OverCite: Finding Overlapping Communities in Citation Network

Tanmoy Chakraborty

Google India PhD Fellow
Indian Institute of Technology, Kharagpur
India

In collaboration with:
Abhijnan Chakraborty
Microsoft Research India, India – 560001

4th International Workshop on Business Applications of Social Network Analysis (BASNA), collocted with ASONAM
Niagara Falls, Canada, August 25-28, 2013

Outline

- Problem definition
- Tripartite Publication Hypergraph
- Our Proposed Algorithm: OverCite
- Dataset
- Evaluation
- Recommendation System
- Conclusion

Outline

Problem definition

Tripartite Publication Hypergraph

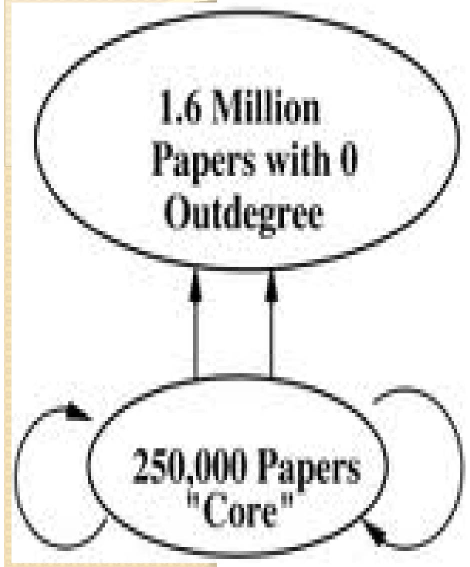
Our Proposed Algorithm: OVERCITE

Dataset

Evaluation

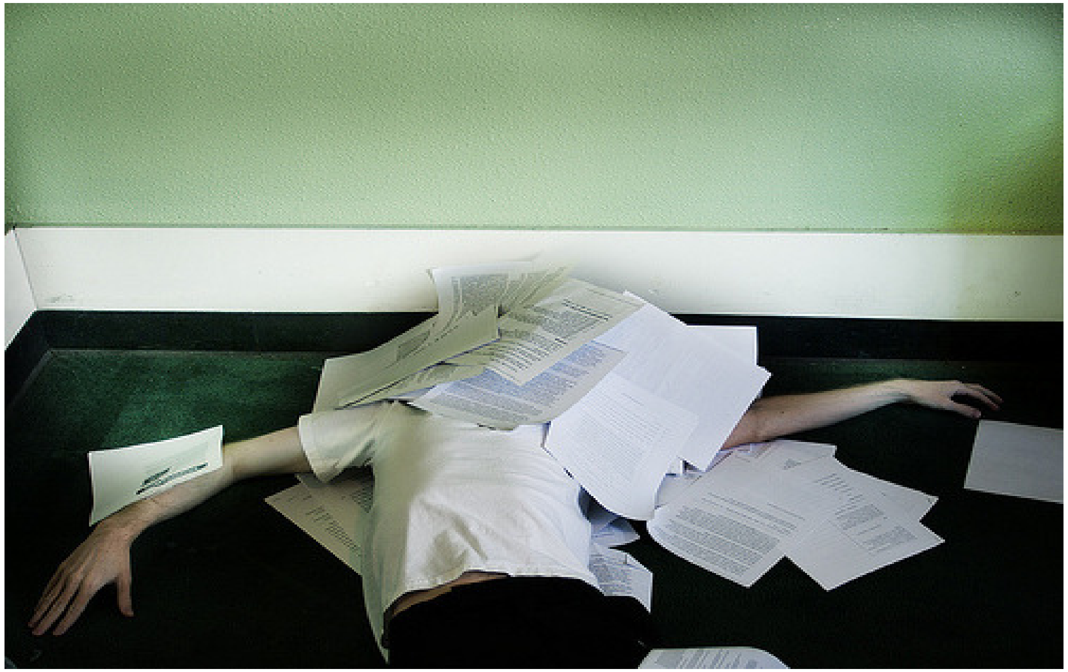
Recommendation System

Conclusion



Exponential growth of papers in every year

Ohh!!
My project has already been published….
How did I forget to read it before?



Motivation

- Use of citation network in paper search system
- Communities in scientific domain => different areas of interest
- Papers in multiple communities can act as interdisciplinary publications
- **Overlapping communities** => enhance the search and recommendation systems

Problem Definition

- Propose an overlapping community detection
- Published papers =>Tripartite Hypergraph structure
- Together: Papers, authors and publication venues
- Show how detected communities can lead to enhance the paper recommendation system

Outline

Problem definition

Tripartite Publication Hypergraph

Our Proposed Algorithm: OVERCITE

Dataset

Evaluation

Recommendation System

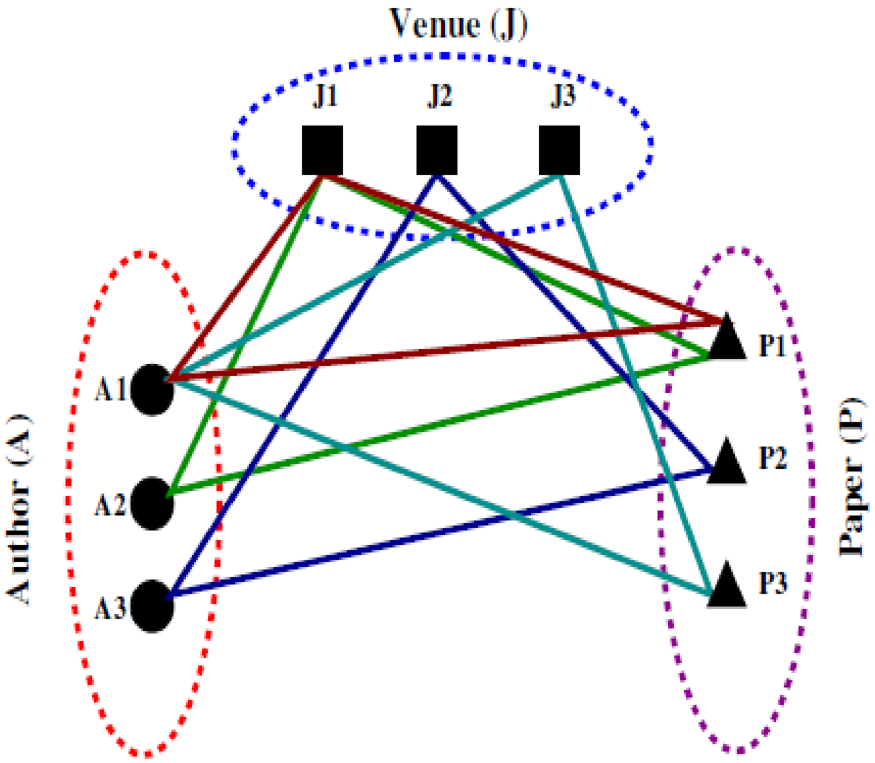
Conclusion

Tripartite Publication Hypergraph

Properties:

1. Uneven partition size:
   $|V_P| >= | V_A| >= | V_J |$

2. **Mapping:**
   - Paper-> Journal: one-to-one
   - Journal -> Paper: one-to-many
   - Author-> Paper: one-to-many
   - Author- > Journal: one-to-many



Outline

Problem definition

Tripartite Publication Hypergraph

Our Proposed Algorithm: OverCite

Dataset

Evaluation

Recommendation System

Conclusion

Limitations of Traditional Algorithms on Citation Network

- **Citation network** is **bounded by fixed time interval =>** Information loss of the older and newer papers.
- **Less cited papers =>** treated as **outliers**
- **Citation and co-authorship networks** are generally **sparse**

OverCite:

Overlapping community Detection in Citation Network

STEP 1: Convert publication hypergraph *H* into weighted line graph *H*$^{/}$

1.1 Nodes in *H* become edges in *H*$^{/}$, edges in *H* become nodes in *H*$^{/}$

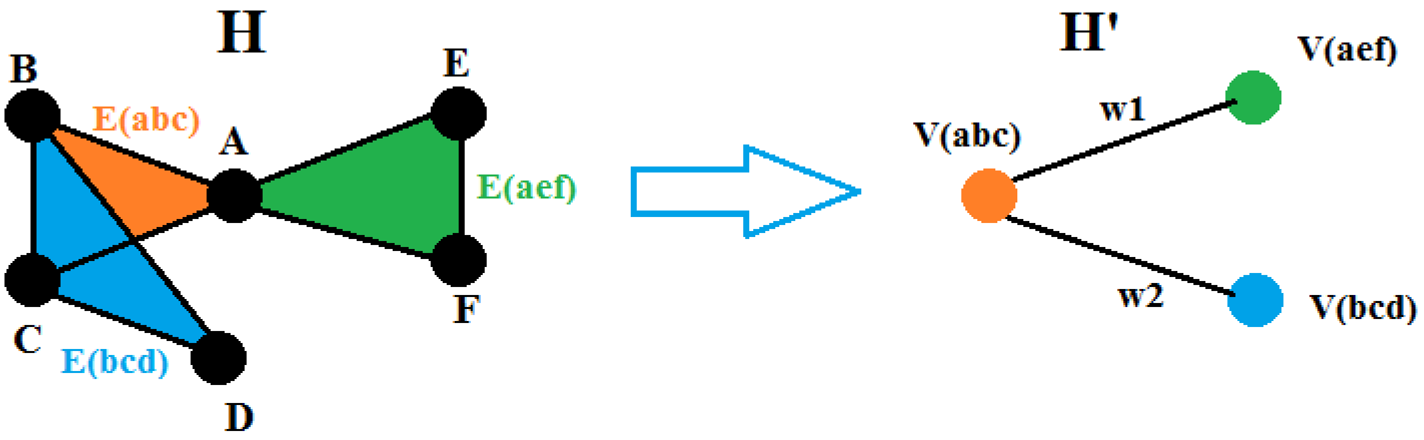
1.2 weights of edges in *H*$^{/}$ is determined by the **similarity measures:**

(a) Hypergraph Neighbourhood Similarity (HNS)

(b) Co-citation Strength (CCS)

(c) Bibliographic-Coupling Strength (BCS)

1.3 Final weight is determined by: α.HNS + β.CCS + γ.BCS (where, 0 <= α, β, γ <= 1)



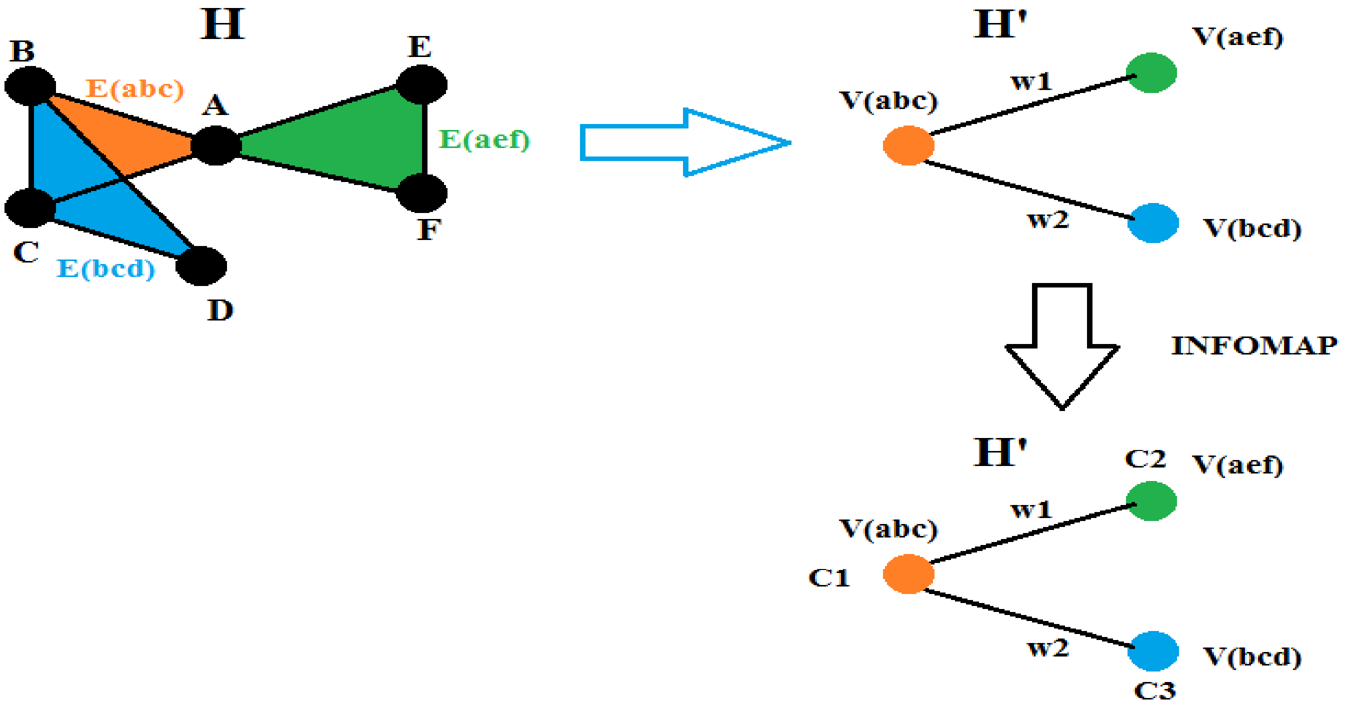
OverCite:

Overlapping community Detection in Citation Network

STEP 2. Any unipartite community detection algorithm can be applied on *H*′

We use Infomap [Rosvall & Bergstrom, PNAS, 2008]

2.1. nodes in *H*′ (edges in *H*) are assigned single community

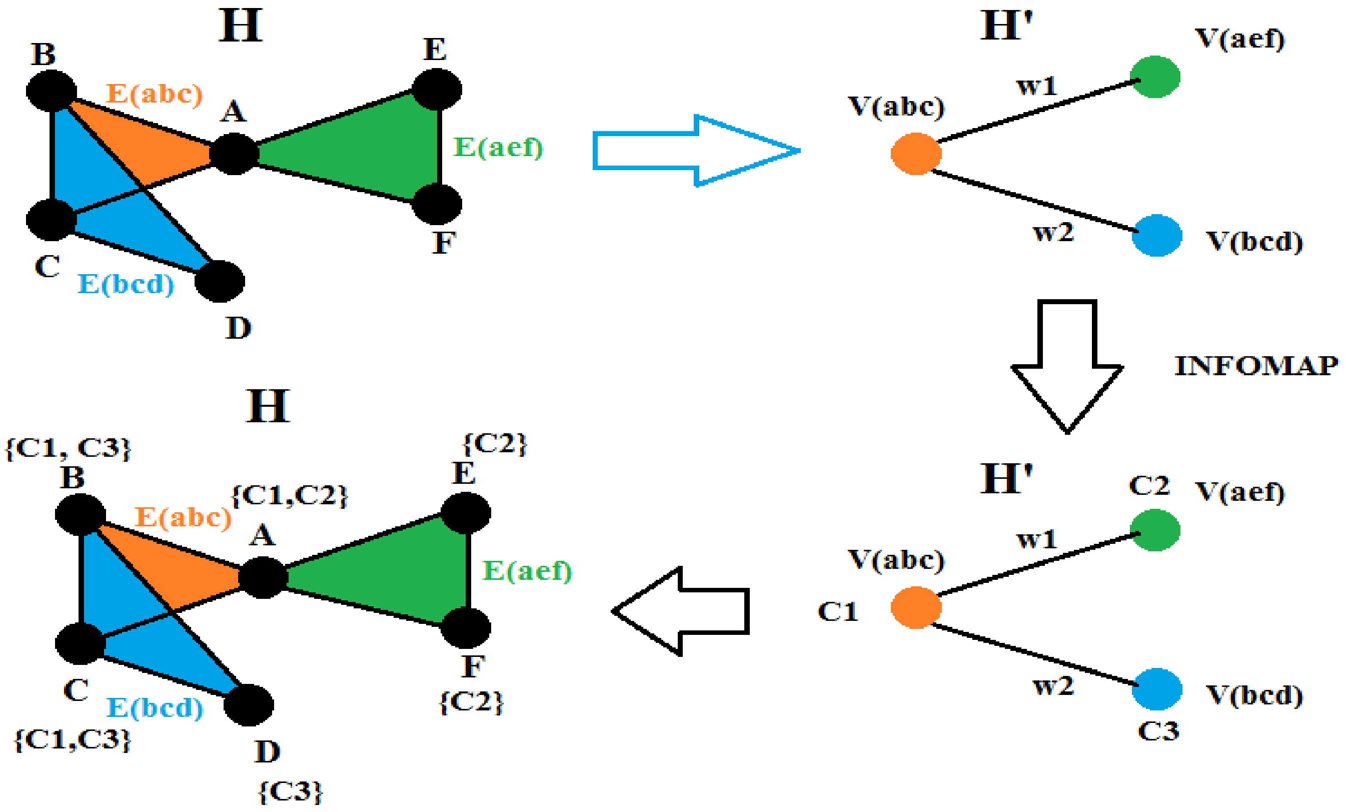


OverCite:

Overlapping community Detection in Citation Network

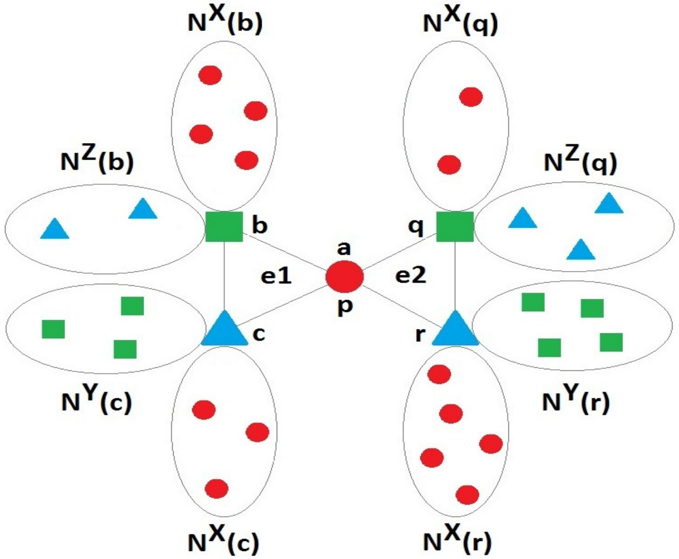
STEP 3. Unfold H' to produce H

3.1 Each node is assigned with community tags of its connected edges



OverCite:
Similarity Measures

1. Hypergraph Neighbourhood Similarity (HNS):



Hyperedges:
$e_1 = (a, b, c)$, $e_2 = (p, q, r)$
where a = p

$$S = N^X(b) \bigcup N^X(c)$$

$$S' = N^X(q) \bigcup N^X(r)$$

$$\alpha(e_1, e_2) = \frac{|S \bigcap S'| + |N^Y(c) \bigcap N^Y(r)| + |N^Z(b) \bigcap N^Z(q)|}{|S \bigcup S'| + |N^Y(c) \bigcup N^Y(r)| + |N^Z(b) \bigcup N^Z(q)|}$$

[Chakraborty et al, ACM HyperText, 2012]

OverCite:
Similarity Measures (Contd.)

2. Co-citation Strength (CCS):

Number of times two papers are cited together in the subsequent literatures.

If e_i=(a,b,c) and e_j=(x,y,z) and *CITE(b)*= set of papers which cite b

$$CCS(e_i, e_j) = \frac{|CITE(b) \cap CITE(y)|}{|CITE(b) \cup CITE(y)|}$$

3. Bibliographic-coupling Strength (BCS):

number of common citations two papers mention in the reference sections.

If e_i=(a,b,c) and e_j=(x,y,z) and *REF(b)*= set of papers cited by paper b

$$BCS(e_i, e_j) = \frac{|REF(b) \cap REF(y)|}{|REF(b) \cup REF(y)|}$$

Outline

Problem definition

Tripartite Publication Hypergraph

Our Proposed Algorithm: OverCite

Dataset

Evaluation

Recommendation System

Conclusion

Dataset

- Large **DBLP dump** used in Arnetminer project
  [Tang et al., SIGKDD, 2008]

- Bibliographic information during **1960-2008**
  - paper name
  - Author(s)
  - Publication venue
  - year of publication
  - Abstract
  - References

# of valid papers	702,973
# authors	495,311
Avg. number of papers/author	3.52
Avg. number of authors/paper	2.609
# unique venue name	1,705

- Missing

  Field information of each paper

Tagging Dataset

➢ Field Tagging

- Automated crawling of Microsoft Academic Search

[http://academic.research.microsoft.com/]

24 Fields

AI	Bioinformatics	NLP
Algorithm	Graphics	WWW
Networking	Comp. Vision	Education
Database	Data Mining	OS
Dist Comp.	Prog. Lang.	Embedded Sys.
Architecture	Security	Simulation
Software Engg.	IR	HCI
Machine Learning	Scientific Comp.	Multimedia

11.23% papers belong to multiple fields

Publicly available: http://cnerg.org
http://cse.iitkgp.ac.in/~tanmoyc/

Dataset: Ground-Truth Communities

- Each field servers as scientific community
- Total 24 fields => **ground-truth communities**
- Papers belonging to multiple fields => **overlapping nodes**

[Chakraborty et al., ASONAM, 2013]

Outline

Problem definition

Tripartite Publication Hypergraph

Our Proposed Algorithm: OverCite

Dataset

Evaluation

Recommendation System

Conclusion

Evaluation Metrics

1. Rand Index

[Rand, Journal of the American Statistical Association, 1971]

2. Omega Index:

Overlapping version of Rand Index

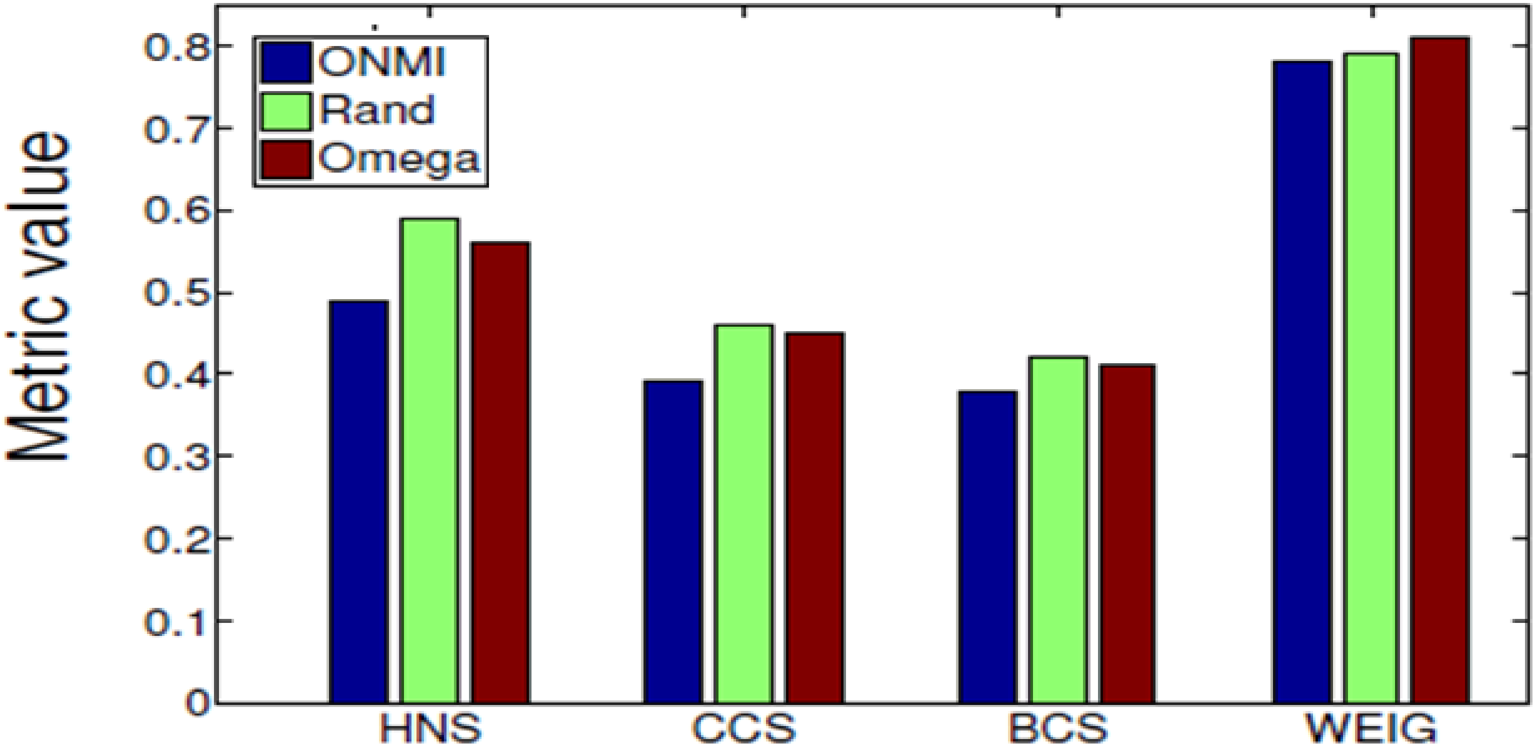
[Collins & Den, Multivariate Behavioural Research, 1988]

3. ONMI:

Overlapping Normalized Mutual Information

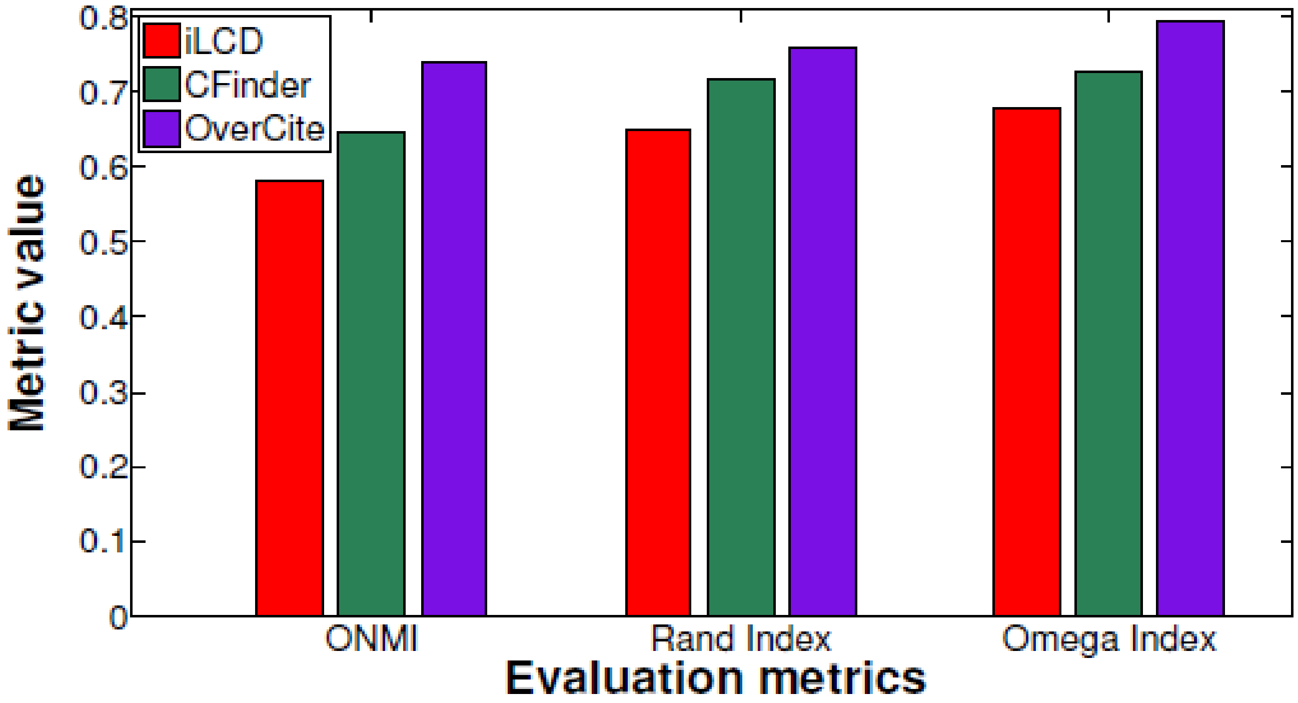
[McDaid, D. Greene, CoRR, 2011]

Significance of Similarity Measures



Best parameter selection **α= 0.45 , β= 0.32, γ=0.23**

Comparing Different algorithms



iLCD: Overlapping community in dynamic networks

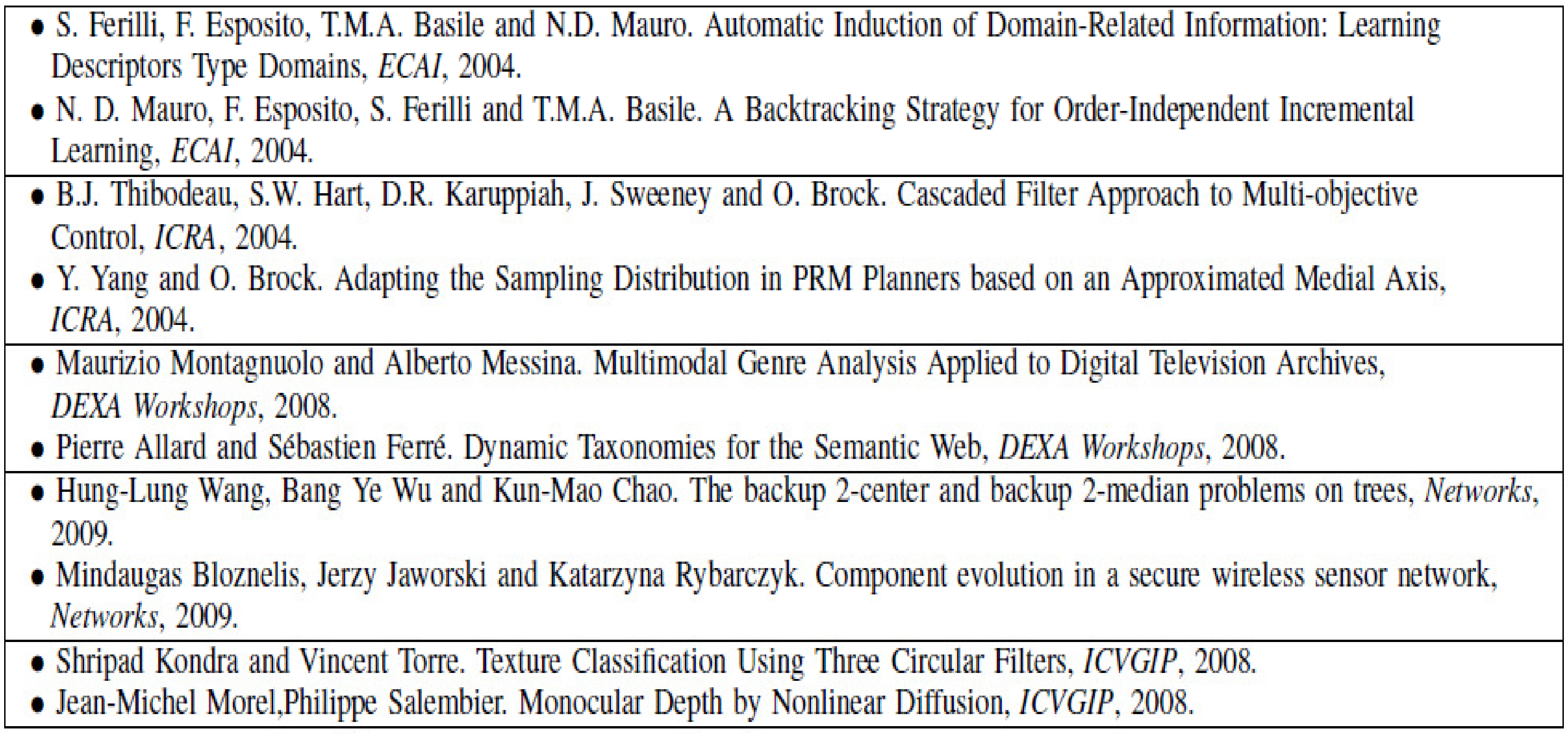
[Cazabet et al, *SocialCom,* 2010]

CFinder: Clique percolation Algorithm

[Palla et al., *Nature* , 2005]

Exploring Communities detected by OverCite

Selected pair of papers assigned into same community by **OverCite**, but not by others



• S. Ferilli, F. Esposito, T.M.A. Basile and N.D. Mauro. Automatic Induction of Domain-Related Information: Learning Descriptors Type Domains, *ECAI*, 2004. • N. D. Mauro, F. Esposito, S. Ferilli and T.M.A. Basile. A Backtracking Strategy for Order-Independent Incremental Learning, *ECAI*, 2004.
• B.J. Thibodeau, S.W. Hart, D.R. Karuppiah, J. Sweeney and O. Brock. Cascaded Filter Approach to Multi-objective Control, *ICRA*, 2004. • Y. Yang and O. Brock. Adapting the Sampling Distribution in PRM Planners based on an Approximated Medial Axis, *ICRA*, 2004.
• Maurizio Montagnuolo and Alberto Messina. Multimodal Genre Analysis Applied to Digital Television Archives, *DEXA Workshops*, 2008. • Pierre Allard and Sébastien Ferré. Dynamic Taxonomies for the Semantic Web, *DEXA Workshops*, 2008.
• Hung-Lung Wang, Bang Ye Wu and Kun-Mao Chao. The backup 2-center and backup 2-median problems on trees, *Networks*, 2009. • Mindaugas Bloznelis, Jerzy Jaworski and Katarzyna Rybarczyk. Component evolution in a secure wireless sensor network, *Networks*, 2009.
• Shripad Kondra and Vincent Torre. Texture Classification Using Three Circular Filters, *ICVGIP*, 2008. • Jean-Michel Morel,Philippe Salembier. Monocular Depth by Nonlinear Diffusion, *ICVGIP*, 2008.

✓ Either authors or conferences are same in each pair

Outline

Problem definition

Tripartite Publication Hypergraph

Our Proposed Algorithm: OverCite

Dataset

Evaluation

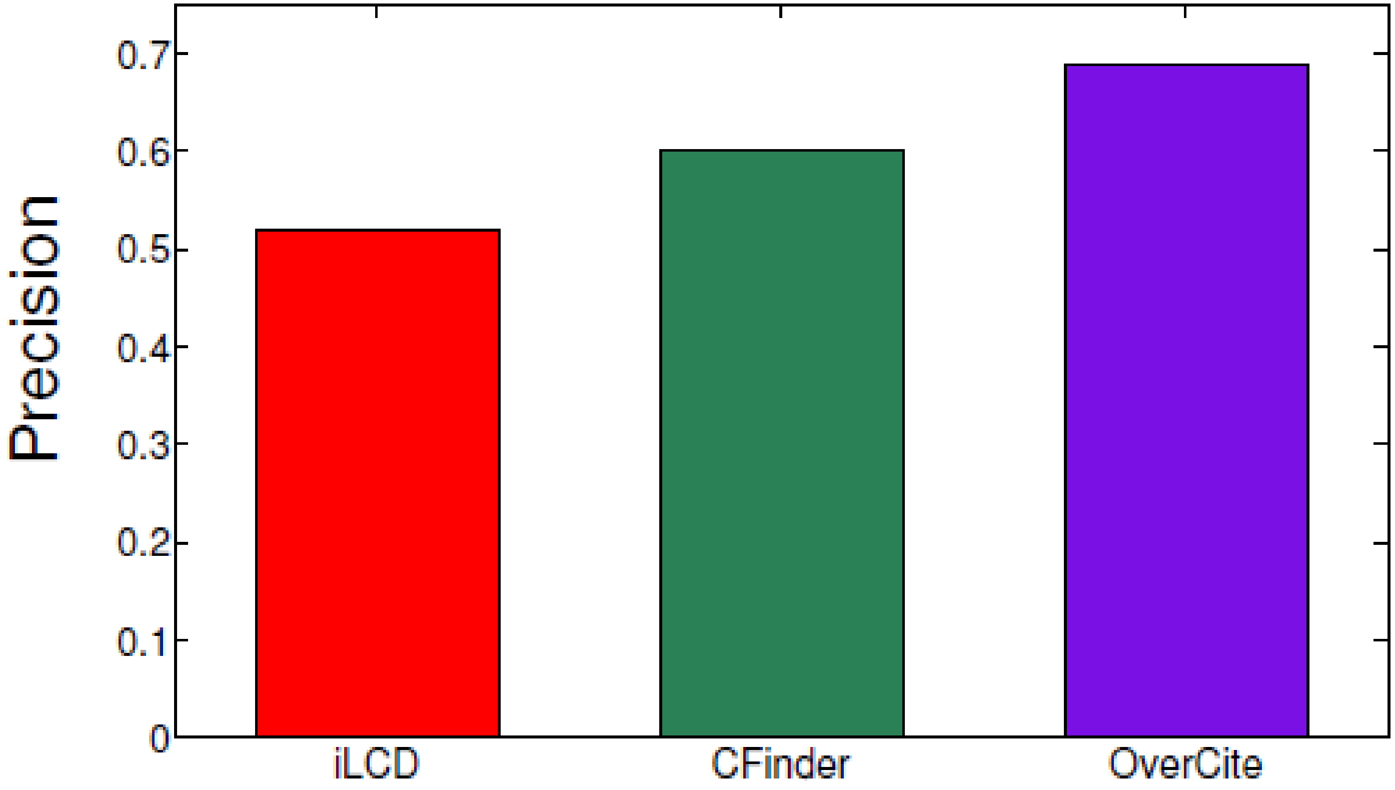
Recommendation System

Conclusion

Experimental Setup

- 38 students of CSE dept. were selected for evaluation
- 270 unique papers were searched.
- For each searched paper, system recommends other relevant papers purely based on communities identified by iLCD, CFinder and OverCite
- Volunteers were asked to tag each recommended paper as **Relevant** or **Non-relevant**
- Total 3612 relevance judgments were received

Recommendation Results



Outline

- Problem definition
- Tripartite Publication Hypergraph
- Our Proposed Algorithm: OverCite
- Dataset
- Evaluation
- Recommendation System

Conclusion

Conclusion

- Publications are represented by Tripartite Hypergraph Structure

- Edge based clustering instead of node based

- Both graph-based and citation-based similarity measure

- simple recommendation systems performs well over others

Future works

- Applying this approach to other domains like Facebook, Folksonomies etc.

- Finding relationship between performance with no of partitions of the hypergraph

- Incorporate **collaborative filtering** to improve recommendation system

Acknowledgements

- Financial Support: Google India Pvt. Ltd.

- Travel support:

  Dept. of Science & Technology, Govt. of India

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

- Experimental Support

  Ayushi Dalmia, HIT, Kolkata, India

- Technical support:

  All the members of **CNeRG, IIT-Kgp**

http://cse.iitkgp.ac.in/~tanmoyc/

http://cnerg.org/