Heterogeneous Information Network Clustering Methods

Reportor:Wenbao Li

Data Mining Lab,Big Data Research Center

Background

- Graph(network) clustering has attracted increasing research interest.
- **For homogeneous network:** Spectral clustering,symmetric Non-negative Matrix Factorization,Markov clustering,Ncut,Mcut,...
- **However,heterogeneous information network clustering are concentrated until recently.**

Background

- **Heterogeneous Information network**:
  - Is an information network composed of multiple types of objects.
  - Consists of some partial **attributes** within types of objects and **links** between different types of objects.
  - Examples:
    - DBLP(author,paper,conference,term)
    - Social Network(people,groups,books,blogs,posts,etc)
    - Movies(movie,actor,director,)
    - Newsgroup(news,writer,group)

Part One

Related Work

- ☑ RankClus (Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)
- ☑ NetClus (Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)
- ☐ ENetClus (Manish Gupta,Charu C. Aggarwal,Jiawei Han,Yizhou Sun,ASONAM'11)
- ☑ GenClus (Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)
- ☑ PathSelClus (Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)
- ☑ CGC (Wei Cheng,Xiang Zhang,Zhishan Guo,Yubao Wu,KDD'13)
- ☐ SI-Cluster (Yang Zhou,Ling Liu,KDD'13)
- ☐ ComClus (Ran Wang,Chuan Shi,Philip S. Yu,Bin Wu,PAKDD'13)

RankClus

Wenbao Li

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

- Idea:Iteratively clustering and ranking which **map the target type into a new K-dimensional feature space according to which the clustering is performing.**
- Advantage:
  - improve the performance of clustering and ranking simultaneously.
  - avoiding to calculate the pairwise similarity of target objects.

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

Some Definitions

- Bi-type Information Network

DEFINITION 1. ***Bi-type Information Network.*** *Given two types of object sets X and Y, where $X = \{x_1, x_2, \ldots, x_m\}$, and $Y = \{y_1, y_2, \ldots, y_n\}$, graph $G = \langle V, E \rangle$ is called a bi-type information network on types X and Y, if $V(G) = X \cup Y$ and $E(G) = \{\langle o_i, o_j \rangle\}$, where $o_i, o_j \in X \cup Y$.*

For convenience, we decompose the link matrix into four blocks: W_{XX}, W_{XY}, W_{YX} and W_{YY}, each denoting a sub-network of objects between types of the subscripts. W thus can be written as:

$$W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix}$$

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

Some Definitions

- Ranking Function

DEFINITION 2. ***Ranking Function.*** *Given a bi-type network* $G = \langle\{X \cup Y\}, W\rangle$, *if a function* $f : G \rightarrow (\vec{r}_X, \vec{r}_Y)$ *gives rank score for each object in type* X *and type* Y, *where*

$$\forall x \in X, \vec{r}_X(x) \geq 0, \sum_{x \in X} \vec{r}_X(x) = 1, \ \ and$$

$$\forall y \in Y, \vec{r}_Y(y) \geq 0, \sum_{y \in Y} \vec{r}_Y(y) = 1,$$

we call f *a ranking function on network* G.

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)
Some Definitions

- Conditional Rank and Within-Cluster rank

DEFINITION 3. ***Conditional rank*** *and* ***within-cluster rank.*** *Given target type* X, *and a cluster* $X' \subseteq X$, *sub-network* $G' = \langle \{X' \cup Y\}, W' \rangle$ *is defined as a vertex induced graph of* G *by sub vertex set* $X' \cup Y$. *Conditional rank over* Y, *denoted as* $\vec{r}_{Y|X'}$, *and within-cluster rank over* X', *denoted as* $\vec{r}_{X'|X'}$, *are defined by the ranking function* f *on the sub-network* G': $(\vec{r}_{X'|X'}, \vec{r}_{Y|X'}) = f(G')$. *Conditional rank over* X, *denoted as* $\vec{r}_{X|X'}$, *is defined as the propagation score of* $\vec{r}_{Y|X'}$ *over network* G:

$$\vec{r}_{X|X'}(x) = \frac{\sum_{j=1}^{n} W_{XY}(x,j)\vec{r}_{Y|X'}(j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_{Y|X'}(j)}.$$

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)
Some Definitions

- Target type:the type we are going to cluster.
- Attribute type:the other types.

- Assumptions: $W_{XX} = 0$

RankClus(**Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09**)

Flow:

① Give an initial partition of target object X

② Compute the conditional ranking $\vec{r}_{X|X_k}, \vec{r}_{Y|X_k},$ **A**

③ Estimate the parameter $\Theta_{m\times K} = \{\pi_{i,k}\}(i=1,2,\ldots,m; k=1,2,\ldots,K)$ **B**

④ Form a new feature space $\Theta_{m\times K} = \{\pi_{i,k}\}(i=1,2,\ldots,m; k=1,2,\ldots,K)$

⑤ Calculate the center of each cluster according to the new feature space(**mean**).

⑥ According the new feature space ,assign each target object into the **nearest** cluster **C**

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

A.Ranking Score——Ranking function

- Simple Rank

$$\begin{cases} \vec{r}_X(x) = \frac{\sum_{j=1}^{n} W_{XY}(x,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)} \\ \vec{r}_Y(y) = \frac{\sum_{i=1}^{n} W_{XY}(i,y)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)} \end{cases}$$

- Authority Rank
  - Give ranking scores according some authority rules.
    - Rule 1: Highly ranked authors publish *many* papers in highly ranked conferences.
    - Rule 2: Highly ranked conferences attract *many* papers from *many* highly ranked authors.

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

normalization

$$\vec{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i)\vec{r}_X(i). \quad \vec{r}_Y(j) \leftarrow \frac{\vec{r}_Y(j)}{\sum_{j'=1}^{n} \vec{r}_Y(j')},$$

$$\vec{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_Y(j). \quad \vec{r}_X(i) \leftarrow \frac{\vec{r}_X(i)}{\sum_{i'=1}^{m} \vec{r}_X(i')},$$

$$\begin{cases} \vec{r}_X = \dfrac{W_{XY}\vec{r}_Y}{\|W_{XY}\vec{r}_Y\|} \\ \vec{r}_Y = \dfrac{W_{YX}\vec{r}_X}{\|W_{YX}\vec{r}_X\|} \end{cases}$$

$$\vec{r}_X = \frac{W_{XY}W_{YX}\vec{r}_X}{\|W_{XY}W_{YX}\vec{r}_X\|}$$

$\vec{r}_X$ is the eigenvector of $W_{XY}W_{YX}$.

Similarly, $\vec{r}_Y$ is the primary eigenvector of $W_{YX}W_{XY}$

- Rule 3: The rank of an author is enhanced if he or she co-authors with many authors or many highly ranked authors.

$$\vec{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j)\vec{r}_X(j) + (1-\alpha) \sum_{j=1}^{n} W_{YY}(i,j)\vec{r}_Y(j).$$

Similarly, we can prove that $\vec{r}_Y$ should be the primary eigenvector of $\alpha W_{YX}W_{XY} + (1-\alpha)W_{YY}$, and $\vec{r}_X$ should be the primary eigenvector of $\alpha W_{XY}(I-(1-\alpha)W_{YY})^{-1}W_{YX}$.

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

B. Estimate the assignment parameter

Set $p_k(Y) = \vec{r}_{Y|X_k}$ $p_k(X) = \vec{r}_{X|X_k}$

$p_{x_i}(Y) = p(Y|x_i)$ to generate a link between x_i and y in Y.

$$p_{x_i}(Y) = \sum_{k=1}^{K} \pi_{i,k} p_k(Y), \text{ and } \sum_{k=1}^{K} \pi_{i,k} = 1.$$

$$\pi_{i,k} = p(k|x_i) \propto p(x_i|k)p(k)$$

EM to estimate $\Theta_{m\times K} = \{\pi_{i,k}\}(i = 1,2,\ldots,m; k = 1,2,\ldots,K)$

$$L'(\Theta|W_{XY}, W_{YY}) = p(W_{XY}|\Theta)p(W_{YY}|\Theta)$$

$$= \prod_{i=1}^{m}\prod_{j=1}^{n} p(x_i, y_j|\Theta)^{W_{XY}(i,j)} \prod_{i=1}^{n}\prod_{j=1}^{n} p(y_i, y_j|\Theta)^{W_{YY}(i,j)}$$

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

Cluster Centers and Distance Measure

K-dimensional vector $\vec{s}_{x_i} = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,K})$.

Center of cluster k:

$$\vec{s}_{X_k} = \frac{\sum_{x \in X_k} \vec{s}(x)}{|X_k|}$$

Distance measure:

$$D(x, X_k) = 1 - \frac{\sum_{l=1}^{K} \vec{s}_x(l)\vec{s}_{X_k}(l)}{\sqrt{\sum_{l=1}^{K}(\vec{s}_x(l))^2}\sqrt{\sum_{l=1}^{K}(\vec{s}_{X_k}(l))^2}}.$$

RankClus(Yizhou Sun, Jiawei Han , Peixiang Zhao, Zhijun Yin , Hong Cheng, Tianyi Wu,EDBT'09)

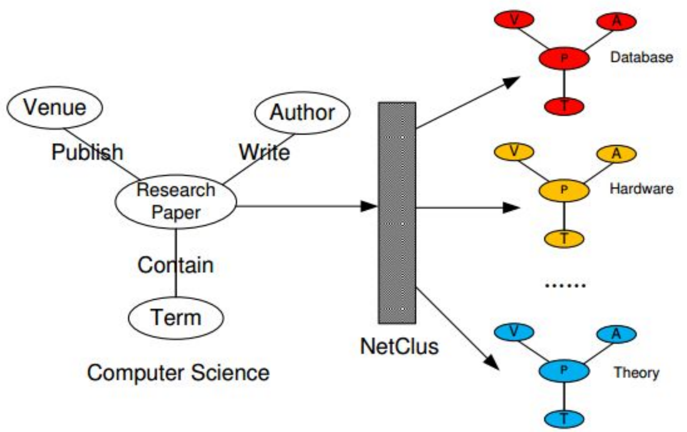
- Extensions to arbitrary multi typed information network
  - One-type:set Y = X.
  - Bi-type with $W_{XX} \neq 0$:Add type Z,Z = X. map to 2K-dimensional feature space.
  - Multi-typed: N types. map to NK-dimensional feature space.

NetClus

Wenbao Li

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

- NetClustering(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)
  - Idea:Find a new K-dimensional feature space by ranking which are determined by a probability generative model.
  - **Advantage:**
- **suit for multi types objects.**
- **cluster attribute type object.**



NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09) Definitions

Information Network

Definition 1. **Information Network.** Given a set of objects from T types $\mathcal{X} = \{X_t\}_{t=1}^{T}$, where X_t is a set of objects belonging to t_{th} type, a weighted graph $G = \langle V, E, W \rangle$ is called an information network on objects $\mathcal{X}$, if $V = \mathcal{X}$, E is a binary relation on V, and $W : E \rightarrow \mathbb{R}^+$ is a weight mapping from an edge $e \in E$ to a real number $w \in \mathbb{R}^+$. Specially, we call such an information network **heterogeneous network** when $T \geq 2$; and **homogeneous network** when $T = 1$.

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)
Definitions

- Star Network Schema

Definition 2. **Star Network Schema.** An information network $G = \langle V, E, W \rangle$ on $T+1$ types of objects $\mathcal{X} = \{X_t\}_{t=0}^{T}$ is called with **star network schema**, if $\forall e = \langle x_i, x_j \rangle \in E, x_i \in X_0 \wedge x_j \in X_t (t \neq 0)$, or vise versa. G is then called a **star network**. Type X_0 is called the **center type**. X_0 is also called the **target type** and $X_t (t \neq 0)$ are called **attribute types**.

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09) Definitions

- Net-Cluster

Definition 3. **Net-cluster.** Given a network G, a net-cluster C is defined as $C = \langle G', p_C \rangle$, where G' is a **sub-network** of G, *i.e.*, $V(G') \subseteq V(G)$, $E(G') \subseteq E(G)$, and $\forall e = \langle x_i, x_j \rangle \in E(G'), W(G')_{x_i x_j} = W(G)_{x_i x_j}$. Function $p_C : V(G') \rightarrow [0, 1]$ is defined on $V(G')$, for all $x \in V(G')$, $0 \leq p_C(x) \leq 1$, which denotes the probability that x belongs to cluster C, *i.e.*, $P(x \in C)$.

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

- Flow:

A. B. C. D.

① Give an initial partition of G,which is K clusters.And induce net-clusters from the partition. $\{C_k^0\}_{k=1}^K$

② Build **ranking-based probabilities generative model** for each net-cluster,i.e. $\{P(x \mid C_k^t)\}_{k=1}^K$

③ Calculate the posterior probabilities for each target object $(p(C_k^t \mid x))$ and then adjust their cluster assignment according to the new measure defined by the posterior probabilities to each cluster

④ Repeat Step 2 and 3 until the cluster does not change significantly,i.e. $\{C_k^*\}_{k=1}^K = \{C_k^t\}_{k=1}^K = \{C_k^{t-1}\}_{k=1}^K$

⑤ Calculate the posterior probabilities for each attribute object $(p(C_k^* \mid x))$ in each net-cluster

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

A. Induce net-clusters

①Initial:random

②Other:according to the definition of net-clusters.

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

B. Probabilistic Generative Model for target objects

①Given an attribute object x and its type T_x ,the probability to visit x in G is

$$p(x|G) = p(T_x|G) \times p(x|T_x, G)$$

②Assumption: $p(x_i, x_j|T_x, G) = p(x_i|T_x, G) \times p(x_j|T_x, G)$

③Generate a paper d_i in the network:

$$p(d_i|G) = \prod_{x \in N_G(d_i)} p(x|G)^{W_{d_i,x}}$$
$$= \prod_{x \in N_G(d_i)} p(x|T_x, G)^{W_{d_i,x}} p(T_x|G)^{W_{d_i,x}}$$

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

C. Posterior Probability for target Objects and Attribute Objects

①Generative probability of a target object:

$$p(d|G_k) = \prod_{x \in N_{G_k}(d)} p(x|T_x, G_k)^{W_{d,x}} p(T_x|G_k)^{W_{d,x}}$$

②Smoothing handling:

$$P_S(X|T_X, G_k) = (1 - \lambda_S)P(X|T_X, G_k) + \lambda_S P(X|T_X, G)$$

③Posterior probability: $p(k|d_i) \propto p(d_i|k) \times p(k).$

EM

$$logL = \sum_{i=1}^{|D|} \log(p(di)) = \sum_{i=1}^{|D|} \log[\sum_{k=1}^{K+1} p(di|k)p(k)]$$

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

Posterior probability for attribute objects

$$p(k|x) = \sum_{d \in N_G(x)} p(k,d|x) = \sum_{d \in N_G(x)} p(k|d)p(d|x)$$

$$= \sum_{d \in N_G(x)} p(k|d) \frac{1}{|N_G(x)|}$$

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

E.:Ranking distribution for Attribute Objects

①Simple Ranking

$$p(x|T_x, G) = \frac{\sum_{y \in N_G(x)} W_{xy}}{\sum_{x' \in T_x} \sum_{y \in N_G(x')} W_{x'y}}$$

②Authority Ranking

I. $P(Y|T_Y, G) = W_{YZ} W_{ZX} P(X|T_X, G)$

II. As the following PROPERTY2

NetClus(Yizhou Sun,Yintao Yu,Jiawei Han,KDD'09)

PROPERTY 2. *Given a three-typed network with star network schema $G = \langle X \bigcup Y \bigcup Z, E, W \rangle$, where Z is the center type, and $\forall z, N_G(z) = \{x, y\}(x \in X, y \in Y)$, authority ranking $P(X)$ and $P(Y)$ are calculated through Equation 5 iteratively, then estimated joint distribution $\hat{P}(X,Y) = \{\hat{p}(x,y) = P(X = x)P(Y = y), x \in X, y \in Y\}$ equals to the joint distribution represented by one rank matrix $\frac{M}{||M||_1}$, such that $||W_{XZ}W_{ZY} - M||_F$ is minimized.*

III. According to the DBLP rules

$$P(C|T_C, G) = W_{CD} D_{DA}^{-1} W_{DA} P(A|T_A G)$$

$$P(A|T_A, G) = W_{AD} D_{DC}^{-1} W_{DC} P(C|T_C, G)$$

where D_{DA} and D_{DC} are the diagonal matrices with the diagonal value equaling to row sum of W_{DA} and W_{DC}.

GenClus

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Idea:cluster with incomplete attributes across objects and consider **different types of links** which may have variable importance.
- Advantage:
  - Based strength-aware of different links
  - Probabilistic clustering model

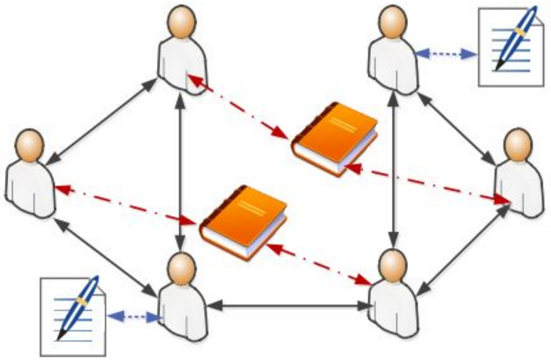


Figure 1: A Motivating Example on Clustering Political Interests in Social Information Networks

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)
Some Definitions

- Heterogeneous IN:G = (V,E,W)
- Mapping function from object to object:

  τ: $V \rightarrow A$ A is object type set.

- Mapping function from link to link type:

  φ: $E \rightarrow R$ R is link type set.

- Relation from type A to type B: $A\mathcal{R}B = B\mathcal{R}^{-1}A$
- Attributes: $\mathcal{X} = \{X_1, \ldots, X_T\}$ $v[X] = \{x_{v,1}, x_{v,2}, \ldots, x_{v,N_{X,v}}\}$

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)
Some Definitions

- Example1:DBLP

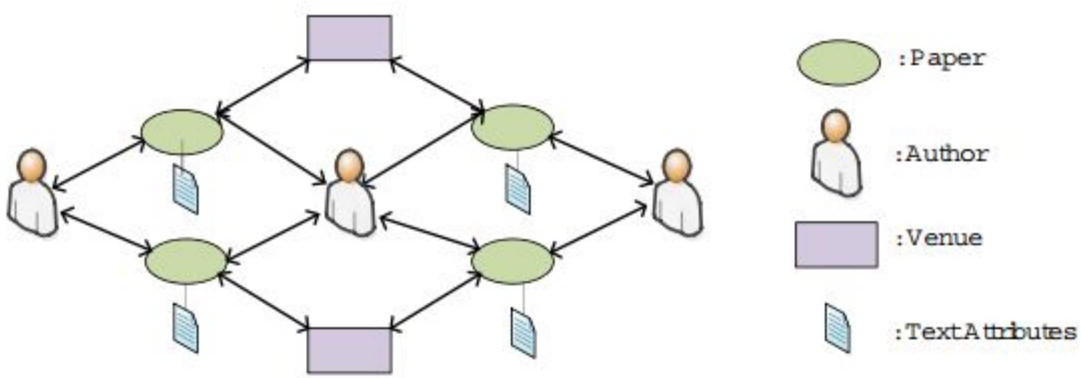


Figure 2: Illustration of Bibliographic Information Network

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)
Some Definitions

- Example2:Weather sensor network

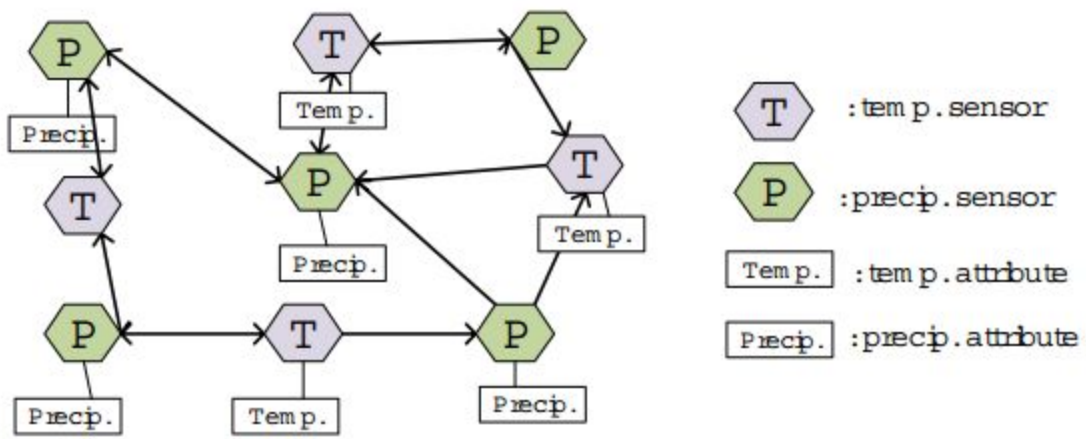


Figure 3: Illustration of Weather Sensor Information Network

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12) Some Definitions

- Formation

Formally, given a network $G = (V, E, W)$, a specified subset of its associated attributes $X \in \mathcal{X}$, the attribute observations $\{v[X]\}$ for all objects, and the number of clusters K, our goal is:

1. to learn a soft clustering for all the objects $v \in V$, denoted by a membership probability matrix, $\Theta_{|V|\times K} = (\boldsymbol{\theta}_v)_{v\in V}$, where $\Theta(v,k)$ denotes the probability of object v in cluster k, $0 \leq \Theta(v,k) \leq 1$ and $\sum_{k=1}^{K} \Theta(v,k) = 1$, and $\boldsymbol{\theta}_v$ is the K dimensional cluster membership vector for object v, and
2. to learn the strengths (importance weights) of different link types in determining the cluster memberships of the objects, $\gamma_{|\mathcal{R}|\times 1}$, where $\gamma(r)$ is a real number and stands for the importance weight for the link type $r \in \mathcal{R}$.

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Clustering Model:

① Two properties:

① attribute generated with high probability

② links beteen objects which have similar clustering probability.

② likelihood function of attribute:

$$p(\{\{v[X]\}_{v \in V_X}\}_{X \in \mathcal{X}}, \Theta | G, \gamma, \beta) = \prod_{X \in \mathcal{X}} p(\{v[X]\}_{v \in V_X} | \Theta, \beta) p(\Theta | G, \gamma) \quad (1)$$

two tasks

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Task One_Modeling Attribute Generation

$$p(\{v[X]\}_{v\in V_X}|\Theta,\boldsymbol{\beta}) = \prod_{v\in V_X}\prod_{x\in v[X]}\sum_{k=1}^{K}\theta_{v,k}p(x|\boldsymbol{\beta}_k) \quad (2)$$

- assume the attribute values have two type:text,numerical

① Text attribute with categorical distribution

$$p(\{v[X]\}_{v\in V_X}|\Theta,\boldsymbol{\beta}) = \prod_{v\in V_X}\prod_{l=1}^{m}(\sum_{k=1}^{K}\theta_{v,k}\beta_{k,l})^{c_{v,l}} \quad (3)$$

② Numerical attribute with Gaussian distribution

$$p(\{v[X]\}_{v\in V_X}|\Theta,\boldsymbol{\beta}) = \prod_{v\in V_X}\prod_{x\in v[X]}\sum_{k=1}^{K}\theta_{v,k}\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \quad (4)$$

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Task One_Modeling Attribute Generation
  - Multiple Attributes

    assume the independence among these attribute,

$$p(\{v[X_1]\}_{v\in V_{X_1}},\ldots,\{v[X_T]\}_{v\in V_{X_T}}|\Theta,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_T) = \prod_{t=1}^{T} p(\{v[X_t]\}_{v\in V_{X_t}}|\Theta,\boldsymbol{\beta}_t) \quad (5)$$

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

Task Two_Modeling Structural Consistency

- The more similar the two objects are in terms of cluster membership,the more likely they are connected by a link.

① Consistency function

$$f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, e, \boldsymbol{\gamma}) = -\gamma(r)w(e)H(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i) = \gamma(r)w(e)\sum_{k=1}^{K} \theta_{j,k} \log \theta_{i,k} \quad (6)$$

② Probability of Θ

$$p(\Theta|G, \boldsymbol{\gamma}) = \frac{1}{Z(\boldsymbol{\gamma})} \exp\{\sum_{e=\langle v_i, v_j\rangle \in E} f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, e, \boldsymbol{\gamma})\} \quad (7)$$

partition function(配分函数)

$$Z(\gamma) = \int_{\Theta} \exp\{\sum_{e=\langle v_i, v_j\rangle \in E} f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, e, \gamma)\} d\Theta$$

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Unified Model(overall goal)

The overall goal of the network clustering problem is to determine the best clustering results Θ, the link type strengths γ and the cluster component parameters β that maximize the generative probability of attribute observations and the consistency with the network structure, described by the likelihood function in Eq. (1).

$$g(\Theta, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \log \sum_{X \in \mathcal{X}} p(\{v[X]\}_{v \in V_X} | \Theta, \boldsymbol{\beta}) + \log p(\Theta | G, \boldsymbol{\gamma}) - \frac{||\boldsymbol{\gamma}||^2}{2\sigma^2} \quad (8)$$

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Algorithm Flow:

① ***Initial:***Set initial strength of different types of links with equally importance.

② ***Clustering optimization step:***Fix the link type weights to th γ best value , γ^* termined in the last iteration.Then optimize the objective function with regard to and , Θ at is Θ

$$[\Theta^*, \boldsymbol{\beta}^*] = \arg\max_{\Theta, \boldsymbol{\beta}} g(\Theta, \boldsymbol{\beta}, \gamma^*).$$ **EM**

GenClus(Yizhou Sun,Charu C.Aggarwal,Jiawei Han,VLDB'12)

- Algorithm Flow:

③ ***Link type strength learning step:*** Fix the clustering configuration parameters $\Theta = \Theta^*$ and $\beta = \beta^*$ corresponding to the values determined in the last step, and use it to determine the best value of γ ,which is consistent with current clustering results.

$$\gamma^* = \arg\max_{\gamma \geq 0} g(\Theta^*, \beta^*, \gamma).$$ **Newton-Raphson**

③ Iteratively repeat step 2 and 3 until convergence is achieved.

PathSelClus

Wenbao Li

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

- Idea:integrating meta-path selection and user-guided clustering to improve both the performance of clustering and learn the weights of different meta-paths.

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

Example

In Figure 2(a), authors are connected via organizations and form two clusters: $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$; in Figure 2(b), authors are connected via venues and form two different clusters: $\{1, 3, 5, 7\}$ and $\{2, 4, 6, 8\}$; whereas in Figure 2(c), a connection graph combining both meta-paths generate 4 clusters: $\{1, 3\}, \{2, 4\}, \{5, 7\}$ and $\{6, 8\}$. ■

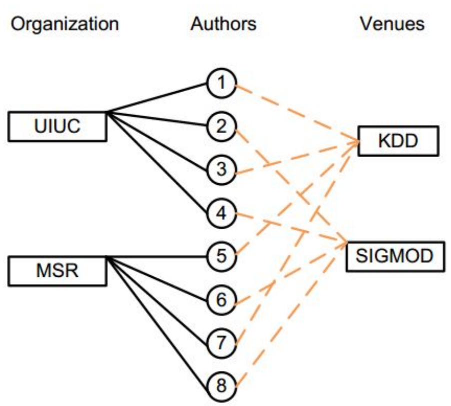


Figure 1: A toy heterogeneous information network containing organizations, authors and venues.

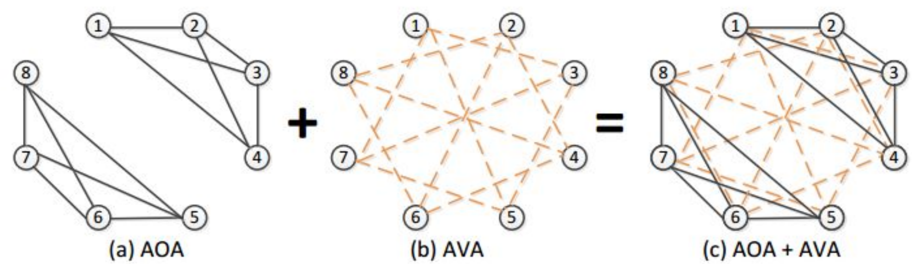


Figure 2: Author connection graphs under different meta-paths.

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

- **Meta-path selection**:M-PS problem is then to determine which meta-paths or their weighted combination to use for a specific clustering task.
- **User-Guided Clustering**:UGU is clustering under the condition of limited object seeds in each cluster given by users.

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

- Input:
  - The target type for clustering,type T.
  - The number of cluster K. say $\mathcal{L}_1, \ldots, \mathcal{L}_K$
  - The object seeds for each cluster, $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_M$
  - A set of M meta-paths starting from type T,
- Output:
  - The weight $\alpha_m \geq 0$ of each meta-path $\mathcal{P}_m$
  - The clustering results $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})$

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

- Modeling the Relation Generation

$$\pi_{ij,m} = P(j|i,m) = \sum_k P(k|i)P(j|k,m) = \sum_k \theta_{ik}\beta_{kj,m} \quad (1)$$

$$P(W_m|\Pi_m, \Theta, B_m) = \prod_i P(\mathbf{w}_{i,m}|\boldsymbol{\pi}_{i,m}, \Theta, B_m) = \prod_i \prod_j (\pi_{ij,m})^{w_{ij,m}} \quad (2)$$

- Modeling the Users Guidance

Dirichlet Distribution $\lambda \mathbf{e}_{k^*} + 1$

$$P(\boldsymbol{\theta}_i|\lambda) \propto \begin{cases} \prod_k \theta_{ik}^{\mathbf{1}_{\{t_i \in \mathcal{L}_k\}}\lambda} = \theta_{ik^*}^{\lambda}, & \text{if } t_i \text{ is labeled and } t_i \in \mathcal{L}_{k^*}, \\ 1, & \text{if } t_i \text{ is not labeled.} \end{cases} \quad (3)$$

Uniform Distribution

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

- Modeling the weights for meta-path selection
- by evaluating W_m the consistency between its relation Π_m matrix .

$$\alpha_m^* = \arg\max_{\alpha_m} \prod_i P(\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m) \quad (4)$$

The posterior of $\boldsymbol{\pi}_{i,m} = \boldsymbol{\theta}_i B_m$ is another Dirichlet distribution with the updated parameter vector as $\alpha_m \mathbf{w}_{i,m} + \mathbf{1}$, according to the multinomial-Dirichlet conjugate:

$$\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m \sim Dir(\alpha_m w_{ij,m}+1, \ldots, \alpha_m w_{i|F_m|,m}+1) \quad (5)$$

$$P(\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m) = \frac{\Gamma(\alpha_m n_{i,m} + |F_m|)}{\prod_j \Gamma(\alpha_m w_{ij,m} + 1)} \prod_j (\pi_{ij,m})^{\alpha_m w_{ij,m}} \quad (6)$$

PathSelClus(Yizhou Sun,Brandon Norick,Jiawei Han,Xifeng Yan,Philip S. Yu,KDD'12)

- Unified Model

$$P(\{\alpha_m W_m\}_{m=1}^{M}, \Pi_{1:M}, \Theta | B_{1:M}, \Phi_{1:M}, \lambda) = \prod_i \Big(\prod_m P(\alpha_m W_m | \Pi_m, \boldsymbol{\theta}_i, B_m) P(\Pi_m | \Phi_m)\Big) P(\boldsymbol{\theta}_i | \lambda) \quad (7)$$

$$J = \sum_i \Big(\sum_m \log P(\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m) + \sum_k \mathbf{1}_{\{t_i \in \mathcal{L}_k\}} \lambda \log \theta_{ik}\Big) \quad (8)$$

$$J = \sum_i \Big(\sum_m \Big(\sum_j \alpha_m w_{ij,m} \log \sum_k \theta_{ik} \beta_{kj,m} + \log \Gamma(\alpha_m n_{i,m} + |F_m|) - \sum_j \log \Gamma(\alpha_m w_{ij,m} + 1)\Big) + \sum_k \mathbf{1}_{\{t_i \in \mathcal{L}_k\}} \lambda \log \theta_{ik}\Big) \quad (9)$$

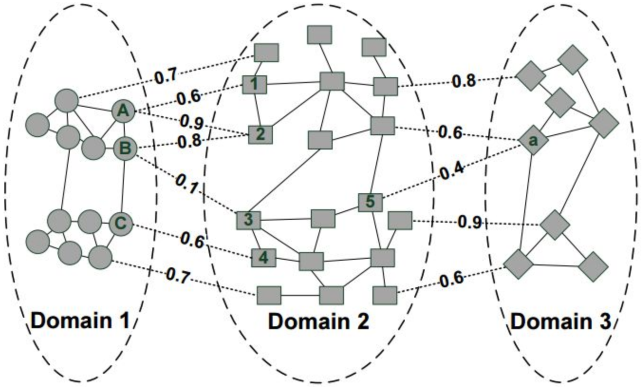
EM optimization

CGC

CGC(Wei Cheng,Xiang Zhang,Zhishan Guo,Yubao Wu,KDD'13)

- Co-Regularized Multi-Domain Graph Clustering

- Idea:Based on NMF,deal with cross-domain with many to many weighted relations.
- Use loss function regularization



CGC (Wei Cheng,Xiang Zhang,Zhishan Guo,Yubao Wu,KDD'13)

- Co-Regularized Multi-Domain Clustering
- $1.Single-Domain: \min L^{(\pi)} = \left\| A^{(\pi)} - H^{(\pi)}(H^{(\pi)})^T \right\|_F^2, \arg\max_j h_{aj}^{(\pi)}$
- $2.A).k_1 = k_2 = ... = k_d = k$

$$\mathcal{J}_{b,l}^{(i,j)} = (\mathbb{E}^{(i,j)}(x_b^{(j)}, l) - \mathbf{h}_{b,l}^{(j)})^2$$

$$\mathbb{E}^{(i,j)}(x_b^{(j)}, l) = \frac{1}{|\mathcal{N}^{(i,j)}(x_b^{(j)})|} \sum_{a \in \mathcal{N}^{(i,j)}(x_b^{(j)})} \mathbf{S}_{b,a}^{(i,j)} \mathbf{h}_{a,l}^{(i)} \quad (3)$$

- Residual of sum of squares loss function

$$\mathcal{J}_{RSS}^{(i,j)} = \sum_{l=1}^{k} \sum_{b=1}^{n_j} \mathcal{J}_{b,l}^{(i,j)} = ||\mathbf{S}^{(i,j)}\mathbf{H}^{(i)} - \mathbf{H}^{(j)}||_F^2 \quad (4)$$

- B).

$$\mathcal{J}_{CD}^{(i,j)} = \sum_{a=1}^{n_j} \sum_{b=1}^{n_j} \left(K(\widetilde{\mathbf{H}}_{a*}^{(i\to j)}, \widetilde{\mathbf{H}}_{b*}^{(i\to j)}) - K(\mathbf{h}_{a*}^{(j)}, \mathbf{h}_{b*}^{(j)}) \right)^2$$

$$= ||\mathbf{S}^{(i,j)}\mathbf{H}^{(i)}(\mathbf{S}^{(i,j)}\mathbf{H}^{(i)})^{\mathrm{T}} - \mathbf{H}^{(j)}(\mathbf{H}^{(j)})^{\mathrm{T}}||_F^2$$

CGC (Wei Cheng,Xiang Zhang,Zhishan Guo,Yubao Wu,KDD'13)

- Co-Regularized Multi-Domain Clustering
  - Joint Matrix optimization

$$\min_{\mathbf{H}^{(\pi)} \geq 0 (1 \leq \pi \leq d)} \mathcal{O} = \sum_{i=1}^{d} \mathcal{L}^{(i)} + \sum_{(i,j) \in \mathcal{I}} \lambda^{(i,j)} \mathcal{J}^{(i,j)}$$

where $\mathcal{J}^{(i,j)}$ can be either $\mathcal{J}_{RSS}^{(i,j)}$ or $\mathcal{J}_{CD}^{(i,j)}$.

SI-Cluster

Wenbao Li

SI-Cluster(Yang Zhou,Ling Liu,KDD'13)

- Idea:define new vertex similarity metric in terms of **self-influence** similarity and c**o-influence** similarity,and then according the similarity calculated from the social graph and associated activity graph,combine into total social influence and do clustering.

ComClus

Wenbao Li

ComClus(Ran Wang,Chuan Shi,Philip S. Yu,Bin Wu,PAKDD'13)

- deals with the hybrid network with heterogeneous and homogeneous network simultaneously.
- applies star schema with self loop to organize the **hybrid** network and uses a probability model to represent the generative probability of objects.

Part Two

Our Ideas and Involved Difficulties

- From the point of multi-view(multi-kernel) clustering

■ do clustering for all types of objects with constraints which are determined by the relations between different types of objects.

$$obj = f(A,P,V) = f_1(A) + f_2(P) + f_3(V) + CONS.(A,P,V)$$

■ But how to model the clustering of single type object and the constraints?

Our Ideas and Involved Difficulties

- From point of regularization(such as CGC)
  - Do clustering for a target type(such as A,author)

$$Obj = f(A) + L_1(A,P) + L_2(A,V)$$

  - How to model the clustering of single type object and the regularization of its related type.

Our Ideas and Involved Difficulties

- From the point of meta-path(PathSelClus)
  - A-P-A
  - A-V-A
  - A-T-A

Over