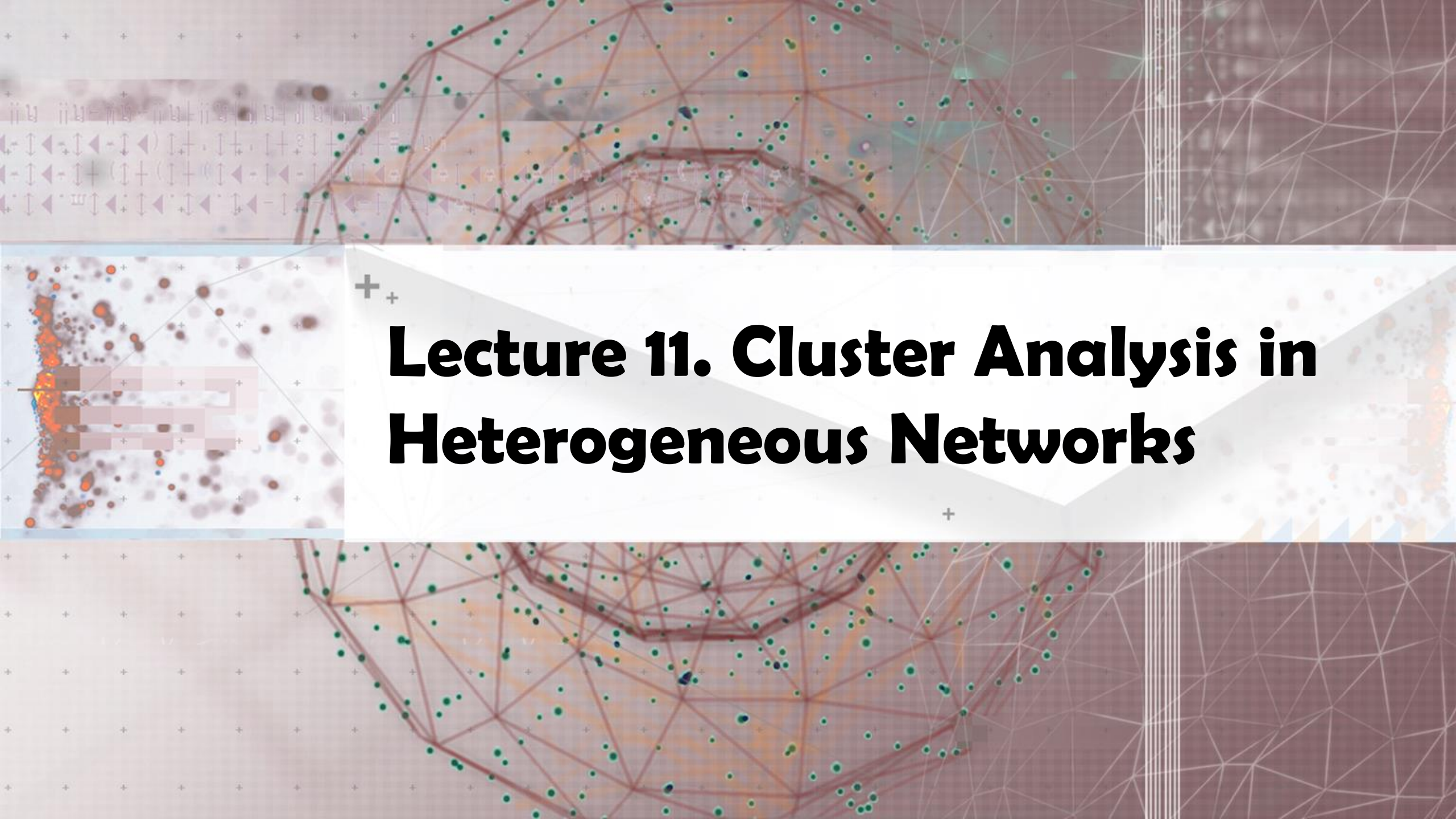
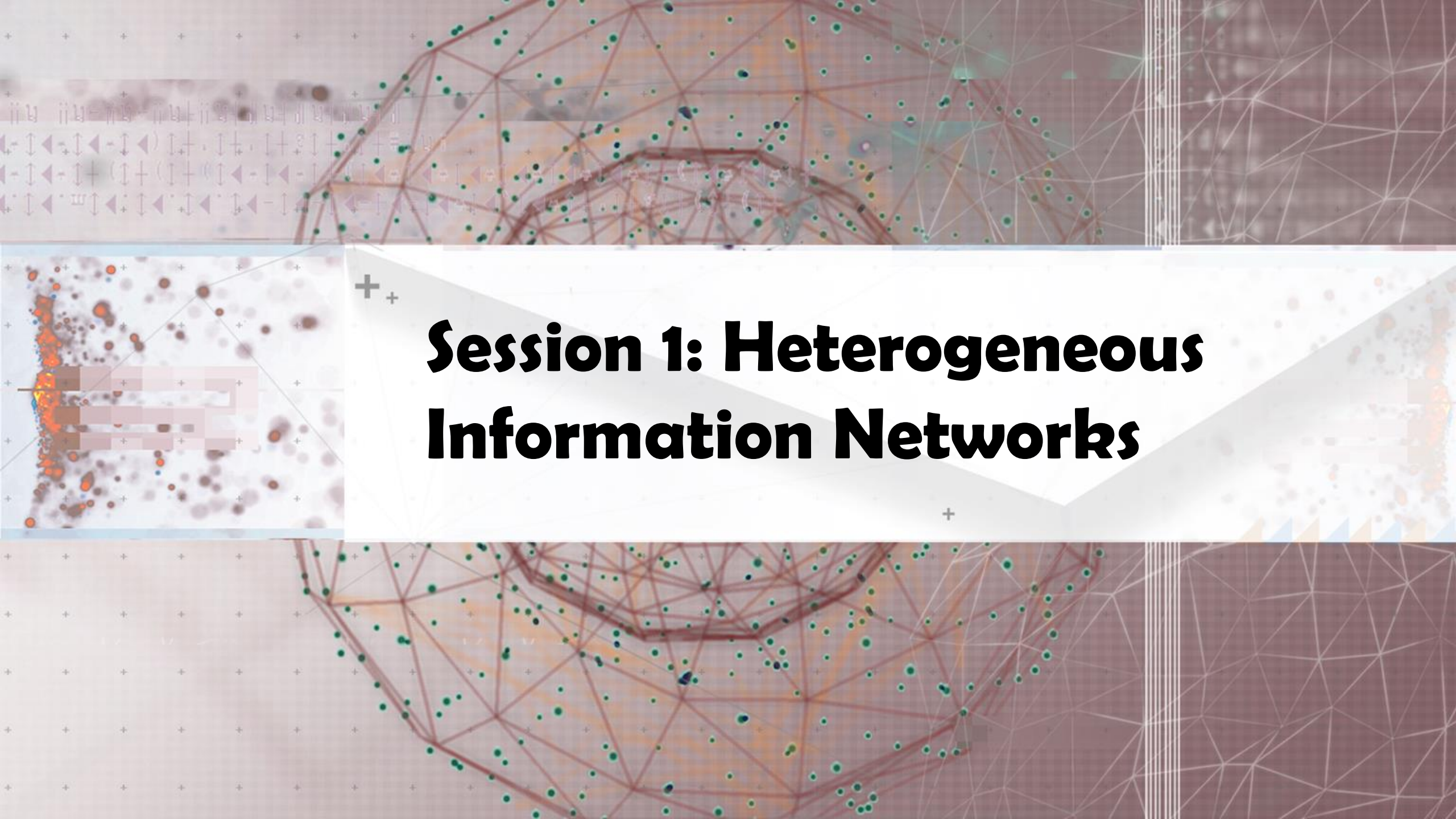


Lecture 11. Cluster Analysis in Heterogeneous Networks

Lecture 11. Cluster Analysis in Heterogeneous Networks

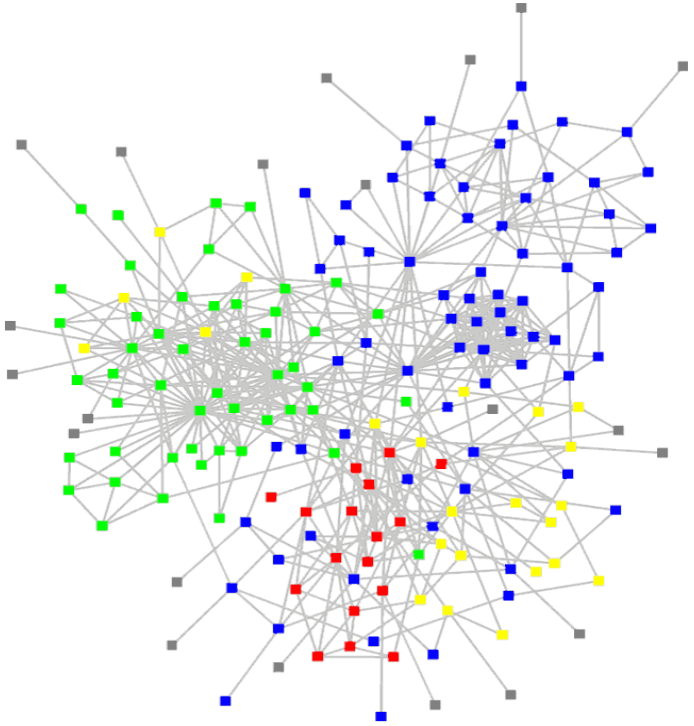
- Heterogeneous Information Networks
- RankClus: Integrated Clustering and Ranking in Heterogeneous Networks
- NetClus: Ranking-Based Clustering with Star Network Schema
- PathSim: Path-Based Similarity Measure for Heterogeneous Networks
- User Guided Meta-Path Selection for Clustering in Heterogeneous Networks
- Summary



Session 1: Heterogeneous Information Networks

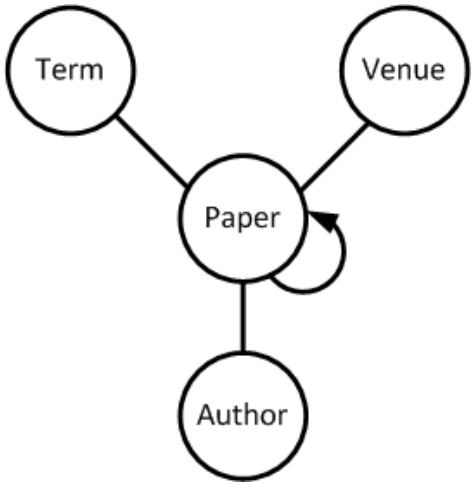
What Are Heterogeneous Information Networks?

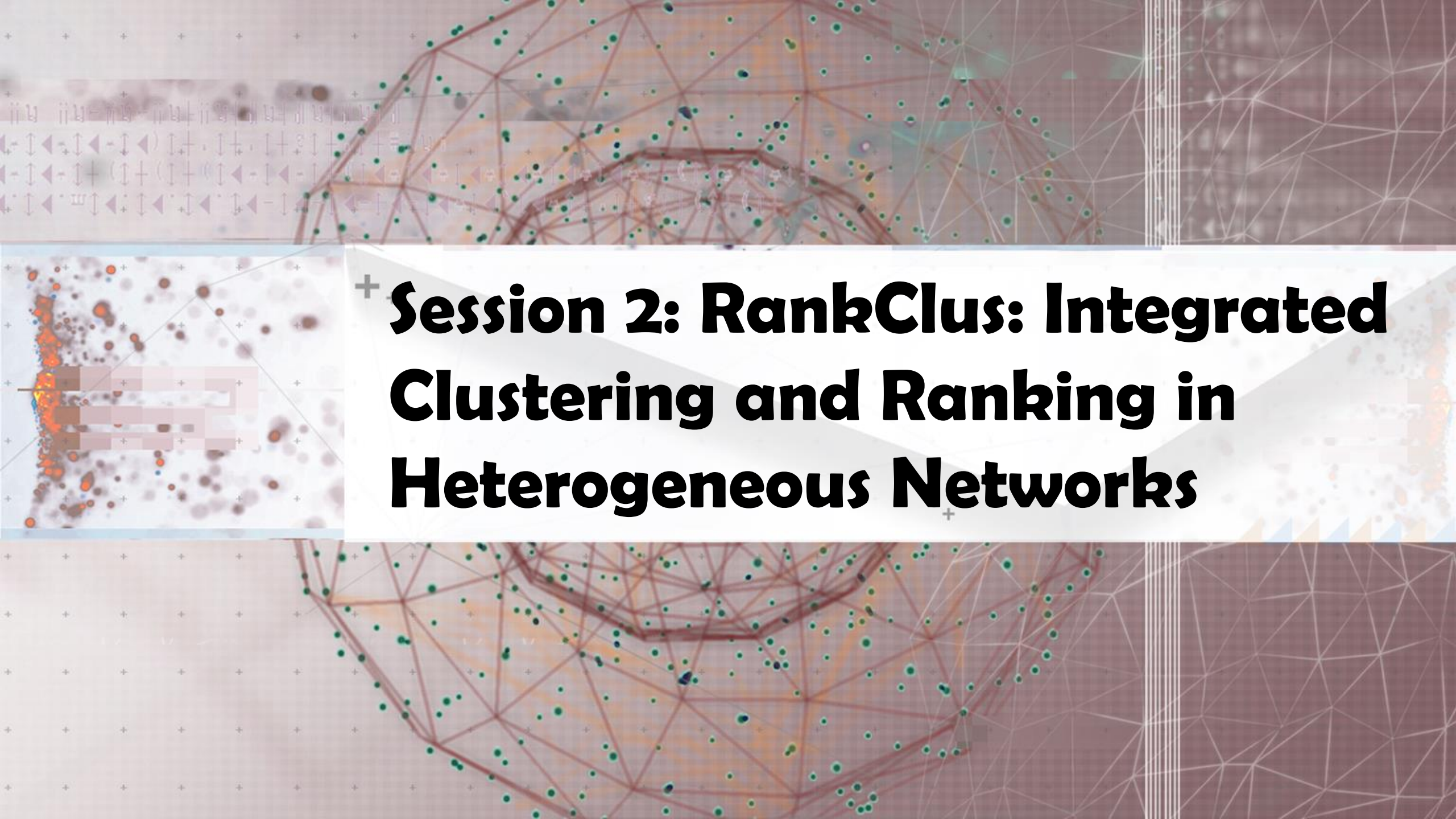
- Information network: A network where each node represents an entity (e.g., actor in a social network) and each link (e.g., tie) a relationship between entities
  - Each node/link may have attributes, labels, and weights
  - Link may carry rich semantic information
- Homogeneous vs. heterogeneous networks
  - Homogeneous networks
    - Single object type and single link type
    - Single model social networks (e.g., friends)
    - WWW: A collection of linked Web pages
  - Heterogeneous, multi-typed networks
    - Multiple object and link types
    - Medical network: Patients, doctors, diseases, contacts, treatments
    - Bibliographic network: Publications, authors, venues (e.g., DBLP > 2 million papers)



Mining Heterogeneous Information Networks

- Homogeneous networks can often be derived from their original heterogeneous networks
  - Ex. Coauthor networks can be derived from author-paper-conference networks by projection on authors
  - Paper citation networks can be derived from a complete bibliographic network with papers and citations projected
- Heterogeneous networks carry richer information than their corresponding projected homogeneous networks
- Typed heterogeneous network vs. non-typed heterogeneous network (i.e., not distinguishing different types of nodes)
  - Typed nodes and links imply more structures, leading to richer discovery
- Mining *semi-structured* heterogeneous information networks
  - Clustering, ranking, classification, prediction, similarity search, etc.





Session 2: RankClus: Integrated Clustering and Ranking in Heterogeneous Networks

Ranking-Based Clustering in Heterogeneous Networks

- Clustering and ranking: Two critical functions in data mining
  - Clustering without ranking? Think about no PageRank dark time before Google
  - Ranking will make more sense within a particular cluster
    - Einstein in physics vs. Turing in computer science
- Why not integrate ranking with clustering & classification?
  - High-ranked objects should be more important in a cluster than low-ranked ones
  - Why treat every object the same weight in the same cluster?
    - But how to get their weight?
- Integrate ranking with clustering/classification in the same process
  - Ranking, as the feature, is conditional (i.e., relative) to a specific cluster
  - Ranking and clustering may mutually enhance each other
  - Ranking-based clustering: RankClus [EDBT'09], NetClus [KDD'09]

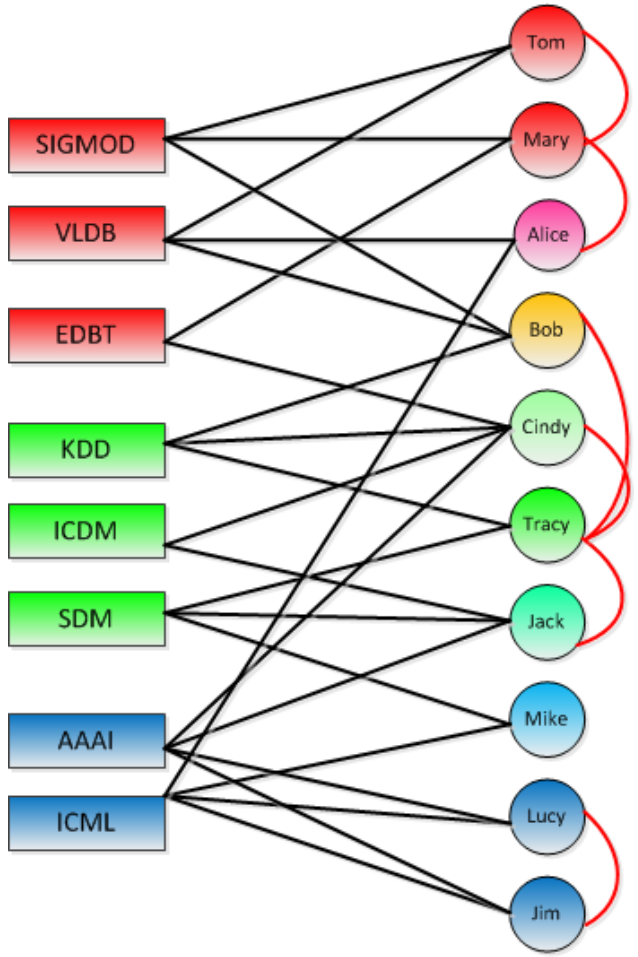
A Bi-Typed Network Model and Simple Ranking

- A bi-typed network model
  - Let X represents type *venue*
    - Y: Type *author*

$$W = \begin{bmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{bmatrix}$$

- The DBLP network can be represented as matrix W
- Our task: Rank-based clustering of heterogeneous network W
- Simple Ranking
  - Proportional to # of publications of an author and a venue
  - Considers only **immediate neighborhood** in the network

$$\begin{cases} \vec{r}_X(x) = \dfrac{\sum_{j=1}^{n} W_{XY}(x, j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i, j)} \\ \vec{r}_Y(y) = \dfrac{\sum_{i=1}^{n} W_{XY}(i, y)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i, j)} \end{cases}$$



A bi-typed network

But what about an author publishing many papers only in very weak venues?

Authority Ranking

- Methodology: **Propagate** the ranking scores in the network over different types
- Rule 1: Highly ranked authors publish *many* papers in highly ranked venues

$$\vec{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i)\vec{r}_X(i)$$

- Rule 2: Highly ranked venues attract *many* papers from *many* highly ranked authors

$$\vec{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_Y(j)$$

- Rule 3: The rank of an author is enhanced if he or she co-authors with *many* highly ranked authors

$$\vec{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j)\vec{r}_X(j) + (1-\alpha)\sum_{j=1}^{n} W_{YY}(i,j)\vec{r}_Y(j)$$

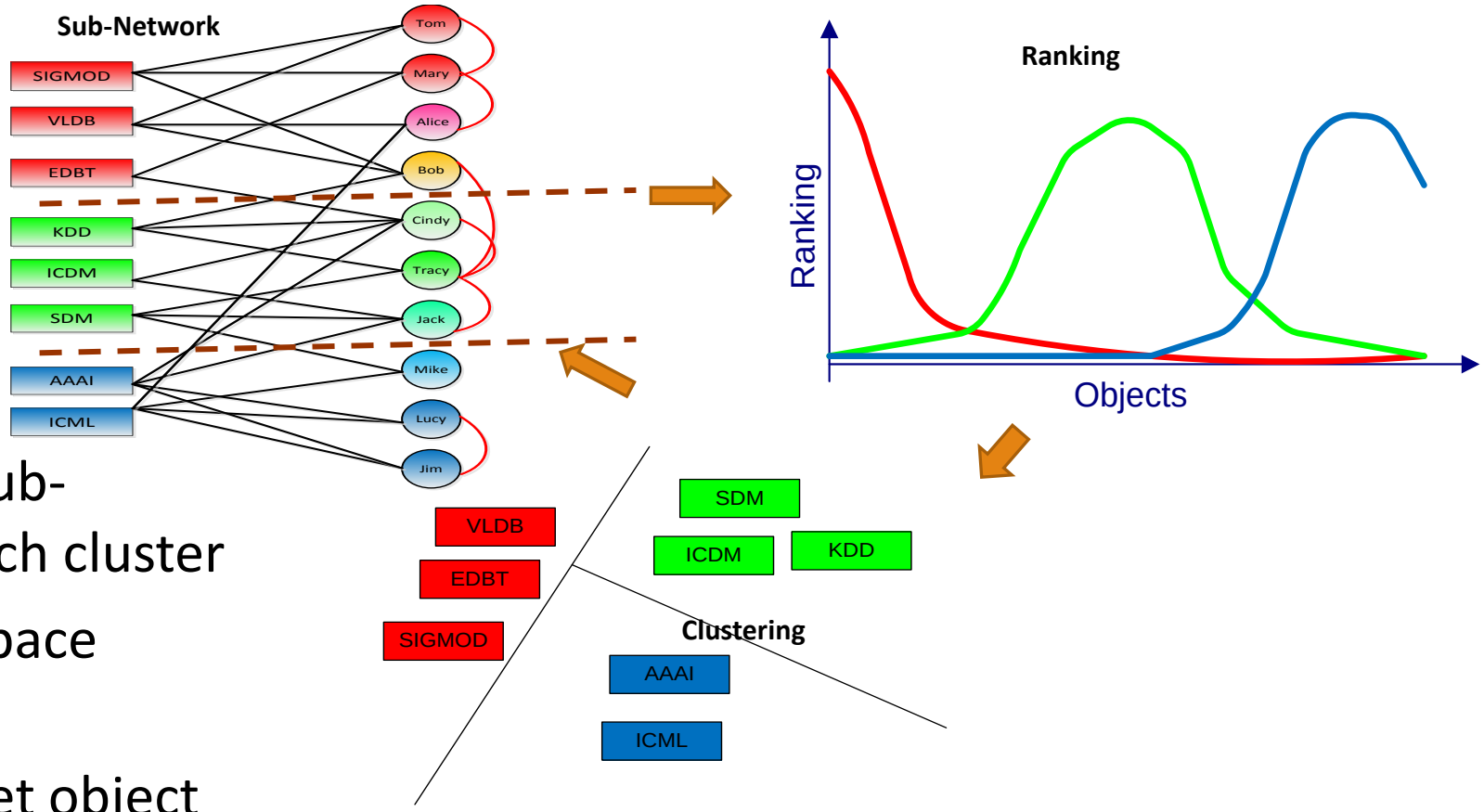
- Other ranking functions are quite possible (e.g., using domain knowledge)
  - Ex. Journals may weight more than conferences in science

From Conditional Rank Distribution to E-M Framework

- Given a bi-typed bibliographic network, how can we use the conditional rank scores to further improve the clustering results?
- Conditional rank distribution as cluster feature
  - For each cluster C_k, *the conditional rank scores,* $r_X|C_K$ *and* $r_Y|C_K$, *can be viewed* as conditional rank distributions of *X and Y, which are the features for cluster* C_k
- Cluster membership as object feature
  - From $p(k|o_i) \propto p(o_i|k)p(k)$, *the higher its conditional* rank in a cluster ($p(o_i|k)$), *the higher possibility an object will belong to that cluster* ($p(k|o_i)$)
  - Highly ranked attribute object has more impact on determining the cluster membership of a target object
- Parameter estimation using the Expectation-Maximization algorithm
  - E-step: Calculate the distribution $p(z = k|y_j, x_i, \Theta)$ *based on* the current value of Θ
  - M-Step: Update Θ according to the current *distribution*

RankClus: Integrating Clustering with Ranking

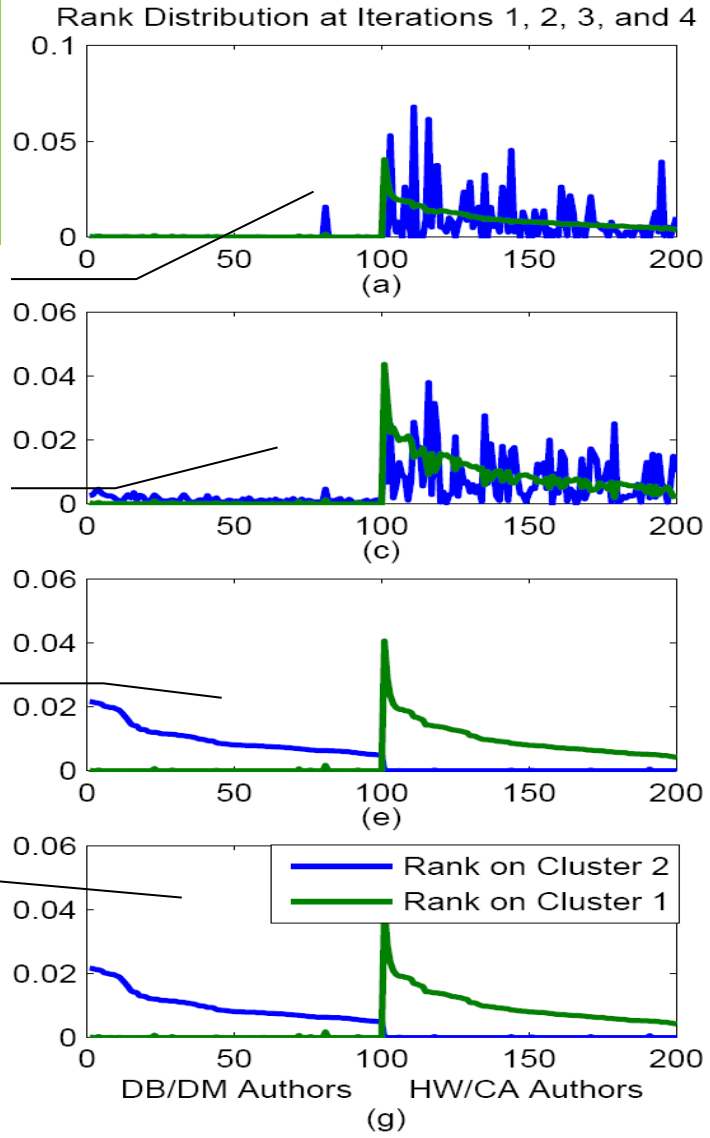
- **An EM styled Algorithm**
  - **Initialization**
    - Randomly partition
  - **Repeat**
    - Ranking
      - Ranking objects in each sub-network induced from each cluster
    - Generating new measure space
      - Estimate **mixture model coefficients** for each target object
    - Adjusting cluster
  - **Until** change < threshold

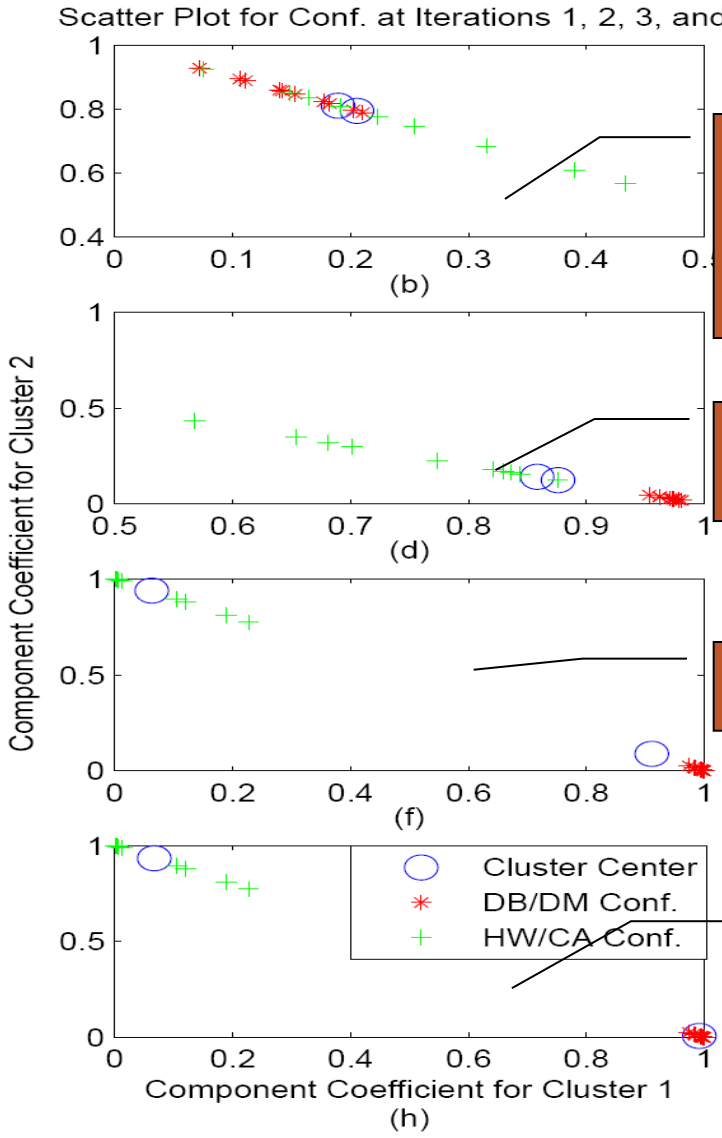


An E-M framework for iterative enhancement

Step-by-Step Running of RankClus

Clustering and ranking of two fields: DB/DM vs. HW/CA (hardware/ computer architecture)





Initially, ranking distributions are mixed together

Improved a little

Improved significantly

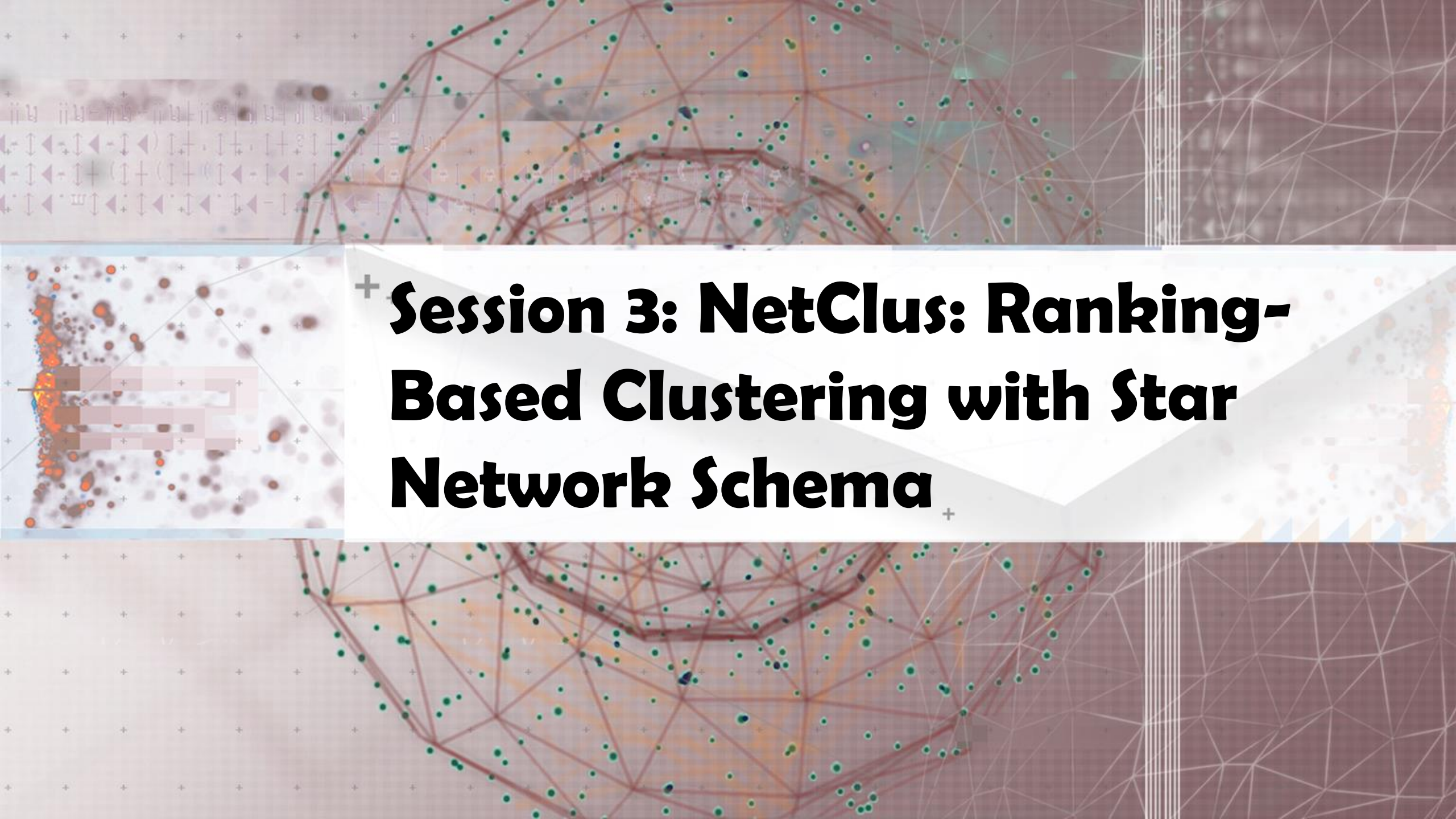
Stable

Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

Stable

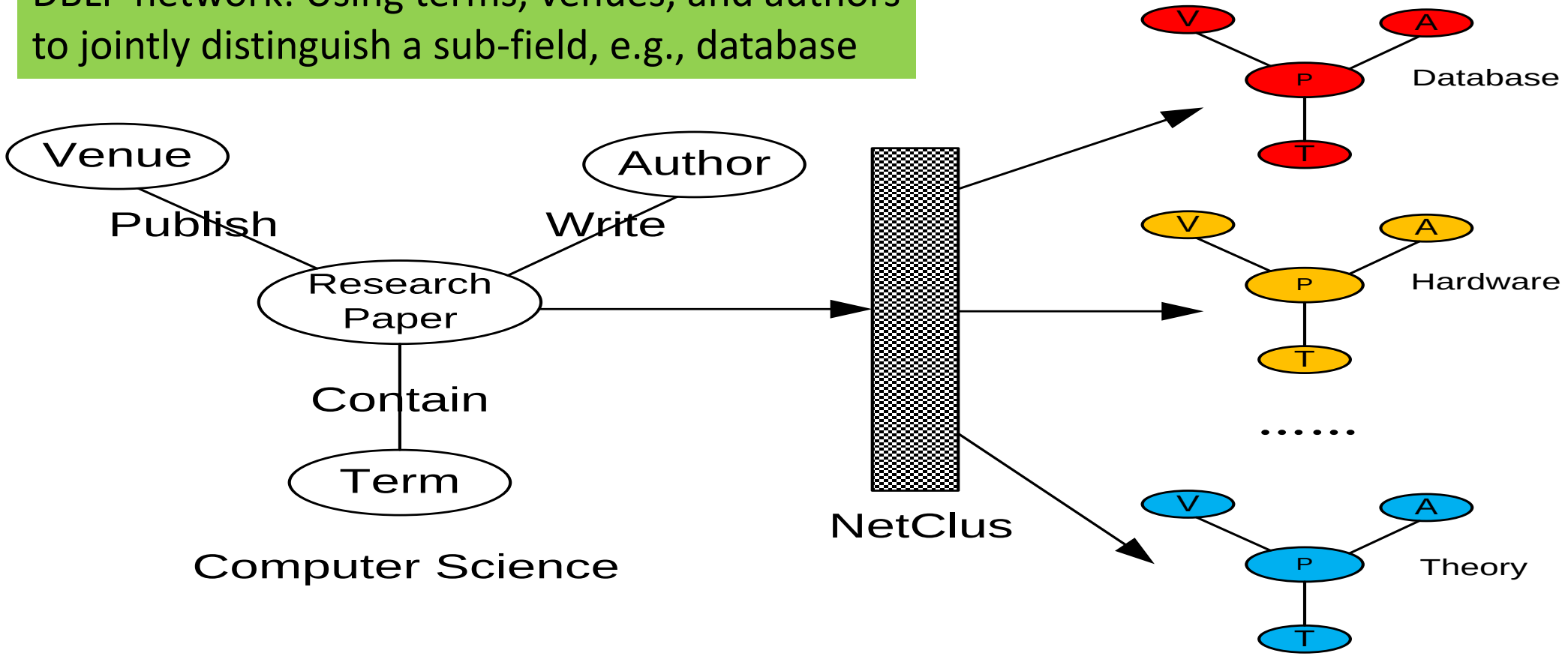


Session 3: NetClus: Ranking-Based Clustering with Star Network Schema

NetClus: Ranking-Based Clustering with Star Network Schema

- Beyond bi-typed network: Capture more semantics with multiple types
- Split a network into multi-subnetworks, each a (multi-typed) net-cluster [KDD'09]

DBLP network: Using terms, venues, and authors to jointly distinguish a sub-field, e.g., database



The NetClus Algorithm

- Generate initial partitions for target objects and induce initial net-clusters from the original network
- Repeat // An E-M Framework
  - Build ranking-based probabilistic generative model for each net-cluster
  - Calculate the posterior probabilities for each target object
  - Adjust their cluster assignment according to the new measure defined by the posterior probabilities to each cluster
- Until the clusters do not change significantly
- Calculate the posterior probabilities for each attribute object in each net-cluster

NetClus: Experiment on DBLP: Database System Cluster

Term	Venue	Author
database 0.0995511	VLDB 0.318495	Surajit Chaudhuri 0.00678065
databases 0.0708818	SIGMOD Conf. 0.313903	Michael Stonebraker 0.00616469
system 0.0678563	ICDE 0.188746	Michael J. Carey 0.00545769
data 0.0214893	PODS 0.107943	C. Mohan 0.00528346
query 0.0133316	EDBT 0.0436849	David J. DeWitt 0.00491615
systems 0.0110413		Hector Garcia-Molina 0.00453497
queries 0.0090603		H. V. Jagadish 0.00434289
management 0.00850744		David B. Lomet 0.00397865
object 0.00837766		Raghu Ramakrishnan 0.0039278
relational 0.0081175		Philip A. Bernstein 0.00376314
processing 0.00745875		Joseph M. Hellerstein 0.00372064
based 0.00736599		Jeffrey F. Naughton 0.00363698
distributed 0.0068367		Yannis E. Ioannidis 0.00359853
xml 0.00664958		Jennifer Widom 0.00351929
oriented 0.00589557		Per-Ake Larson 0.00334911
design 0.00527672		Rakesh Agrawal 0.00328274
web 0.00509167		Dan Suciu 0.00309047
information 0.0050518		Michael J. Franklin 0.00304099
model 0.00499396		Umeshwar Dayal 0.00290143
efficient 0.00465707		Abraham Silberschatz 0.00278185

- NetClus generates high quality clusters for all of the three participating types in the DBLP network
  - Quality can be easily judged by our commonsense knowledge
- Highly-ranked objects: Objects centered in the cluster

Rank-Based Clustering: Works in Multiple Domains

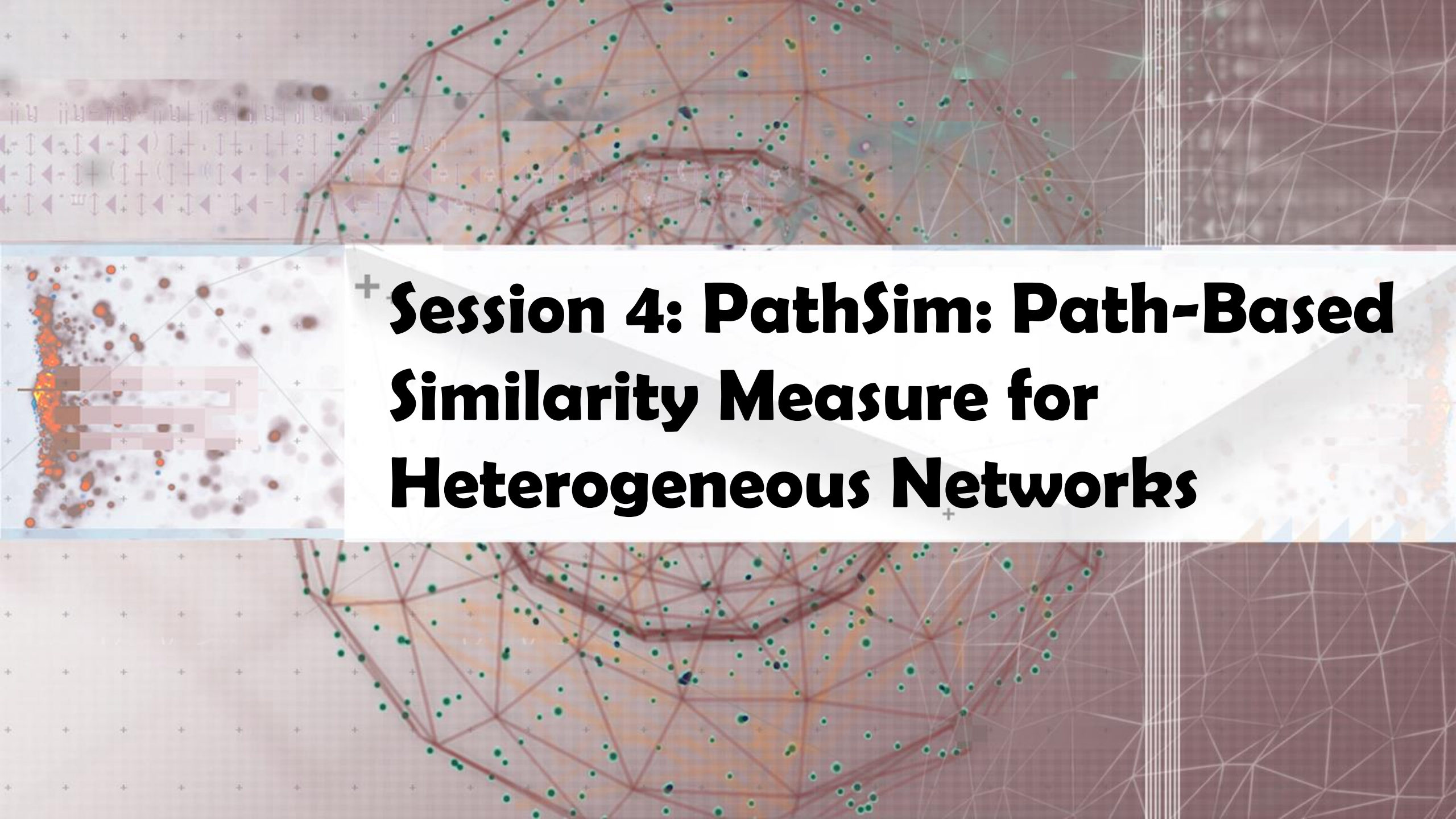
RankCompete: Organize your photo album automatically!

	Top 10 Treatments	Ranking
1	Zidovudine/therapeutic use	0.1679
2	Anti-HIV Agents/therapeutic use	0.1340
3	Antiretroviral Therapy, Highly Active	0.0977
4	Antiviral Agents/therapeutic use	0.0718
5	Anti-Retroviral Agents/therapeutic use	0.0236
6	Interferon Type I/therapeutic use	0.0147
7	Didanosine/therapeutic use	0.0132
8	Ganciclovir/therapeutic use	0.0114
9	HIV Protease Inhibitors/therapeutic use	0.0105
10	Antineoplastic Combined Chemotherapy	0.0103

MedRank: Rank treatments for AIDS from Medline

Use multi-typed image features to build up heterogeneous networks

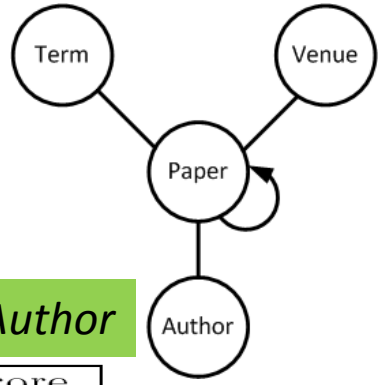
Explore multiple types in a star schema network

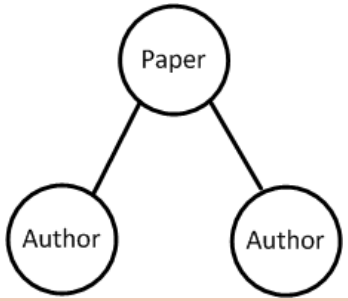


Session 4: PathSim: Path-Based Similarity Measure for Heterogeneous Networks

Similarity Search in Heterogeneous Networks

- Similarity measure/search is the base for cluster analysis
- Who are the most similar to *Christos Faloutsos* based on the DBLP network?
- Meta-Path: **Meta-level description** of a path between two objects
  - **A path** on network schema
    - Denote an existing or concatenated **relation** between two object types
- Different meta-paths tell different semantics





Co-authorship
Meta-path: A-P-A

Meta-Path: *Author-Paper-Author*

Rank	Author	Score
1	Christos Faloutsos	1
2	Spiros Papadimitriou	0.127
3	Jimeng Sun	0.12
4	Jia-Yu Pan	0.114
5	Agma J. M. Traina	0.110
6	Jure Leskovec	0.096
7	Caetano Traina Jr.	0.096
8	Hanghang Tong	0.091
9	Deepayan Chakrabarti	0.083
10	Flip Korn	0.053

Christos' students or close collaborators

Meta-Path: *Author-Paper-Venue-Paper-Author*

Rank	Author	Score
1	Christos Faloutsos	1
2	Jiawei Han	0.842
3	Rakesh Agrawal	0.838
4	Jian Pei	0.8
5	Charu C. Aggarwal	0.739
6	H. V. Jagadish	0.705
7	Raghu Ramakrishnan	0.697
8	Nick Koudas	0.689
9	Surajit Chaudhuri	0.677
10	Divesh Srivastava	0.661

Work in similar fields with similar reputation

Existing Popular Similarity Measures for Networks

- Random walk (RW):
  - The probability of random walk starting at *x* and ending at *y*, with meta-path *P*

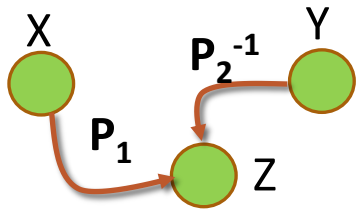
$$s(x, y) = \sum_{p \in P} prob(p)$$

  - Used in Personalized PageRank (P-Pagerank) (Jeh and Widom 2003)
  - Favors **highly visible** objects (i.e., objects with large degrees)

Note: P-PageRank and SimRank do not distinguish object type and relationship type

- Pairwise random walk (PRW):
  - The probability of pairwise random walk starting at (*x, y*) and ending at a common object (say *z*), following a meta-path (P_1, P_2)

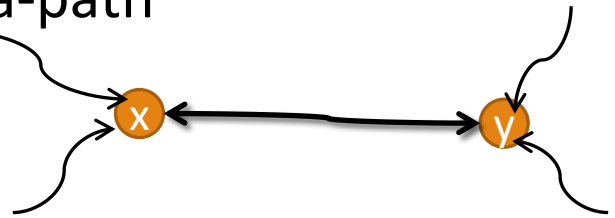
$$s(x, y) = \sum_{(p_1, p_2) \in (P_1, P_2)} prob(p_1) prob(p_2)$$



  - Used in SimRank (Jeh and Widom 2002)
  - Favors ***pure*** objects (i.e., objects with highly skewed distribution in their in-links or out-links)

Which Similarity Measure Is Better for Finding Peers?

- PathSim: Favors peers
  - **Peers**: Objects with strong connectivity and similar visibility with a given meta-path



$$s(x,y) = \frac{2 \times |\{p_{x \rightsquigarrow y} : p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightsquigarrow x} : p_{x \rightsquigarrow x} \in \mathcal{P}\}| + |\{p_{y \rightsquigarrow y} : p_{y \rightsquigarrow y} \in \mathcal{P}\}|}$$

  - Meta-path: APCPA
  - Mike publishes similar # of papers as Bob and Mary
  - Other measures find Mike is closer to Jim

Author\Conf.	SIGMOD	VLDB	ICDM	KDD
Mike	2	1	0	0
Jim	50	20	0	0
Mary	2	0	1	0
Bob	2	1	0	0
Ann	0	0	1	1

Measure\Author	Jim	Mary	Bob	Ann
P-PageRank	**0.376**	0.013	0.016	0.005
SimRank	**0.716**	0.572	0.713	0.184
Random Walk	**0.8983**	0.0238	0.0390	0
Pairwise R.W.	**0.5714**	0.4440	0.5556	0
PathSim (APCPA)	0.083	**0.8**	**1**	0

Comparison of Multiple Measures: A Toy Example

Example with DBLP: Find Academic Peers by PathSim

- Anhai Doan
  - CS, Wisconsin
  - Database area
  - PhD: 2002

- Jignesh Patel
  - CS, Wisconsin
  - Database area
  - PhD: 1998

Meta-Path: ***Author-Paper-Venue-Paper-Author***

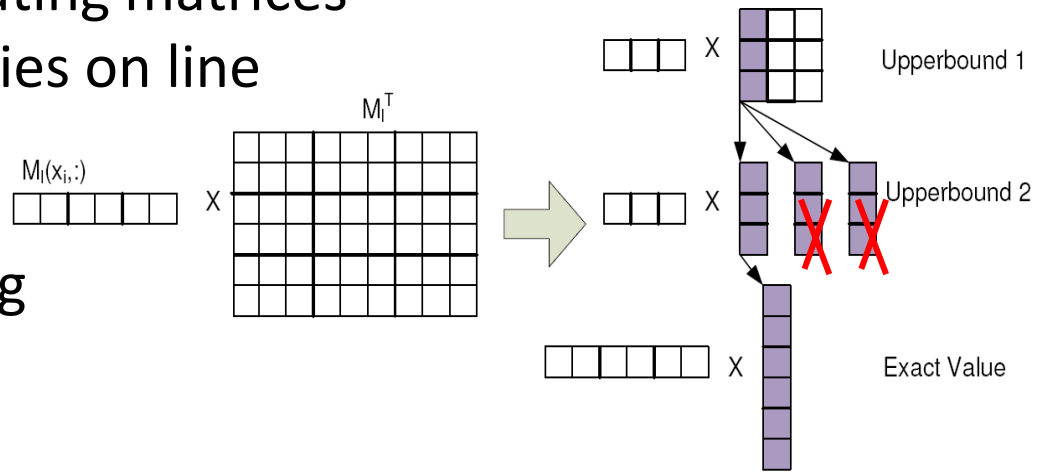
Rank	P-PageRank	SimRank	PathSim
1	AnHai Doan	AnHai Doan	AnHai Doan
2	Philip S. Yu	Douglas W. Cornell	Jignesh M. Patel
3	Jiawei Han	Adam Silberstein	Amol Deshpande
4	Hector Garcia-Molina	Samuel DeFazio	Jun Yang
5	Gerhard Weikum	Curt Ellmann	Renée J. Miller

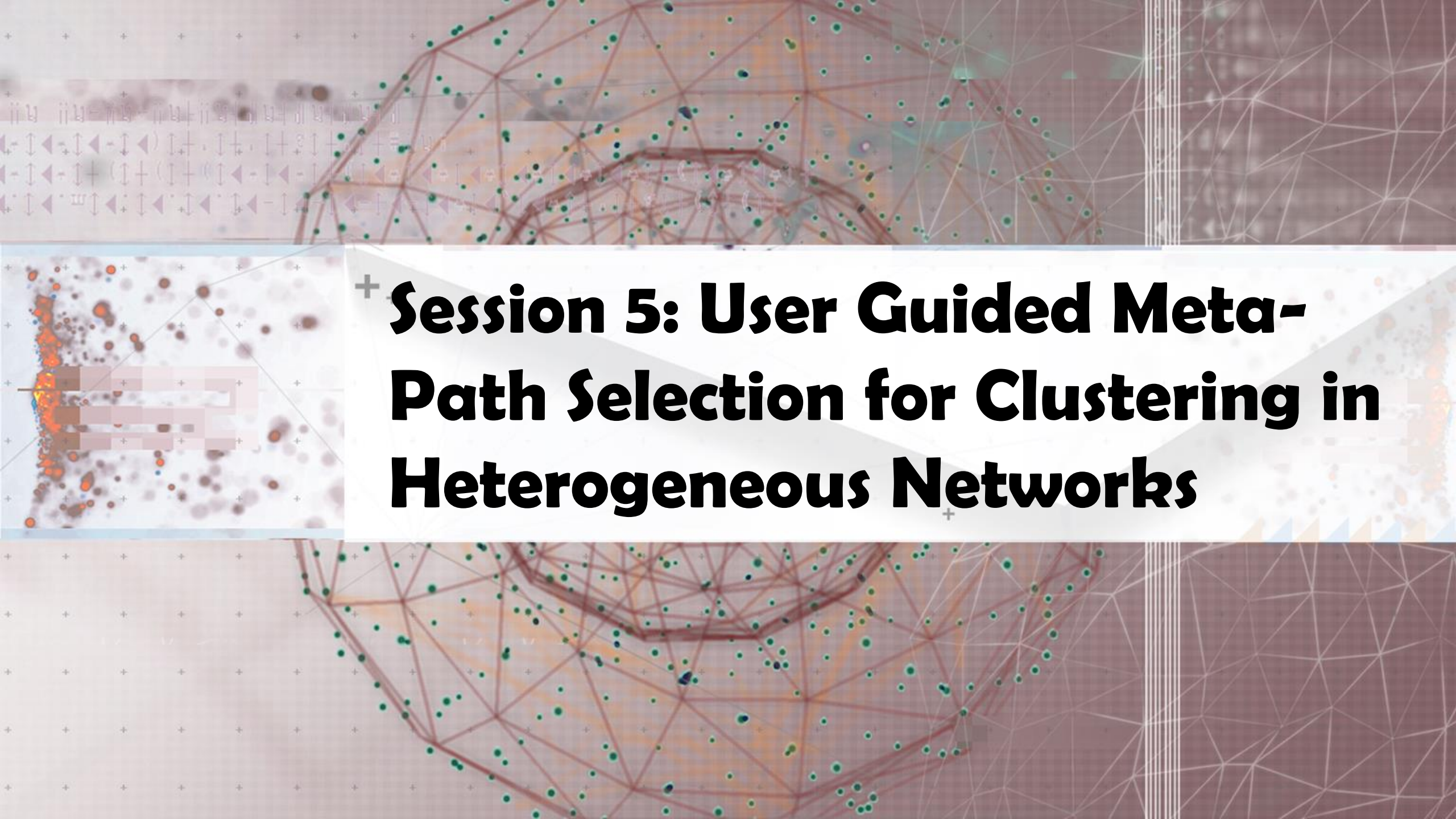
- Amol Deshpande
  - CS, Maryland
  - Database area
  - PhD: 2004

- Jun Yang
  - CS, Duke
  - Database area
  - PhD: 2001

Co-Clustering-Based Pruning Algorithm

- Meta-Path based similarity computation can be costly
- The overall cost can be reduced by storing commuting matrices for short path schemas and computing top-*k* queries on line
- Framework
  - Generate co-clusters for materialized commuting matrices for feature objects and target objects
  - Derive upper bound for similarity between object and target cluster and between object and object
    - Safely prune target clusters and objects if the upper bound similarity is lower than current threshold
  - Dynamically update top-*k* threshold

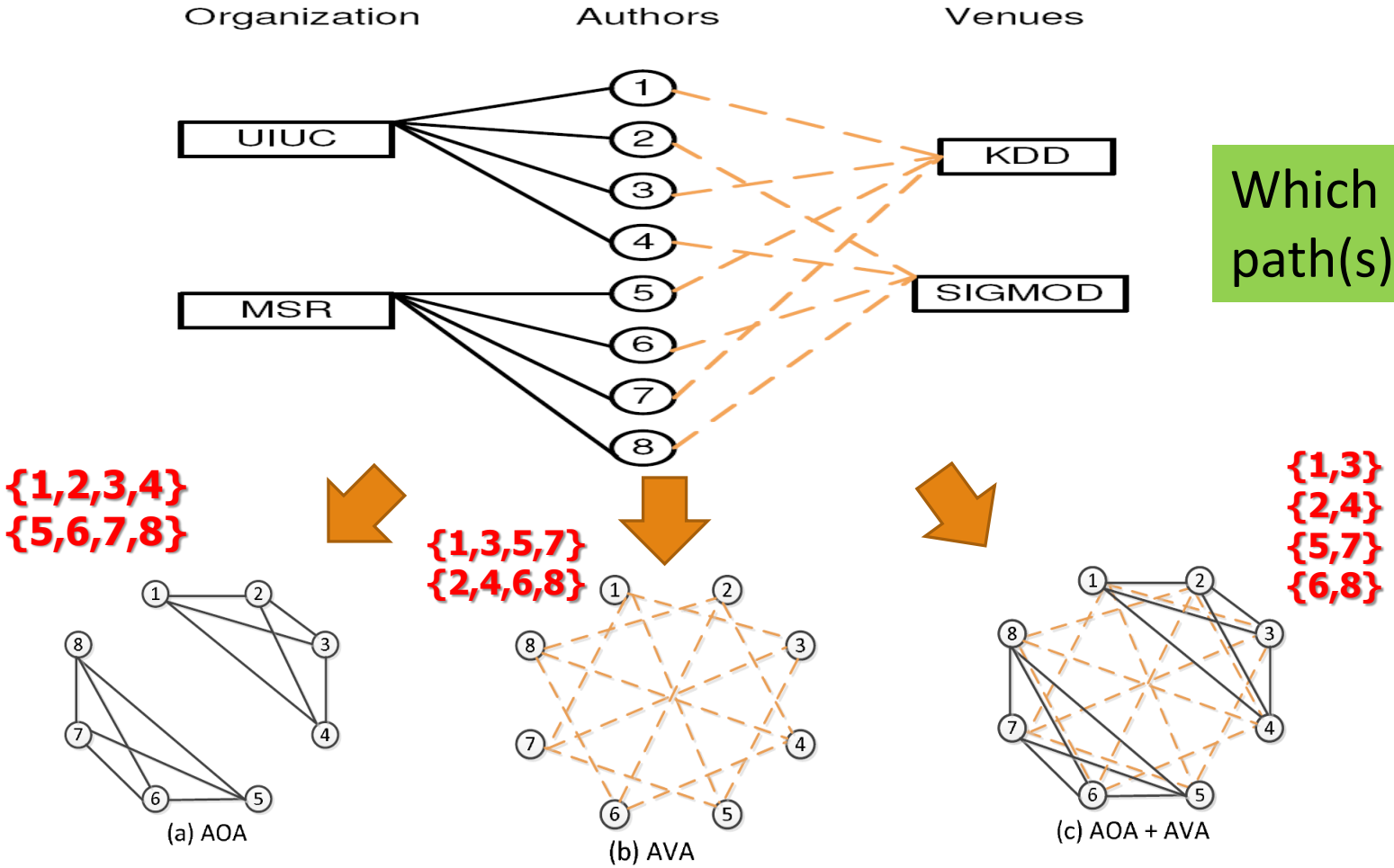




Session 5: User Guided Meta-Path Selection for Clustering in Heterogeneous Networks

Why User Guidance in Clustering?

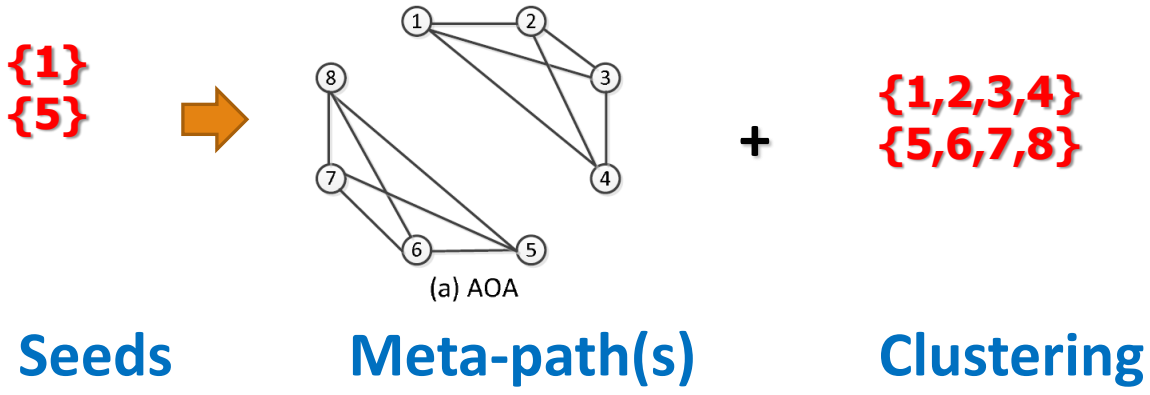
- Different users may like to get different clusters for different clustering goals
  - Ex. Clustering authors based on their connections in the network



Which meta-path(s) to choose?

User Guidance Determines Clustering Results

- Different user preferences (e.g., by seeding desired clusters) lead to the choice of different meta-paths



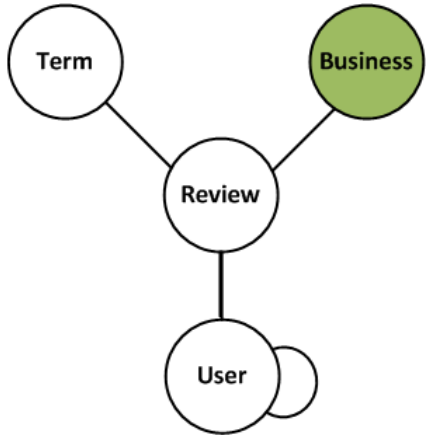
- **Problem: User-guided clustering with meta-path selection**
- Input:
  - The target type for clustering T
  - # of clusters k
  - Seeds in some clusters: $L_1, \ldots, L_k$
  - Candidate meta-paths: $P_1, \ldots, P_M$
- Output:
  - Weight of each meta-path: $w_1, \ldots, w_m$
  - Clustering results that are consistent with the user guidance

PathSelClus: A Probabilistic Modeling Approach

- Part 1: Modeling the Relationship Generation
  - A good clustering result should lead to high likelihood in observing existing relationships
    - Higher quality relations should count more in the total likelihood
- Part 2: Modeling the Guidance from Users
  - The more consistent with the guidance, the higher probability of the clustering result
- Part 3: Modeling the Quality Weights for Meta-Paths
  - The more consistent with the clustering result, the higher quality weight

Effectiveness of Meta-Path Selection

- Experiments on Yelp data
  - Object Types: Users, Businesses, Reviews, Terms
  - Relation Types: UR, RU, BR, RB, TR, RT
- Task: Candidate meta-paths: *B-R-U-R-B, B-R-T-R-B*
  - Target objects: restaurants
  - # of clusters: 6
- Output:
  - PathSelClus vs. others
    - High accuracy
  - Restaurant vs. shopping
    - For restraunts, meta-path B-R-U-R-B weighs only 0.1716
    - For clustering shopping, B-R-U-R-B weighs 0.5864



%S	Measure	PathSelClus	LP	ITC	LP_voting	LP_soft	ITC_voting	ITC_soft
1%	Accuracy	**0.7435**	0.1137	0.1758	0.2112	0.2112	0.2430	0.2022
	NMI	**0.6517**	0.0323	0.0178	0.0578	0.0578	0.2308	0.2490
2%	Accuracy	**0.8004**	0.1264	0.1910	0.2202	0.2202	0.2762	0.2792
	NMI	**0.6803**	0.0487	0.0150	0.0801	0.0801	0.2099	0.2907
5%	Accuracy	**0.8125**	0.2653	0.2200	0.2437	0.2437	0.3049	0.3240
	NMI	**0.6894**	0.1111	0.0220	0.1212	0.1212	0.2252	0.2692

Users try different kinds of food

Session 6: Summary

Summary: Cluster Analysis in Heterogeneous Networks

- Heterogeneous Information Networks
- RankClus: Integrated Clustering and Ranking in Heterogeneous Networks
- NetClus: Ranking-Based Clustering with Star Network Schema
- PathSim: Path-Based Similarity Measure for Heterogeneous Networks
- User Guided Meta-Path Selection for Clustering in Heterogeneous Networks
- Summary

Recommended Readings

- Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, T. Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis", EDBT'09
- Y. Sun, Y. Yu, J. Han, “Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema", KDD'09
- Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, “PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks”, VLDB'11
- Y. Sun and J. Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers, 2012
- Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, X. Yu. “PathSelClus: Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks”, ACM Trans. on Knowledge Discovery from Data (TKDD), 7(3) 2013
- C. Aggarwal. Data Mining: The Textbook. Springer, 2015