Combining Link and Content for Community Detection: A Discriminative Approach

Tianbao Yang[1], Rong Jin[1], Yun Chi[2], Shenghuo Zhu[2]
[1]Michigan State University [2] NEC Laboratories America

Presenter: April Hua LIU

Outline

- Background
- Conditional Link Model
- Discriminative Content Model
- Optimization Algorithms
- Extensions
- Experiments
- Conclusion

Background

- Community detection in network
  - Community:
    - Densely connected in links
    - Common topic in contents
  - Network data
    - Links between nodes: e.g. citation between papers
    - Content describing nodes: e.g. bag-of words for papers

Background(Cont.)

- Most work on community detection
  - Link analysis, but links are sparse and noisy
  - Content analysis, but content can be misleading

- Combing link and content
  - Most are based on generative models
    - Link-model (PHITS)+ topic-model (PLSA)
    - Connected by the community memberships (hidden variable)

Our contribution

- Problems with existing models
  - Community membership is insufficient to model links
    - Our contribution: introduce popularity of nodes

  - Generative model, vulnerable to irrelevant attributes
    - Our contribution: discriminative content model

Notations

$\mathcal{V} = \{1, \dots, n\}$	nodes
$\mathcal{E} = \{(i \rightarrow j) \mid s_{ij} \neq 0\}$	directed links
$\mathcal{LO}(i) \in \mathcal{V}$	link-out space of node i
$\mathcal{LI}(i) \in \mathcal{V}$	link-in space of node i
$\mathcal{O}(i) \in \mathcal{V}$	nodes cited by node i
$\mathcal{I}(i) \in \mathcal{V}$	nodes cites node i
$z_i \in \{1, \dots, K\}$	community of node i
$\gamma_i = (\gamma_{i1}, \dots, \gamma_{iK})$	community membership of node i
$x_i \in \mathbb{R}^d$	content vector of node i

Conditional link model

- Popularity-based conditional link model(PCL)
  - Model conditional link probability: Pr(j|i)
    - Probability of linking node i to node j
    - Popularity of node i : $b_i \geq 0$
      - Large b_i→high probability cited by other nodes

$$\Pr(j|i) = \sum_{k=1}^{K} \boxed{\Pr(z_i = k|i)}\, \Pr(j|z_i = k)$$

$$= \sum_{k=1}^{K} \boxed{\gamma_{ik}} \frac{\gamma_{jk} b_j}{\sum_{j \in \mathcal{LO}(i)} \gamma_{jk} b_j}$$

Analysis of PCL model

- PCL model

$$\Pr(\mathrm{j}|\mathrm{i}) = \sum_{k=1}^{K} \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j \in \mathcal{LO}(i)} \gamma_{jk} b_j}$$

$$\Pr(\mathrm{j}|\mathrm{i}) = \sum_{k=1}^{K} \gamma_{ik} \frac{\gamma_{jk} b_{jk}}{\sum_{j \in \mathcal{LO}(i)} \gamma_{jk} b_{jk}}$$

$$\Pr(j|i) = \sum_{k=1}^{K} \Pr \; \ldots \; z = k) = \sum_{k} \gamma_{ik} \beta_{jk}$$

PHITS model

Maximum Likelihood Estimation

- The log-likelihood:

$$\log \mathcal{L} = \sum_{(i \to j) \in \mathcal{E}} \hat{s}_{ij} \log \sum_k \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j' \in \mathcal{LO}(i)} \gamma_{j'k} b_{j'}}$$

- We find optimal γ, b by maxmizing the log-likelihood

$$\max_{\gamma, b} \quad \sum_{(i \to j) \in \mathcal{E}} \hat{s}_{ij} \log \sum_k \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j' \in \mathcal{LO}(i)} \gamma_{j'k} b_{j'}}$$

$$s.t. \quad \sum_k \gamma_{ik} = 1, \gamma_{ik} \geq 0, b_i \geq 0$$

Discriminative Content (DC) model

- A discriminative model that determines community memberships by node contents

$$\Pr(z_i = k) = y_{ik} = \frac{\exp(w_k^T x_i)}{\sum_l \exp(w_l^T x_i)}$$

Where $w_k \in \mathbb{R}^d$ weights different content features

PCL + DC

$$\Pr(j|i) = \sum_{k=1}^{K} \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j \in \mathcal{LO}(i)} \gamma_{jk} b_j} \qquad \gamma_{ik} = \frac{\exp(w_k^T x_i)}{\sum_l \exp(w_l^T x_i)}$$

Optimization Algorithm

- We maximize the log-likelihood over the free parameters w and b

$$\{w, b\}^* = \arg\max_{w,b} \log \mathcal{L} = \sum_{i=1}^{n} \sum_{j \in \mathcal{LO}(i)} \widehat{s}_{ij} \log \Pr(j|i; w, b)$$

- EM algorithm

Experiments

- Data sets

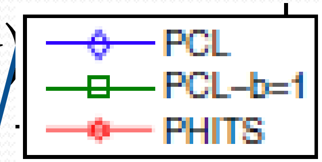
Data set	#nodes	#links	Content	Labels	K	Description
Political Blog	1490	19090	No	Yes	2	Blog network
Wikipedia	105	799	No	No	20	Webpages hyperlinks
Cora	2708	5429	Yes	Yes	7	Paper citation
Citeseer	3312	4732	Yes	Yes	6	Paper citation

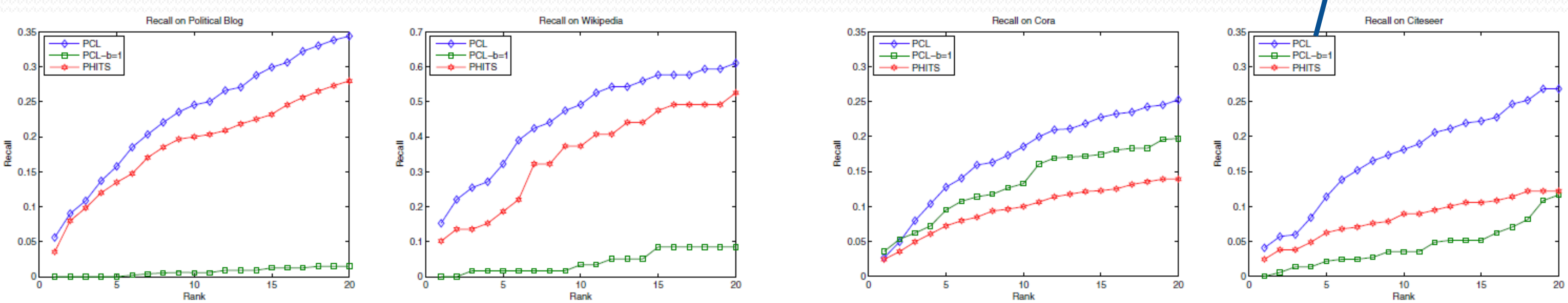
Experiments

- Performance Metrics
  - Supervised metrics
    - *normalized mutual information* (NMI)
    - *pairwise F-measure* (PWF)
  - Unsupervised metrics
    - *modularity* (Modu)
    - *normalized cut* (Ncut)

Experiments: link prediction

- Baselines: PHITS, PCL-b=1 (constant popularity)
- Recall measure





(a) Recall on Political Blog (b) Recall on Wikipedia (c) Recall on Cora (d) Recall on Citeseer

- PCL performs better than PHITS
- Modeling popularity better than without modeling

Experiments

- Community detection on two paper citation data sets

Table 1: Partition Measure on Cora and Citeseer dataset

		Cora				Citeseer			
	Algorithm	NMI	PWF	Modu	NCut	NMI	PWF	Modu	NCut
Link	PHITS	0.0570	0.1894	0.3929	3.2466	0.0101	0.1773	0.4588	2.2370
	LDA-Link	0.0762	0.2278	0.2189	4.5687	0.0356	0.2363	0.2211	3.7457
	PCL	0.0884	0.2055	0.5903	1.9391	0.0315	0.1927	0.6436	1.1181
	NCUT	0.1715	0.2864	0.2701	**0.2732**	0.1833	0.3252	0.6577	**0.1490**
Content	PLSA	0.2107	0.2864	0.2682	4.2686	0.0965	0.2298	0.2885	3.2294
	LDA-Word	0.2310	0.2774	0.2970	3.7820	0.1342	0.2880	0.3022	3.0165
	NCUT(RBF kernel)	0.1317	0.2457	0.1839	4.7775	0.0976	0.2386	0.2133	3.7078
	NCUT(pp kernel)	0.1804	0.2912	0.2487	4.6612	0.1986	0.3282	0.4802	1.8118
Link + Content	PHITS-PLSA	0.3140	0.3526	0.3956	3.2880	0.1188	0.2596	0.3863	2.7397
	LDA-Link-Word	0.3587	0.3969	0.4576	2.8906	0.1920	0.3045	0.5058	2.0369
	LCF	0.1227	0.2456	0.1664	4.8101	0.0934	0.2361	0.2011	3.6721
	NCUT(RBF kernel)	0.2444	0.3062	0.3703	1.6585	0.1592	0.2957	0.4280	1.7592
	NCUT(pp kernel)	0.3866	0.4214	0.5158	0.7903	0.1986	0.3282	0.4802	1.8118
	PCL-PLSA	0.3900	0.4233	0.5503	2.1575	0.2207	0.3334	0.5505	1.6786
	PHITS-DC	0.4359	0.4526	0.6384	1.5165	0.2062	0.3295	0.6117	1.2074
	PCL-DC	**0.5123**	**0.5450**	**0.6976**	1.0093	**0.2921**	**0.3876**	**0.6857**	0.7505

Experiments

- Link model: PCL is better than PHITS
- On combining link with content:
  - **PCL** + content-model performs better than link-models + content model
  - Link-models + **DC** performs better than link-model + topic-models
  - **PCL** + **DC** performs better than the other combination models

Conclusion

- A conditional link model capture popularity of nodes
- A discriminative model for content analysis
- A unified model to combine link and content
  - Link structure → noisy estimation of community memberships $\tilde{y}$ (PCL)
  - $\tilde{y}$ used as supervised information → high-quality memberships y (DC)
- Encouraging empirical results

Thanks
Q&A?