Social Network Mining Based on Academic Literatures

Yikai Wang
School of Data Science
Fudan University
15300180076@fudan.edu.cn

Dan Wu
School of Data Science
Fudan University
15307130399@fudan.edu.cn

Zheng Wei
School of Data Science
Fudan University
15302010015@fudan.edu.cn

I. Task Review

A. Task 1

Design clustering algorithms or community mining algorithms to cluster all the papers in the data set. Use visualize tools to show all fields (ie, communities, identify corresponding community research topics), and highlight the most influential scholars in each field.

B. Task 2

Realization of demonstrating the ego-network to any input scholar (refer to the function example on the ArnetMiner website).

C. Task 3

Use the data provided by DBLP and ArnetMiner to analyze and model more social relationships among scholars, such as predicting the cooperation or citation relationship between two scholars, and predicting which conference will a scholar publish papers on in the future.

II. Division of Work

1) Yikai Wang:
   In brief, he applies several methods in network representation learning and uses and expands a hierarchical representation learning algorithm(HARP) for Networks. For homogeneous networks, he mainly uses deepwalk, node2vec and LINE. For heterogeneous networks, he uses metapath2vec++. For task 1, he uses both homogeneous NRL and heterogeneous NRL methods to extract features for clustering and uses KMeans and Birch to complete the clustering task. For task 3, he uses heterogeneous NRL methods to generate the vector representation of scholar cooperation network, scholar citation network and scholar-conference network.
2) Dan Wu:
   a) Revise the preprocessing part of word2vec model in task 1. (Stop words deleted, all words are transformed into lower case.)
   b) Build a LDA model to get the key words of each group of the papers in task 1.
   c) Use PageRank index to calculate the influence of each paper and scholar. Visualize the citation and cooperative relationship in task 1.
   d) Visualize the ego-network in task 2.
   e) Build a baseline model for link prediction in task 3.
3) Zheng Wei:
   In this project, his work is mainly divided into two part: For task1, he uses the word vector file to train the sentence vector of each paper, and he is responsible for generating small data sets for testing. For task3, he uses the MDP model to train the transition matrix and predicts the relationship between scholars and conferences. Based on this, he further predict the relationship between scholars and scholars.

III. Methodology

A. Network Representation Learning

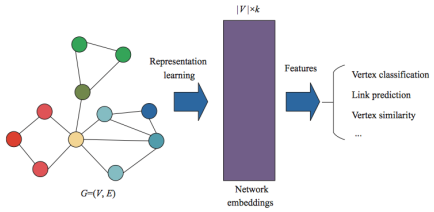


Fig. 1. Network representation learning flowchart

Network is an important form of expressing the relationship between objects and objects. A key issue for the analysis of networks is to study how to reasonably represent feature information in the network. With the development of machine learning technology, feature learning for nodes in the network has become an emerging research task. Network representation learning algorithm transforms network information into low-dimensional dense real vector.

There are two types of networks in NRL, one supposes that all nodes in the network have the same type, and the other supposes that all nodes have different types. The former network is called homogeneous network and the other is called heterogeneous network.

Formally, let $G = (V, E)$ be a graph, where V is the set of nodes and E is the set of edges. The goal of network representation learning is to develop a mapping function $\Phi : V \rightarrow R^{|V|*d}, d \ll |V|$, which defines the latent

representation of each node $v \in V$. Here we briefly introduce some homogeneous network learning methods and detailed introduce a heterogeneous network learning method and an optimization algorithm on all these methods.

Methods for homogeneous network learning

1) DeepWalk:
   The idea of DeepWalk [1] is borrowed from word2vec. It uses short random walks to generate paths of a node, then just like word2vec, it uses the path to generate a probability of its neighbors, which could give each node a vertices in networks.
2) Node2vec:
   Node2vec [2] is an improvement for DeepWalk. In short, it regards the random walk as a searching problem. It provides a way of balancing the exploration-exploitation tradeoff that in turn leads to representations obeying a spectrum of equivalences from homophile to structural equivalence.
3) LINE:
   LINE [3] uses another way to realize network embedding. The main idea of this method is learning a low-dimensional embedding with preserving both the first-order proximity and the second-order proximity between the vertices.

Methods for heterogeneous network learning

Different with homogeneous networks, heterogeneous networks involve more than one node types and relationships between the same type of nodes and/or different types of nodes. Thereafter, these networks cannot be handled by representation learning models that specifically designed for homogeneous networks. In this project, we use a heterogeneous skip-gram model, *metapath2vec++* [4], to model the heterogeneous neighborhood of a node.

Specifically, in *metapath2vec++*, we enable skip-gram to learn effective node representations for a heterogeneous network $G = (V, E, T)$ with $|T_V| > 1$ by maximizing the probability of having the heterogeneous context $N_t(v), t \in T_V$ given a node v:

$$argmax_\theta \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} logp(c_t|v;\theta) \tag{1}$$

where $N_t(v)$ denotes v's neighborhood with the t^{th} types of nodes and $p(c_t|v;\theta)$ is commonly defined as a softmax function, that is: $p(c_t|v;\theta) = \frac{exp\{X_{c_t} * X_v\}}{\sum_{u_t \in V_t} exp\{X_{u_t} * X_v\}}$, where X_v is the v^{th} row of X, representing the embedding vector for node v. V_t is the node set of type t in the network.

However, the method will cost a terrible long time. To solve the problem, negative sampling was introduced by Mikolov et al [5]. Using this method, we just need to sample a small set of nodes from the network in order to construct softmax. Specifically, given a negative sample size M, we use following method to update Equation 1:

$$log\sigma(X_{c_t} * X_v) + \sum_{m=1}^{M} E_{u_t^m \sim P_t(u_t)}[log\sigma(-X_{u_t^M} * X_v)] \tag{2}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ and $P(u)$ is the pre-defined distribution from which a negative node u^m is drew from for M times. In our model, we regard different types of nodes homogeneously and do not distinguish them when drawing negative nodes. The gradients of equation 2 are derived as follows:

$$\begin{aligned} \frac{\partial O(X)}{\partial X_{u_t^m}} &= (\sigma(X_{u_t^m} X_v - I_{c_t}[u_t^m]))X_v \\ \frac{\partial O(X)}{\partial X_v} &= \sum_{m=0}^{M} (\sigma(X_{u_t^m} X_v - I_{c_t}[u_t^m]))X_{u_t^m} \end{aligned} \tag{3}$$

Having defined the skip-gram model, the problem we need to solve is how to effectively transform the heterogeneous network into skip-gram. Similar to homogeneous network, we can use random walk to generate paths of multiple types of nodes. Different with homogeneous network, in heterogeneous network, we need to consider the type of the nodes in the random walk.

Formally, a meta-path scheme P is defined as a path that is denoted in the form of $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \dots V_t \xrightarrow{R_t} V_{t+1} \dots \xrightarrow{R_{l-1}} V_l$, wherein $R = R_1 \circ R_2 \circ \dots \circ R_{l-1}$ defines the composite relations between node types V_1 and V_l. Thus we could show how to use meta-paths to guide heterogeneous random walkers. Given a heterogeneous network $G = (V, E, T)$ and a meta-path scheme P, the transition probability at step i is as follows:

$$p(v^{i+1}|v_t^i, P) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) = t+1 \\ 0 & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) \neq t+1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases} \tag{4}$$

where $v_t^i \in V_t$ and $N_{t+1}(v_t^i)$ denote the V_{t+1} type of neighborhood of node v_t^i. Further more, meta-paths are commonly used in a symmetric way, which means its first node type is the same as the last one. That is:

$$p(v^{i+1}|v_t^i) = p(v^{i+1}|v_1^i), \text{if } t = l \tag{5}$$

The complete algorithm is expressed as follows:

HARP

The network embedding method we introduced above is very useful in some situations. However, there are two main disadvantages for these methods:

1) all the models do not involve high-order network structural information
2) their stochastic optimization can fall victim to poor initialization

Thus, we change the traditional problem into hierarchical representation learning problem. The main idea is we seek to find a graph $G_s = (V_s, E_s)$ which captures the essential structure of G, but is much smaller(i.e. $|V_s| \ll |V|, |E_s| \ll |E|$). It is trivial that G_s is easier to embed. Since we have much less relationships, which means the mapping could be much smoother. Further more, since the G_s is much smaller, the models that focus on local structure now could have a better performance on global structure's representation.

Algorithm 1: The metapath2vec++ Algorithm.

Input: The heterogeneous information network $G = (V, E, T)$, a meta-path scheme P,# walks per node w, walk length l, embedding dimension d,neighborhood size k

Output: The latent node embeddings $X \in R^{|V|*d}$

initialize X;
for $i = 1; i \leq w$ **do**
  for $v \in V$ **do**
    MP=MetaPathRandomWalk(G, P, v, l);
    X=HeterogeneousSkipGram(X, k, MP);
return X;
MetaPathRandomWalk(G, P, v, l) MP[1]=v;
for $i = 1; i < l$ **do**
  draw u according to Eq. 4;
  MP[i+1]=u;
return MP;
HeterogneousSkipGram(X, k, MP) **for** $i = 1; i \leq l$ **do**
  v = MP[i];
  for $j = max(0, i - k); j \leq min(i + k, l); j \neq i$ **do**
    $c_t = MP[j]$;
    $X^{new} = X^{old} - \eta \frac{\partial O(X)}{\partial X}$(Eq. 3;

Algorithm 2: HARP(G,Embed())

Input: network $G = (V, E)$, arbitrary network embedding algorithm EMBED()

Output: node embeddings $\Phi \in R^{|V|*d}$

$G_0, G_1, \ldots, G_L \leftarrow$GRAPHCOARSENING(G);
Initial $\Phi^{'}_{G_L}$ by assigning zeros;
$\Phi_{G_L} \leftarrow EMBED(G_L, \Phi^{'}_{G_L})$ **for** $i = L - 1; i \geq 0$ **do**
  $\Phi^{'}_{G_i} \leftarrow PROLONGATE(\Phi_{G_{i+1}}, G_{i+1}, G_i)$;
  $\Phi_{G_i} \leftarrow EMBED(G_i, \Phi^{'}_{G_i})$
return Φ_{G_0};
GraphGoarsening($G(V, E)$) $L \leftarrow 0$;
$G_0 \leftarrow G$;
while $|V_L| \geq threshold$ **do**
  $L \leftarrow L + 1$;
  $G_L \leftarrow$EDGECOLLAPSE(STARCOLLAPSE(G));
return $G_0, G_1, \ldots, G_L$;

Our method for multilevel network representation learning, HARP [6], consists of three parts-graph coarsening, graph embedding and representation refinement.

1) Graph Coarsening:
   Given a graph G, graph coarsening algorithms create a hierarchy of successively smaller graphs $G_0, G_1, \ldots, G_L$, where $G_0 = G$. The coarser graphs preserve the global structure of the original graph and have much fewer nodes and edges.
2) Graph Embedding on the Coarsest Graph:
   Using provided network embedding algorithms to to graph embedding.(DeepWalk, Node2vec, LINE, etc.)
3) Graph Representation Prolongation and Refinement:
   We prolong and refine the graph representation from the coarsest to the finest graph. For each graph G_i, we use the graph representation of G_{i+1} as its initial embedding and refine the graph embedding.

The complete algorithm is expressed as follows:

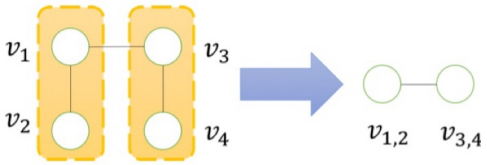
In the algorithm, **Edge Collapse** is an algorithm for preserving first-order proximity when coarsening the network. For edges E, it will select $E^{'} \subset E$, satisfying no two edges in the subset are incident to the same vertex. That is, for each $(u_i, v_i) \in E^{'}$, it merges the pairs into a single node w_i, and merge the edges incident to u_i and v_i. **Star Collapse** is an algorithm for preserving second-order proximity when coarsening the network. It will merges nodes with the same neighbors into super-nodes. For example, in the Figure 2,$(v_1, v_2), (v_3, v_4)$ and (v_5, v_6) are merged into super-nodes as they share the same neighbors (v_7).

It is worth mentioning that in the model of HARP, they just consider homogeneous networks. However, without much modify, it can be used in heterogeneous networks, too. Specifically, in the algorithm, we first do star collapse and then do edge collapse. For a heterogeneous network, since the edges are mostly between two different types of nodes, we could find that similar nodes with the same type will be regard as the same super-node, which is just what we hope to happen.

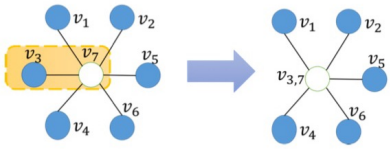
B. MDP

The Markov decision-making process is based on the historical state and makes predictions or decisions about the future state. The sequential nature of the recommendation process was noticed in the past [7] Taking this idea one step farther, we suggest that recommendation is not simply a sequential prediction problem, but rather, a sequential decision problem. At each point the Recommender System makes a decision: which recommendation to issue. This decision should take into account the sequential process involved and the optimization criteria suitable for the recommender system, such as the profit generated from selling an item. Thus, we suggest the use of Markov decision processes(MDP) [8] , a well known stochastic model of sequential decisions. [9] For task three, we must predict whether a scholar will send a paper or a cooperative relationship with other scholars in the future. This must be closely related to the behavior of the scholar in previous years. We will build transition matrices based on the state of scholars in previous years and make predictions about the future state of scholars. Task 3 can be divided into two subtasks: scholar-conference and scholar-scholar. We describe the algorithms for these two tasks separately below.

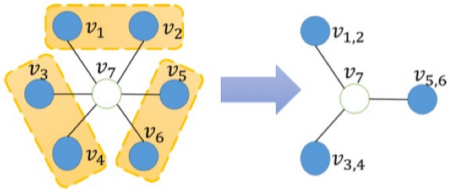
1) Scholar-Conference: We predict whether scholars will issue papers or not, which is determined by the situation of papers published by scholars in previous years. We will first select a parameter order to determine the order of the training



(a) Edge Collapsing.



(b) Edge Collapsing fails to collapse stars.



(c) Star Collapsing.

Fig. 2. Illustration of graph coarsening algorithms

transfer matrix, that is, for the forecast year, the year of the previous order. Let the number of conferences be n, which also represents the vector dimension of each year, a vector: $[c_1, c_2, \ldots, c_{21}]$. The value at each position represents the number of papers the author sent at the conference that year. Since we only care about whether the user has presented the paper at the relevant meeting, we set all the positive numbers in the vector to 1, which means that the author published the paper at the meeting. For an author, we will train a transfer matrix that specifically corresponds to it. The dimension of this transfer matrix is (n, n). It's both rows and columns represent the references' transfer, for example, if row is 3 and column is 2, it shows the third kind of conferences transfer the second kind of conferences. Each item in transfer matrix is a tuple and it contains four value, respectively $1 \to 1, 1 \to 0, 0 \to 1$, and $0 \to 0$.

After all train data has been employed to construst the transfer matrix, we need to do some calculation operations on the matrix. We should know the probability that the scholar issue paper or not. Of course, no matter what situation, the function of calculation is below:

$$Prob(issued) = \frac{count < issued \quad number >}{count < total \quad number >} \tag{6}$$

So, each item of transfer matrix will only contain two value standing for probability. Now, we can predict the solution of next year using this year's data. Defind V as this year's column vector, M as the transfer matrix, and V_next as next year's column vector. We can get:

$$V'_{next} = V' \times M \tag{7}$$

It is worth noting that above times operation is abnormal operation. we take i-th item from it as example:

$$V'^{(i)}_{next} = V' * M[column = i] \tag{8}$$

It looks like no strange, but we should know that item of $M[column = i]$ has two value. So, when we calculate the probability, we need to choose corrsponding value. In short, namely issued match 1 outherwise 0.

2) Scholar-Scholar: We predict the cooperation relationship of scholars via their historical cooperation, it's intuitive. However, considering the above prediction for scholars and conferences, it will play the role of cornerstone for consistency. So, prediction will be influenced by two factors, historical cooperation and scholar-conference's prediction.

Define s_1, s_2 as the two scholar and v_1, v_2 as the two corresponding vector, $coo(s1, s2)$ as a bool variable representing that whether they has historical cooperation. We think if two scholar has cooperated ago, then they will has more probability to continue work together. So, if there are 1 existing in the same index of their next year's vector, we claim they will cooperate, and otherwise not. As for two scholars without cooperation, we will calculate the number of value 1 in same index of next year's vector, and use the value to be divided by smaller value in sum of scholars' vector. Our calculation shows as the below:

$$P_{(s_1,s_2)} = \begin{cases} I_{\{sum(v_1==v_2)>0\}} & if \quad coo(s_1, s_2), \\ \frac{sum(v_1==v_2)}{min(sum(v_1),sum(v_2))} & otherwise. \end{cases} \tag{9}$$

Of course, we need to set a threshold of probability to make the more in line with historical data, this part will be discussed in detail. Intuitively, it should take 0.5.

C. Pagerank

There are ome examples of metrics used to evaluate the publication record of a scientist are the number of publications, total number of citations, the number of citations per paper, such as the i10-index and the h-index. The h-index is perhaps the most sophisticated and nuanced measure among these, since it accounts for both the quality and the quantity of a scientists' research publications. However, many shortcomings of h-index have been pointed out creating considerable debate over the use of h-index [10] [11] [12], and many variants of h-index have been proposed to address these. However, a fundamental issue not addressed by all these metrics is that they still treat all citations equally. Yet, it is perfectly clear that a citation by a paper from a highly regarded journal, such as Nature, should be treated differently from a citation by a workshop paper or a technical report. If this does not happen, locally famous authors whose research does not have global

impact but gets cited by their colleagues in their country or research circle can get rewarded.

Considering this problem, we use the pagerank-index to calculate the influence of papers and scholars, which is designed to address the drawbacks of existing indices by utilizing the underlying citation network dynamics. The index is calculated by running the well-known pagerank algorithm [4] on the evolving citation network to give scores to papers. We use the original form of pagerank algorithm to come up with pagerank values for each publication, and the score of an author is calculated as a weighted sum of the scores of the papers he/she has written. The process has three stages: (i) computing the page-rank value of each paper in the system (each node in the citation network) (ii) assigning weighted proportions of such values to each author in the system (each node in the collaboration network) (iii) computing the author pagerank-index as a percentile. This process is illustrated by Figure 3 . In the first stage, publications are ranked using the page-rank

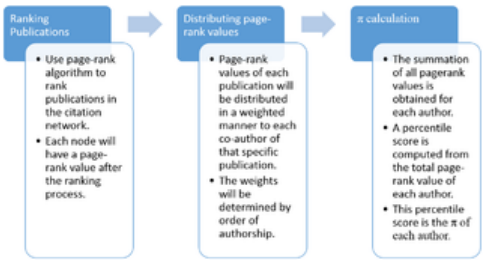


Fig. 3. The process of computing pagerank index.

algorithm as described by Larry Page and Sergei Brin [13] which can be interpreted to mimic the behavior of a random surfer in the world wide web. We use page-rank algorithm as described by Eqs 10 where P_i^t is the pagerank of node i at time t, A_{ij} is the adjacency matrix of the network, $k_{out}(j)$ is the number of outgoing links from node *j *and α is a reset parameter. N is the number of nodes in the network.

$$P_i^t = \frac{1-\alpha}{N} + \alpha \sum_j \frac{A_{ij} P_j^{(t-1)}}{k_{out}(j)} \tag{10}$$

The second stage of computation involves distributing this page-rank value among the respective authors of each publication. The page-rank value of the publication could be distributed equally between the authors. However, we find this distribution to be typically flawed because equal weight would be given to the first author and the last author. However the usual practice is to order the authors by the contributions they made to a certain publication. Hence, in order to maintain fairness and objectivity, we distribute the page-rank values proportionately as shown in Eqs 11 and 12, where W_d^s is the weight of the pagerank value assigned to a particular author (scientist) s from document d, N_a^d is the number of authors of document d, whereas R_s is the 'position' of author (scientist) s in the list of authors in document d. Further,ρ_d is the pagerank value of document (node) d at steady state, andρ_d^s is the pagerank value assigned to author s from it.

$$W_d^s = \frac{N_a - R_s + 1}{0.5 N_a (N_a + 1)} \tag{11}$$

$$\rho_d^s = W_d^s \cdot \rho_d \tag{12}$$

The final stage involves aggregating the page-rank values received by each author from each of their publications respectively, to come up with a single page-rank summation value for each author node in the collaboration network, as shown in Eqs 13.

$$\Omega_s = \sum_d \rho_d^s \tag{13}$$

D. Networkx

As a baseline model, We use preferential attachment model. Preferential attachment has received considerable attention as a model of the growth of net-works [14]. The basic premise is that the probability that a new edge involves node x is proportional to $|\Gamma(x)|$, the current number of neighbors of x. Newman [15] and Barabasi et al. [16] have further proposed, on the basis of empirical evidence, that the probability of co-authorship of x and y is correlated with the product of the number of collaborators of x and y. This corresponds to the measure $score(x, y) := |\Gamma(x)| * |\Gamma(y)|$.

E. Sen2Vec

As we all know, Word2vec can use a fixed-dimensional vector to represent a word. Converting words into vectors is not enough for our task. [17] However, because the data set obtained by our task is about the information of the paper, the information of the paper is often composed of short sentences such as title and abstract, so we need to use the word vector to construct the vectors of these short sentences. A straightforward and plain idea is: For each sentence, we calculate the mean of the sum, of all word vector from each word belonging to it. In this way, we can use corresponing vector to do some operations.

IV. Experiment

A. Data Set

Our data are selected from Arnetminer [18]. We choose all papers published on 21 conferences until 2017. A summary of our data are shown in Table I:

B. Task 1

1) Clustering: The first part of task 1 is clustering. We need to cluster all the papers with the information involving authors' names, references, title, abstract and year. We tried several methods to make use of the information. Specifically, word2vec uses title and abstract to generate a vector. DeepWalk and LINE uses the citation relationship of papers. Metapath2vec++ regards papers and authors as node and generate vectors. The accuracy of our methods are shows in Table III. From the result we could find that the information

Field	Conferences	Number of Papers	Number of scholars
Distributed & Parallel Computing	PPOPP,PACT,IPDPS,ICPP	12309	20135
Natural Language Processing	ACL,EACL,COLING, EMNLP	11894	12224
Data Mining	ICDE,SIGMOD,KDD,ICDM	18454	26174
Computer Education	AIED,ITS,ICALT	5624	8564
Machine Learning	IJCAI,ICML,NIPS	17200	20320
Operating Systems & Simulations	MASCOTS,SOSP,OSDI	2100	4502

TABLE I
A SUMMARY OF DATA.

Real Subjects	Subjects Extraction
Distributed&Parallel System	*performance*, **parallel,** *memory*, **distributed**, applications
Natural Language Processing	**language**, based, *model*, **text**, using
Machine Learning	**learning**, **classification**, data, *training, feature*
Data Mining	*algorithm*, *model*, time, problem, algorithms
Operating Systems / Simulations	based, **systems**, paper, user, information
Computer Education	data, query, mining, database, queries

TABLE II
THE COMPARISON BETWEEN REAL SUBJECTS AND THE SUBJECTS LDA EXTRACTS. THE BOLDFACED WORDS MATCH THE REAL SUBJECTS. ITALICS ARE RELEVANT TO THE SUBJECT BUT NOT ACCURATE ENOUGH.

between papers and authors are more important the relationship between papers and papers. Further, our HARP model could improve the accuracy of about 5 percent. It also shows that KMeans is better than Birch in some sense. And word2vec seems useless from our result.

Model/Cluster Algorithm	KMeans	Birch
word2vec	44.25%	45.13%
DeepWalk	53.08%	53.85%
LINE	50.44%	38.82%
metapath2vec++	66.39%	59.69%
HARP(metapath&DeepWalk)	**71.66%**	61.24%
HARP(metapath&line)	67.05%	53.49%
DeepWalk+word2vec	52.26%	52.24%
LINE+word2vec	51.21%	36.77%
(metapath2vec++)+word2vec	49.88%	58.12%
HARP(metapath&DeepWalk)+word2vec	51.72%	55.39%
HARP(metapath&line)+word2vec	60.55%	53.15%

TABLE III
ACCURACY OF MODELS IN CLUSTERING TASK.

2) Subject Extraction: We treat the combination of the title and abstract (if any) of a paper as one document. All the documents construct a corpus, then train a LDA model on the corpus with the parameter, number of subjects, equals to 6. Compare the outcome of LDA with the real subjects of the corpus. Outcome is shown in TableII.In the table, boldfaced words are relatively accurate subjects, the italics are words that relevant to this subject but not general enough. From table II we can infer that, as a traditional method for subject extraction, LDA works well for the top 3 subjects. However, the model the last subject should be knowledge discovery in database.

3) Influence Calculation (In top10Author.xlsx): Use the PageRank index to calculate the influence of the authors, then get the top 8 authors. Compare the score we calculate and the h-index, i10 score google provides.

Name	Influence score	h-index	i10 score
Philip S. Yu	68.4411	142	930
Jiawei Han	65.4823	159	701
Michael I. Jordan	64.1964	147	463
Rakesh Agrawal	64.0368	100	264
Andrew Y. Ng	61.737	107	215
Christopher D. Manning	59.6019	103	288
Wei Wang	42.5574	**117**	**1299**
Christos Faloutsos	42.126	**118**	**451**

TABLE IV
THE COOPERATION OF THE INDEX WE CALCULATE AND THE INDEX GOOGLE PROVIDES.

We can infer that the order of scores is similar. Considering the difference in the scoring method and the inevitable drawbacks that we only use the papers published on the top summits, it is reasonable that there is some difference between the three ratings.

C. Task 2

To get a glimpse at a scholar's relationship network, the ego-network becomes significant. We preprocess the whole

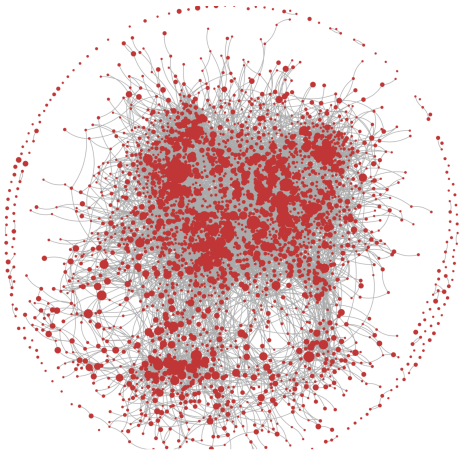


Fig. 4. A citation network in the NLP field. Each node corresponds to a paper. The edges represent the referential relation from one paper to another. Some small groups can be figured out.

Fig. 5. A cooperation network. Each node corresponds to a scholar. The edges represent the cooperation relation from one scholar to another. Colors of the nodes infer the different fields of the scholars' major research topics. An apparent clustering can be found in the cooperation network.

dataset to get a dictionary where the key is the author's name, the value is a list containing his co-authors. At the first, the relationship is thought more important than the influence of the author. Therefore, in the ego-network we use the time of cooperation as the size of node.

However, it is unreasonable to represent the time of cooperation, an attribute of the relationship, by a character of the node. Then we proposed another method to represent a more reseaonable ego-network with the size of node representing the influence of a scholar, the length of the edge reflecting the time of cooperation.

D. Task 3

1) Introdution: In this task, we need to do prediction on scholar-conference and scholar-scholar. Overall, we used a total of three methods, one of them is baseline method and the other are improved. Considering the difference of methods, we use corresponding evaluation criterion to match them respectively and following, we will do introduction on them one by one in detail.

2) Dataset of task3: The data set is divided into training set and test set, as the publish year earlier or not ealier than 2012.

3) Evaluation:

a. Calculate the accuracy, precision, recall and F1-score.
Accuracy: The proportion of correctly sampled samples to all samples, ie:

$$accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{14}$$

Precision: The proportion of correctly sampled positive samples to the total number of positive samples divided by the classifier, ie:

$$precision = \frac{TP}{TP + FP} \tag{15}$$

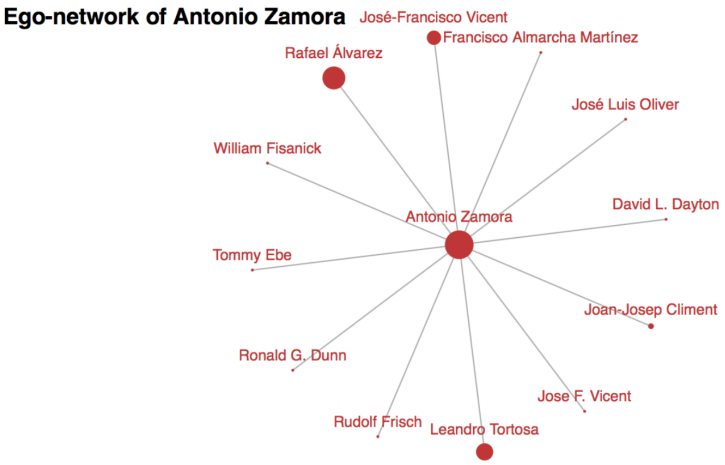


Fig. 6. The ego-network of Antonio Zamora. The size of each node represents the time of cooperation.

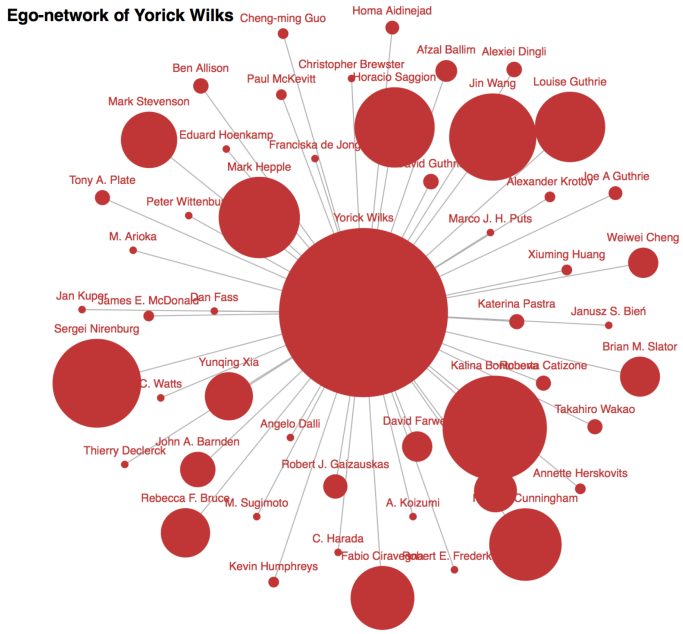


Fig. 7. The ego-network of Yoricks Wilks. The size of node represents the influence of a scholar, the length of edge reflecting the time of cooperation.

Recall: The proportion of correctly sampled positive samples to the number of positive samples, ie:

$$recall = \frac{TP}{TP + FN} \tag{16}$$

F1-score It is the harmonic mean of precision and recall, ie:

$$F1 = \frac{2 * P * R}{P + R} \tag{17}$$

b. Sampling AUC

Provided the rank of all non-observed links, the AUC values can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. To save time, we calculate the score of each non-observed link instead of giving the ordered list. At each time, we randomly pick a missing link and a nonexistent link to compare their scores. If among n independent comparisons, there are n_1 times the missing link having a higher score and n_2 times they have the same score, the AUC value is:

$$AUC = \frac{n_1 + 0.5n_2}{n} \tag{18}$$

If all the scores are generated from an independent and identical distribution, the AUC value should be about 0.5. Thereafter, the degree to which the value exceeds 0.5 indicates how much better the algorithm performs than pure chance.

In our experiments, we use the cosine distance between two vectors as the score of two nodes.

4) Methodology:

a. Baseline model
Treat a scholar as a node, the co-author relationship as the edges, calculate the preferential attachment score. Sort the prediction edges by the score from high to low, collect the top 16642 edges, which is of the same size as testing set. There are only 20 predictions that hits those in the testing set, that is to say, the precision, recall and the F1-score are 0.0012. The link prediction precision of the baseline model on the co-author relationship is low mainly due to 2 problems, the method of scoring, the occurrence of existed edges in the prediction.

b. Improved MDP model
The judging criteria used by us is to calculate the accuracy, precision, recall and F1-score. It is mainly calculated by the below confusion matrix:

	Predicted T	Predicted F
Labeled T	TP	FN
Labeled F	FP	TN

TABLE V
CONFUSION MATRIX

As for predicting scholar-conference, we set order to 5 and 10 respectively, we need to compare two thing: one is whether a scholar issue paper and the another is whether a scholar issue a paper on corrsponding conference. The result shows the below:

Order	Previous 5	Previous 10
Accuracy	0.865	0.844
Precision	0.299	0.252
Recall	0.219	0.24
F1-score	0.237	0.228

TABLE VI
ISSUE OR NOT

From above table, We can find the result of order taken 5 is better than order 10 overall. It maybe shows that the situation in the last five years has a greater impact on this matter. We can also find the distance between accuracy and other indicators is so large. I guess that the reason mainly is that the value of TN is too large and predited number of 0 has excessive proportion.

Now, let's look at the another result as below:
We can easily find that this is very similar to the above result, although overall there is no obvious difference in the results of the different values of the two orders, but when the order takes 5, the result is always better.
Let's see the prediction result of scholar-scholar as the below and il's worth nothing that we only take order as 10: Similarly, we can find that accuracy is very high, and other indicators are very low. The difference between them is larger than before. I guess the reason should be exactly the same.

Order	Previous 5	Previous 10
Accuracy	0.991	0.99
Precision	0.193	0.162
Recall	0.158	0.173
F1-score	0.162	0.156

TABLE VII
HIT OR NOT

Order	Previous 10
Accuracy	0.998
Precision	0.134
Recall	0.006
F1-score	0.012

TABLE VIII
SCHOLAR - SCHOLAR

I Scholar-Conference
With regard to the forecast results between scholars and conferences, I think it should be noted that, as mentioned above, the proportion of TN is too large. The result of the resulting accuracy will be high, but it is imaginary and has no practical significance. From F1 we can see that this model still needs to improve its generalization ability. Of course, it is already acceptable for us to make such a result on such a small data set. If our data set gets more, I think our transfer matrix will perform better and the result will be better.
In addition, we also found that the situation in the past five years has had a greater impact on a scholar's dissertation than the situation in the past decade. This seems to be in line with our intuition. After all, academic innovation is often a matter of recent years. We often seldom rely on our advantages of a decade ago to get more academic development.

II Scholar-Scholar One of the points that needs to be clarified about the predictions between scholars and scholars is that this result is greatly affected by the predictions of scholars and conferences. It can be said that the former is the vassal of the latter. So this result seems a little worse is expected. We had thought about building transition matrices for scholars and scholars. However, the mammoth matrix made us daunting. But further thinking that because the relationship between the two is inseparable, in order to ensure consistency, we use the relationship between scholars and conferences as the cornerstone of the relationship between scholars and scholars. It seems that this guarantee

can be more assured, of course, the price will be the result will Looks worse.

c. NRL model

For this task, we construct several networks. A trivial idea is that we use the whole network, which involves edges between papers and papers(reference), papers and authors, papers and conferences. Another idea is just consider the relationship we want to survey. In the cooperation task, we just construct a network which involves edges between papers and authors. In the citation task, we just construct a network which involves edges between papers and papers, papers and authors. And in publish task, we just construct a network with edges involve relationship between authors and conferences.

To quantify the accuracy of prediction algorithms. We use a method called area under the receiver operating characteristic curve(AUC) [19]. The result is shown in Table IX

Model/task	cooperation	citation	publish
metapath2vec++	75.39%	66.92%	77.86%
HARP(DeepWalk)	83.38%	**87.80 %**	87.76%
HARP(LINE)	83.73%	76.93%	67.32%
HARP(DeepWalk)(whole net)	**85.72%**	85.30%	**88.50%**
HARP(LINE)(whole net)	56.86%	56.43%	54.81%

TABLE IX
AUC OF TASK3

We are excited to find that our method do make sense. Specifically, we could find HARP has a significant improvement in the performance of the model using DeepWalk. When using LINE, the model seems to have difficulty in learning the structure of the network. Also, to our surprise, the model using the whole net has a better performance in almost all the task. Maybe because when having more information of the network, the method could learn better of the structure other than being disturbed by redundant information. All in all, we could say that **HARP with deepwalk of metapath is a good way to embed the structure of network**.

The reason why it works maybe because when using the HARP model, we suppose that the neighbors of the same node have similar representation. For instance, in the network based on the cooperation, the neighbors of one paper involves its authors, which is just the nodes we want to have a similar vector. And in the network based on reference, there are many meta-paths is a structure of author-paper-reference paper-author, which in edge collapsing would be regard as the same super-node. On the other hand, in the model we coarsen the network again and again and finally regard the network as one very huge node. Due to this, we could learn more about the global structure of the network, rather than just local structure we can learned in traditional representation methods.

The reason why DeepWalk make sense but LINE seems to be out of work is when we use DeepWalk, we focus on the probability of length from one node(super-node) to another node(super-node). This relationship could make sense even when the super-nodes don't have too much similarity. However, LINE focuses on the similarity of super-nodes with its super-neighbors. When we have a big super-node, it is possible that there are very little similarity between these super-nodes so LINE could not learn a good representation for them.

REFERENCES

[1] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 701–710. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623732

[2] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," *CoRR*, vol. abs/1607.00653, 2016. [Online]. Available: http://arxiv.org/abs/1607.00653

[3] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.

[4] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 135–144.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[6] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "Harp: Hierarchical representation learning for networks," *arXiv preprint arXiv:1706.07845*, 2017.

[7] A. Zimdars, D. M. Chickering, and C. Meek, "Using temporal data for making recommendations," pp. 580–588, 2001.

[8] J. Rust, "Structural estimation of markov decision processes," *Handbook of econometrics*, vol. 4, pp. 3081–3143, 1994.

[9] G. Shani, D. Heckerman, and R. I. Brafman, "An mdp-based recommender system," *Journal of Machine Learning Research*, vol. 6, no. Sep, pp. 1265–1295, 2005.

[10] R. Costas and M. Bordons, "The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level," *Journal of informetrics*, vol. 1, no. 3, pp. 193–203, 2007.

[11] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.

[12] L. Zhivotovsky and K. Krutovsky, "Self-citation can inflate h-index," *Scientometrics*, vol. 77, no. 2, pp. 373–375, 2008.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[14] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.

[15] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical review E*, vol. 64, no. 2, p. 025102, 2001.

[16] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3-4, pp. 590–614, 2002.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 990–998.

[19] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.