Mining Query-Based Subnetwork Outliers in Heterogeneous Information Networks

Honglei Zhuang[1], Jing Zhang[2], George Brova[1], Jie Tang[2], Hasan Cam[3], Xifeng Yan[4], Jiawei Han[1]

[1]University of Illinois at Urbana-Champaign
[2]Tsinghua University
[3]US Army Research Lab
[4]University of California at Santa Barbara

- Suppose we are given travel information of users, including:
  - Flight info,
  - Hotel booking info,
  - Car rental info,
  - …
- **How can an analyst identify terrorists ring from the massive information?**
- This scenario can be naturally extended to a more general problem: *query*-based *subnetwork* outlier detection.

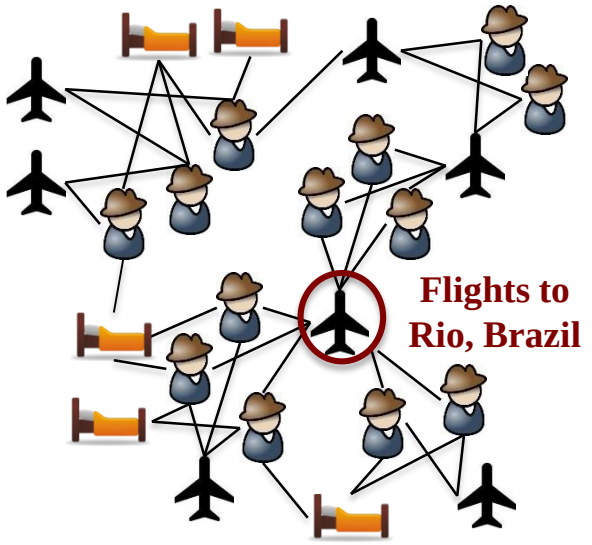
Querying Subnetwork Outliers

Input: A travel information network, a query

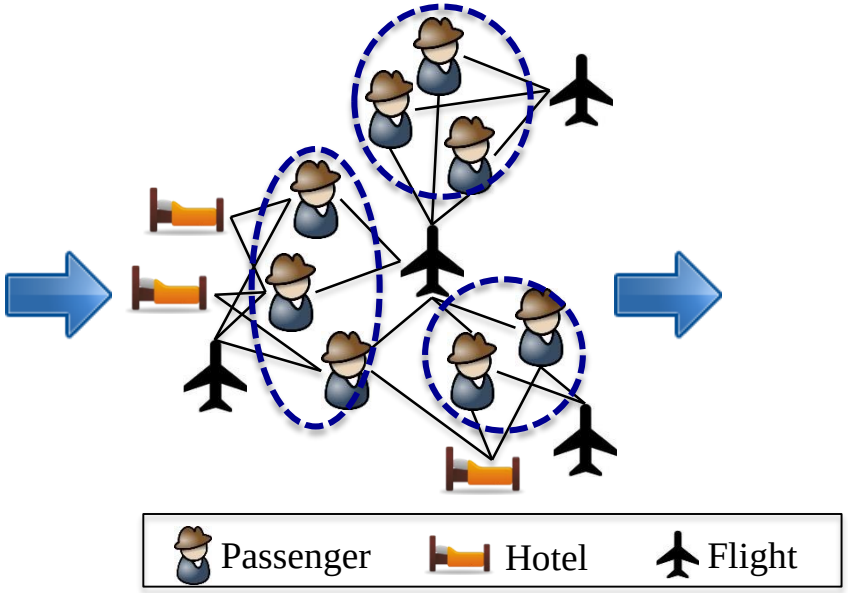
- **User poses a query**: “Analyze passenger groups flying to Rio, Brazil”

Querying Subnetwork Outliers

Input: A travel information network, a query

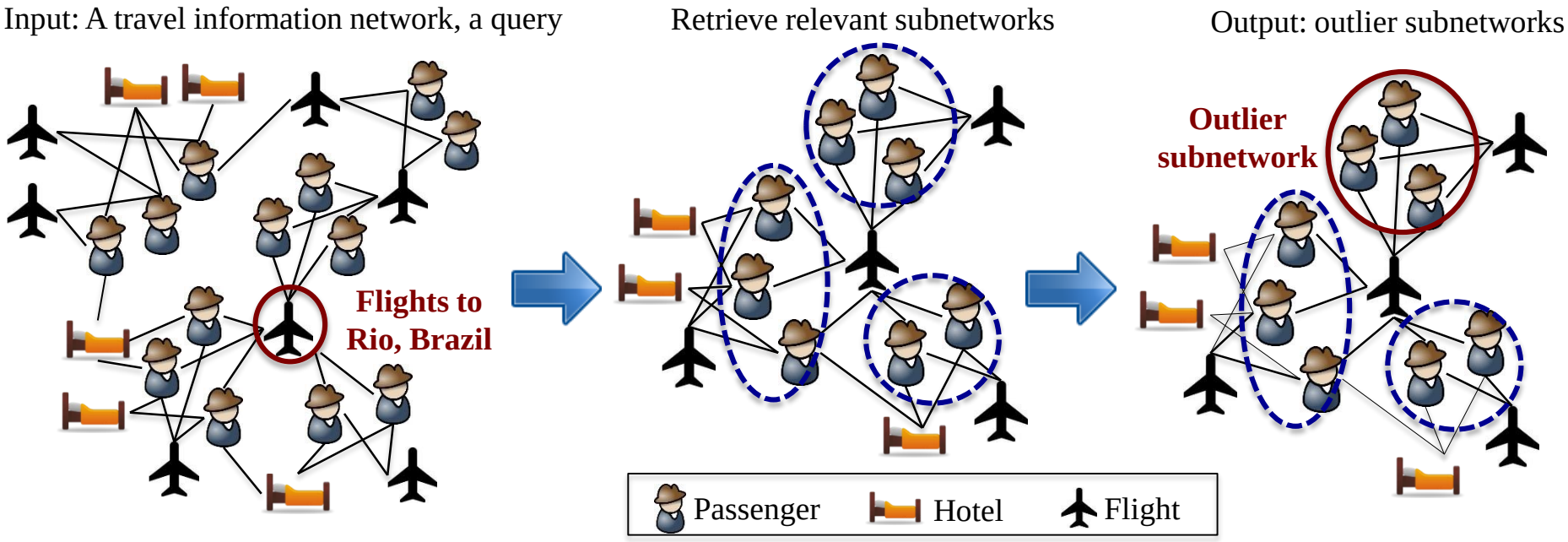


Retrieve relevant subnetworks



- **User poses a query**: “Analyze passenger groups flying to Rio, Brazil”
- **Retrieve candidate subnetworks**: connected and relevant to query

Querying Subnetwork Outliers



- **User poses a query**: "Analyze passenger groups flying to Rio, Brazil"
- **Retrieve candidate subnetworks**: connected and relevant to query
- **Identify outlier subnetworks:** deviating significantly from others

Problem Definition

- Input:
  - A heterogeneous information network G
  - A query consisting of
    - A set of queried vertices (entities) $V_q \subset V$
      - e.g. “Flight 123”
    - Relationship from queried vertices to desired vertices P_Q
      - e.g., “passengers on the flight”
    - How they form subnetworks P_S
      - e.g., “traveling together”

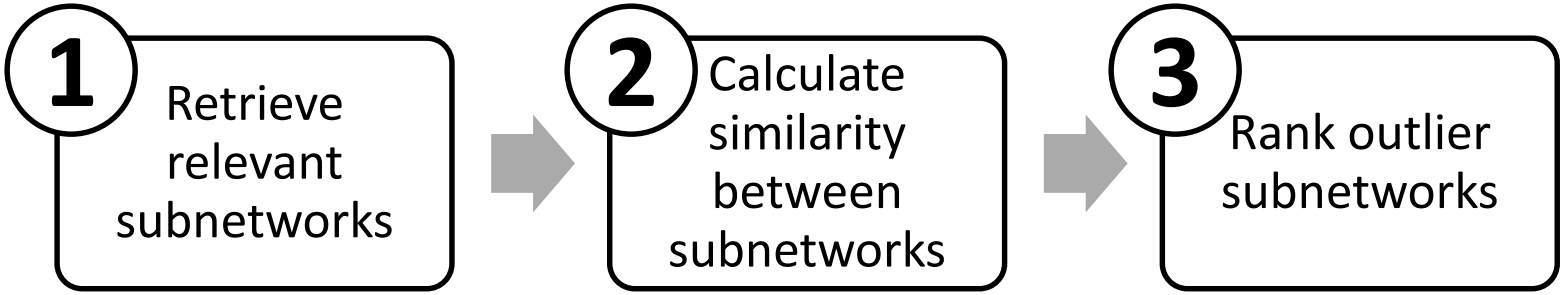
P_Q ↗

P_S ← **meta-path**

- Output:
  - Outlier subnetworks $S_\omega = \{S_{\omega 1} \subset V, \ldots, S_{\omega k} \subset V\}$

Methodology

- General Framework



1 Retrieving relevant subnetworks

- Can be handled by IR techniques
- *Not* our focus of this work
- Applying a simple retrieving strategy based on frequent pattern mining

② Similarity Measure

- **Intuition:** two subnetworks are similar when their members are from similar distribution over communities
- **Basic idea:**
  - Calculate individual similarity by meta-path based similarity measure PathSim*
  - Similarity measures (*w.l.o.g*, $|S_1| \geq |S_2|$)

$$\sigma_{BM}(S_1, S_2) = \frac{1}{|S_1|} \max_{M} \sum_{(v_1^i, v_2^j) \in M} PathSim(v_1^i, v_2^j)$$

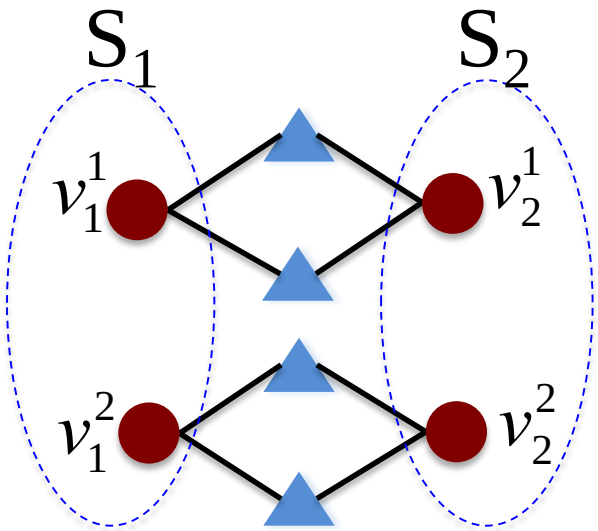
  - where M is a set of pairs of vertices from two subnetworks, satisfying

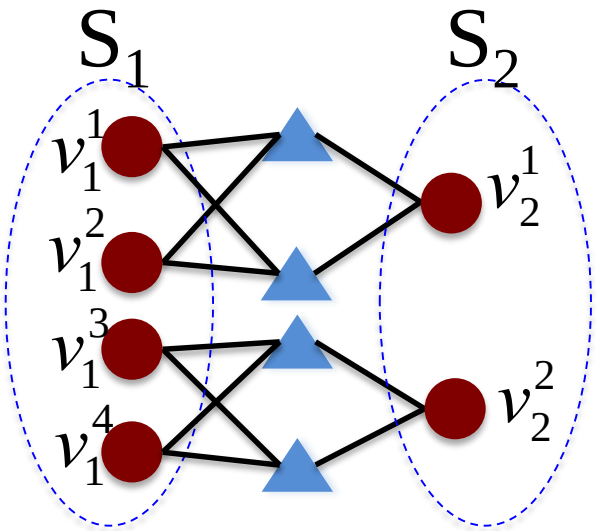
$$\forall v_1^i \in S_1, \left|\left\{v_1^i \mid (v_1^i, v_2^j) \in M\right\}\right| = 1 \quad \forall v_2^j \in S_2, 1 \leq \left|\left\{v_2^j \mid (v_1^i, v_2^j) \in M\right\}\right| < 1 + \frac{|S_1|}{|S_2|}$$

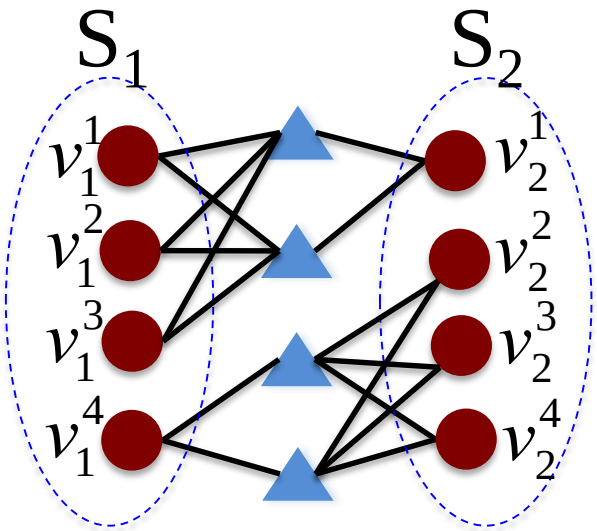
* Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta-path based top-k similarity search in heterogeneous information networks. In VLDB, pages 992–1003, 2011.

② Similarity Measure (cont')

- Example







Desired	1.0
AvgSim	0.5
*MatchSim	1.0
BMSim	*1.0*

Desired	1.0
AvgSim	0.5
*MatchSim	0.5
BMSim	*1.0*

Desired	<1
AvgSim	0.375
*MatchSim	0.5
BMSim	*0.5*

* Z. Lin, M. R. Lyu, and I. King. Matchsim: a novel neighbor-based similarity measure with maximum neighborhood matching. In CIKM, pages 1613–1616, 2009.

③ Subnetwork Outliers

- **Intuition:**
  - Clustering subnetworks by either assigning a subnetwork with an "*exemplar*" subnetwork, or classifying the subnetwork as an *outlier*
- **Basic Ideas:**
  - Calculate the outlierness by

$$\Omega(S_i) = -\max_{j \neq 0} \left[a_{i \leftarrow j} + \sigma(i, j) \right]$$

  - Automatically weighting multiple similarity measures instantiated by different meta-paths

*B. J. Frey and D. Dueck. Clustering by passing messages between data points. Science, 315(5814):972–976, 2007.

③ Subnetwork Outliers

- **Intuition:**
  - Clustering subnetworks by either assigning a subnetwork with an "*exemplar*" subnetwork, or classifying the subnetwork as an *outlier*
- **Basic Ideas:**
  - Calculate the outlierness by

$$\Omega(S_i) = -\max_{j \neq 0} \left[a_{i \leftarrow j} + \sigma(i, j) \right]$$

How good j is an exempler **Similarity between i and j**

  - Automatically weighting multiple similarity measures instantiated by different meta-paths

*B. J. Frey and D. Dueck. Clustering by passing messages between data points. Science, 315(5814):972–976, 2007.

Data Sets

	#Vertices	#Edges	#Types	Labels
Synthetic	1,000	about 33,000	2	Inserted outliers
Bibliography	3,701,765	24,639,131	4	Labeled for 5 queries
Patent	2,317,360	11,051,283	6	N/A

- Synthetic + 2 real world data sets are employed
- Bibliography data set are constructed based on DBLP
- Patent data set are constructed based on US Patent data

Experimental Results

- Performance

Data set	Synthetic			Bibliography		
Measure	**P@5**	**MAP**	**AUC**	**P@5**	**MAP**	**AUC**
Ind	60.00	66.61	85.00	28.00	24.82	59.91
NB	75.00	75.76	93.68	28.00	30.20	67.87
Proposed	**84.00**	**92.04**	**99.50**	**44.00**	**45.05**	**79.55**

- Baselines
  - *Ind*: sum of individual outlierness
  - *NB*: topic modeling with an "outlier" topic

Case Study

- Query: outlier author subnetworks related to "topic modeling"

Proposed Method \ Ind	Ind \ Proposed Method
Sanjeev Arora, Rong Ge, Ankur Moitra Theory group	**Tu Bao Ho, Khoat Than** Data mining group
Giovanni Ponti, Andrea Tagarelli Name ambiguity problem for Giovanni Ponti – could be an economics researcher or a data mining researcher	**Zhixin Li, Huifang Ma, Zhongzhi Shi** Machine learning and data mining group

Summary

- Study a novel problem of *query*-based *subnetwork* outlier detection in heterogeneous information networks

- Propose a framework to tackle the problem
  - Formalize the query
  - Propose a subnetwork similarity
  - Rank outlier subnetworks

Thanks

12/16/2014