GIN: A Clustering Model for Capturing Dual Heterogeneity in Networked Data

Jialu Liu Chi Wang Jing Gao Quanquan Gu
Charu Aggarwal Lance Kaplan Jiawei Han

May 1, 2015

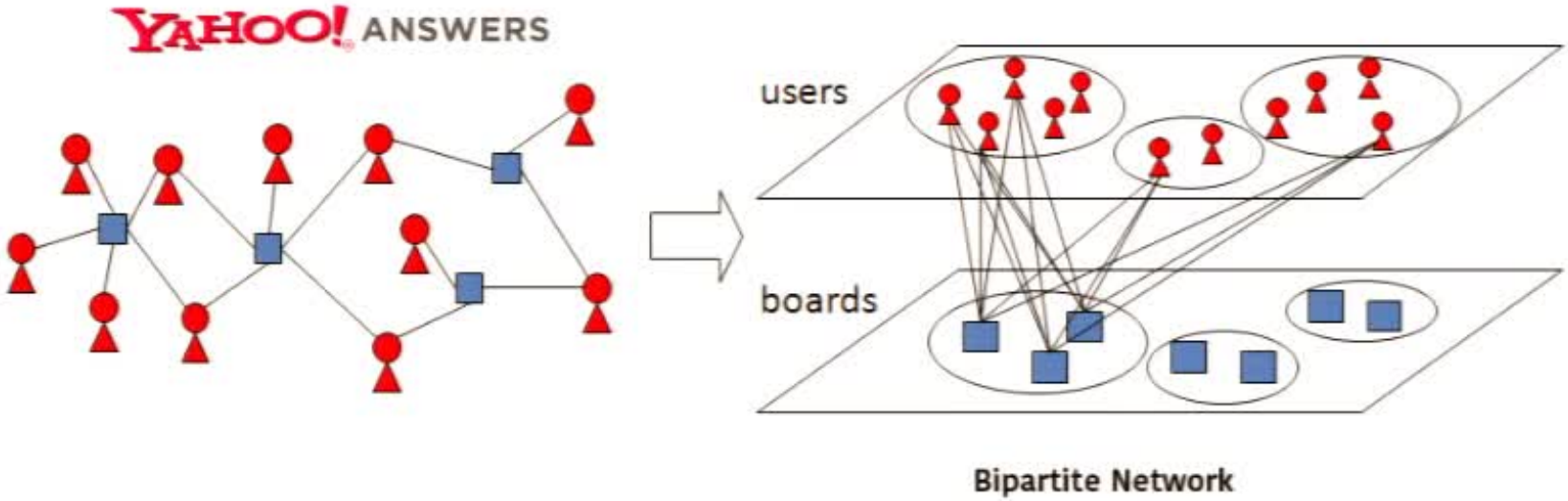
Outline

Networked Data

Many real-world data can be represented as a network (or graph), which is composed of nodes interconnected with each other via meaningful links.

Node Heterogeneity

In real networks, there will likely be multiple types of nodes.

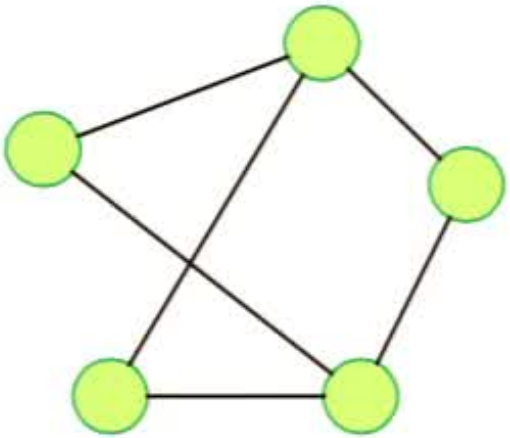


Bipartite Network

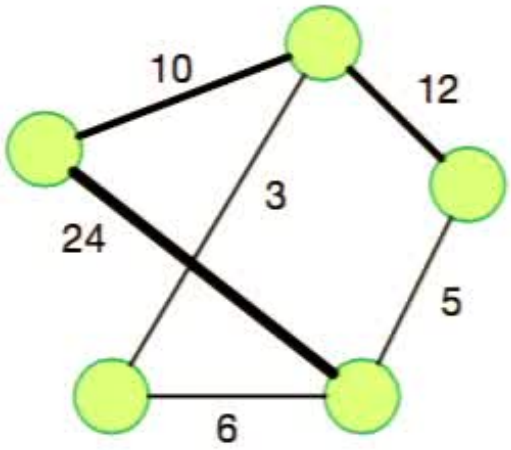
From Murata et.al. "Analysis of Online Question-Answering Forums as Heterogeneous Networks"

Link Heterogeneity

Meanwhile, links can be categorized into different types.



Binary/Unweighted Links



Weighted Links

Besides link weights, links can be directed or undirected.

Dual Heterogeneity

In this work, we work on heterogeneous networks that contain interconnected multi-typed nodes and links. Specifically, links are *undirected* but are allowed to be either *binary* or *weighted*.

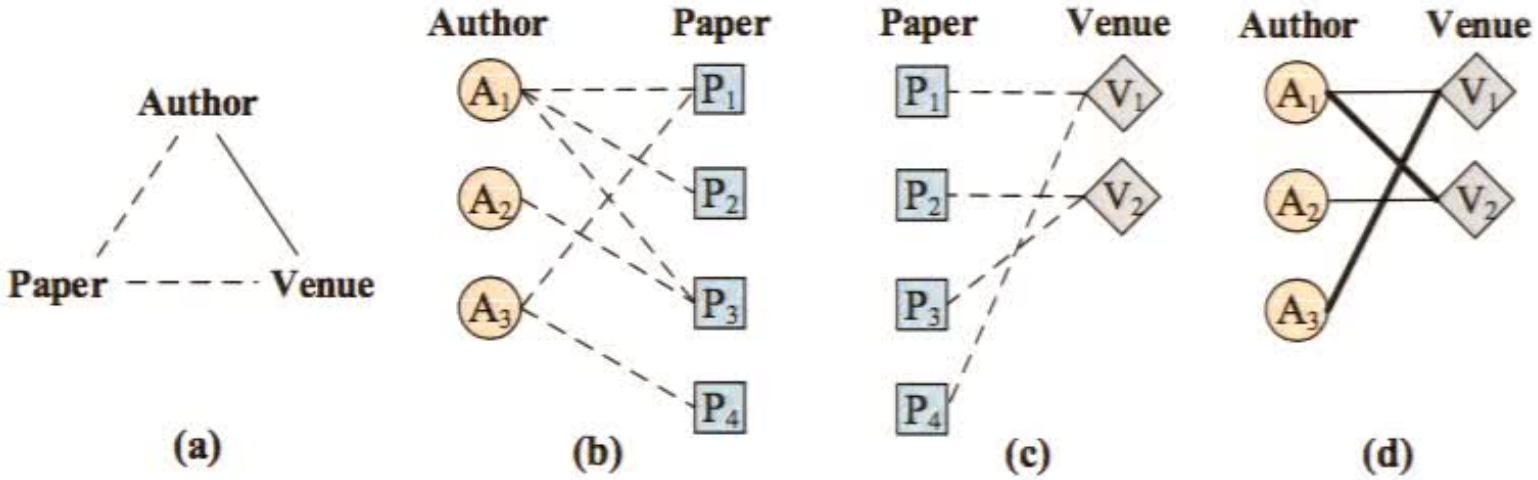


Figure: Dashed line – binary links, Solid line – weighted links.

Task and Novelty

Network Clustering: We aim to find a clustering solution given a *general* heterogeneous network, in which each cluster consists of *multiple types* of nodes and links.

Novelty compared with previous works:

- We are considering heterogeneity in both nodes and links;
- The algorithm does not have requirement on the network schema;
- The algorithm shows that sampling unobserved links (negative sampling) improves performance.

Subnetworks

A subnetwork in heterogeneous network is either a homogeneous network or a bipartite network.

A network with the number of object types $T = 1$ is called *homogeneous network*.

It is called *bipartite network* when $T = 2$ and links only exist between two object types.

Task and Novelty

Network Clustering: We aim to find a clustering solution given a *general* heterogeneous network, in which each cluster consists of *multiple types* of nodes and links.

Novelty compared with previous works:

- We are considering heterogeneity in both nodes and links;
- The algorithm does not have requirement on the network schema;
- The algorithm shows that sampling unobserved links (negative sampling) improves performance.

Symbols

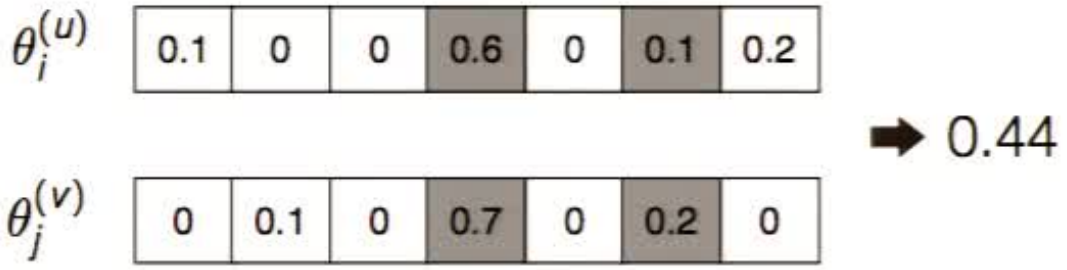
- We use G to denote a heterogeneous network and $G^{(uv)}$ to represent its subnetwork (can be homogeneous or bipartite network depending on whether object type u equals v).
- $G^{(uv)}$ can be either unweighted or weighted. That is to say, link $e_{ij}^{(uv)}$ between nodes $x_i^{(u)}$ and $x_j^{(v)}$ with weight $W_{ij}^{(uv)}$ can be binary or take any non-negative values.

Subnetworks with Binary Links

Suppose the probability of a link between nodes $x_i^{(u)}$ and $x_j^{(v)}$ is $P(e_{ij}^{(uv)} = 1)$.

Specifically, we factorize $P(e_{ij}^{(uv)} = 1)$ into $\sum_{k=1}^{K} \theta_{ik}^{(u)} \theta_{jk}^{(v)}$ where $\{\theta_{ik}^{(u)}\}_{k=1}^{K}$ is a vector with length K indicating the cluster membership of node $x_i^{(u)}$.

This factorization implies that two nodes get connected more easily if they share the same cluster distribution.



The underlying generative process for link $e_{ij}^{(uv)}$ is as follows:

$$e_{ij}^{(uv)} \sim Bernoulli(\sum_{k} \theta_{ik}^{(u)} \theta_{jk}^{(v)}).$$

For the whole set of binary links $E^{(uv)}$, the following likelihood can be derived to estimate parameters:

$$\prod_{i<j} \Big(P(e_{ij}^{(uv)} = 1) \Big)^{W_{ij}^{(uv)}} \underbrace{\Big(P(e_{ij}^{(uv)} = 0) \Big)^{1 - W_{ij}^{(uv)}}}_{\text{Unobserved Links}} \qquad (1)$$

Subnetworks with Weighted Links

Similar to the Bernoulli setting in the previous subsection, we first model the existence of a link between a given pair of nodes.

In addition to the cluster membership vector $\theta_i^{(u)}$, we incorporate a scale parameter $\sigma_i^{(u)}$ for each node $x_i^{(u)}$ in consideration of the weighted setting.

Then we can come up with the following generative process for weighted links:

$$\begin{aligned} &\text{(a) } e_{ij}^{(uv)} \sim \textit{Bernoulli}\Big(\sum_k \theta_{ik}^{(u)}\theta_{jk}^{(v)}\Big) \\ &\text{(b) If } e_{ij}^{(uv)} = 1, \quad \omega_{ij}^{(uv)} \sim \textit{Poisson}\Big(\sigma_i^{(u)}\sigma_j^{(v)}\sum_k \theta_{ik}^{(u)}\theta_{jk}^{(v)}\Big) \end{aligned} \tag{2}$$

where discrete random variable $\omega_{ij}^{(uv)}$ is the weight of the link.

Objective Function

We first define two sets of subnetworks belonging to the same heterogeneous network G: $\mathcal{B}$ and $\mathcal{W}$. They represent subnetworks having binary and weighted links respectively, satisfying that $\mathcal{B} \cup \mathcal{W} = G$ and $\mathcal{B} \cap \mathcal{W} = \varnothing$.

$$
\begin{aligned}
&\prod_{G^{(uv)} \in \mathcal{B}} \prod_{i<j} \left(\sum_k \theta_{ik}^{(u)} \theta_{jk}^{(v)} \right)^{W_{ij}^{(uv)}} \left(1 - \sum_k \theta_{ik}^{(u)} \theta_{jk}^{(v)} \right)^{1 - W_{ij}^{(uv)}} \\
&\times \prod_{G^{(uv)} \in \mathcal{W}} \prod_{W_{ij}^{(uv)} = 0} \left(1 - \sum_k \theta_{ik}^{(u)} \theta_{jk}^{(v)} \right) \\
&\times \prod_{W_{ij}^{(uv)} > 0} \left(\sum_k \theta_{ik}^{(u)} \theta_{jk}^{(v)} \right) \frac{\left(\sigma_i^{(u)} \sigma_j^{(v)} \sum_k \theta_{ik}^{(u)} \theta_{jk}^{(v)} \right)^{W_{ij}^{(uv)}}}{W_{ij}^{(uv)}!} \\
&\times e^{-\sigma_i^{(u)} \sigma_j^{(v)} \sum_k \theta_{ik}^{(u)} \theta_{jk}^{(v)}}.
\end{aligned}
\tag{4}
$$

Complete Log-likelihood

To directly optimize the previsou expression is difficult. We apply EM algorithm by using $\phi_{ijk_1k_2}^{(uv)}$ to denote the posterior probability of an unobserved link generated from different cluster assignments of two end nodes, i.e., $k_1 \neq k_2$. Meanwhile, we use $\psi_{ijk}^{(uv)}$ to denote the posterior probability of a link resulted from the same cluster assignments of two end nodes.

$$
\begin{aligned}
\mathcal{L}(\Theta, \Sigma) = & \sum_{G^{(uv)} \in \mathcal{B}} \sum_{W_{ij}^{(uv)}=1} \sum_{k} \psi_{ijk}^{(uv)} \log \theta_{ik}^{(u)} \theta_{jk}^{(v)} \\
& + \sum_{G^{(uv)} \in \mathcal{W}} \sum_{W_{ij}^{(uv)}>0} \left(W_{ij}^{(uv)} + 1 \right) \sum_{k} \psi_{ijk}^{(uv)} \log \theta_{ik}^{(u)} \theta_{jk}^{(v)} \\
& + \sum_{G^{(uv)} \in G} \sum_{W_{ij}^{(uv)}=0} \sum_{k_1 \neq k_2} \phi_{ijk_1k_2}^{(uv)} \log \theta_{ik_1}^{(u)} \theta_{jk_2}^{(v)} \\
& + \sum_{G^{(uv)} \in \mathcal{W}} \sum_{W_{ij}^{(uv)}>0} W_{ij}^{(uv)} \log \sigma_i^{(u)} \sigma_j^{(v)}.
\end{aligned}
\tag{5}
$$

Update Functions

Expectation Step:

$$\phi^{(uv)}_{ijk_1k_2} = \frac{\theta^{(u)}_{ik_1}\theta^{(v)}_{jk_2}}{\sum_{l_1\neq l_2}\theta^{(u)}_{il_1}\theta^{(v)}_{jl_2}}.$$

$$\psi^{(uv)}_{ijk} = \frac{\theta^{(u)}_{ik}\theta^{(v)}_{jk}}{\sum_{l}\theta^{(u)}_{il}\theta^{(v)}_{jl}}.$$

Maximization Step:

$$\theta^{(u)}_{ik} \propto \sum_{G^{(uv)}\in\mathcal{B}}\sum_{W^{(uv)}_{ij}=1}\psi^{(uv)}_{ijk} + \sum_{G^{(uv)}\in\mathcal{W}}\sum_{W^{(uv)}_{ij}>0}\left(W^{(uv)}_{ij}+1\right)\psi^{(uv)}_{ijk}$$

$$+ \sum_{G^{(uv)}\in G}\sum_{W^{(uv)}_{ij}=0}\sum_{l\neq k}\phi^{(uv)}_{ijkl}.$$

Complete Log-likelihood

To directly optimize the previsou expression is difficult. We apply EM algorithm by using $\phi^{(uv)}_{ijk_1k_2}$ to denote the posterior probability of an unobserved link generated from different cluster assignments of two end nodes, i.e., $k_1 \neq k_2$. Meanwhile, we use $\psi^{(uv)}_{ijk}$ to denote the posterior probability of a link resulted from the same cluster assignments of two end nodes.

$$
\begin{aligned}
\mathcal{L}(\Theta, \Sigma) = & \sum_{G^{(uv)} \in \mathcal{B}} \sum_{W^{(uv)}_{ij}=1} \sum_{k} \psi^{(uv)}_{ijk} \log \theta^{(u)}_{ik} \theta^{(v)}_{jk} \\
& + \sum_{G^{(uv)} \in \mathcal{W}} \sum_{W^{(uv)}_{ij}>0} \left(W^{(uv)}_{ij} + 1 \right) \sum_{k} \psi^{(uv)}_{ijk} \log \theta^{(u)}_{ik} \theta^{(v)}_{jk} \\
& + \sum_{G^{(uv)} \in G} \sum_{W^{(uv)}_{ij}=0} \sum_{k_1 \neq k_2} \phi^{(uv)}_{ijk_1k_2} \log \theta^{(u)}_{ik_1} \theta^{(v)}_{jk_2} \\
& + \sum_{G^{(uv)} \in \mathcal{W}} \sum_{W^{(uv)}_{ij}>0} W^{(uv)}_{ij} \log \sigma^{(u)}_{i} \sigma^{(v)}_{j}.
\end{aligned}
\tag{5}
$$

Efficiency Issue

$$\phi^{(uv)}_{ijk_1k_2} = \frac{\theta^{(u)}_{ik_1}\theta^{(v)}_{jk_2}}{\sum_{l_1\neq l_2}\theta^{(u)}_{il_1}\theta^{(v)}_{jl_2}} \qquad O(k^2)$$

$$\Rightarrow \sum_{l\neq k}\phi^{(uv)}_{ijkl} = \frac{\sum_{l\neq k}\theta^{(u)}_{ik}\theta^{(v)}_{jl}}{\sum_{l_1\neq l_2}\theta^{(u)}_{il_1}\theta^{(v)}_{jl_2}} = \frac{\theta^{(u)}_{ik} - \theta^{(u)}_{ik}\theta^{(v)}_{jk}}{1-\sum_l \theta^{(u)}_{il}\theta^{(v)}_{jl}} \qquad O(k)$$

$$\theta^{(u)}_{ik} \propto \sum_{G^{(uv)}\in\mathcal{B}}\sum_{W^{(uv)}_{ij}=1}\psi^{(uv)}_{ijk} + \sum_{G^{(uv)}\in\mathcal{W}}\sum_{W^{(uv)}_{ij}>0}\left(W^{(uv)}_{ij}+1\right)\psi^{(uv)}_{ijk}$$
$$+ \sum_{G^{(uv)}\in G}\sum_{W^{(uv)}_{ij}=0}\left[\sum_{l\neq k}\phi^{(uv)}_{ijkl}\right]. \qquad (6)$$

Sampling Unobserved Links

For the unobserved links, the spatial/time complexity increases significantly if we need to go over all of them. To alleviate such burden we sampled a potential neighbourhood for each node. This also downweights the third term of $\theta_{ik}^{(u)}$

$$\theta_{ik}^{(u)} \propto \sum_{G^{(uv)} \in \mathcal{B}} \sum_{W_{ij}^{(uv)}=1} \psi_{ijk}^{(uv)} + \sum_{G^{(uv)} \in \mathcal{W}} \sum_{W_{ij}^{(uv)}>0} \left(W_{ij}^{(uv)} + 1\right) \psi_{ijk}^{(uv)} + \sum_{G^{(uv)} \in G} \sum_{W_{ij}^{(uv)}=0} \sum_{l \neq k} \phi_{ijkl}^{(uv)} \quad \downarrow \tag{7}$$

We keep all the non-zero links and sample ηM unobserved links to make its size proportional to the total number of links M (we choose $\eta = 0.1$ in the experiments).

Outline

Datasets

Four real world data sets were used.

- The **DBLP** data set is a collection of CS publications. We use a subset that belong to four research areas.
- The **4Groups** data set contains co-author and author-term relationships where researchers are selected from four data mining and machine learning research groups.
- The **Flickr** data set is a network containing three types of objects: image, user and tag. Links exist between image-user and image-tag.
- The **NSF** data set describes NSF Research Awards Abstracts from 1990 to 2003. We use documents associated with terms and investigators that belong to the largest 10 programs.

The important statistics of four datasets are summarized in the following table.

Data set	DBLP	4Groups	Flickr	NSF
#Nodes	70,536	1,618	4,076	30,995
#Links	332,388	5,568	14,396	1,883,682
Sparsity	6.7e-5	2.1e-3	8.7e-4	2.0e-3
#Clusters	4	4	8	10
#Objects	4	2	3	3
#Subnet.	3	2	2	2
Link Cat.	Binary	Weighted	Binary	Fused

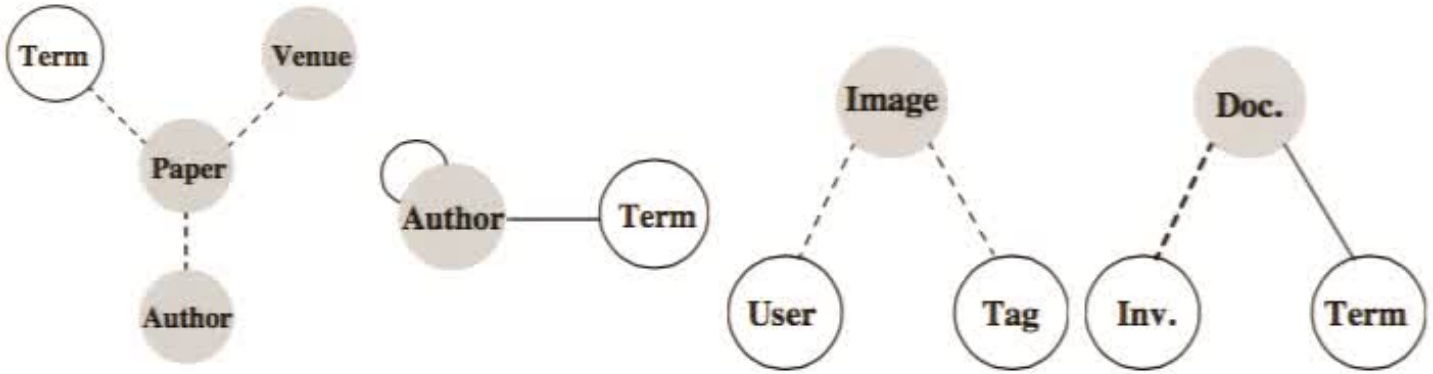


Figure: Network schemas of all data sets in which circles of labelled object types are in grey. Dashed (resp., solid) lines refer to binary (resp., weighted) links.

Compared Algorithms

We compared with the following algorithms:

- **GIN:** A Generative Model for Heterogeneous Information Networks. This is the proposed algorithm.
- **NetClus:** It is a rank-based algorithm integrating ranking and clustering together for networks with star schema.
- **SCIN:** Spectral Clustering for Heterogeneous Information Networks. We derived this algorithm by extending spectral clustering to the heterogeneous networks.
- **SC:** Standard Spectral Clustering, a spectral-based algorithm which is designed to segment graphs and is shown to be effective on networks.
- **PHIN:** A Poisson Model for Homogeneous Information Networks. This generative model is recently proposed to cluster homogeneous network data.

Performance

Clustering accuracy on the four data sets:

Data set	DBLP				4Groups	Flickr	NSF
Object	Author	Paper	Venue	Average	Author	Image	Doc.
GIN	**93.01**	**84.75**	**100.00**	**92.85**	**97.16**	**48.44**	**74.48**
NetClus	89.90	80.00	**100.00**	89.72	-	44.94	70.42
SCIN	86.26	81.00	90.00	86.16	89.89	42.12	72.29
SC	46.03	41.00	30.00	45.84	56.14	37.74	44.62
PHIN	75.71	63.00	60.00	75.35	62.28	43.97	61.95
#Labels	4,236	100	20	-	99	1,028	10,606

We chose Flickr and NSF data sets and conducted a thorough study since they have more clusters than the others.

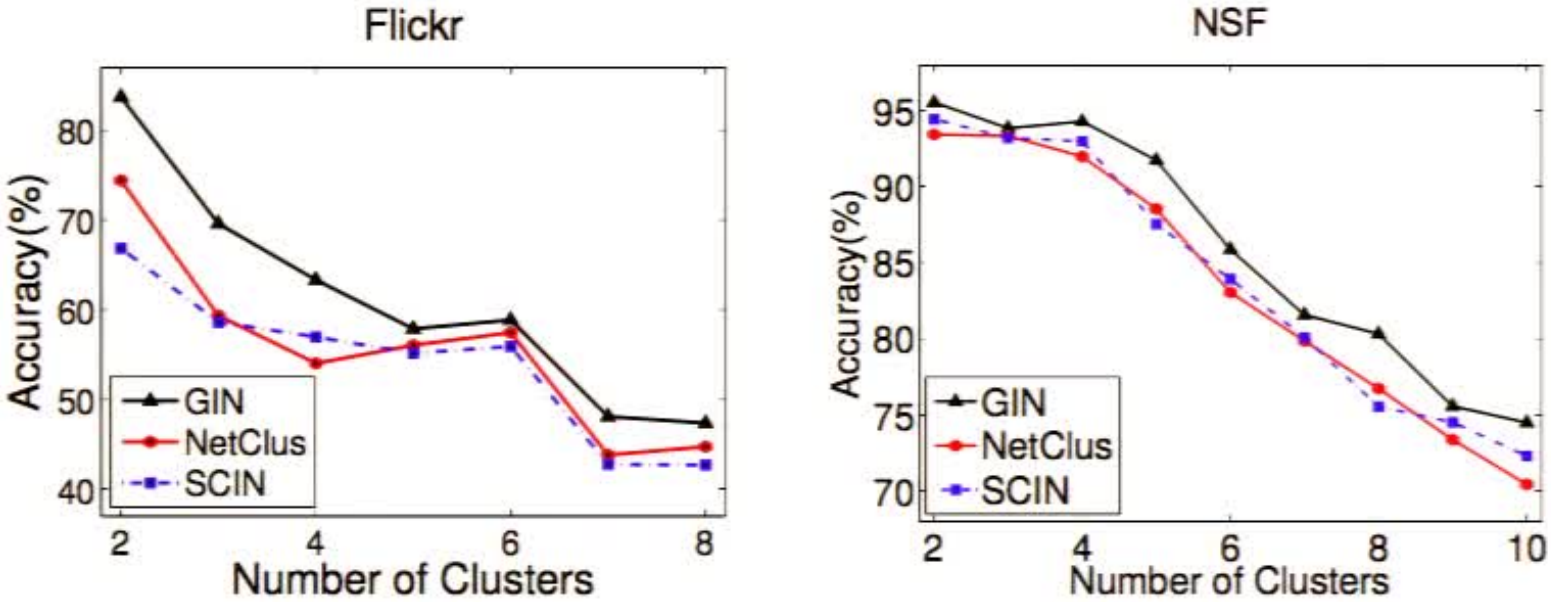


Figure: Clustering performance on Flickr and NSF.

Parameter Study

One parameter in our model is the sample size (ηM) of non-linked node pairs. We have tested various values of η in the range of $[10^{-3}, 10^{0}]$.

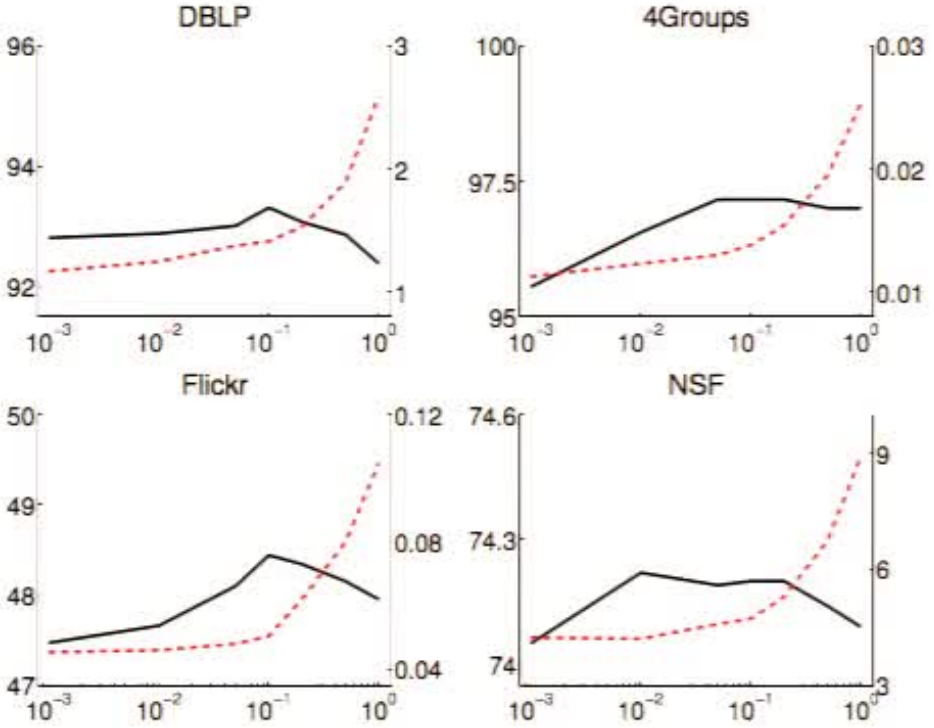


Figure: Accuracy and running time (in seconds) v.s. sample proportion η. Dashed (resp., solid) lines refer to running time (resp., accuracy).

Conclusions

We have proposed a general clustering approach to model heterogeneous information network.

- It models binary and weighted links as well as multi-typed nodes.
- Subnetworks are separately modelled and then unified (schema-free).
- It samples non-observed links which is shown to improve performance.
- Time efficient $O(MK + NK + \eta MK)$.