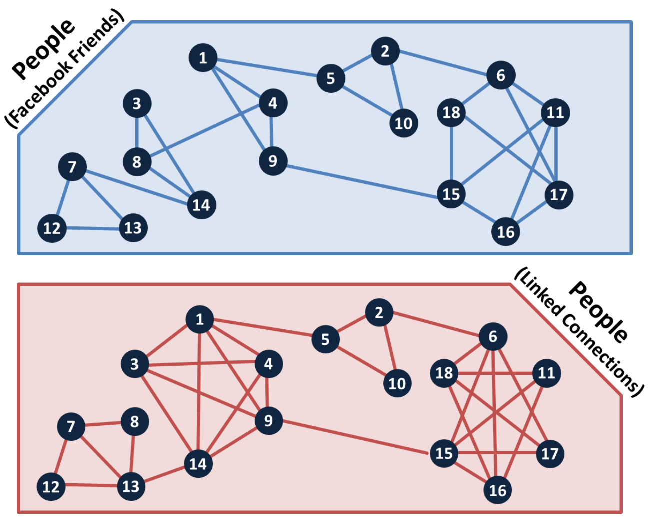


AND Composition using *Decoupling Approach*

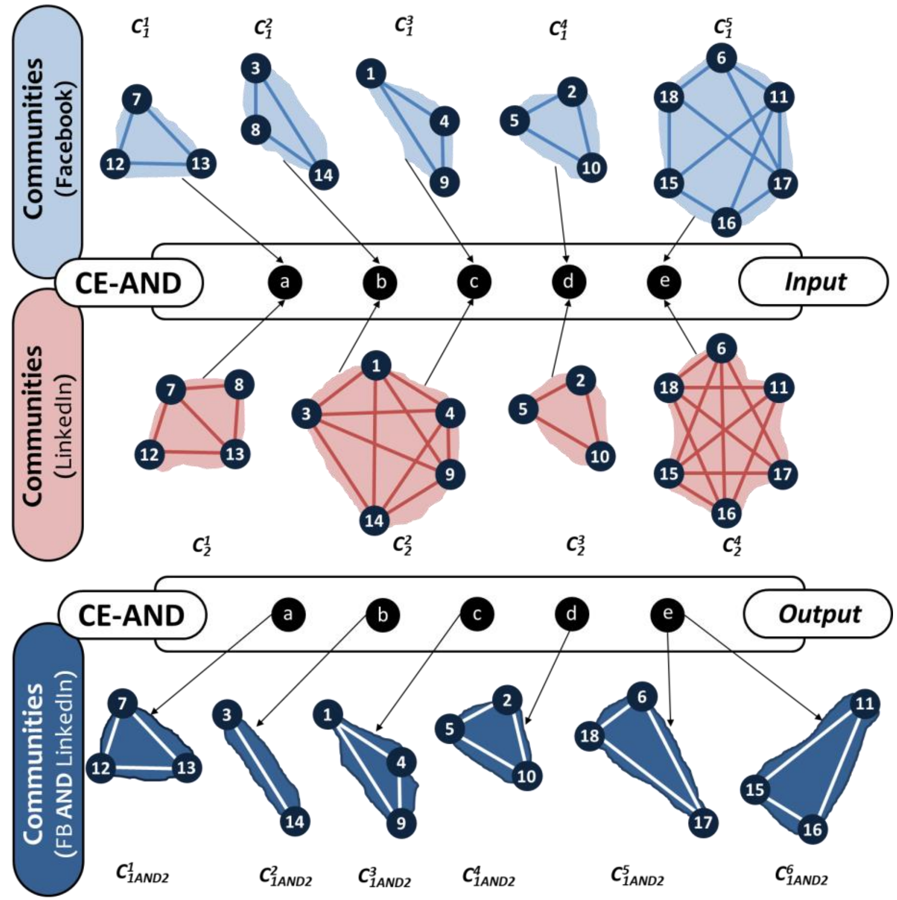
- **Correctness Criterion:** Generate the same communities obtained by ANDing layers into a single graph (**termed C-SG-AND)**
- **Specification: $C(G_1)$ AND $C(G_2)$ AND … AND $C(G_k)$**
  - **$C(G_i)$:** Communities of G_i
- **Approaches for *2 layer Composition***
  - **CV-AND: Node-based intersection** of **layer wise communities**
    - **Accuracy is not very high** as *community topology not considered*. Works well in the presence of **cliques**.

AND Composition using *Decoupling Approach*

(CE-AND Algorithm - *Illustration*)



Layer-wise
Communities



- **Accuracy improves significantly**
- **Additional computation required!**

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)

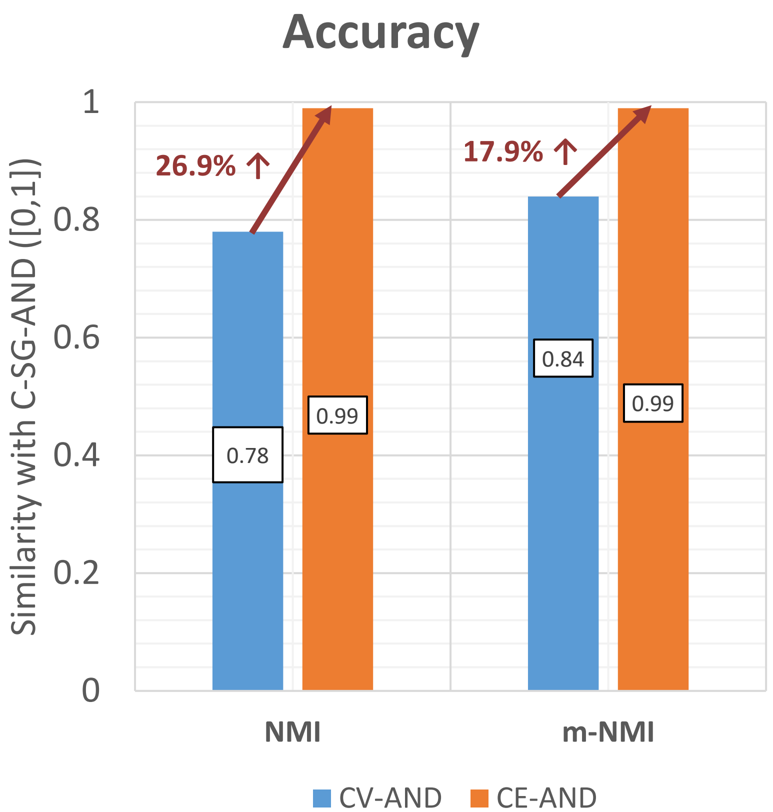
Cost(CV-AND) < Cost(C-SG-AND), Cost(CE-AND) < Cost(C-SG-AND)

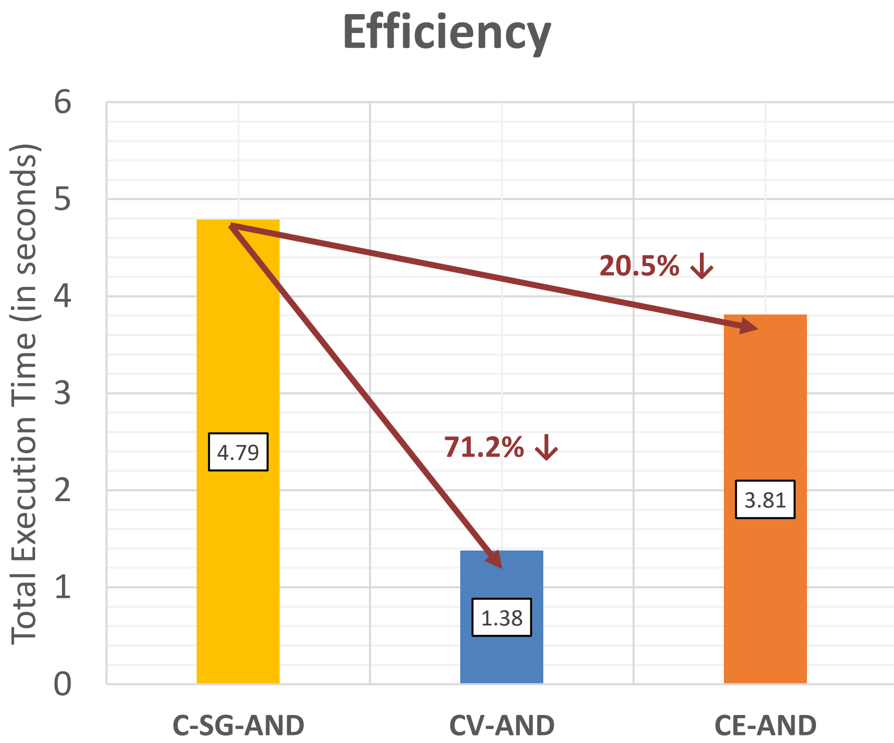
- Cost benefit amortized over large analysis space (2^N)

Experimental Results

- **Setup (Accident HoMLN)**
  - **Nodes**: **5000 Accidents from UK in 2014**
  - **Layers**: 2 nodes connected if the accidents occurs within 10 miles of each other and had similar **Light** (Layer L) or **Weather** (Layer W) or **Road Surface** (Layer R) condition
  - **4 AND Composition Analysis**
    - L ***AND*** W, W ***AND*** R, L ***AND*** R, (L ***AND*** W) ***AND*** R
  - **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*

Trade-off between Accuracy and Efficiency



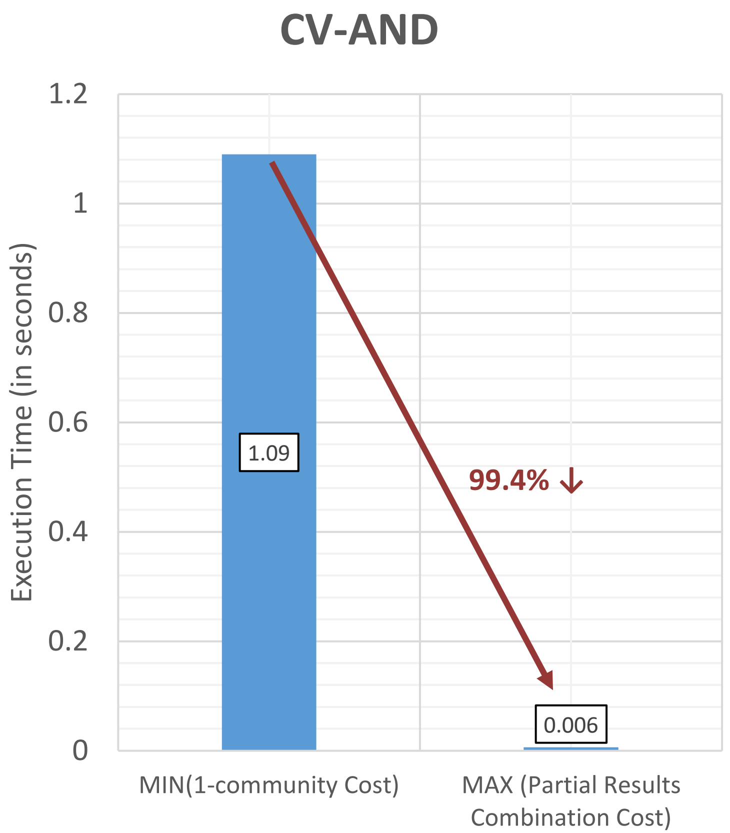


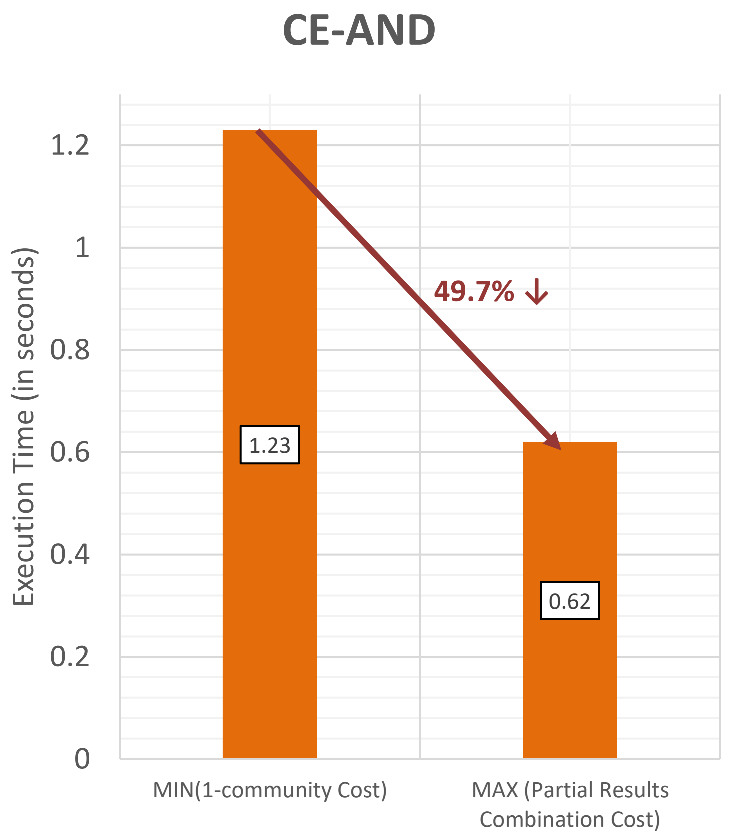
Efficiency improves with more analysis (~O(2^N)) for large N

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

Community Detection in HoMLN

Brief Introduction to Community

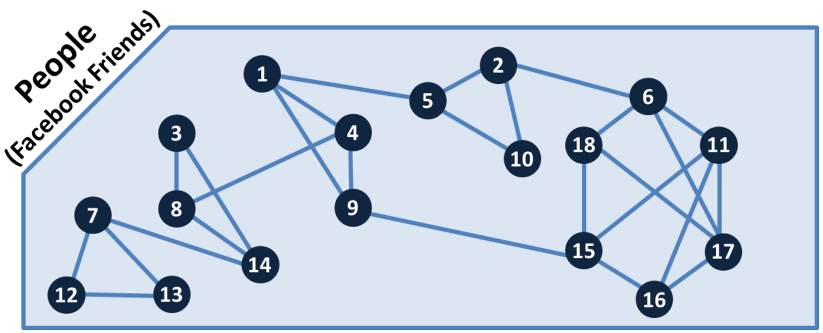
Boolean AND Composition

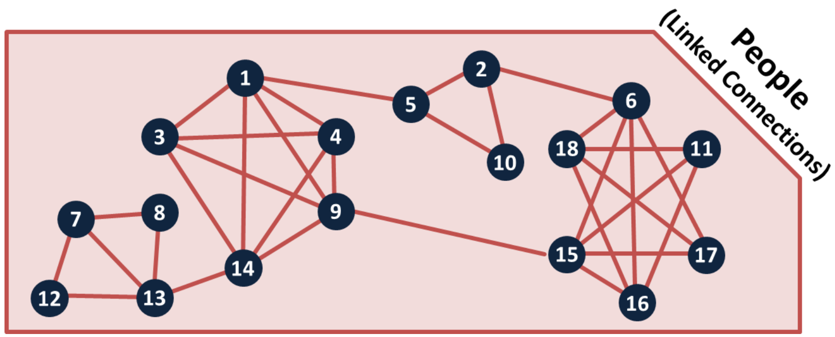
Boolean OR Composition

Case Studies

OR Composition using *Single Graph*

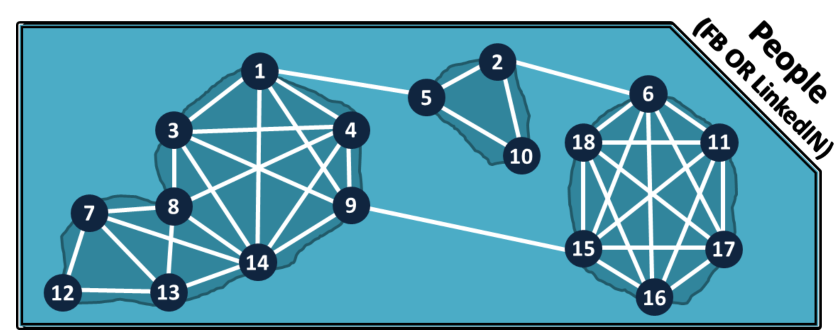
- Single Graph Approach for Communities **(C-SG-OR)**
  - **Combine** the required layers using **OR operator**
  - Apply **existing community detection algorithms**
- **Specification: $C(G_1 \text{ OR } G_2 \text{ OR} \ldots \text{OR } G_k)$**
  - G_i: Original MLN layer or NOT layer
- Communities used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





Groups of people connected by *some* medium?

Combine using **OR**

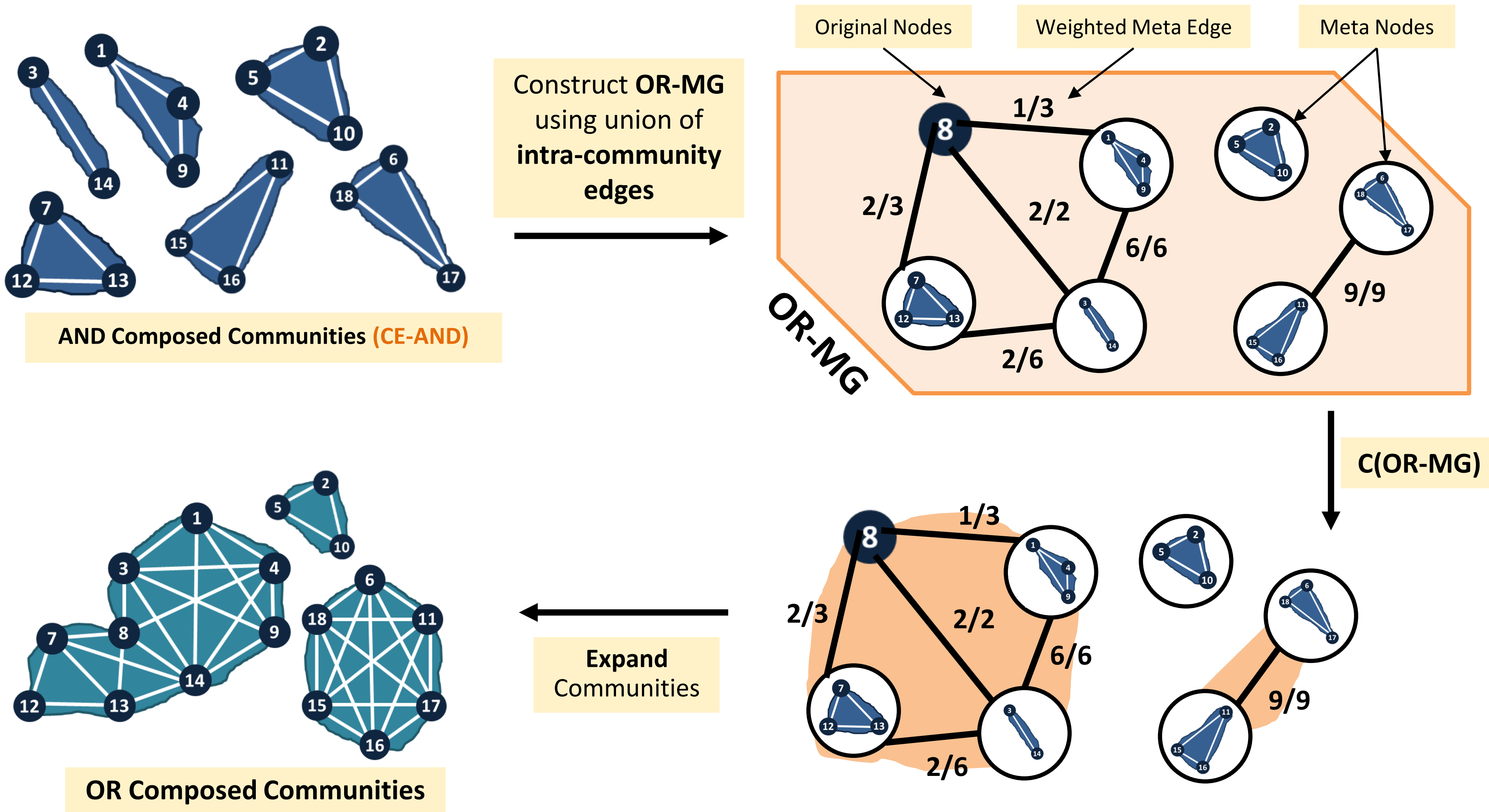


Communities
C(FB OR LinkedIn)

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*)



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)

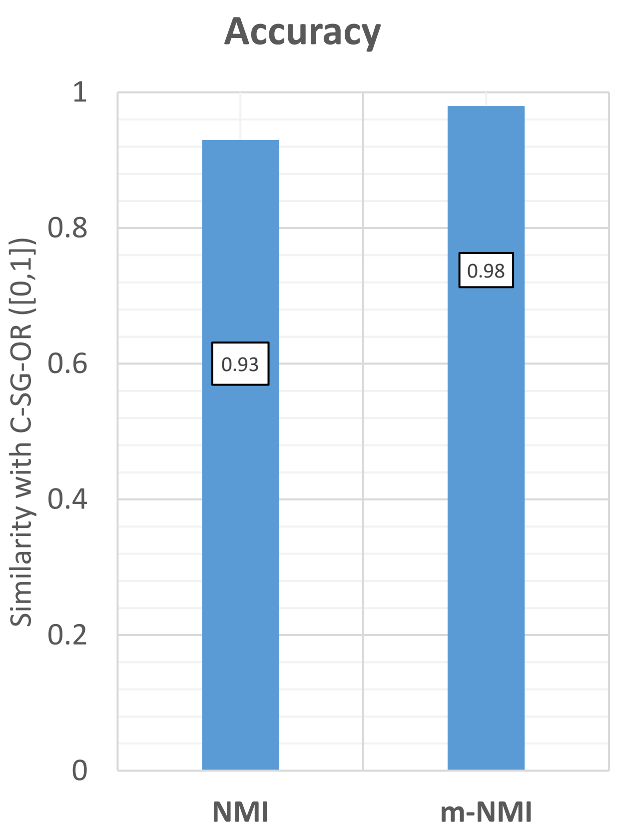
Cost(CE-OR) < Cost(C-SG-OR)

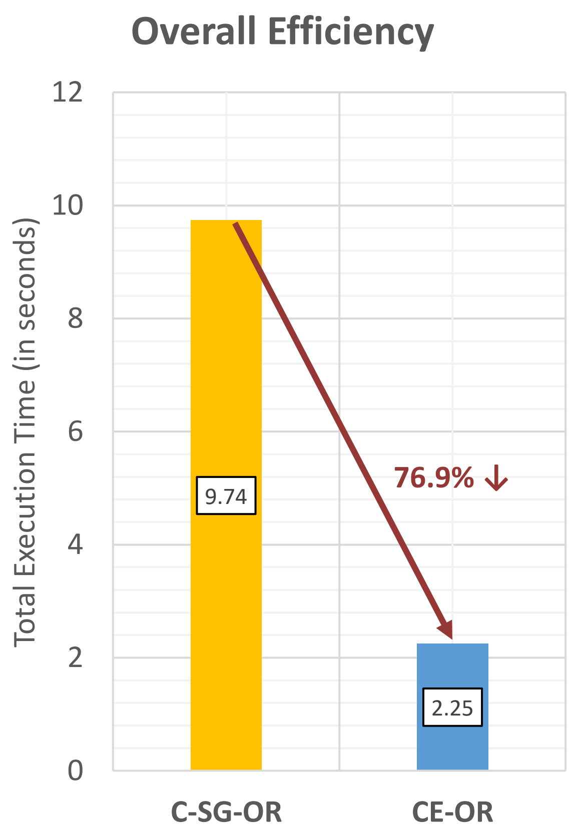
- Cost benefit amortized over large analysis space (2^N)

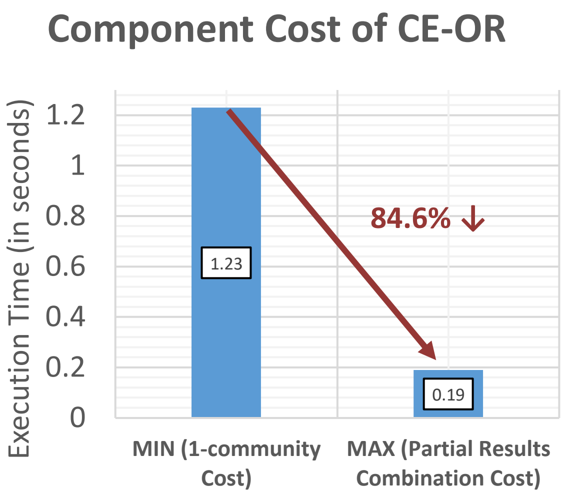
Experimental Results

- **Setup (Accident HoMLN)**
  - **5000 Accidents with 3 layers (L, W, R)**
  - **4 OR Composition Analysis**
    - L ***OR*** W, W ***OR*** R , L ***OR*** R, (L ***OR*** W) ***OR*** R
  - **Community Detection Algorithm: Infomap**
  - **CV-AND** used for AND composition
- **Accuracy Metrics**
  - **NMI and m-NMI**

Accuracy and Efficiency







Significant Savings in Worst Case Component Cost Comparison validates efficiency

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

OR Composition Decoupling Process leads to more savings as compared to AND, as **density(OR layer) > density (AND layer)**

Community Detection in HoMLN

Brief Introduction to Community

Boolean AND Composition

Boolean OR Composition

Case Studies

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)

*One percent data has been used

IMDb-Actors HoMLN

Which are the **largest groups** of **co-actors** that lead to the **most popular movie ratings**?

C(CoActing) CE-AND C(AvgRating)

- For the **most popular average actor rating, [6-7)**, the **largest co-actor groups**
  - Hollywood (876 actors), Indian (44 actors), Hong Kong (12 actors) and Spanish (9 actors)
- Prominent Actors in the **Hollywood** Group
  - **Al Pacino, Robert De Niro, Tom Cruise, Will Smith, …**
- Prominent Actors in the **Indian** Group
  - **Amitabh Bachchan, Shah Rukh Khan, …**
- Prominent Actors in the **Hong Kong** Group
  - **Jackie Chan, …**

IMDb-Actors HoMLN

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)

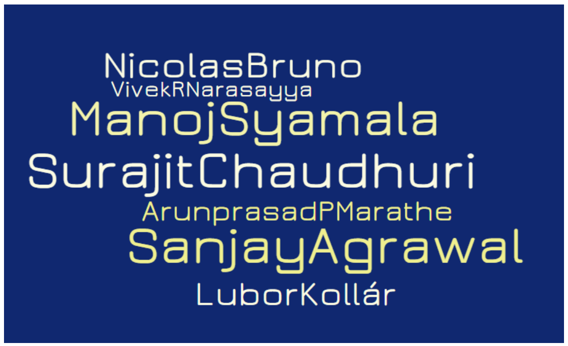
Actor/Actresses	Prominent Genres
Willem Dafoe, Russell Crowe	Action, Crime
Hilary Swank, Kate Winslet	Drama
Tom Hanks, Reese Witherspoon, Cameron Diaz	Comedy, Romance
Johnny Depp, Tom Cruise	Adventure, Action
Leonardo DiCaprio, Ryan Gosling	Crime, Romance
Nicolas Cage, Antonio Banderas	Action, Thriller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

DBLP-CoAuthors HoMLN

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)



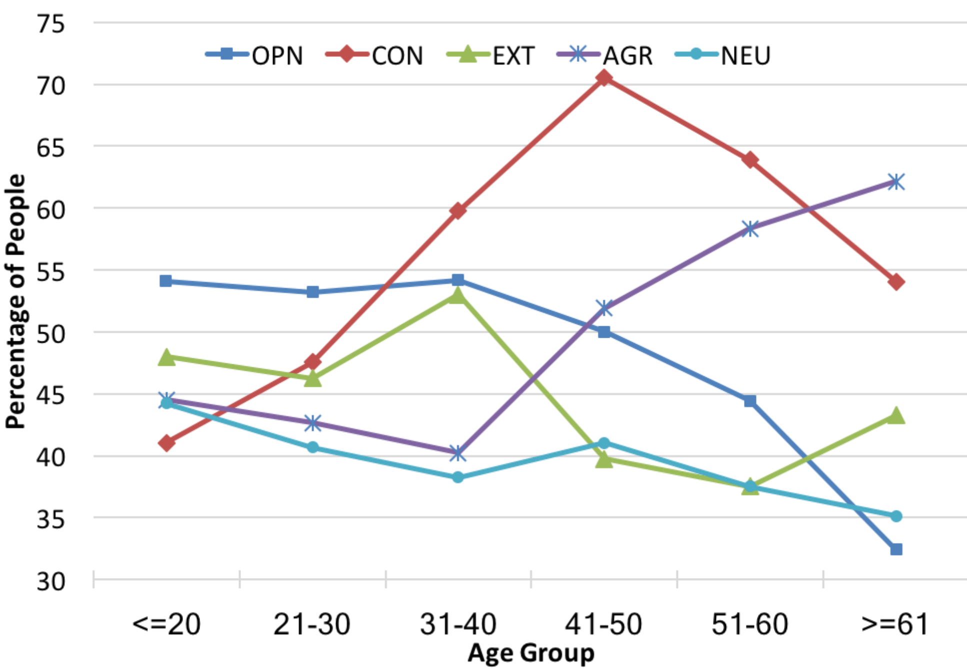
Widely accepted collaboration groups with high quality work

- **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*

Facebook HoMLN

How do the **personality traits (Big 5) 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)



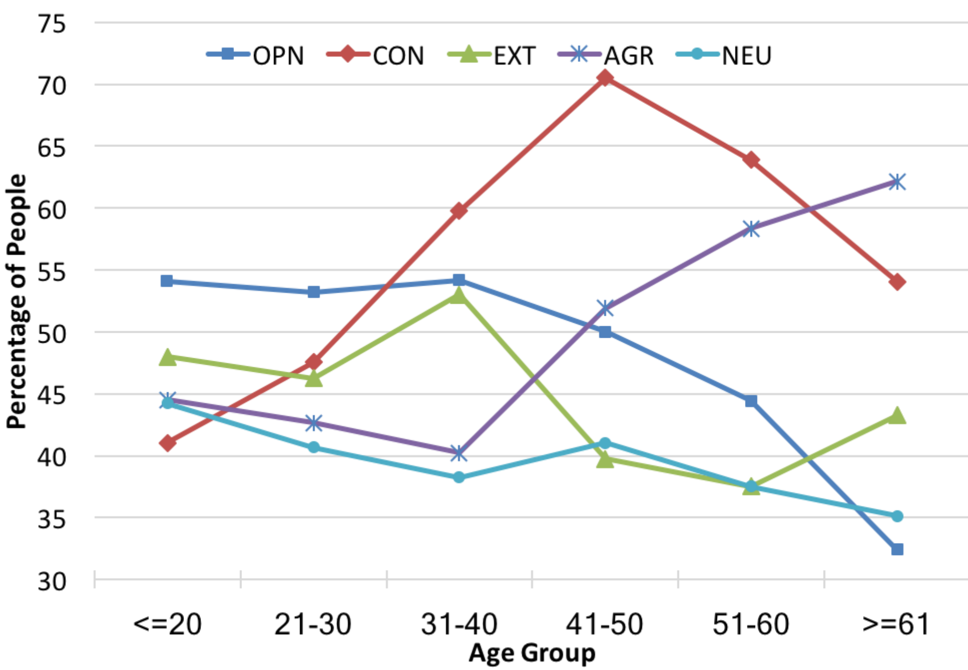
Openness (OPN)

- 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 (Big 5) 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)



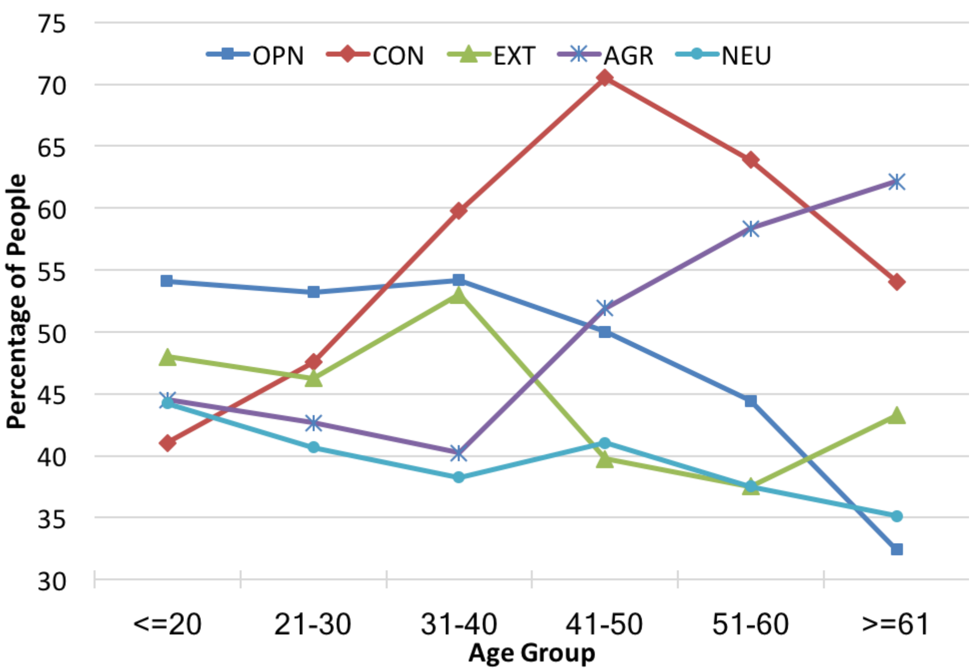
Conscientiousness (CON)

- 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 (Big 5) 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)



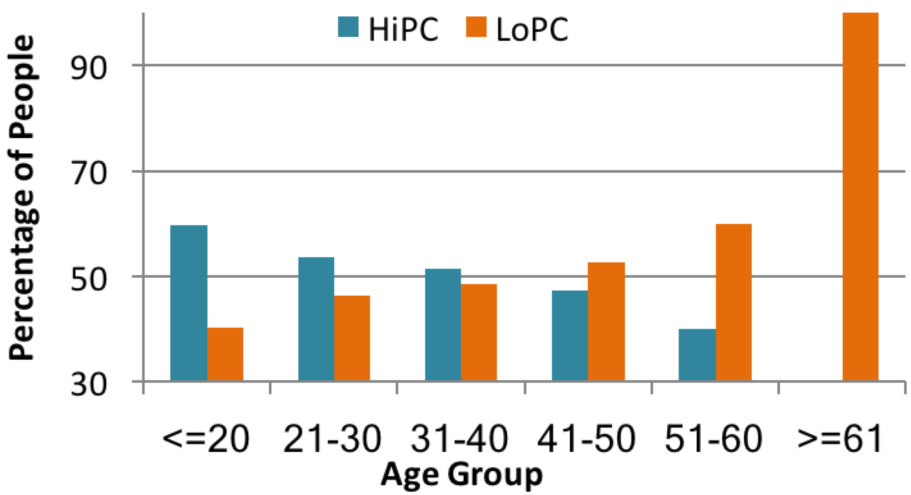
Neuroticism (NEU)

- 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

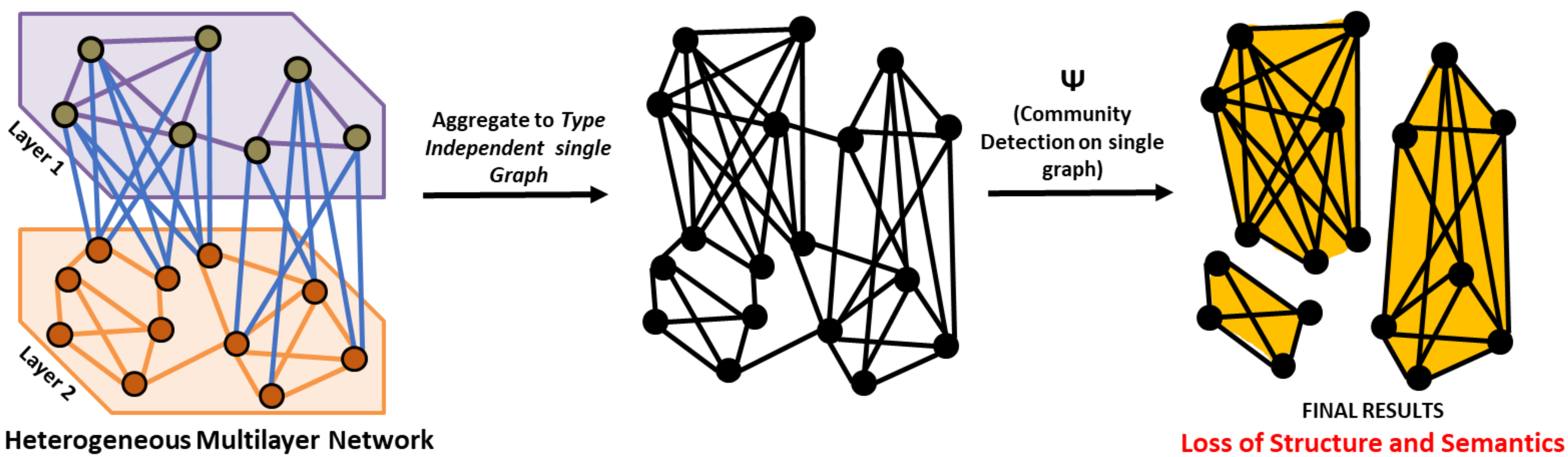
Heterogeneous Community Definition

Maximal Weighted Bipartite Coupling (MWBC) Composition

Case Studies

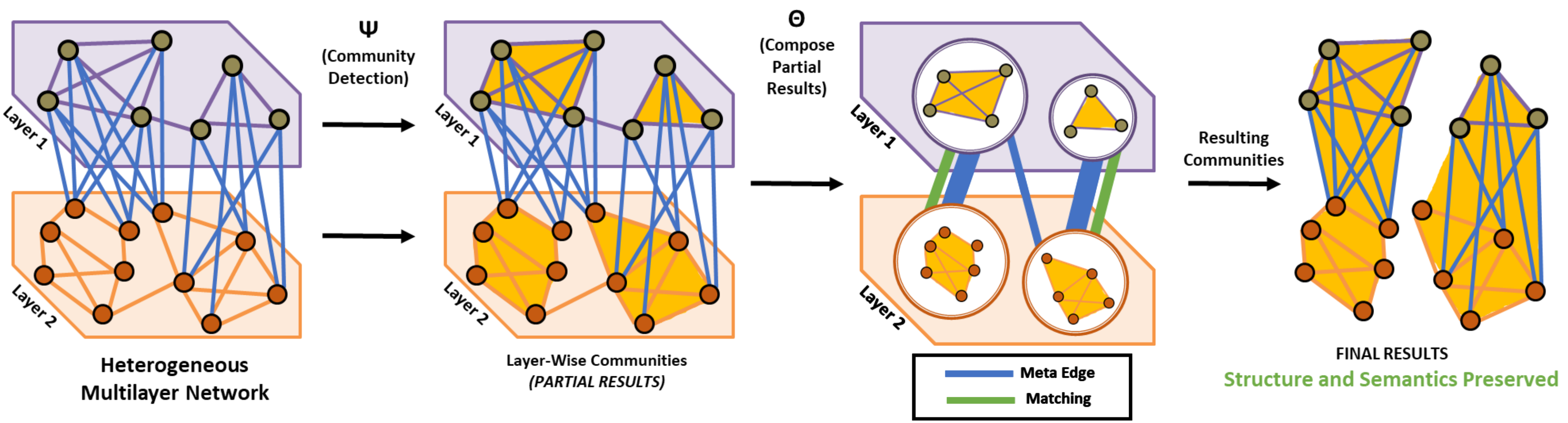
Community Notion in a HeMLN

- **Currently no structure-preserving definition**
  - There are type-independent, projection-based definitions
  - They have some undesirable properties **(loss of information/semantics, distortion of data, …)**



Community Notion in a HeMLN

- **Structure Preservation Required for Semantics**
  - Preserve **layer community structure (including types)**
  - Preserve **inter-layer edges (including relationships)**
  - So, *combined communities* are indeed **HeMLN**
  - **Drill-down analysis possible** with structure-preservation
- **Detection must be Computationally Efficient**



serial k-community

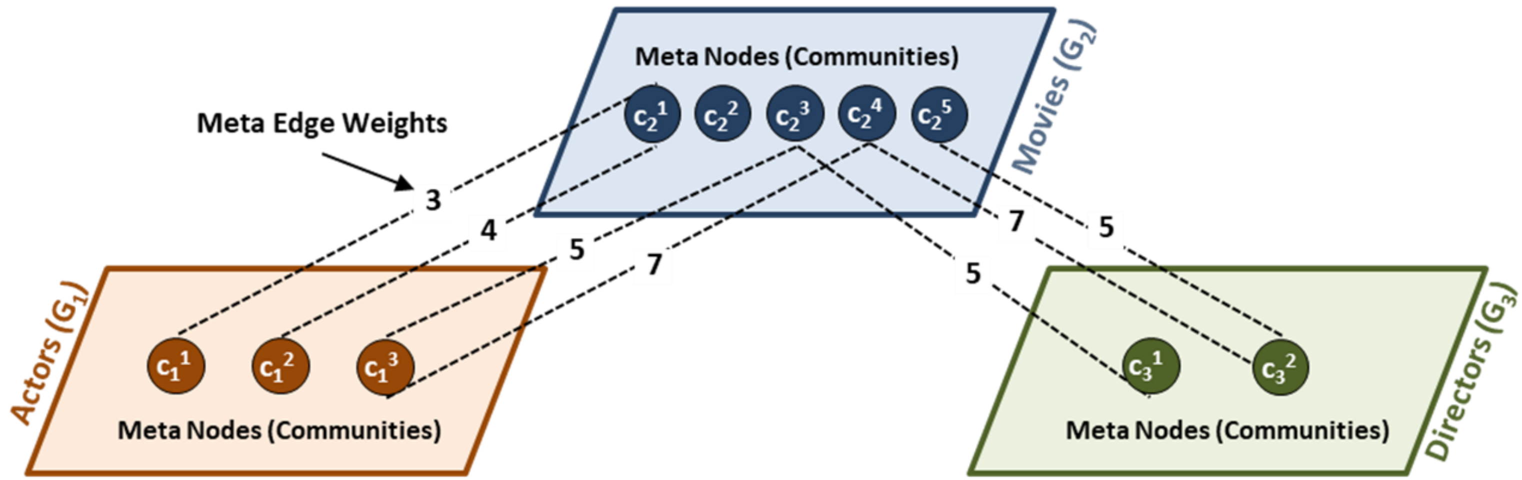
- **A set,** where each element represents
  - **k strongly knit groups of entities** (one community from each of k layers) that also have
  - ***progressively* strong coupling** among them (in specified coupling order)
- **1-community:** Layer-wise community as a set
- **Composition Function:**
  - **Maximum weighted bipartite coupling** on bipartite graph between two layers using **meta nodes** (corresponding to a community)
  - Couplings with a **choice of weight metrics**

serial k-community

- **Input Specification: Analysis expression including Ordering**
  - $C(G_{n1})\ \Theta_{n1,n2}\ C(G_{n2})\ \Theta_{n2,n3} \ldots \Theta_{ni,nk}\ C(G_{nk})$
    - **Acyclic/Cyclic expressions**
  - **Weight metric** specific to analysis requirement
- **Output: A set of HeMLN community, where each has the form**
  - $< c_{n1}^{m1}, c_{n2}^{m2}, \ldots, c_{nk}^{mk} ; x_{n1,n2}, x_{n2,n3}, \ldots, x_{ni,nk} >$
  - First Component: Ordering of **k community ids** from distinct layers in the specification
  - Second Component: Ordering of at **least (k-1) expanded meta edge sets** between communities

Toy Example

Order dependence: Different specification orders give different results



2- and 3-community Specification and Result Representation

$$G_1\,\Theta_{1,2}\,G_2 = \{<c_1^1, c_2^1; x_{1,2}>, <c_1^2, c_2^1; x_{1,2}>, <c_1^3, c_2^4; x_{1,2}>\}$$

$$(G_1\,\Theta_{1,2}\,G_2)\,\Theta_{2,3}\,G_3 = \{<c_1^1, c_2^1, 0;\ x_{1,2}, \Phi>, <c_1^2, c_2^1, 0;\ x_{1,2}, \Phi>, <c_1^3, c_2^4, c_3^2; x_{1,2}, x_{2,3}>\}$$

2- and 3-community Specification and Result Representation

$$G_2\,\Theta_{2,3}\,G_3 = \{<c_2^3, c_3^1; x_{2,3}>, <c_2^4, c_3^2; x_{2,3}>, <c_2^5, c_3^2; x_{2,3}>\}$$

$$(G_2\,\Theta_{2,3}\,G_3)\,\Theta_{2,1}\,G_1 = \{<c_2^3, c_3^1, c_1^3; x_{2,3}, x_{2,1}>, <c_2^4, c_3^2, c_1^3; x_{2,3}, x_{2,1}>, <c_2^5, c_3^2, 0;\ x_{2,3}, \Phi>\}$$

Partial 3-community element

Total 3-community element

Community Detection in HeMLN

Heterogeneous Community Definition

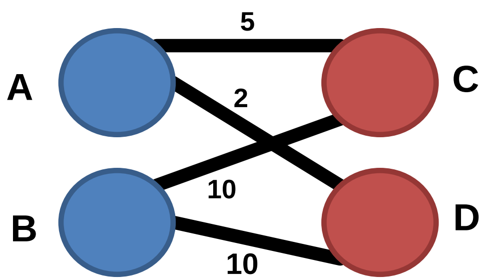
Maximal Weighted Bipartite Coupling (MWBC) Composition

Case Studies

Need For Maximum Weighted Bipartite Coupling

Traditional Maximum Bipartite Matching (Edmonds, 1965)

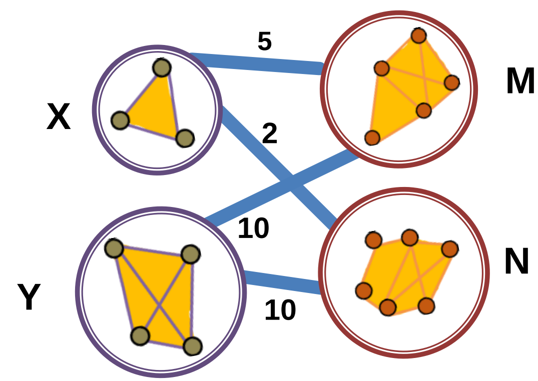
- **Simple nodes** (hiring, dating)
- Weighted Edges supported
- **One to One matching supported,** ties not resolved



TMM Matches: **A – C, B – D**

REQUIREMENT

- Nodes are **Communities**
- Meta edges need to reflect **participating community characteristics**
- **One to Many matching possible in case of ties**



MWBC Matches: **X – M, Y – M, Y – N**

Proposed Weight Metrics for Meta Edge (u, v)

Number of Inter-Community Edges

- $\omega_e = number\ of\ interlayer\ edges\ between\ cu\ and\ cv$
- **Intuition:** *Maximize number of interactions* between the participating communities

Density and Edge Fraction

- $\omega_d = (c_u\ density) * \frac{\omega_e}{|c_u\ nodes| * |c_v\ nodes|} * (c_v\ density)$
- **Intuition**: *Stronger intra-community and inter-community interaction* **that includes participants**

Hub Participation

- $\omega_h = (c_u\ ratio\ of\ hubs\ participating) * \frac{\omega_e}{|c_u\ nodes| * |c_v\ nodes|} * (c_v\ ratio\ of\ hubs\ participating)$
- **Intuition**: *Participation of influential nodes* **within and between** participating communities

Experimental Results

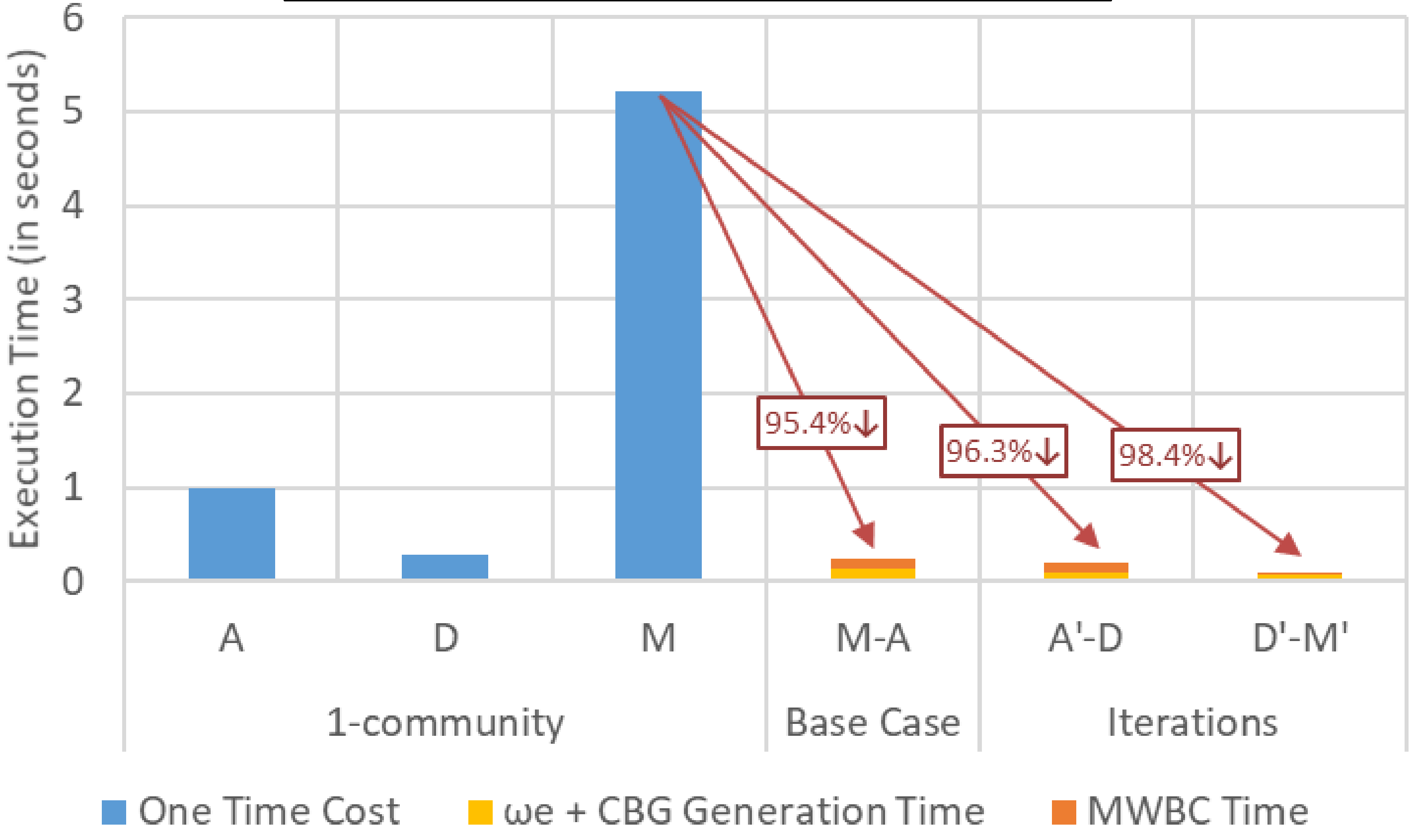
➢ Setup (IMDb HeMLN)

- **Nodes**: 9485 Actors (Layer A), 4510 Directors (Layer D), 7951 Movies (Layer M)
- **Intra-layer edges:** Pearson correlation based similar genres (A and D), Same rating range (M)
- **Inter-layer edges:** acts-in-a-movie (A-M), directs-a-movie (D-M), directs-an-actor (D-A)
- **1-community Detection Algorithm: Louvain**
  - Layer A: 63 communities, Layer D: 61 communities, Layer M: 10 communities

Efficiency

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

M $\Theta_{M,A}$ A $\Theta_{A,D}$ D $\Theta_{D,M}$ M, ω_e



Community Detection in HeMLN

Heterogeneous Community Definition

Maximal Weighted Bipartite Coupling (MWBC) Composition

Case Studies

Real Life HeMLNs

IMDb HeMLN

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

Based on initial set of top 500 actors

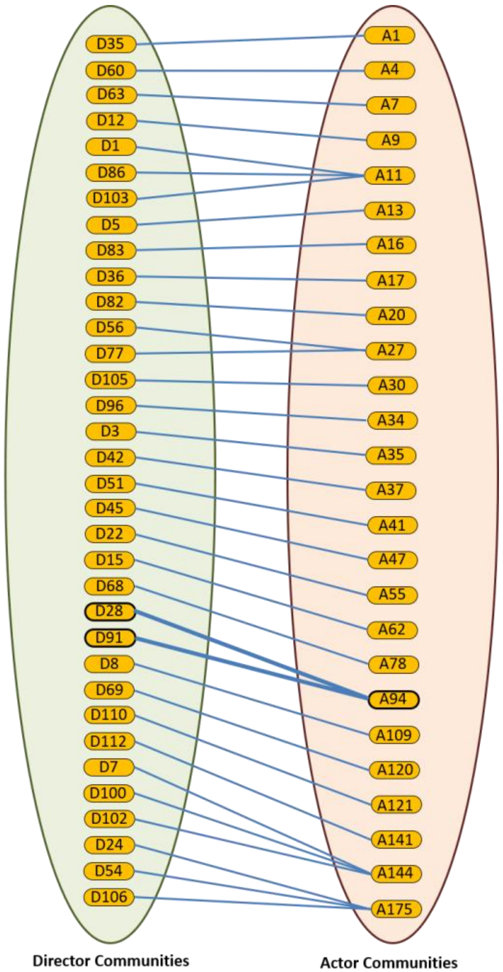
DBLP HeMLN

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

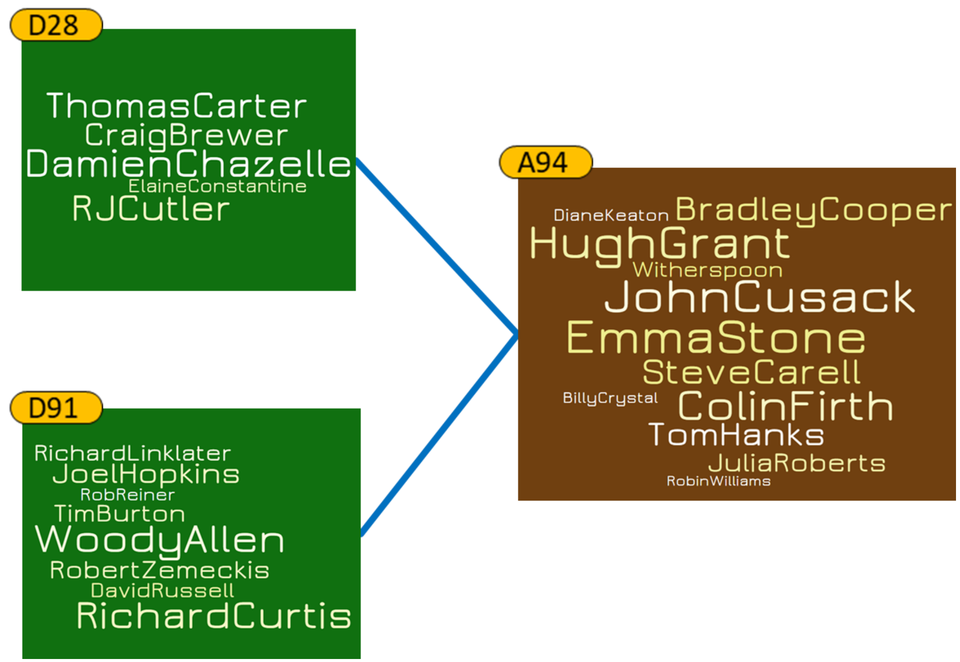
Based on publications in VLDB, SIGMOD, ICDM, KDD, DASFAA, DaWaK from 2001 to 2018

IMDb HeMLN

For each **director group** which are the **actor groups** whose **majority** of the **most versatile members** interact?



C(Directors) $\Theta_{D,A}$ C(Actors), ω_h

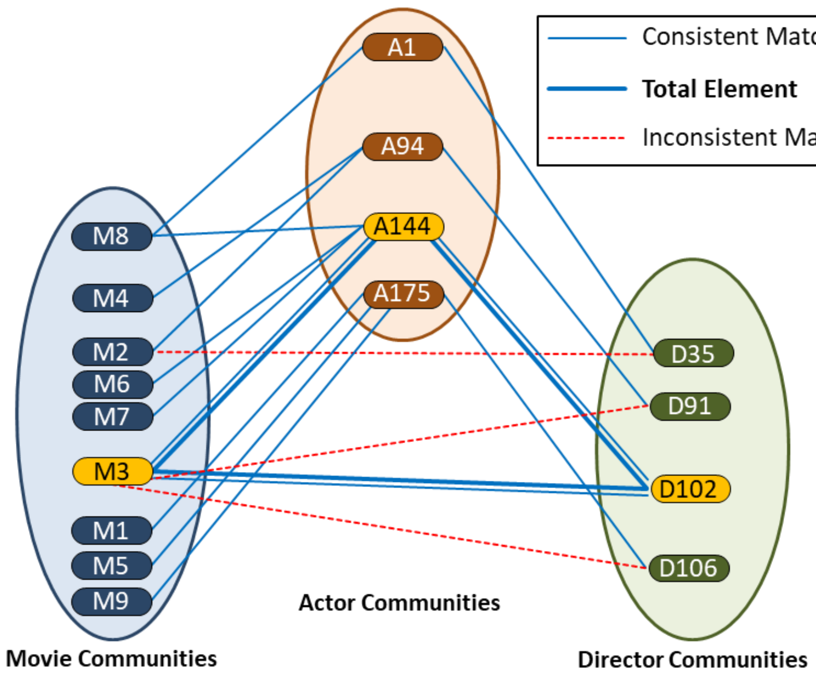


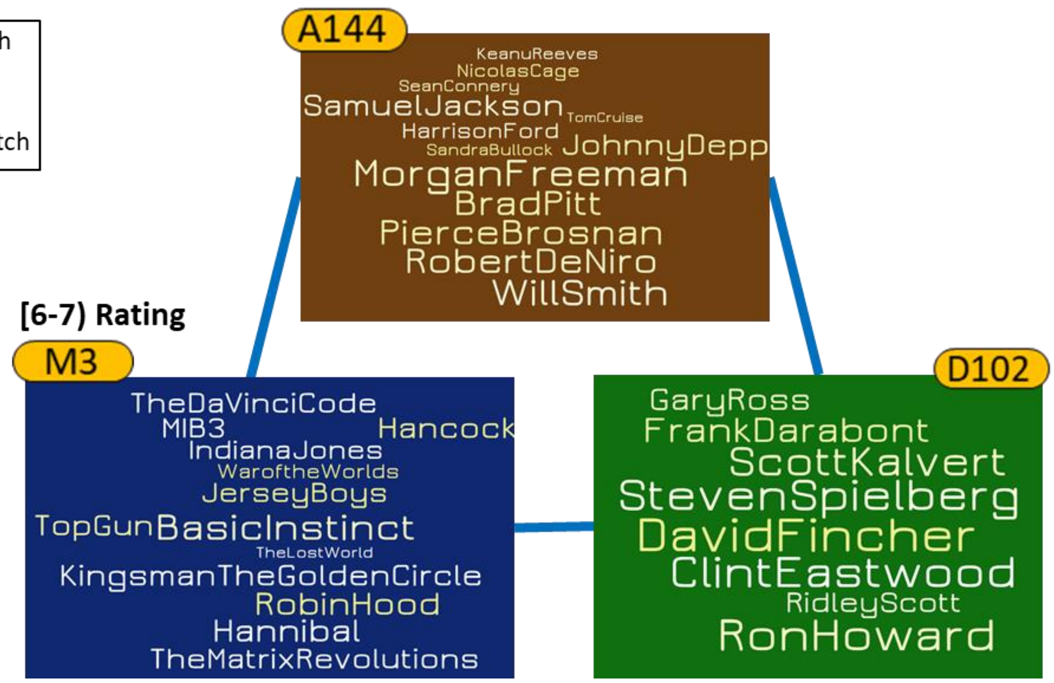
- Academy award winners like **Damien Chazelle** and **Woody Allen** pair up with the actor group with members like **Diane Keaton, Emma Stone** and **Hugh Grant**
- Dominant Genre: **Romance, Comedy and Drama**

IMDb HeMLN

For the ***most popular* actor groups**, for each **movie rating** class, find the **director groups** with which they have ***maximum interaction*** and who also make **movies with similar ratings**

$$C(\text{Movies})\ \Theta_{M,A}\ C(\text{Actors})\ \Theta_{A,D}\ C(\text{Directors})\ \Theta_{D,M}\ C(\text{Movies}),\ \omega_e$$



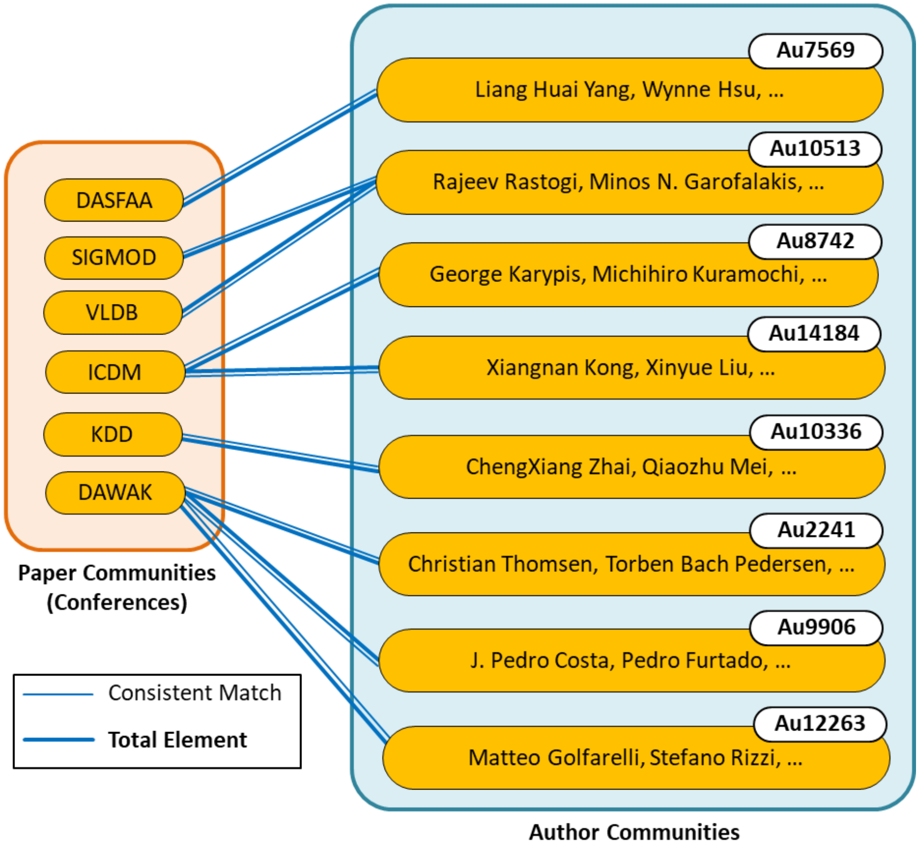


Dominant Genre: Action, Drama

DBLP HeMLN

For each **conference**, which is the ***most cohesive* group of authors** who ***publish frequently***?

C(Papers) $\Theta_{P,Au}$ C(Authors), ω_d

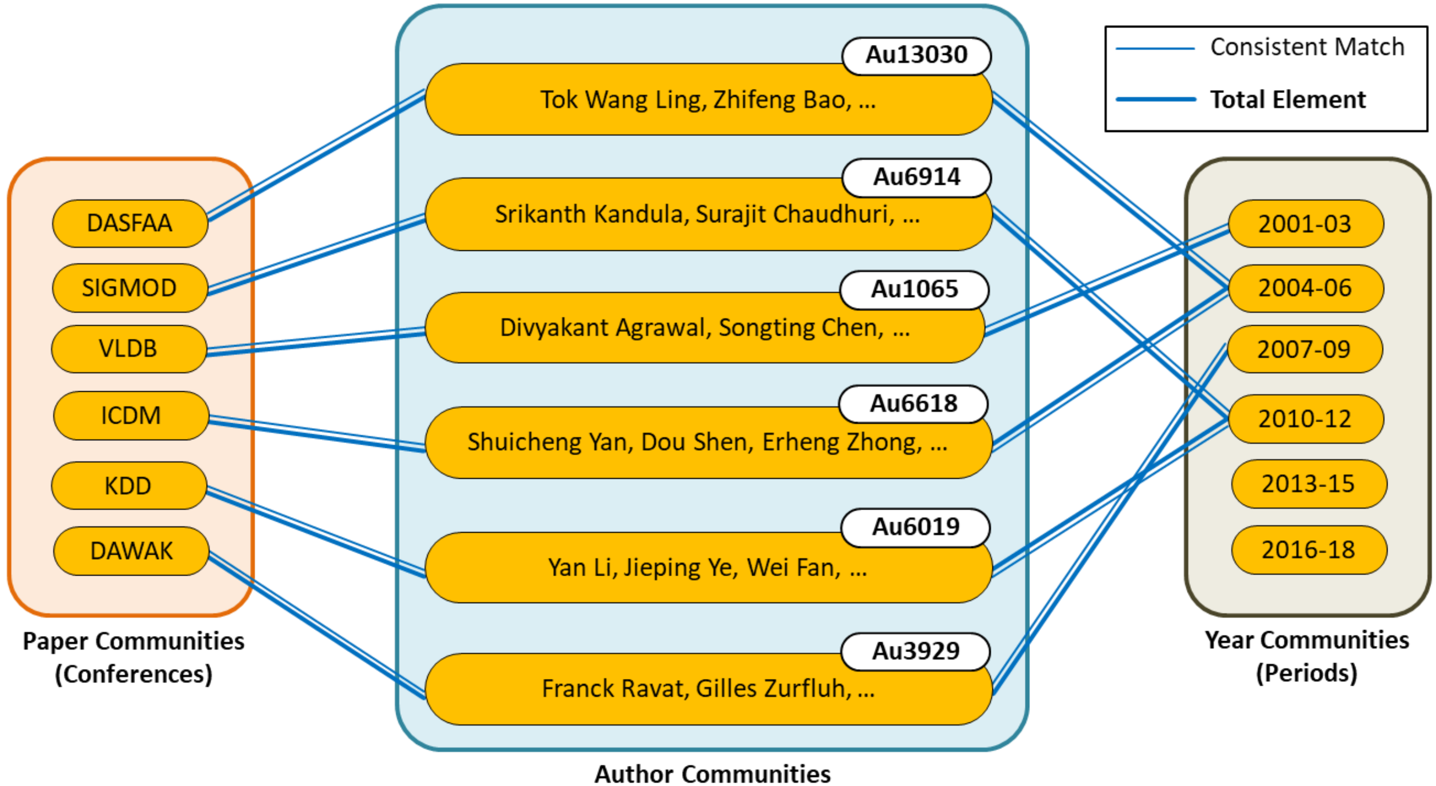


- **ICDM** and **DaWaK** have *multiple author communities* that are equally important
- **George Karypis** and **Michihiro Kuramochi** are members of one of the frequently publishing co-author groups for **ICDM (4 papers)**
  - **Validating fact:** George Karypis recipient of *IEEE ICDM 10-Year Highest-Impact Paper Award (2010)* and *IEEE ICDM Research Contributions Award (2017)*
- Co-authors **Rajeev Rastogi** and **Minos N. Garofalakis** are strongly associated with **SIGMOD (7 papers)** and **VLDB (4 papers)** in the past 18 years

DBLP HeMLN

For the ***most popular*** **collaborators** from each **conference**, which are the **3-year period(s)** when they were **most active**?

$C(Papers)\ \Theta_{P,Au}\ C(Authors)\ \Theta_{Au,Y}\ C(Years),\ \omega_e$



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

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

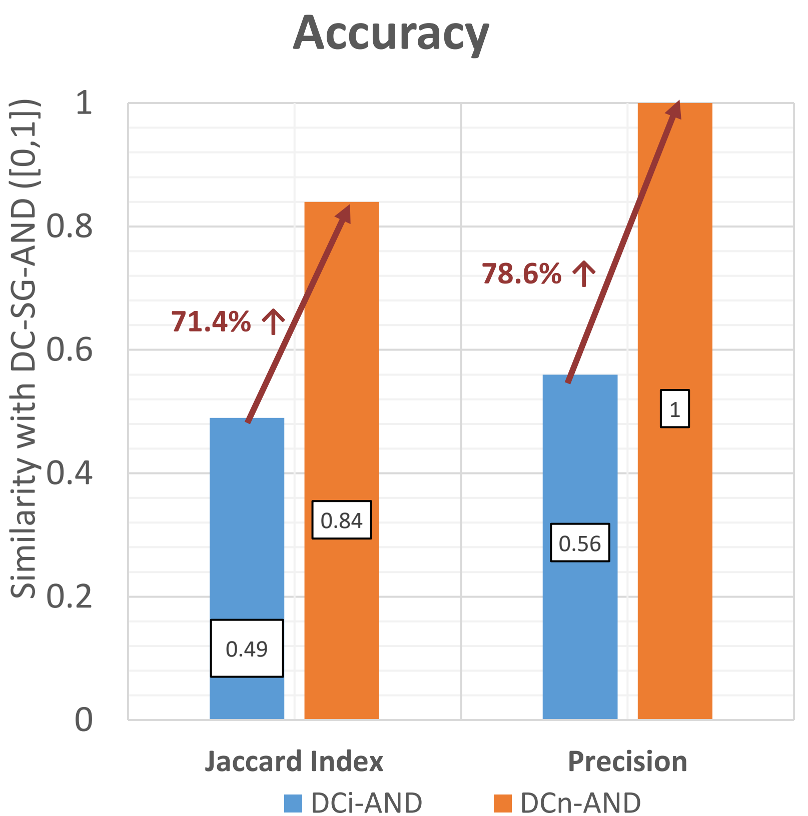
Degree Centrality Heuristics

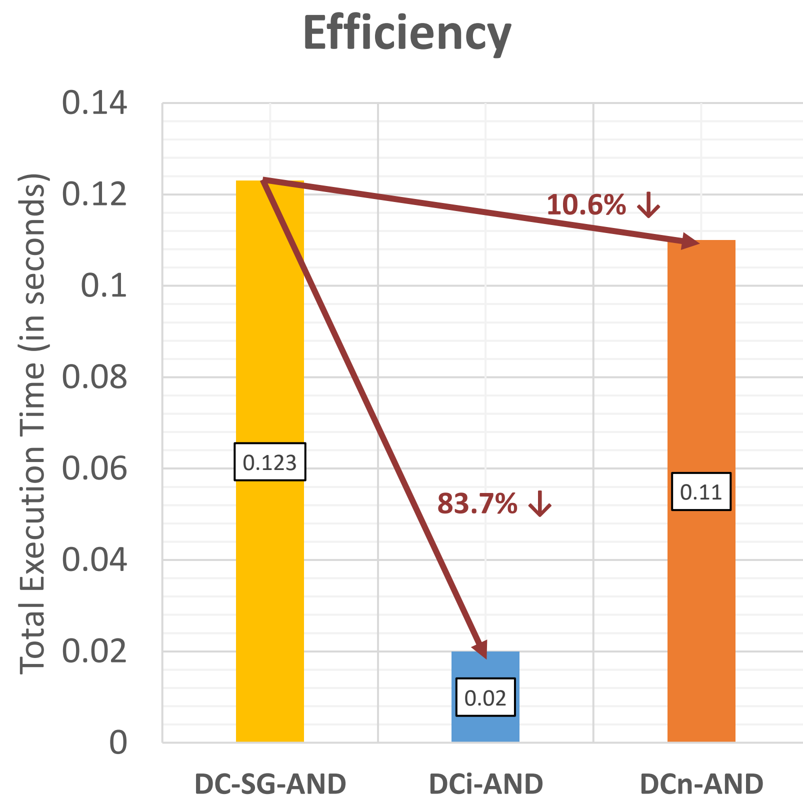
- **DCi-AND: Intersect** the layer-wise hubs
  - **Layers have similar topology:** High Accuracy, Low Overhead
  - **In general, low accuracy** due to presence of *false positives and negatives*
- **DCn-AND:** Check if the **common hubs** have **enough shared neighbors**
  - **Additional Overhead**
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





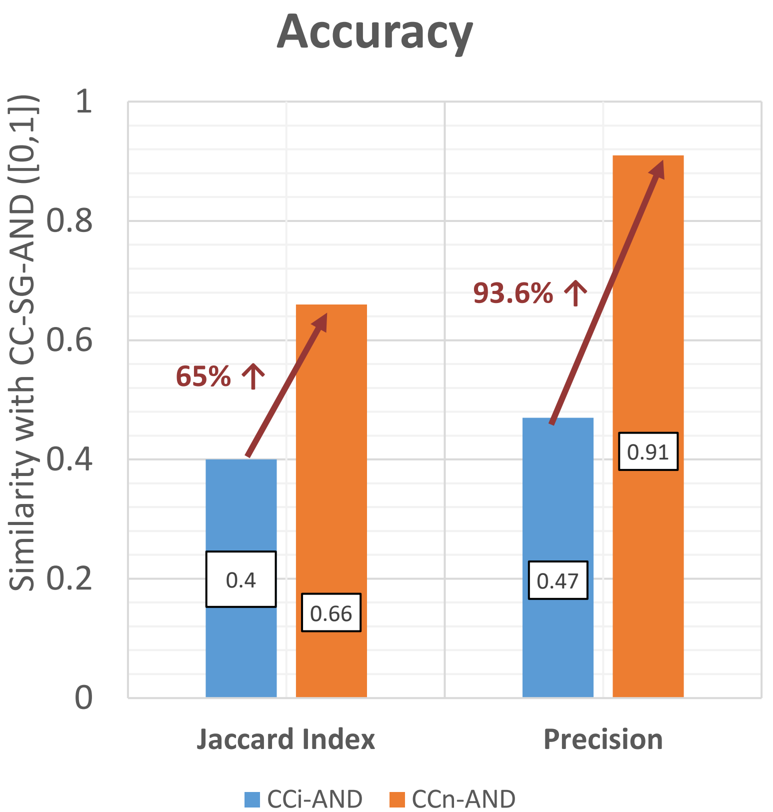
Elimination of False Positive increases the **Precision, Decreases Efficiency**

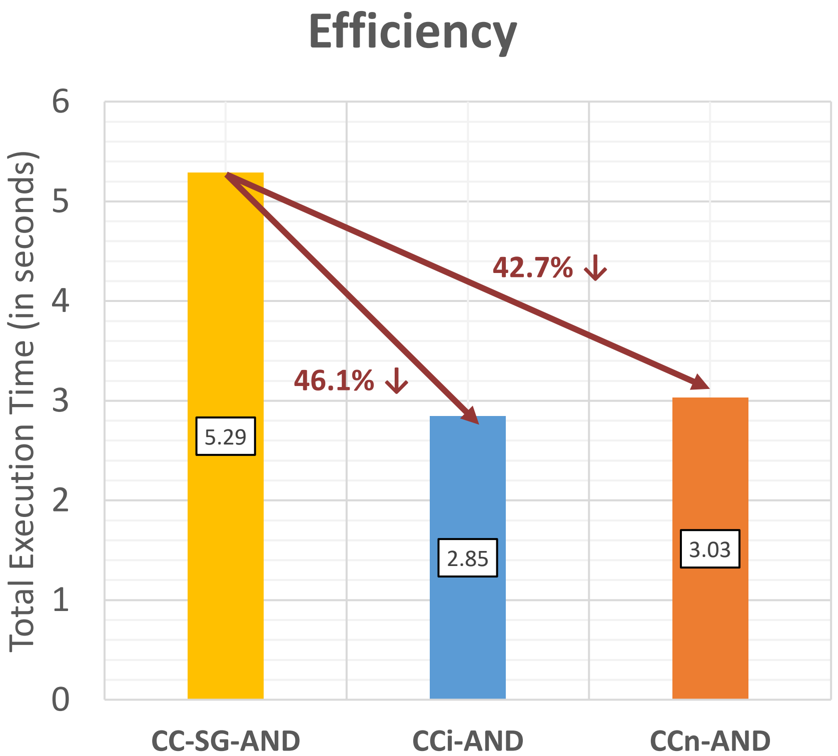
For **large N** (number of MLN layers), **denser layers, more analysis: Efficiency is higher**

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

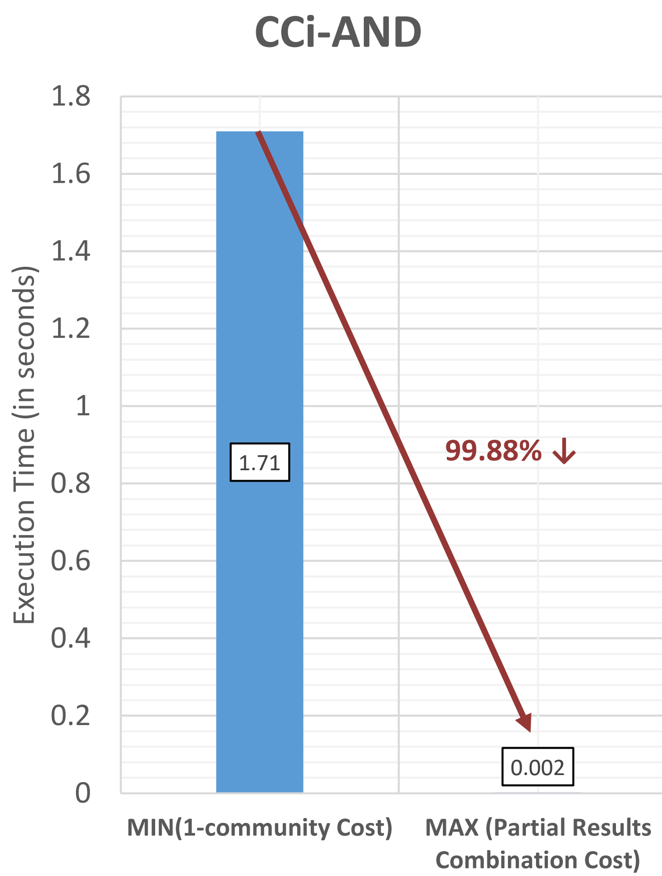


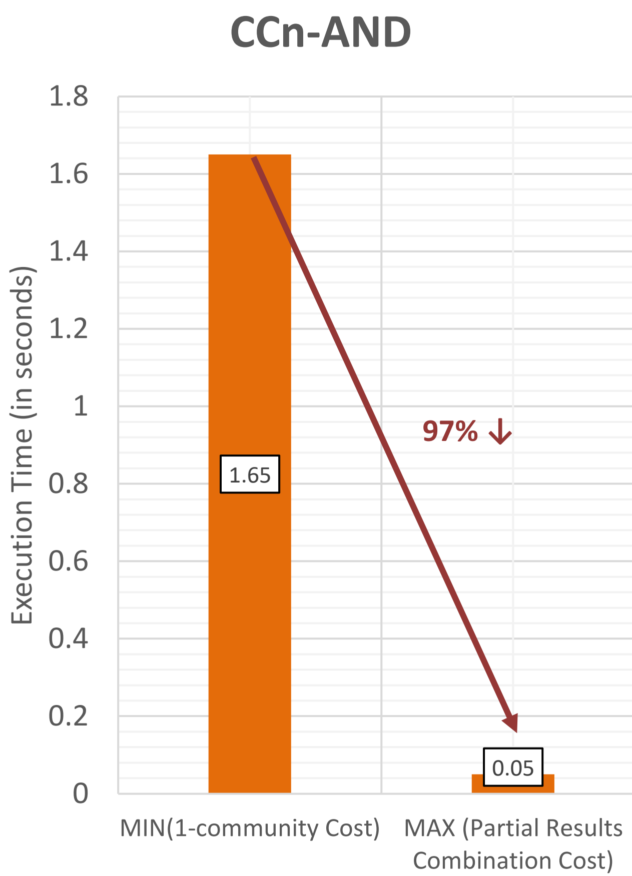


Decrease in False Positives increases the **Precision, Decreases Efficiency**

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

Real Life HoMLNs

US Airline HoMLN

	#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**

US Airline HoMLN

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))

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*

Related Reading

Publications

Publications

- **Abhishek Santra**, Sanjukta Bhowmick: Holistic Analysis of Multi-source, Multi-feature Data: Modeling and Computation Challenges. **BDA 2017**
- **Abhishek Santra**, Sanjukta Bhowmick, Sharma Chakravarthy: Efficient Community Re-creation in Multilayer Networks Using Boolean Operations. **ICCS 2017**
- **Abhishek Santra**, Sanjukta Bhowmick, Sharma Chakravarthy: HUBify: Efficient Estimation of Central Entities Across Multiplex Layer Compositions. **ICDM Workshops 2017**
- Xuan-Son Vu, **Abhishek Santra**, Sharma Chakravarthy, Lili Jiang: Generic Multilayer Network Data Analysis with the Fusion of Content and Structure. **CICLing 2019**
- **Abhishek Santra**, Kanthi Sannappa Komar, Sanjukta Bhowmick, Sharma Chakravarthy: Structure- And Semantics-Preserving Community Definitionand Its Computation For Heterogeneous Multilayer Networks. **TKDE 2020** *(In Preparation)*
- **Abhishek Santra**, Sanjukta Bhowmick, Sharma Chakravarthy: Efficient Community Detection in Boolean Composed Multilayer Networks. **TKDD 2020** *(In Preparation)*
- **Abhishek Santra**, Kanthi Sannappa Komar, Sanjukta Bhowmick, Sharma Chakravarthy: Data-Driven Aggregate Analysis of MLNs: Modeling, Computation, and Versatility. **DASFAA 2020** *(Under Review)*
- Sharma Chakravarthy, **Abhishek Santra**, Kanthi Sannappa Komar: Humble Data Management to Big Data Analytics/Science: A Retrospective Stroll. **BDA 2018**

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 (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

Questions?

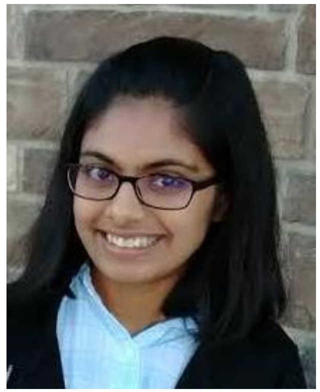
UNIVERSITY OF TEXAS ARLINGTON

UNT UNIVERSITY OF NORTH TEXAS

UMEÅ UNIVERSITY

Sharma Chakravarthy
Professor

Abhishek Santra
PhD Candidate



Kanthi Komar
MS Thesis Alumna

Sanjukta Bhowmick
Associate Professor

Lili Jiang
Assistant Professor

Xuan-Son Vu
PhD Candidate

For more information visit:

http://itlab.uta.edu

References

- 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.
- Chakravarthy, S., Santra, A. and Komar, K.S., 2018, December. Humble data management to big data analytics/science: A retrospective stroll. In International Conference on Big Data Analytics (pp. 33-54). Springer, Cham.
- Vu, X.S., Santra, A., Chakravarthy, S. and Jiang, L., 2019. Generic multilayer network data analysis with the fusion of content and structure. In International Conference on Computational Linguistics and Intelligent Text Processing