Mining Research Topic-related Influence between Academia and Industry

Dan He
PKDD 2011, Athens, Greece
danhe@cs.ucla.edu

Background

- Mining influence in networks, especially social networks has attracted tremendous attentions.

Background

- Mining influence in networks, especially social networks has attracted tremendous attentions.
- Topic-level influence between academia and industry is useful in many cases:
  - Students whose career goal is to be a researcher in a company's research lab often seek advisor with high influence in industry.
  - Funding agencies may choose to grant certain awards to researchers who work closely with industry companies, on certain topics.
  - Companies may want to collaborate with academia researchers who have tight industry connection.

Background

- Mining influence in networks, especially social networks has attracted tremendous attentions.
- Topic-level influence between academia and industry is useful in many cases:
  - Students whose career goal is to be a researcher in a company's research lab often seek advisor with high influence in industry.
  - Funding agencies may choose to grant certain awards to researchers who work closely with industry companies.
  - Companies may want to collaborate with academia researchers who have tight industry connection.
- To our knowledge, our work is the first one on mining influences between academia and industry.

Properties of co-authorship network

- Nodes: researchers.
- Edges: co-authorship between the pair of researchers.
- Edge weights: number of publications co-authored by the pair of researchers.
- Unlike social networks, there are no temporal "actions": the pair of authors publish at the same time.

Outline

- Topic-specific influence model in co-authorship network
- Three models for influence between academia and industry
  - Simple Additive model
  - Weighted Additive model
  - Clustering-based Additive model
- Evaluations of the three models
- Conclusions

Outline

- Topic-specific influence model in co-authorship network
- Three models for influence between academia and industry
  - Simple Additive model
  - Weighted Additive model
  - Clustering-based Additive model
- Evaluations of the three models
- Conclusions

Models for pairwise influences in co-authorship network

- Naïve Method:
  - Based on the number of co-authored publications
  - Influence(A→B) = $\frac{\#co-publication(A,B)}{\#publication(B)}$
  - Problem of the method: Considers A,B as independent to other authors, which is usually not true.

Models for pairwise influences in co-authorship network

- TAP Method (Tang et al., KDD 2009):
  - Based on topical affinity propagation (TAP)
  - A topical factor graph is built on top of the co-authorship network
  - A likelihood function for the graph is maximized by TAP.
  - Influences are computed via TAP.
  - Models influences among multiple researchers on specific topics.

Outline

- Topic-specific influence model in co-authorship network
- Three models for influence between academia and industry
  - Simple Additive model
  - Weighted Additive model
  - Clustering-based Additive model
- Evaluations of the three models
- Conclusions

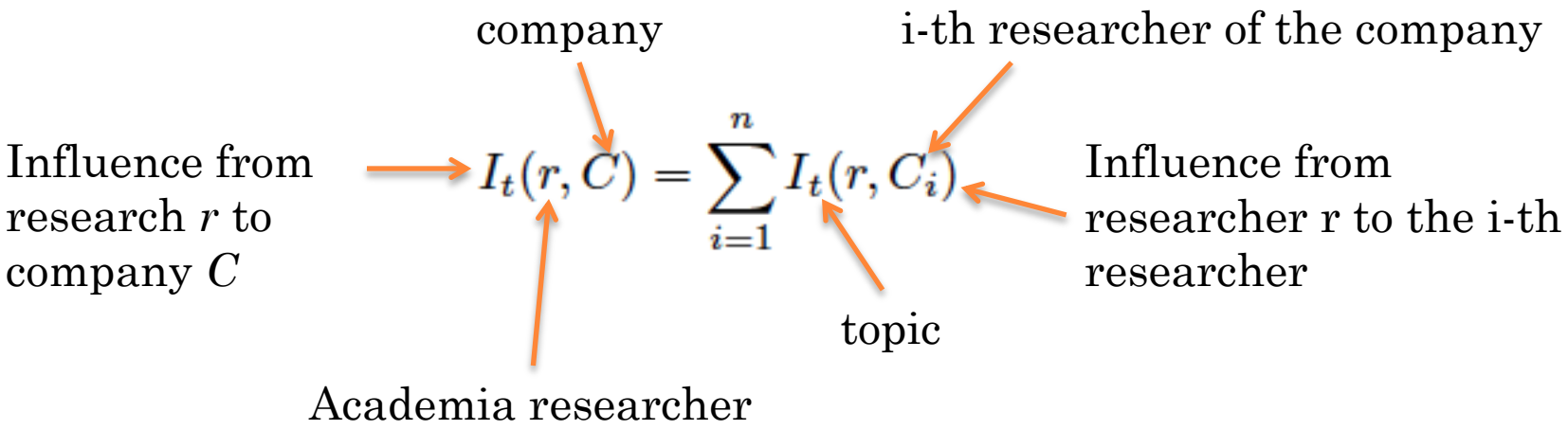
Prerequisites

- Assume we already obtain the topic-specific influences between researchers in the co-authorship network, via the TAP method.
- The influences are directed.
- The influences between two nodes are usually not symmetric.

$$Influence(A \rightarrow B) \neq Influence(B \rightarrow A)$$

Simple Additive Model

- Basic Idea: Simply sum the influences on topic t from an academia researcher r to all the researchers in a company C



Simple Additive Model

- Problem of the model: not all the researchers are of equal importance to the company.
  - A manager should have higher weight than his/her group members do
  - A senior researcher should have higher weights than a junior researcher does

Weighted Additive Model

- Basic Idea: Each industry researcher is weighted according to their internal influences in their company.

i-th researcher of the company

Internal influence of C_i

Weight of C_i → $$W_t(C_i) = \frac{\sum_{j=1}^{n} I_t(C_i, C_j)}{\sum_{j=1,k=1}^{j=n,k=n} I_t(C_k, C_j)}$$

$$I_t(r, C) = \sum_{i=1}^{n} I_t(r, C_i) \times W_t(C_i)$$

Total influence in the company

Weighted Additive Model

- Basic Idea: Each industry research is weighted according to their internal influences in their company.
- Problem of the model: there might be researchers often publish together.
  - These researchers will have high influence to each other.
  - They may have high internal influences, leading to biased high weights.

Weighted Additive Model

- Basic Idea: Each industry research is weighted according to their internal influences in their company.
- Problem of the model: there might be researchers often publish together.
  - These researchers will have high influence to each other.
  - They may have high internal influence, leading to biased high weights.
- Question: How to normalize the weights of these researchers to address the bias?

Clustering-based Additive Model

- Basic Idea: Put the researchers that collaborate a lot into a cluster.
  - Consider each cluster as a *super researcher*.
  - Average the influences in a cluster.
  - Conduct the weighted additive model on the cluster level instead of on the researcher level.
  - The weights of more correlated researchers, or researcher who collaborate often, can be adjusted appropriately.

Clustering the researchers

- We do not know a reasonable number of clusters for clustering algorithms such as K-means, EM, spectral clustering etc.
- Modularity based clustering do not require a number of clusters, but it ignores the distances between the nodes.

Clustering the researchers

- We do not know a reasonable number of clusters for clustering algorithms such as K-means, EM, spectral clustering etc.
- Modularity based clustering do not require a number of clusters, but it ignores the distances between the nodes.
- We propose a clustering algorithm which is based on distances between the nodes and does not require pre-knowledge on the number of clusters.

Clustering the researchers

- Similarity of two researchers: Jaccard's coefficient on the number of publications.

$$\frac{\#co-publication}{\#total\quad publication}$$

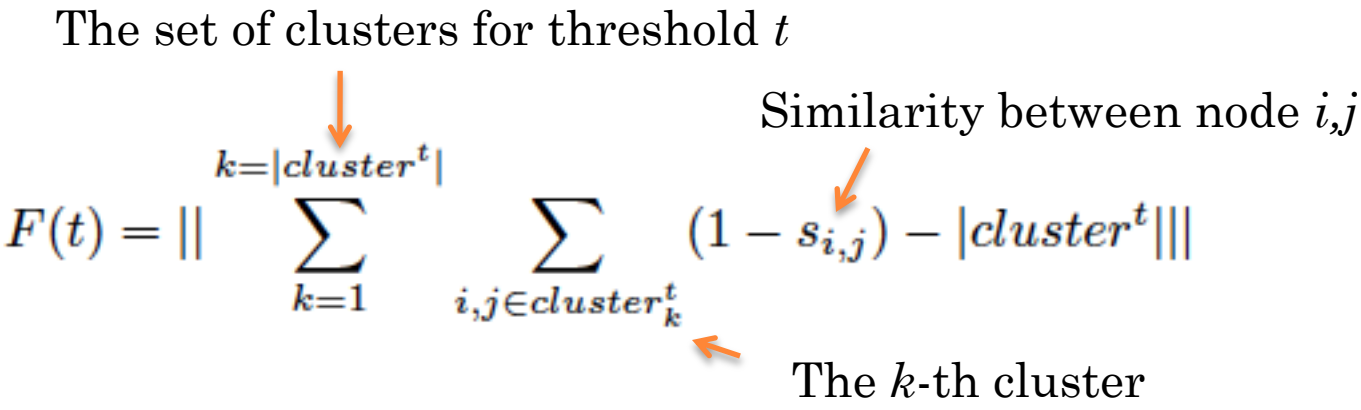
- We use number of publications instead of influence since the similarity needs to be symmetric.

Clustering the researchers

- The nodes in the same cluster are fully-connected, namely each cluster is a clique.
- Given a threshold *t*, an edge connects two nodes only when the similarity of the two nodes is no less than *t*
- We can generate a set of clusters, by searching for cliques in the graph. The nodes in the same clique are of similarity no less than *t* to each other.
- We want to select an appropriate *t* automatically.

Clustering the researchers

- We propose an objective function for *t*



- Find a *t* that minmizes *F(t)*
- Motivation of the objective function:
  - Minimize the sum of distances between the nodes in the clusters.
  - Also take into consideration the number of clusters

Clustering the researchers

- Find t to minimize $F(t)$:
  - Compute pairwise similarity for all the nodes in the co-authorship network.
  - We prove at least one of the pairwise similarities is able to minimize $F(t)$.
  - We try all pairwise similarities and find the one that minimizes $F(t)$.
  - If we assume every researcher has only constant number of collaborators h, this step takes $O(nh)$ complexity, where n is the total number of nodes. Since h is a constant and usually speaking $h << n$, $O(nh) \approx O(n)$.

Weighted Additive Model on Clusters

- Compute the weight of the researchers in the cluster according to the normal weighted additive model:

$$W_t(L_i) = \frac{\sum_{j=1}^{n} I_t(L_i, L_j)}{\sum_{j=1,k=1}^{j=n,k=n} I_t(L_k, L_j)}$$

- Compute the influence from a researcher *r* to a cluster *L* based on the weights of the researchers in the cluster:

$$I_t(r, L) = \sum_{i=1}^{n} I_t(r, L_i) \times W_t(L_i)$$

Weighted Additive Model on Clusters

- Compute the influence between pair of clusters based on the weights of researchers in the clusters

$$I_t(L,K) = \sum_{i=1,j=1}^{i=|L|,j=|K|} I_t(L_i,K_j) \times W_t(L_i) \times W_t(K_j)$$

- Compute the weights of the clusters by considering each cluster as a "virtual researcher"

$$W_t(C_L) = \frac{\sum_{K \in cluster(C)} I_t(C_L,C_K)}{\sum_{L \in cluster(C), K \in cluster(C)} I_t(C_L,C_K)}$$

Weighted Additive Model on Clusters

- Compute the influence of an academia researcher r to a company C according to the normal weighted additive model, by considering each cluster as a "virtual researcher".

$$I_t(r, C) = \sum_{L \in cluster(C)}^{n} I_t(r, C_L) \times W_t(C_L)$$

Clustering-based additive model

- As we will show later in the experiments, when there are bias towards the researchers who often publish together, the simple additive model and the weighted additive model are not able to address the bias.
- The clustering-based additive model is able to better address the bias.

Outline

- Topic-specific influence model in co-authorship network
- Three models for influence between academia and industry
  - Simple Additive model
  - Weighted Additive model
  - Clustering-based Additive model
- Evaluations of the three models
- Conclusions

Experiments on Real Data

- Experiments Set up:
  - The same data set as used in the work of Tang et al, KDD 2009, with 8 different topics.
  - Researchers are affiliated with either companies or universities
  - One researcher only has one affiliation.

Topics	#Researchers	#Companies	#Universities
Data Mining	679	33	99
Machine Learning	976	48	97
Database System	1127	66	116
Information Retrieval	657	49	87
Web Services	400	27	48
Semantic Web	671	38	35
Bayesian Network	554	24	45
Web Mining	348	25	47

Experiments on Real Data

- Experiments Set up:
  - The researchers with missing affiliation are excluded.
  - Influences on the 8 topics among the researchers are computed via the TAP method.

Topics	#Researchers	#Companies	#Universities
Data Mining	679	33	99
Machine Learning	976	48	97
Database System	1127	66	116
Information Retrieval	657	49	87
Web Services	400	27	48
Semantic Web	671	38	35
Bayesian Network	554	24	45
Web Mining	348	25	47

SAMPLE RESULTS

- The top-5 most influential researchers to Microsoft and IBM from the three models on the topic "Data Mining"

Data Mining (Microsoft)			Data Mining (IBM)		
simple additive	*weighted additive*	*clustering-based additive*	*simple additive*	*weighted additive*	*clustering-based additive*
Jiawei Han	Huan Liu	Clark Glymour	Jiawei Han	Jiawei Han	Jiawei Han
Huan Liu	Clark Glymour	Huan Liu	Philip S. Yu	Philip S. Yu	Philip S. Yu
Xifeng Yan	Michail Vlachos	Michail Vlachos	Michail Vlachos	Michail Vlachos	Michail Vlachos
Clark Glymour	Xuanhui Wang	Padhraic Smyth	Tao Tao	Tao Tao	Tao Tao
Philip S. Yu	Padhraic Smyth	Bing Liu	Ricardo Vilalta	Ricardo Vilalta	Ricardo Vilalta

SAMPLE RESULTS

- The top-5 most influential researchers to Microsoft and IBM from the three models on the topic "Data Mining"
- The ranks may not be accurate due to missing affiliation information.

Data Mining (Microsoft)			Data Mining (IBM)		
simple additive	*weighted additive*	*clustering-based additive*	*simple additive*	*weighted additive*	*clustering-based additive*
Jiawei Han	Huan Liu	Clark Glymour	Jiawei Han	Jiawei Han	Jiawei Han
Huan Liu	Clark Glymour	Huan Liu	Philip S. Yu	Philip S. Yu	Philip S. Yu
Xifeng Yan	Michail Vlachos	Michail Vlachos	Michail Vlachos	Michail Vlachos	Michail Vlachos
Clark Glymour	Xuanhui Wang	Padhraic Smyth	Tao Tao	Tao Tao	Tao Tao
Philip S. Yu	Padhraic Smyth	Bing Liu	Ricardo Vilalta	Ricardo Vilalta	Ricardo Vilalta

Sample results

- The top-5 most influential researchers to Microsoft and IBM from the three models on the topic "Machine Learning"

Machine Learning (Microsoft)		
simple additive	*weighted additive*	*clustering-based additive*
Brendan J. Frey	Aaron Hertzmann	Aaron Hertzmann
Aaron Hertzmann	Michael I. Jordan	Michael I. Jordan
Andrew McCallum	Andrew McCallum	Andrew McCallum
Michael I. Jordan	Brendan J. Frey	William T. Freeman
William T. Freeman	William T. Freeman	Yoav Freund
Machine Learning(IBM)		
simple additive	*weighted additive*	*clustering-based additive*
Inderjit S. Dhillon	Inderjit S. Dhillon	Inderjit S. Dhillon
Adam R. Klivans	Manfred K. Warmuth	Manfred K. Warmuth
Nader H. Bshouty	Roni Khardon	Roni Khardon
Joydeep Ghosh	Geoffrey J. Gordon	Geoffrey J. Gordon
Manfred K. Warmuth	Gerald Tesauro	Gerald Tesauro

Evaluation on the Clustering Algorithm

- Our clustering algorithm found clusters of researchers who collaborated quite often.
- We show 4 pairs of researchers who co-authored many papers. According to the total number of their publications, we guess one of them is relatively senior and the other is relatively junior.
- We validate the titles of the senior researchers and it is true that all of them are manager/principal

Senior	Junior	Co-authored
47	23	11
64	16	9
62	18	11
94	28	22

Experiments on Simulated Data

- Due to lack of ground-truth and missing affiliation, we can not effectively validate our models on real data.

Experiments on Simulated Data

- We validate our models on simulated data:
  - Generate a company with 200 researchers
  - Randomly select 20 of them as manager
  - Assign the remaining researchers to one of the 20 groups randomly
  - Generate 400 researchers in academia
  - Ground-truth of two types of most influential academia researchers to the company:
    - *Influence many researcher:* Academia researcher who influences many researchers of the company
    - *Influence important researchers:* Academia researcher who influences important researchers, such as manager, of the company

Experiments on Simulated Data

- We select 10 *influence many* researchers and 10 *influence important* researchers
- We also select 20 *influence same group* researchers, who influence researchers in the same group of the company.
- Rank the academia researchers according to their influences to the company by the three models.
- We expect the top-20 most influential researchers are the 10 *influence many* researchers + the 10 *influence important* researchers

Experiments on Simulated Data

- We conduct two sets of experiments:
  - The researchers in the same group have LOW influence to each other (no bias)
  - The researchers in the same group how HIGH influence to each other (with bias)

Experiments on Simulated Data

- We conduct two sets of experiments:
  - The researchers in the same group have LOW influence to each other (no bias)
  - The researchers in the same group how HIGH influence to each other (with bias)
- What we observe:
  - Simple Additive model tends to assign high ranks to the *influence many* researchers

Experiments on Simulated Data

- We conduct two sets of experiments:
  - The researchers in the same group have LOW influence to each other (no bias)
  - The researchers in the same group how HIGH influence to each other (with bias)
- What we observe:
  - Simple Additive model tends to assign high ranks to the *influence many* researchers
  - Weighted Additive model tends to assign high ranks to the *influence important* researchers

Experiments on Simulated Data

- What we observe:
  - Clustering-based additive model
    - Is similar to the weighted additive model, if researchers in the same group have low influence to each other (no bias)

Experiments on Simulated Data

- What we observe:
  - Clustering-based additive model
    - Is similar to the weighted additive model, if researchers in the same group have low influence to each other (no bias)
    - Tends to assign higher ranks to the *influence important* researchers and lower ranks to the *influence same group* researchers, if researchers in the same group have high influence to each other (with bias)

Experiments on Simulated Data

- What we observe:
  - Clustering-based additive model
    - Is similar to the weighted additive model, if researchers in the same group have low influence to each other (no bias)
    - Tends to assign higher ranks to the *influence important* researchers and lower ranks to the *influence same group* researchers, if researchers in the same group have high influence to each other (with bias)
  - When the researchers in the same group have high influence to each other, clustering-based additive model is able to better distinguish the *influence same group* researchers from the *influence important* and the *influence many* researchers.

Outline

- Topic-specific influence model in co-authorship network
- Three models for influence between academia and industry
  - Simple Additive model
  - Weighted Additive model
  - Clustering-based Additive model
- Evaluations of the three models
- Conclusions

Take Home Message

- We addressed the problem of mining research topic-specific influence between academia and industry.
- We proposed three models to learn the influence:
  - Simple additive model
  - Weighted additive model
  - Clustering-based additive model
- The influence from industry to academia can be obtained using the same models but influence with reverse directions.

Thanks

- Questions?