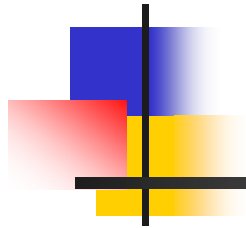


# Combining Link and Content for Community Detection: A Discriminative Approach



Tianbao Yang<sup>1</sup>, Rong Jin<sup>1</sup>, Yun Chi<sup>2</sup>, Shenghuo Zhu<sup>2</sup>

<sup>1</sup>Michigan State University, <sup>2</sup>NEC Laboratories America

Presenter: Jieping Ye  
Arizona State University



# Outline

---

- Background
- Conditional Link Model
- Discriminative Content Model
- Optimization Algorithms
- 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

---

- 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 + topic-model, connected by the community memberships



# Our contribution

---

- Problems with existing link models
  - Links determined only by community memberships
  - *Our contribution: introduce popularity of nodes*
- Problems with existing content analysis
  - Generative model, vulnerable to irrelevant attributes
  - *Our contribution: discriminative content model*



# Notations

---

$\mathcal{V} = \{1, \dots, n\}$	nodes
$\mathcal{E} = \{(i \rightarrow j)   s_{ij} \neq 0\}$	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 $i$
$\mathbf{x}_i \in \mathbb{R}^d$	content vector of node $i$



# Popularity-based Conditional Link (PCL) model

---

- Model conditional link probability:  $\Pr(j|i)$ 
  - Probability of creating a link from node  $i$  to node  $j$
  - Popularity of node  $i$ :  $b_i$ 
    - Large popularity  $\rightarrow$  high probability receiving a link

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

# Popularity-based Conditional Link (PCL) model

- Model conditional link probability:  $\Pr(j|i)$ 
  - Probability of creating a link from node  $i$  to node  $j$
  - Popularity of node  $i$ :  $b_i$ 
    - Large popularity  $\rightarrow$  high probability receiving a link

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

$\Pr(z_i = k|i) = \gamma_{ik}$

# Popularity-based Conditional Link (PCL) model

- Model conditional link probability:  $\Pr(j|i)$ 
  - Probability of creating a link from node  $i$  to node  $j$
  - Popularity of node  $i$ :  $b_i$ 
    - Large popularity  $\rightarrow$  high probability receiving a link

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



# Analysis of PCL model

---

## PHITS model

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

## PCL model

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



# Analysis of PCL model

---

## PHITS model

$$\begin{aligned}\Pr(j|i) &= \sum_{k=1}^K \Pr(z = k|i) \Pr(j|z = k) = \sum_k \gamma_{ik} \beta_{jk} \\ &= \sum_{k=1}^K \gamma_{ik} \frac{\Pr(j) \gamma_{jk}}{\sum_{j=1}^n \Pr(j) \gamma_{jk}}\end{aligned}$$

## PCL model

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



# Analysis of PCL model

## PHITS model

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

## PCL model

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

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



# Discriminative Content (DC) model

---

- A discriminative model that determines community memberships by node contents

$$\Pr(z_i = k|i) = y_{ik} = \frac{\exp(w_k^T x_i)}{\sum_{l=1}^K \exp(w_l^T x_i)}$$

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



# Discriminative Content (DC) model

- A discriminative model that determines community memberships by node contents

$$\Pr(z_i = k|i) = y_{ik} = \frac{\exp(w_k^T x_i)}{\sum_{l=1}^K \exp(w_l^T x_i)}$$

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

- **PCL** + **DC**

$$\Pr(j|i) = \sum_{k=1}^K \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_j \gamma_{jk} b_j} \quad \gamma_{ik} = \frac{\exp(w_k^T x_i)}{\sum_{l=1}^K \exp(w_l^T x_i)}$$



# Optimization Algorithm

---

- Find optimal parameters by maximizing the log-likelihood of data

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

where  $\hat{s}_{ij}$  is the normalized link from node  $i$  to  $j$

- EM algorithm
  - E step: bound log-likelihood from below
  - M step: maximize lower bound over  $w, b$



# Experiments

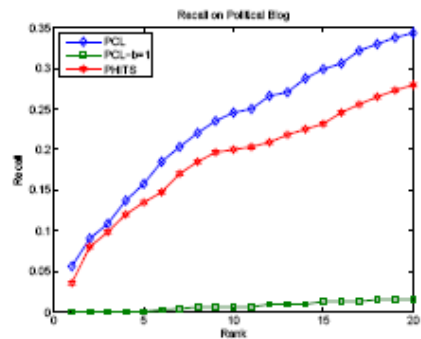
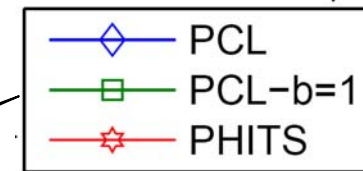
---

- Data sets

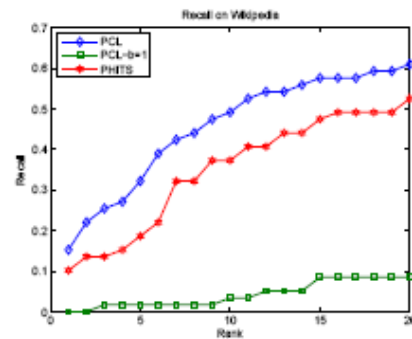
<b>Data set</b>	<b>#nodes</b>	<b>#links</b>	<b>content</b>	<b>labels</b>	<b>K</b>	<b>description</b>
Political Blog	1490	19090	no	yes	2	Blog network
Wikipedia	105	799	no	no	20	Webpages hyperlinks
Cora	2708	5429	yes	yes	7	Paper citation
Citeseer	3312	4732	yes	yes	6	Paper citation

# Experiments: link prediction

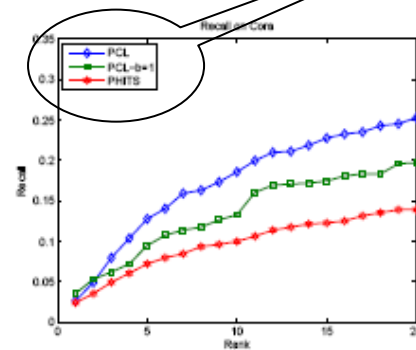
- Baselines: PHITS, PCL-b=1 (constant popularity)
- Recall measure



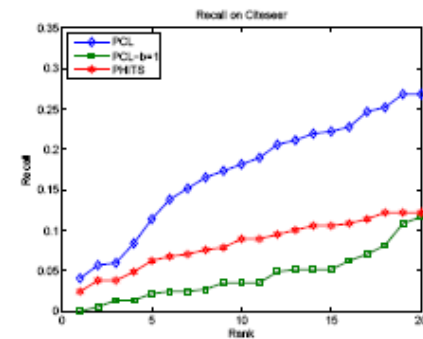
(a) Recall on Political Blog



(b) Recall on Wikipedia



(c) Recall on Cora



(d) Recall on Citeseer

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



# 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	<b>0.2732</b>	0.1833	0.3252	0.6577	<b>0.1490</b>
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	<b>0.5123</b>	<b>0.5450</b>	<b>0.6976</b>	1.0093	<b>0.2921</b>	<b>0.3876</b>	<b>0.6857</b>	0.7505



# Experiments

---

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



# Conclusion

---

- A unified model to combine link and content
- A conational link model capture popularity of nodes
- A discriminative model for content analysis
- Encouraging empirical results



Thanks

---

Questions?