Holistic Analysis of Multi-Source, Multi-Feature Data: Modeling and Computation Challenges

Abhishek Santra[1] and Sanjukta Bhowmick[2]

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

Special Thanks to **Sharma Chakravarthy** (Professor, UT Arlington)

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?**

Multi Source, Multi Feature Data

- **Data Characteristic**: Multiple relationships existing among same or different type(s) of entities

Multi Source, Multi Feature Data

- **Data Characteristic**: Multiple relationships existing among **same** or different **type**(s) **of entities**

Interaction among a set of people

Most popular or socially active group of people?

Most influential set of people?

Multi Source, Multi Feature Data

- **Data Characteristic**: Multiple relationships existing among **same** or different **type**(s) **of entities**

Airline connectivity among a set of US cities

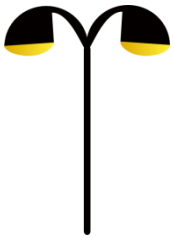
Airline connectivity among a set of Indian cities

Highly central cities (hubs)?

Multi Source, Multi Feature Data

- **Data Characteristic**: Multiple relationships existing among **same** or different **type**(s) **of entities**

Similarity among a set of traffic accidents



Light Conditions

Weather Conditions

Road Surface Conditions

Accident Prone Regions?

Most dominant feature? Order of features?

Multi Source, Multi Feature Data

- **Data Characteristic**: Multiple relationships existing among same or **different type(s) of entities**

Connectivity among scientists, cities and conferences

Collaboration

Direct Flights

Research Domains

Attendance

Residences

Venues

Best city to hold a workshop?

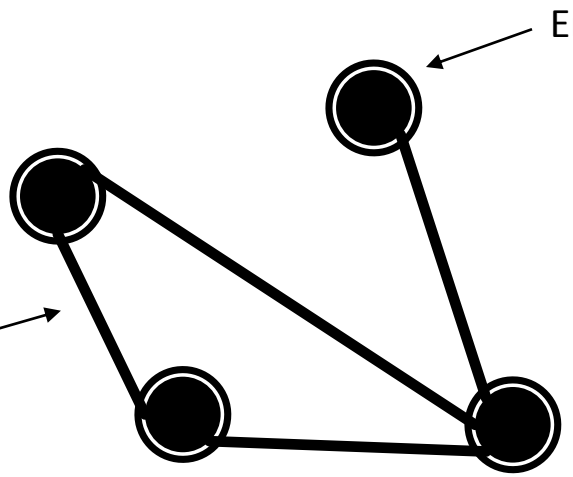
Which research domains are prevalent in which state?

Multi Source, Multi Feature Data

- **Data Characteristic**: Multiple relationships existing among same or different type(s) of entities
  - Explicit relationships (friends, flights …) or based on similarity metric (numeric, location, video, time …)
- **Holistic or Flexible Analysis:** Study effect of *different combinations of features or perspectives*
- Two Challenges
  - **Modeling**
  - **Processing/Computations**

Traditional Approach: Monoplex Modeling

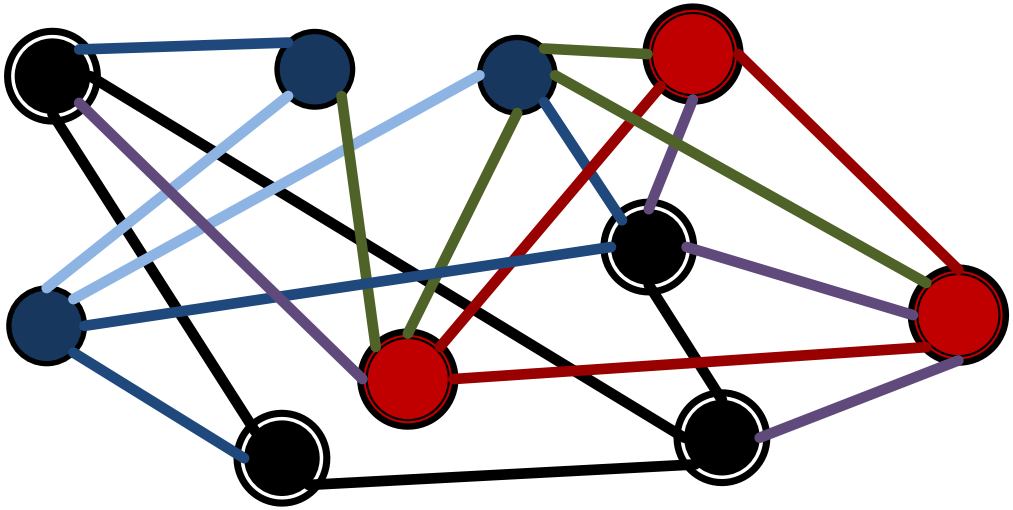
- Single Edge Monoplex

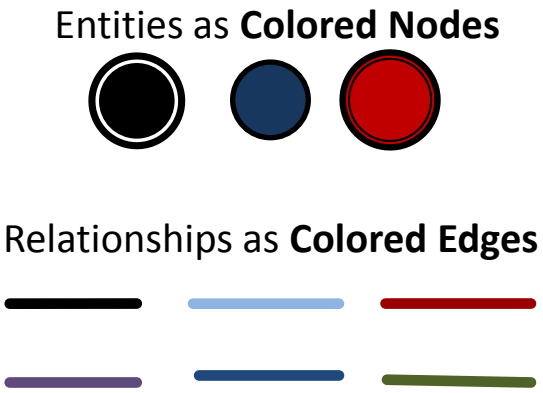


- Drawbacks
  - Need to be generated for every **feature combination**
  - Repeated Dataset Scanning
  - Repeated Similarity Metric Computation

Traditional Approach: Monoplex Modeling

- Multi Edge Monoplex





- Drawbacks
  - Loss of partial analysis results w.r.t. different feature combinations
  - Repeated Graph Traversals to extract any feature combination subgraph
  - Convoluted Representation

Multiple Monoplexes = Multiplex Modeling

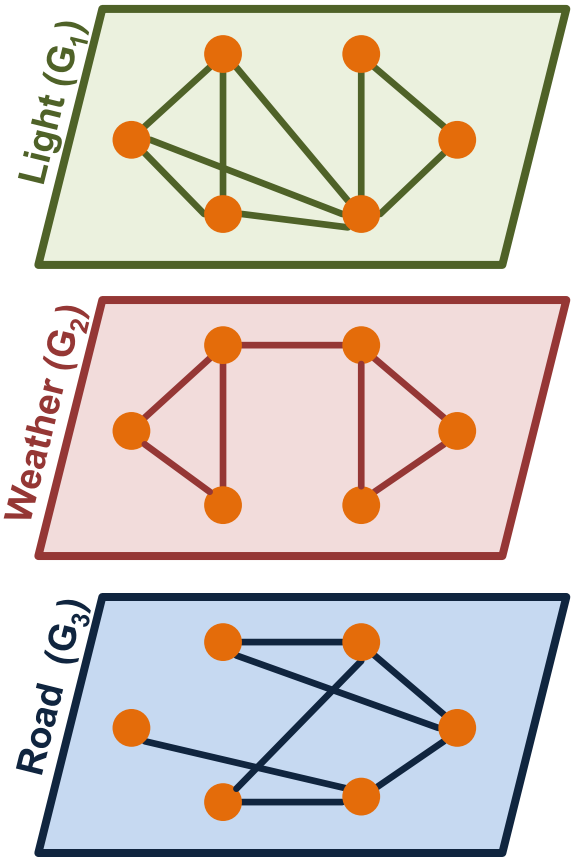
- Non-Suitability of Monoplex Modeling
  - Computations will obscure information
  - Computationally Expensive
  - Convoluted Representations
- Use **Multiplexes or Multilayer Networks**
  - A ***network of networks***
  - **Each layer**/network represents a **single perspective or feature**
  - Differentiated into **3 types based on type of entities**

Multiplex Modeling
(Same Entities, Different Relationships)

Homogeneous Multiplex

Multiple relationships among **same type of entities**

- **Similarity of disasters** (**accidents**, storms etc.) based on factors
- **Interaction among people** via various media (social media, calls etc.)
- **Connectivity among cities** based on different airlines



Accident Multiplex

Multiplex Modeling
(Different Entities, Different Relationships)

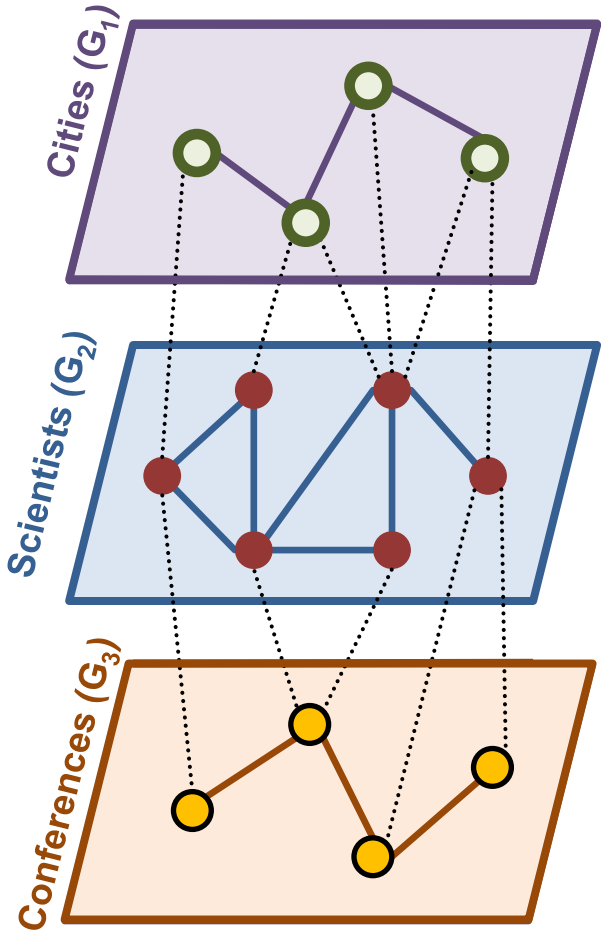
Heterogeneous Multiplex

Multiple relationships among **different types of entities**

- Residence, venue and attendance connectivity among ***city airline***, ***scientist collaboration*** and ***similar conference*** networks

Hybrid Multiplex

- Any layer of the heterogeneous multiplex can be expanded to a homogenous multiplex
  - Different friendship networks among scientists



City-Scientist-Conference Multiplex

Benefits of Multiplexes

- **Flexible analysis**
  - **Mixing required layers** in **arbitrary ways**
    - **Homogeneous (Boolean, Linear etc.)**
    - **Heterogeneous (Projection, Type Independent etc.)**
- **Parallel algorithms** can be leveraged
- Ease of **handling the dataset incrementally**
  - **Addition** of new entities (nodes), relationship with existing entities (edges) and features/perspectives (layers)
- Efficient maintenance of the **latest dataset snapshot**
  - **Deletion** of obsolete entities (nodes), relationships (edges) and features/perspectives (layers)

Multi Feature Computations using Multiplexes

- **Multiplex-based analysis is at a nascent stage**
  - Layers considered **individually or all layers aggregated together**, in specific **sub-disciplines**
  - **Hardly any work on mining and querying multiplexes**
- Existing algorithms for monoplexes can be leveraged
  - Need to **generate, store and analyze each layer combination**
  - **N individual layers ⇒ $O(2^N)$ layer combinations!**
  - **Exponential overall computational costs (storage and time), if N is large**

Multiplex based Holistic Analysis

(Need of the Hour)

- Develop **cost effective** and **robust algorithms** for **flexibly analyzing any layer combination** to
  - Compute **network characteristics**,
  - **Mine interesting hidden patterns**, and
  - **Query**
- **Specific computations and related challenges** vary with the type of multiplex under consideration

Homogeneous Multiplex Computations

(Requirements)

- Devise efficient techniques for **generating interesting network structures** with respect to **any combination of multiplex layers**
- **Communities:** Tightly connected group of nodes
  - **Effectiveness of accident prevention techniques**
    - Variation of **accident prone regions** over time
- **Hubs:** Highly central nodes
  - **Maximize the reach of an advertisement**
    - Most **influential people** across social media

Homogeneous Multiplex Computations

(Requirements)

- Develop methods to compute the **relative ordering and correlation among different feature (or layer) combinations** based on their importance
  - **Allocation of funds for accident prevention**
    - Ordering factors based on impact on accident occurrence
- Identify metrics to **quantify the importance of a layer (or layer combination)**
  - Alternatives: Density, Number of Influential Nodes (Hubs), Core-Periphery Structure, Local and Global Clustering Coefficient

Homogeneous Multiplex Computations

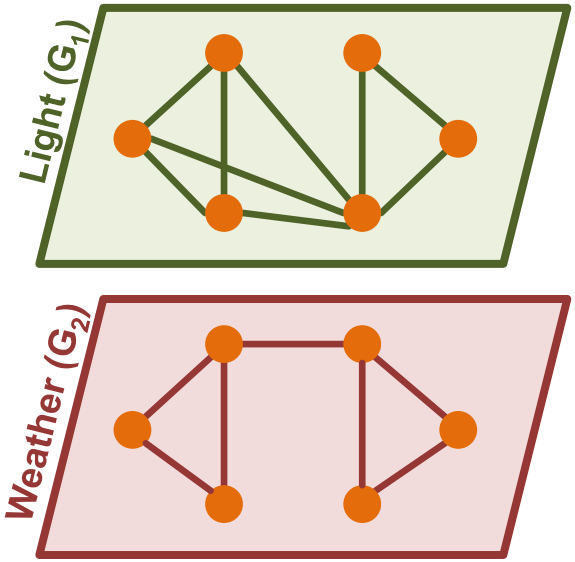
(Challenges)

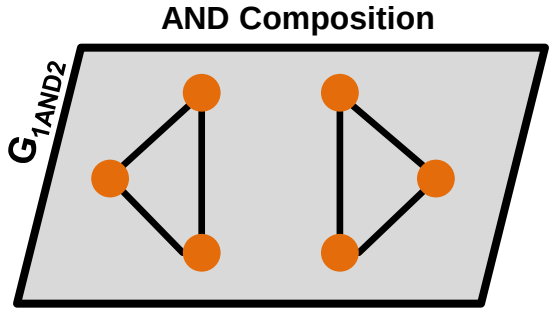
- Need to reduce the **exponential complexity**
- Formulation of **efficient aggregation functions** that can **combine the results from N individual layers to compute the results of any layer combination**
  - **Diverse formats of layer-wise results**: Substructures (communities), Real Numbers (density, clustering coefficients), Sets (high centrality nodes, nodes in inner core)
  - **Analyze performance** of aggregation functions
    - **Accuracy** using NMI, ARI, Purity and Jaccard Index
    - Obtain **Confidence Intervals** for different characteristics of layer-wise results
    - **Savings in computational time and storage space** by eliminating the need to generate, store and analyze each layer combination

Homogeneous Multiplex Computations

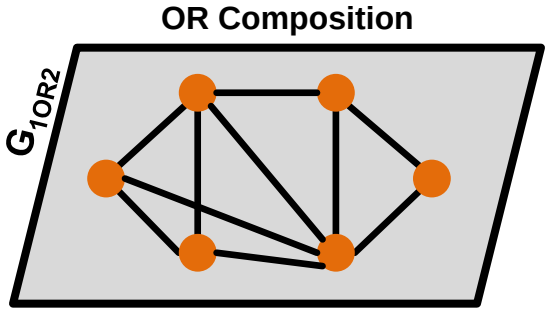
(Preliminary Work – Boolean Composition)

- Proposed Multiplex Layer Compositions though **Boolean Operations – AND, OR, NOT**

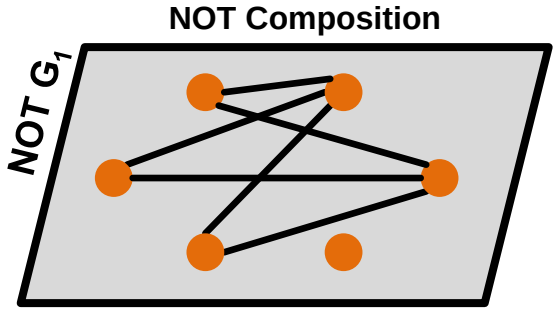




Relationships present in *all* layers



Relationships present in *at least one* layer

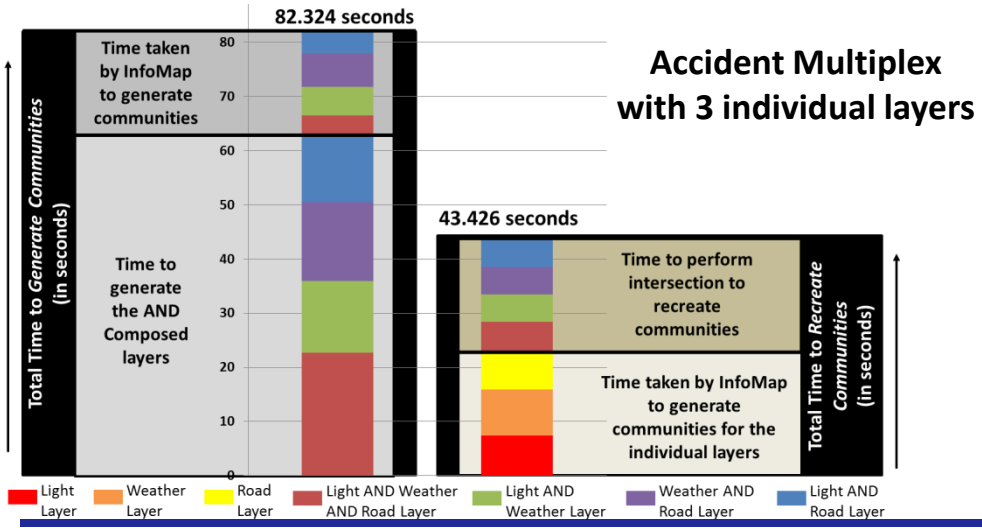


Relationships *not* present in a layer

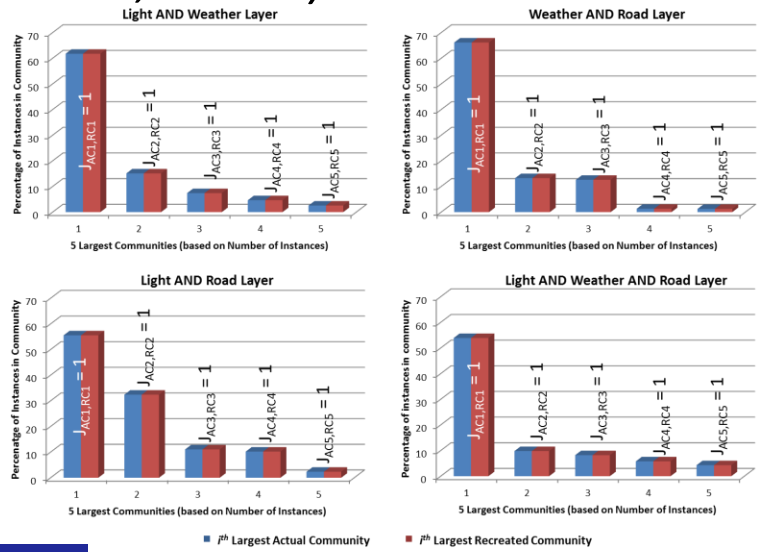
Homogeneous Multiplex Computations

(Preliminary Work – Recreating Communities)

- Proposed Multiplex Layer Compositions though **Boolean Operations – AND, OR, NOT**
- Proposed an accurate **intersection-based community re-creation technique for any AND-composed multiplex layer using layer-wise communities***
- **Reduced the overall computation time by over 40%** *(with real-life multi-feature datasets – traffic accidents, storms)*



Accident Multiplex with 3 individual layers



*with individual layers having self-preserving communities

Homogeneous Multiplex Computations

(Related Publications)

- Scalable Holistic Analysis of Multi-Source, Data-Intensive Problems Using Multilayered Networks – *CoRR abs 2016*
- Efficient Community Re-creation in Multilayer Networks Using Boolean Operations – *ICCS 2017*
- HUBify: Efficient Estimation of Central Entities across Multiplex Layer Compositions – *ICDM-W 2017*

Homogeneous Multiplex Computations

(Preliminary Work – Estimating Hubs)

- Hubs defined as the high degree or closeness centrality vertices
- Proposed **efficient heuristics to estimate the high centrality vertices for any AND-composed multiplex layer using the layer-wise hub sets**
- Overall average **accuracy of at least 70-80%, Reduced the overall computation time by over 30%** *(with real-life multi-feature datasets – traffic accidents, IMDb)*

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

Performance of degree centrality based heuristic (DC1) on IMDb Multiplex for co-actors with Comedy (m_1), Action (m_2) and Drama (m_3) individual genre-based layers

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

Performance of closeness centrality based heuristic (CC1) on Accident Multiplex for similar accidents with Light (a_1), Weather (a_2) and Road Surface (a_3) Conditions individual layers

Homogeneous Multiplex Computations

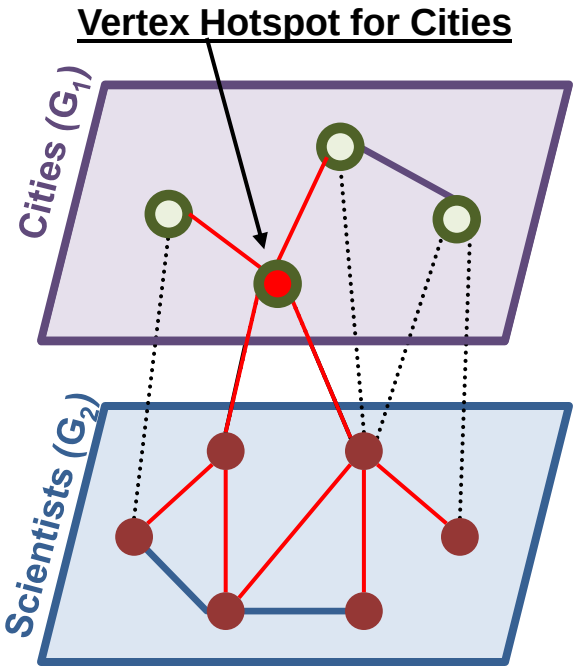
(Related Publications)

- Scalable Holistic Analysis of Multi-Source, Data-Intensive Problems Using Multilayered Networks – *CoRR abs 2016*
- Efficient Community Re-creation in Multilayer Networks Using Boolean Operations – *ICCS 2017*
- HUBify: Efficient Estimation of Central Entities across Multiplex Layer Compositions – *ICDM-W 2017*

Heterogeneous Multiplex Computations

(Requirements and Challenges)

- Devise efficient ways to **compute high centrality vertices (or communities) across multiple connected layers**
  - **Best city to organize a workshop**
- **Aggregation functions** need to take into account results of the **bipartite graph formed by the inter-layer edges**
- Layers may be connected not just sequentially one after another (i.e. layer A connected to layer B connected to layer C) but can be connected in different directions (i.e. **layers A, B and C can be connected to each other in a triangle**)

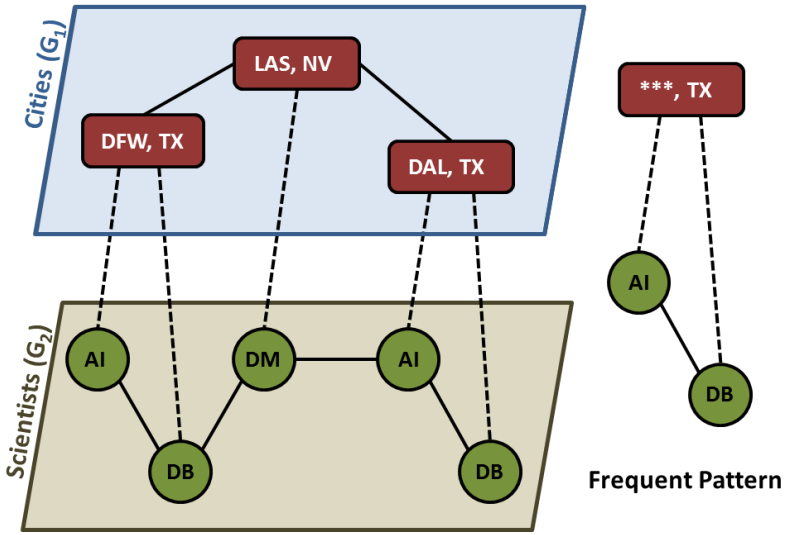


Vertex Hotspot for Cities with respect to the Scientists

Heterogeneous Multiplex Computations

(Requirements and Challenges)

- Develop algorithms for **mining on multiplexes**
  - Define the notion of **subgraphs and patterns in a multiplex**
  - Establish metrics like **MDL** and **frequency**
  - Define **Exact and Similar (or in-exact) substructures**
  - Devise strategies to **partition a multiplex**



Example of a
frequent city-collaboration pattern

Heterogeneous Multiplex Computations

(Requirements and Challenges)

- Develop **query processing algorithms for queries on multiplexes**
  - **Node degree based:** Cities where scientists attending *most number of conferences* reside
  - **Community based:** Cities where scientists belonging to *largest group of collaborators* reside
  - **Path based:** Best possible city where a well-connected group of collaborators can meet up by taking the *minimum number of flights*

Heterogeneous Multiplex Computations

(Requirements and Challenges)

- Develop **query processing algorithms for queries on multiplexes**
  - Determine the **order (in parallel or as a partial order) to process layers** for efficiency
  - Generating **metric to evaluate alternate query plans**
  - Evaluate the suitability of an **index-based or substructure expansion-based approach**
  - Identify **query processing requirements** in terms of the graph properties

Heterogeneous Multiplex Computations

(Related Publications)

- Heterogeneous Community Detection using Network Composition – *KDD 2018 (under preparation)*

Conclusions

- **Multiplexes have modeling and computational advantages** for holistically analyzing multi source, multi feature data
- **Computational challenges** identified in this paper are being addressed by the research community
- **Efficient techniques** as solutions to these challenges will **enrich the data analytics repertoire** making it **easier to analyze such a class of big data problems**

Questions?

Sharma Chakravarthy
Professor
sharma@cse.uta.edu

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

Sanjukta Bhowmick
Associate Professor
sbhowmick@unomaha.edu

For more information visit:

http://itlab.uta.edu