Clustering in heterogeneous network and higher order structure

Jing Ma, Mao-Chuang Yeh, Yu-Wei Lin

Abstract

Cluster identification in large scale information network is a highly attractive issue in the network knowledge mining. Traditionally, community detection algorithms are designed to cluster object population based on minimizing the cutting edge number. Recently, researchers proposed the concept of higher-order clustering framework to segment network objects under the higher-order connectivity patterns. However, the essences of the numerous methodologies are focusing on mining the homogeneous networks to identify groups of objects which are closely related to each other, indicating that they ignore the heterogeneity of different types of objects and links in the networks. In this study, we propose an integrated framework of heterogeneous information network structure and higher-order clustering for mining the hidden relationship, which include three major steps: (1) Construct the heterogeneous network, (2) Convert HIN to Homogeneous network, and (3) Community detection.

1. Introduction

Network plays an important role in assisting mining essential knowledge from large scale unstructured data across various research area, such as biology, sociology, business, etc. Networks can be generated into numerous structures based on the needs. For example, biological gene network are used to identifying gene-disease (Özgür et al, 2008). Different analytics are then selected to reach to goals.

Traditionally, the networks are generated in the form of homogeneity which have only one type of node and one type of link. Based on the homogeneous network structure, researchers developed analytic methods for different purposes, such as node degree, centrality, community detection, etc. Clustering is one of the most significant network analytics for network knowledge mining, which is widely studied. Typical cluster models include connective models, centroid models, distribution models, etc. Moreover, Benson et al, 2016, proposed the concept of higher-order community detection. Instead of cutting edges, they designed an algorithm divided the whole graph into subgraphs by cutting the “motifs”, instead of cutting edges (Benson et al, 2016).

In addition to homogeneous network structure, Sun et al. (2011) claim that mapping data into homogeneous network structure will cause a lot of information lost, such as semantic lost. To distinguish different types of objects and links in the networks, in 2011, they proposed the structure of heterogeneous information networks, which is the network structure carrying the heterogeneity of nodes and relations. Corresponding clustering model are designed to meet the goal of better clustering. For example, both GNetMine and RankClus are demonstrated that they can cluster the objects population
in the networks much more accurate in comparison of homogeneous cluster models (Sun et al, 2009; Ji et al, 2010). However, their models are still highly relying on edge cuts based clustering, which ignore the potential of higher-order clustering in heterogeneous information network, such as heterogeneous motif cut. Therefore, this research proposes an integrated framework of heterogeneous information network structure and higher-order clustering for mining the hidden relationship.

In this paper, we proposed a higher order clustering methodologies, which include three major steps: (1) Construct the heterogeneous network, (2) Convert HIN to Homogeneous network, and (3) Community detection. The experiment results shows that our methodology is better performance than traditional clustering methods in terms of accuracy and efficiency. Moreover, our method do not need any training data, indicating that it is more generalizable then any state of art methods.

2. Network model

2.1 Definition of Information Network

An information network is a set of items (nodes or vertices) containing information, with connections (edges or links) disseminating information. The form of the information network is wildly used to portray the world. There are all sorts of information networks such as social networks, the World Wide Web, biological networks, organizational networks, health networks, research publication networks, and many others.

Formally, we define the information network as a graph $G = (V, E)$, which consists of a set of vertices $V = \{v_1, v_2, ..., v_n\}$ and a set of edges $E = \{e_1, e_2, ..., e_m\}$ where each element, $e_i \in \{0, 1\}$, represents a link existence from node $v_i$ to $v_j$.

2.2 Definition of Heterogeneous Information Network

Heterogeneous information network is a kind of information network comprising multiple types of objects and multiple types of links. Different from the traditional network definition, we introduce additional definition that object type mapping function $\tau: V \rightarrow A$ and a link type mapping function $\phi: E \rightarrow R$, where each object $v \in V$ belongs to one particular object type $\tau(v) \in A$, each link $e \in E$ belongs to a particular relation $\phi(e) \in R$. Namely, if there exist more than one types of objects $|A|$ or relations $|R|$ within an information network, we can assert that it is a heterogeneous information network.

2.3 Definition of Meta-Path

A meta path is a sequence of relations between object types, which defines a new composite relation between its starting type and ending type (Sun et al. 2011.) Formally, we can define the meta-path as follows. Based on the given the network schema $TG = (A, R)$. A meta path $MP$ can be presented in $A_1 \ A_2 \ ... \ A_n$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_{n-1}$ between types $A_1$ and $A_n$, and $\circ$ denotes the composition operator on relations.
3. Theoretical foundation

3.1 kmeans
k-means clustering is aimed at dividing the n observations into k groups in which each observation belongs to the cluster with the nearest mean. More detailed description as following:

- Give a set of observations \((x_1, x_2, \ldots, x_n)\), where each observation is a d-dimensional real vector. The result from k-means should be \(S = \{S_1, S_2, \ldots, S_k\}\), where k should be less or equal than n. The objective is to minimize the within-cluster sum of squares, the equation can be presented as:

\[
\min \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2
\]

where \(\mu_i\) is the mean of points in \(S_i\).

The standard algorithm steps are: first initial “means” (for example k=3, then randomly select three “means”). Then k clusters are generated according to the association to each nearest mean. The third step is to calculate the centroid which is the actual mean of each cluster. Final step is to repeat step 2 and 3 until the convergence has been reached.

3.2 pathSim
The previous similarity measures methods are biased to either highly visible objects or highly concentrated objects, for instance, the random walk based favors the objects with high degree, and the pairwise favors pure objects. While the PathSim can capture the peer similarity, like finding the similar author with similar research field and reputation.

Given a symmetric meta path \(P\), PathSim between two objects of the same type \(x\) and \(y\) is:

\[
s(x, y) = \frac{2 \times |\{p_{x \sim y} : p_{x \sim y} \in P\}|}{|\{p_{x \sim x} : p_{x \sim x} \in P\}| + |\{p_{y \sim y} : p_{y \sim y} \in P\}|}
\]

where \(P_{x \sim y}\) is a path instance between \(x\) and \(y\), \(P_{x \sim x}\) is that between \(x\) and \(x\), and \(P_{y \sim y}\) is that between \(y\) and \(y\).

The definition here indicates that given a meta-path \(P\), \(s(x, y)\) is defined in terms of two parts: (1) their connectivity defined by the number of paths between them following \(P\); and (2) the balance of their visibility, where the visibility of an object according \(P\) is defined as the number of path instances between the object itself following \(P\).

3.3 GNetMine
GNetMine is proposed to model the link structure in information networks with arbitrary network schema and arbitrary number of object/link types (Ji et al. 2010). The methodology is based on the assumption that classification of networked data can be essentially viewed as a process of knowledge propagation, where information is propagated from labeled objects to unlabeled ones through links until a stationary state...
is achieved. Specifically, GNetMine is a novel graph-based regularization framework to address the classification problem on heterogeneous information networks.

This algorithm framework is based on the consistency assumption: the class assignments of two linked objects are likely to be similar. And the class prediction on labeled objects should be similar to their pre-assigned labels. This algorithm ensure the consistency is preserved over each relation graph corresponding to each type of links separately.

### 3.4 Network motif

Network motifs can be defined as subgraphs of a network that appear much more frequently than those found in randomized networks (Shen-Orr et al. 2002). There are several kinds of motif shown in the figure 1.

![Network motif examples](image)

**Figure 1:** Some types of motifs: (a) three-vertex feedback, (b) three chain, (c) feed-forward loop, (d) bi-parallel, (e) four-vertex feedback, (f) bi-fan, (g) feedback with two mutual dyads, (h) fully connected triad and (i) uplinked mutual dyad (Costa et al. 2008).

### 3.5 Intensity and coherence of motifs in weighted complex networks

To detect those motifs that are likely to be important in the weighted networks, Onnela et al. (2005) proposed to defined the intensity (IM) of motif M as the geometric mean of its arc weights, which can be formally defined as following equations:

\[
I(g) = \left( \prod_{(ij) \in E_g} w_{ij} \right)^{1/|E_g|}
\]
where intensity $l(g)$ of a subgraph $g$ with vertices $v_g$ and links $l_g$ is based on the geometric mean of its weights on the network

$$I_M = \sum_{g \in M} l(g),$$

where $IM$ of a motif $M$ in the network is the sum of its subgraph intensities.

Moreover, Onnela et al. (2005) also defined the coherence (QM) of motif $M$ as the ratio of the geometric to the corresponding arithmetic mean.

$$Q(g) = I(g)|I_g| \left/ \sum_{(ij) \in l_g} w_{ij} \right.$$

Based on these measures, motif scores were defined as:

$$ZI_M = \frac{(I_M - \langle i_M \rangle)}{\left(\langle I_M^2 \rangle - \langle i_M \rangle^2 \right)^{1/2}},$$

$$ZQ_M = \frac{(Q_M - \langle q_M \rangle)}{\left(\langle Q_M^2 \rangle - \langle q_M \rangle^2 \right)^{1/2}},$$

where $ZIM$ and $ZQM$ are the motif intensity and coherence score respectively; $i_M$ and $q_M$ are the total intensity and coherence of motif $M$ in one realization of the random regime, respectively.

Finally, In a weighted graph, the clustering around a node $i$ can be calculated as the geometric average of subgraph node weights (Onnela et al., 2005)

$$\tilde{C}_i = \frac{2}{k_i(k_i - 1)} \sum_{j,k} (\tilde{w}_{ij} \tilde{w}_{jk} \tilde{w}_{ki})^{1/3}$$

### 3.6 Motif-based embedding for graph clustering

Instead of firstly focusing on identifying the most significant Motif $M$, then do clustering based on geometric average of subgraph node weights, Lim and Lee (2016) proposed a motif-based embedding methodology, which was demonstrated to be more effective in detecting communities than existing graph embedding methods, spectral embedding and force-directed embedding, both theoretically and experimentally.

Their method is based on motif-based weighting method which can reflect motif substructures in a given graph.

For instance, in figure 2, node $i$ and $j$ are two arbitrary nodes in the homogeneous network. Firstly, 1st order weight between node $i$ and $j$, $wij$, are calculated based on the number of linkage between two nodes. Secondly, 2nd order weight between node $i$ and $j$, $mij$, are computed by counting the number of third nodes connecting both node $i$ and $j$. The total weight between node $i$ and $j$ is the sum of $wij$ and $mij$. 
4. Methodology

Our goal is to utilize higher order clustering concept to detect communities in Heterogeneous Information Network. We first provide an overview of the proposed method by defining the major components which include three major steps: (1) Construct the heterogeneous network, (2) Convert HIN to Homogeneous network, and (3) Community detection.

4.1 Meta-Path selection

In order to apply the higher order structure in previous chapter to the heterogeneous network (with entity types: paper (P), venue (V), author (A), and term (T)), we first consider mapping the heterogeneous network into homogeneous network of only author nodes according to 3 meta path which are A-P-A, A-P-V-P-A and A-P-T-P-A. The result of this mapping is a undirected, homogeneous graph. The weight of each meta-path which represents the weight of edge of this graph can be adjusted according to the semantic meaning of each meta-path.

In this study, we implemented the meta path APVPA to convert the heterogeneous graph to homogeneous graph. The reason of using this meta path is because the limited dataset we used, which only have correct label of research areas for part of the object (the dataset we used is the dataset provided in hw1, which consist of author, paper, conference, term. And this dataset labeled part of these objects belonging to database, data mining, machine learning, or artificial intelligence.). When authors prepared to submit a paper, he/she already decided which area this paper may belong to, thus we guess using the APVPA to construct the graph will get the best result in terms of this dataset and the truth it is. We tried using APA and APTPA but the result accuracies are around 60%. The reason is using APA to construct a homogeneous author network
actually is a co-author network. The authors who worked together get clustered, while people working on different areas (e.g., one working on AI and the other working on data mining) can also be co-author. Thus using the co-author network cannot achieve a high accuracy. For the APTPA, it constructed a very sparse graph and the clustered authors who wrote the papers for similar topic. A paper can contain many terms and an author can write many papers, thus it is hard to find the author cluster based only on the term similarity.

Because we already got the clustering results of APTPA and APA are not desirable. Here we guess the reason is that the homogeneous graph constructed by these two meta paths are pretty sparse first. And combined with the semantic meaning using the APTPA meta path, it seems like to get the similar topic clustered together, and using the APA meta path. Thus here we will only talk about APVPA. Thus, we construct the homogeneous graph based on meta path APVPA. The constructed graph is a weighted undirected homogeneous network with all node is author nodes and can be presented as a adjacent matrix W.

4.2 Motif selection
One of the most challenge part in this study is to re-define the motif structure. As mentioned, the easiest way to start with is to see how it works in homogeneous network. Therefore, we first project the heterogeneous network to homogeneous network with edges as the possible meta paths. In the homogeneous network of authors, we basically have follows 11 different motifs. There may be some isolated points in graph after the process of one motif. But, if we consider all motifs (M1 to M11), there seems no isolated point in our data.

![Figure 3. Triangle motif structures](image)
Based on above 11 possible cliques for motifs, we can construct 11 corresponding modified author networks based on counting each given motif. If we need to consider more than one motifs at same time we can just add all these required motif graphs to give the combined results.
In this study, we select M1 as out target motif for motif based clustering in the Heterogeneous network.

4.3 Normalized Adjacent matrix Construction
There are various method to convert HIN to Homogeneous network and preserve HIN information, including Pathcount, PathSim, RateSim, motif count, and geometric motif intensity.

1. **Path count**: Path count measures the number of path instances between two objects following a given meta path. Path count can be calculated by the products of adjacency matrices associated with each relation in the meta path.
2. **RateSim**: We normalize paper number published by each author on each venue by the total number of publications of the author.
3. **PathSim**: We normalize the adjacent matrix based on the method we mentioned in last chapter called pathSim which is to capture the similarity of two different authors.
4. **Motif count**: Motif count measures the number of motif between the two nodes.
5. **Geometric motif intensity**: Motif count can be normalized by calculating the geometric mean.

4.4 Community detection
k-means clustering is then applied to split the objects within the network into k groups that each node belongs to the group with the nearest mean.

5. Dataset
The data input is a heterogeneous information network of academic publications, with 4 types: author, conference, paper and term. The dataset we use contains 14376 papers, 20 conferences, 14475 authors and 8920 terms. Within the dataset, 4057 authors, 100 papers and all 20 conferences are manually labeled to four classes, representing four different research areas: database, data mining, information retrieval and artificial intelligence.

6. Experiment
To simplify the analysis, we first assume we can classify the venues into 4 classes and it’s not hard to do this with accuracy 1.0. Next, we construct a weighted graph based on path count, PathSim, and RateSim over the paths between any two authors through the 4 classes. “A-class-A” means we follow the meta path “A-P-V-Class-V-P-A” since we classified the venue into 4 classes to construct the homogeneous network using path count method. The “A-PathSim-A” means based on previous meta path “A-P-V-Class-V-P-A” we normalized the result adjacent matrix with PathSim algorithm. Similarly, for “A-RateSim-A” we normalized the result adjacent matrix with our defined RateSim
algorithm which was talked previously. Then we do the k-mean clustering over the normalized graph. The results can be seen as following:

Table 1. Results of using Path Count, PathSim and RateSim method and follow the meta path “A-P-Class-P-A”

<table>
<thead>
<tr>
<th></th>
<th>A-Class-A (Path Count)</th>
<th>A-PathSim-A (PathSim on 4 classes)</th>
<th>A-RateSim-A (RateSim on 4 classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.9357</td>
<td>0.9332</td>
<td>0.9342</td>
</tr>
</tbody>
</table>

The results indicate we achieve pretty high accuracy around 93% for these three method. In addition, we tried not cluster the venue first, the results shown in following table. To explanation, “A-V-A” means we follow the meta path “A-P-V-P-A” to construct the homogeneous author graph. The “A-PathSim-A” means based on previous meta path “A-P-V-P-A” we normalized the result adjacent matrix with PathSim algorithm. Similarly, for “A-RateSim-A” we normalized the result adjacent matrix with our defined RateSim algorithm which was talked previously. Then we do the k-mean clustering over the normalized graph. Compared with previous results, the accuracy decreased slightly, especially for normalized ones (RateSim and PathSim) but generally still achieve a high accuracy above 90%. Compared with the results we got for author from hw1 GNetMine (around 93%), here we didn’t involve any training data but still achieve a good result.

Table 2. Results of using Path Count, PathSim and RateSim method and follow the meta path “A-P-V-P-A”

<table>
<thead>
<tr>
<th></th>
<th>A-V-A (Path Count)</th>
<th>A-PathSim-A (PathSim on venues)</th>
<th>A-RateSim-A (RateSim on venues)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.932</td>
<td>0.9058</td>
<td>0.9086</td>
</tr>
</tbody>
</table>

From table 3, each number is the accuracy score of different combination of motif and normalization methods. The results show that no matter order 2 or 3 cliques, the performance of utilizing geometric means to normalize adjacent matrix is better than motif count in terms of accuracy scores. Moreover, the order 3 cliques performed better than order 2 cliques, inferring that we need to preserve the complete structure of motif while clustering.
Table 3. Results of using motif and geometric motif intensity and follow the meta path “A-P-V-P-A”

<table>
<thead>
<tr>
<th></th>
<th>geometric</th>
<th>weight multiplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy of clique of order 3</td>
<td>0.9288</td>
<td>0.9105</td>
</tr>
<tr>
<td>Accuracy of wedge or path of 2</td>
<td>0.9103</td>
<td>0.9039</td>
</tr>
</tbody>
</table>

The 1st order path vs 2nd order path:

The graph with first order meta path and the graph with the second order meta path can not help each other to improve accuracy:

- 1st vs 2nd (weight multiplication):
  1st: homogenous author graph with APVPA(20 venues)
  2nd: the 2nd order of APVPA meta path (weight multiplication)
  total graph = a * 1st + b * 2nd with a+b = 1
  results showed only a=1 achieves best accuracy = 0.932
  the accuracies of others are 0.928

![Figure 4. The accuracy changes with the weight between 1st and 2nd order path (weight multiplication).](image)
• 1\textsuperscript{st} vs 2\textsuperscript{nd} (geometric mean):
  
  1\textsuperscript{st} : homogenrous author graph with APVPA(20 venues)
  2\textsuperscript{nd} : the 2nd order of APVPA meta path (geometric mean)
  total graph = a \times 1\textsuperscript{st} + b \times 2\textsuperscript{nd} with a+b =1
  results show only a=1 gives best accuracy = 0.932
  the accuracies of others are 0.9305

\begin{figure}
\centering
\includegraphics[width=0.8\textwidth]{figure5}
\caption{The accuracy changes with the weight between 1st and 2nd order path (geometric mean).}
\end{figure}

**The application of motif:**
As we know the accuracies of homogeneous graph with meta path APVPA is 0.9320, and accuracy of generated graphs after motif calculation are less than this number. Although the generated graph has less accuracy, we expect the generated graph still can help to improve accuracy of homogeneous graph.

• generated graph by TVV motif (clique of 3 edges respectively of term, venue, and Venue) + homogeneous graph with APVPA (20 venues) edge:
  a. TVV motif seems help a little: From 0.9320 to 0.9330.
  b. the accuracy for only TVV motif graph is 0.8878.
  c. there is no difference between geometrix mean and authority multiplication for TCC motif
Figure 6. The accuracy changes with the weight between generated graph and original graph path.

- Motif VVV + Motif TVV: They indeed help each other to get better accuracy: the accuracy increases from 0.9288 (and 0.8878) to 0.9330 (for geometric mean).

Figure 7. The accuracy changes with the weight between two generated graphs of motif VVV and motif TVV. Are calculated by geometric mean.
Figure 8. The accuracy changes with the weight between two generated graphs of motif VVV and motif TVV (Evaluated by weight multiplication.

- Motif VVV + Motif TVV + Motif PVV:
  The accuracy increases from 0.9105 (and 0.8878) to 0.9155 (for weight multiplication. The maximum accuracy for combination of three is 0.9330.
- Adding the graph of all meta path seems not help the accurate getting better.
- the graph cut accuracy for paper homogeneous graph with meta path PAP, PVP, and PTP are really bad.

8. Conclusions and Future Works

In this paper, we study the problem of clustering in heterogeneous network and higher order structure. A higher order clustering methodology is proposed, which include three primary stages: (1) Construct the heterogeneous network, (2) Convert HIN to Homogeneous network, and (3) Community detection. Our method is flexible and generalizable since we do not need any training data.

There are several interesting insights from the experiment results:
- the meta path APVPA (or APCPA) contains most classification information.
- each motif contains special semantic meanings and respects different aspect of view. Usually, it needs to consider more than one motifs the prove accuracy: motif VVV and motif TVV indeed help each other to improve accuracy.
- the experiment results show that our methodology is better performance than traditional clustering methods in terms of accuracy, which indicate the power of higher order clustering.
- no training data needs for GraphCut method and it achieve a high accuracy compared with GNetMine

For the future works, we may also need to systematically identify the importance of different motifs in the weighted networks by respectively calculating the motif intensity and coherence scores.
REFERENCES
Han, J., Sun, Y., Yan, X., & Yu, P. S. (2010, July). Mining heterogeneous information networks. In Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD’10), Washington, DC.
Han, J. (2012). A Meta Path–Based Approach for Similarity Search and Mining of Heterogeneous Informa) on Networks.


