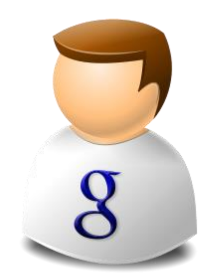
Latent Information Aware Social Network Labeling

Li Xiong, Tong He

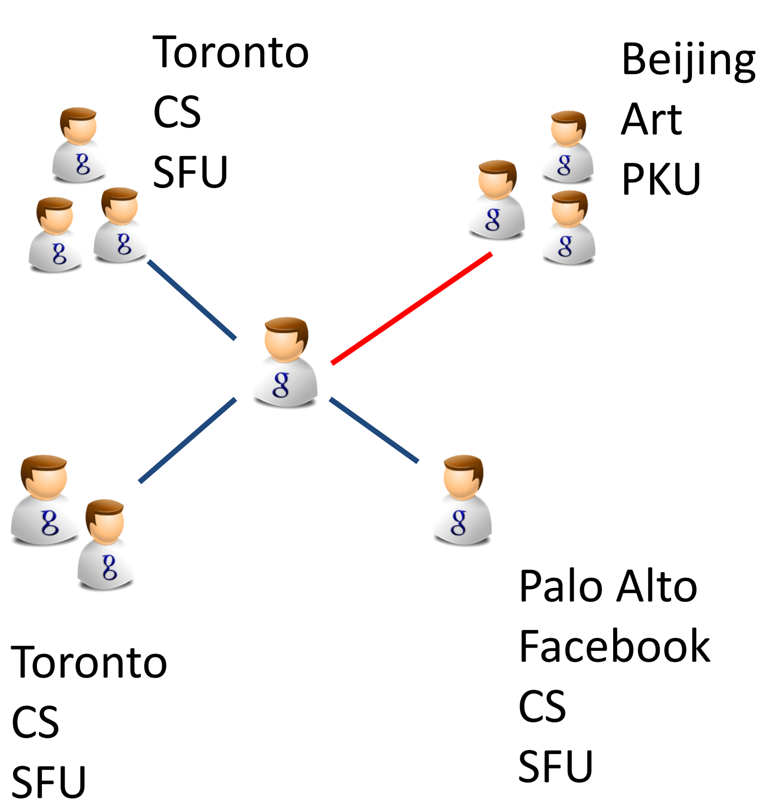
Motivation



Profile

Place lived: Toronto
Employer: ?
Major: CS
School: ?
Hobbies, Politics, …

- Profile is often incomplete
- Profile is useful
  - Personalized recommendation
  - Easy friends searching
  - Local ads targeting



Question: How can we fill in missing labels?

- Make use of social network information and assumption of *homophily [3]*
- Friendship can be explained by explicit labels or latent information
- Infer missing labels by maximizing the explanation of friendships

Problem Formulation

Given a graph $g = (V, E, T)$, where $T = \{t_1, t_2, \dots, t_k\}$ is a set of label types, V is a set of vertices, and $E = \{(u, v) \mid u, v \in V\}$. $L(t)$ denotes a set of labels w.r.t. label type t. Each vertex has a set of labels either fully labeled or partially labeled.

For a user u, define f_{utl} as the a probability distribution on label l with label type t.

For an edge (u, v), define $r(u,v) = \sum_t \boldsymbol{f}_{ut} \cdot \boldsymbol{f}_{vt}$ as the similarity of labels and $C(u,v)$ as the explanation from latent information.

References

[1] Joint Inference of Multiple Label Types in Large Networks. ICML 2014
[2] *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU 2002.
[3] What is Twitter, a social network or a news media?. 2010

Proposed Model

Key ideas:
- Explain friendship with both explicit labels and latent information
- Try to explain as many friendships as possible

Model Specification
Find optimal $\boldsymbol{f}$ to maximize

$$\underbrace{\prod_{u \sim v}}_{\text{Explain all friendship}} DegreeExplanation(\boldsymbol{f}_u, \boldsymbol{f}_v, C(u,v))$$

- $C(u,v)$: explanation for edge (u,v) from latent information
- $DegreeExplanation(\boldsymbol{f}_u, \boldsymbol{f}_v, C(u,v))$
  $= \sigma(a \cdot is\_reason(\boldsymbol{f}_u, \boldsymbol{f}_v) + is\_reason(C(u,v))$
- $is\_reason(\boldsymbol{f}_u, \boldsymbol{f}_v) = r(u,v)$
- $\sigma(x) = 1/(1+\exp(-x))$
- a is the coefficient which indicates to what extent this edge can be explained by shared explicit labels

Is (u, v) explained by explicit sharing labels or latent information?

Why we need $\boldsymbol{C}(\boldsymbol{u}, \boldsymbol{v})$:
- 60% edges on Google+ dataset share no explicit label
- One way to explain such edges is using latent information

Learning and Inference

Optimization objective:

$$argmax_{\boldsymbol{f}} \prod_{u \sim v} \sigma(a \cdot r(u,v) + C(u,v))$$

Take Neg. log

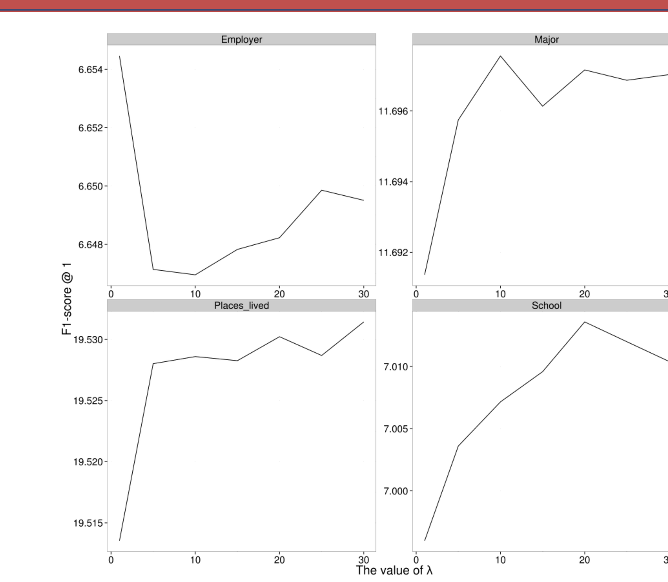
$$argmin_{\boldsymbol{f}} \; -\sum_{u \sim v} \log\{\sigma(a \cdot r(u,v) + C(u,v))\}$$

Add l_2-regularization on $C(u,v)$:

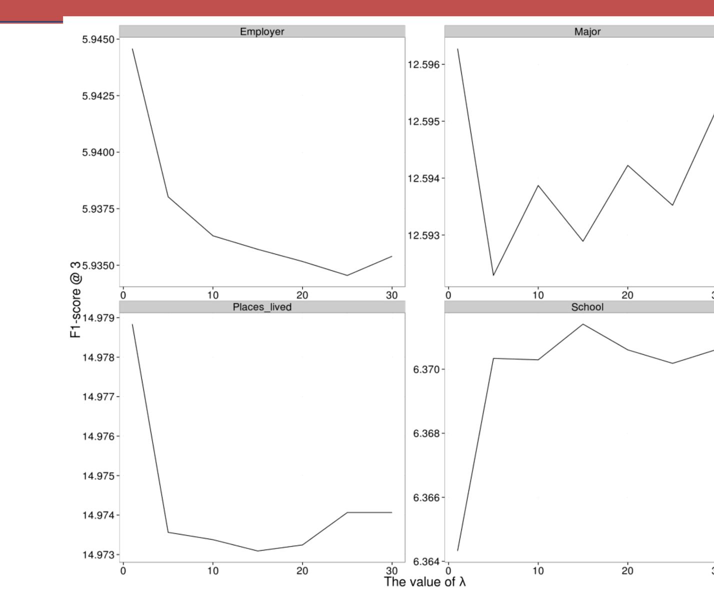
$$argmin_{\boldsymbol{f}} \; -\sum_{u \sim v} \log\{\sigma(a \cdot r(u,v) + C(u,v))\} + \frac{\lambda}{2}\|\boldsymbol{C}\|_2$$

EM based learning algorithm:
- Update $C(u,v)$ and $\boldsymbol{f}$ alternatively until converge
  1. Update $C(u,v)$ by setting $\partial L / \partial C(u,v) = 0$
  2. Update $\boldsymbol{f}$ using proximal gradient ascent

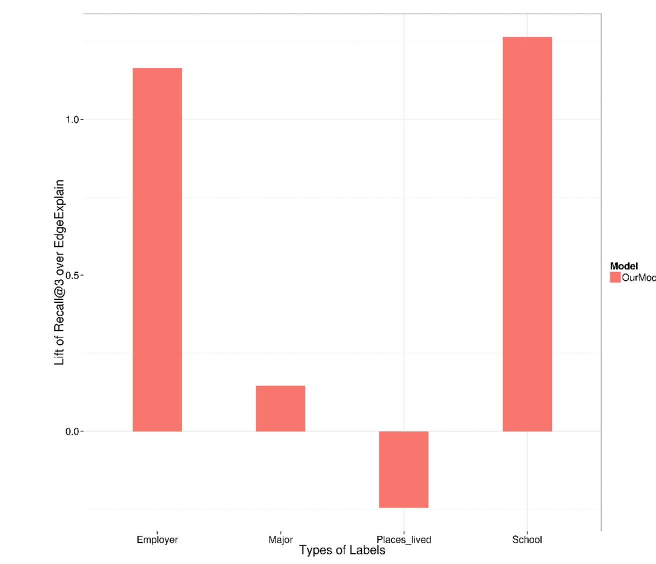
Experiments



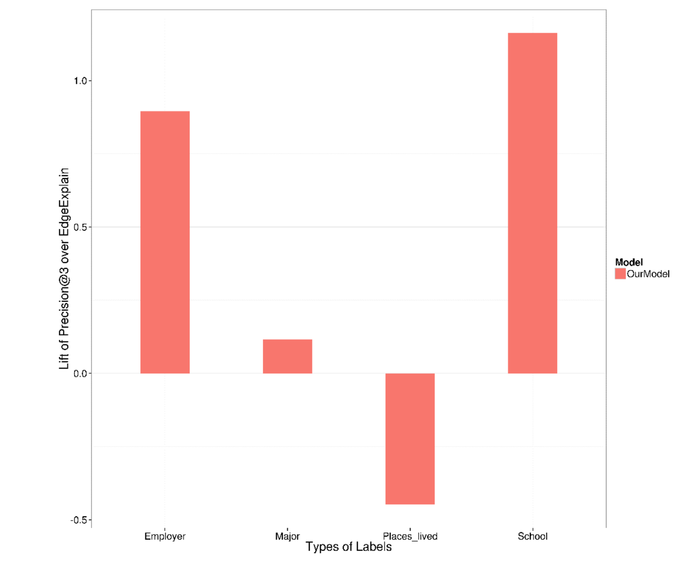
F1-score@1 for different λ



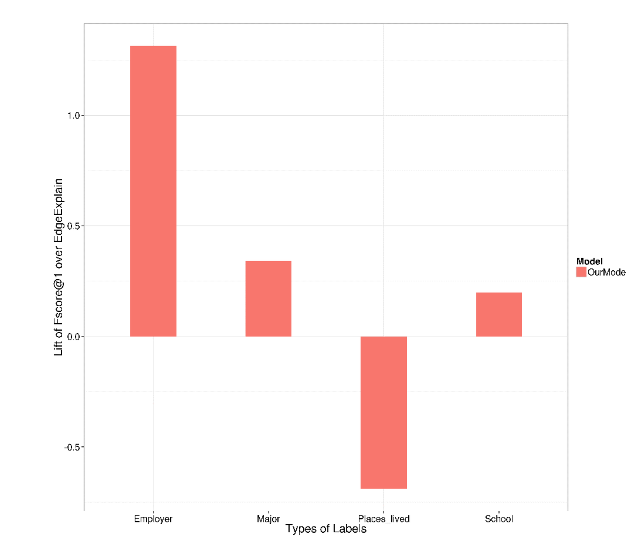
F1-score@3 for different λ



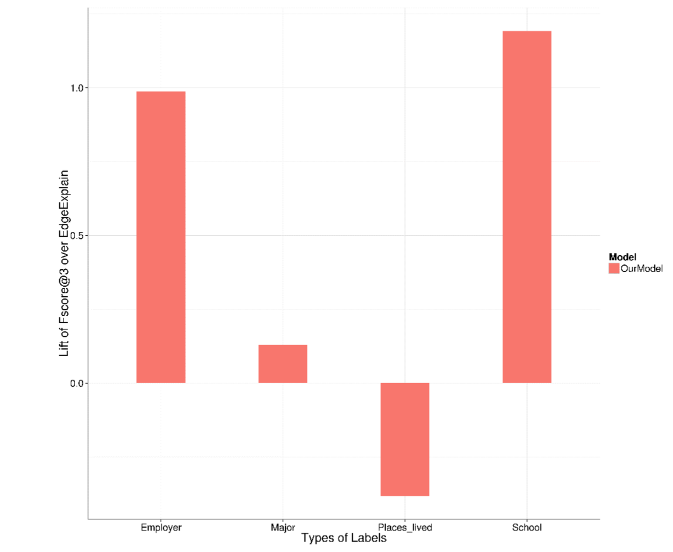
Percentage Lift of Recall@3



Percentage Lift of Precision@3



Percentage Lift of F1-score@1



Percentage Lift of F1-score@3

Summary

Our work

Assumption: Friendship can be explained by both explicit labels and latent information
Model: Use explicit labels and latent information to explain as many edges as possible
Results: F1 score 1% lift than [1]

Future work

- Try other bigger social network datasets
- Introduce correlation among label types
- Introduce different weights for label types