Targeted Network Recruitment on a Budget

Fabricio Murai[1], Bruno Ribeiro[2], Don Towsley[1], Krista Gile[1]

[1]University of Massachusetts Amherst, MA, USA

[2]Carnegie Mellon University, Pittsburgh, PA, USA

Motivation

Recruiting individuals of a given type (e.g., specific political affiliation) in a social network is a fundamental problem. In most real world applications, only **partial data** about **node attributes** and **topology** is available. This information is usually obtained **from the already recruited nodes**. There is often a **penalty** for trying to recruit the "wrong" nodes.

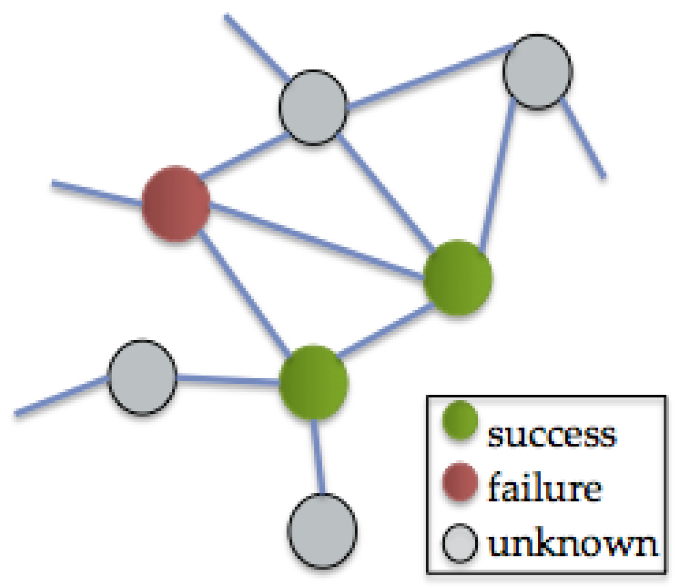

Figure 1: Snapshot of recruitment process.

Hence, the question we investigate is

Can we learn on-the-fly from the observed data to better recruit the "right" nodes?

Contributions

The main contributions of this work are:

- **highly-scalable method** to identify target nodes in partially observed networks
- generalization of **Bayesian** Sequential Analysis of the **logistic regression** [1] equation to consider **structural features**
- **evaluation** of proposed method against baseline and state-of-the-art techiniques

Datasets

Below is a short description of each network. Basic statistics are listed in the table.

- **DBLP**: scientific collaboration network where two authors are connected if they have published together. Binary node attributes that indicate whether an scholar has published in a specific venue. Undirected network, no weights considered.
- **YouTube**: online social network where two YouTube users are connected if they appear as friends. Binary node attributes indicate whether an user is subscribed to an user-defined group. Undirected network.

Dataset	nodes	edges	targets
DBLP	317.08K	1.05M	7.56K (2.4%)
YouTube	1.13M	2.99M	2.22K (0.2%)

Task

Target population: the largest ground-truth community in the network.

Task: recruit the maximum number of target individuals given a budget. Node attributes and neighbors ids are revealed upon recruitment.

Proposed Method and Learning Model

1. **Compute features** of nodes that can be sampled.

   Features = statistics of the **vicinity** of each node. E.g.: degree, fraction of nodes of each type, fraction that exhibit a given attribute and structural properties.
2. **Rank nodes** given its features and a model.

   Logistic Regression models probability of being a target given the features.
3. **Select "best" node.**

   Greedily select first in ranking.
4. **Update model** considering all observations.

   Never-ending learner: update model after each sample. Feasible via **Bayesian formulation**. Also, mechanism to decide relevant features online.

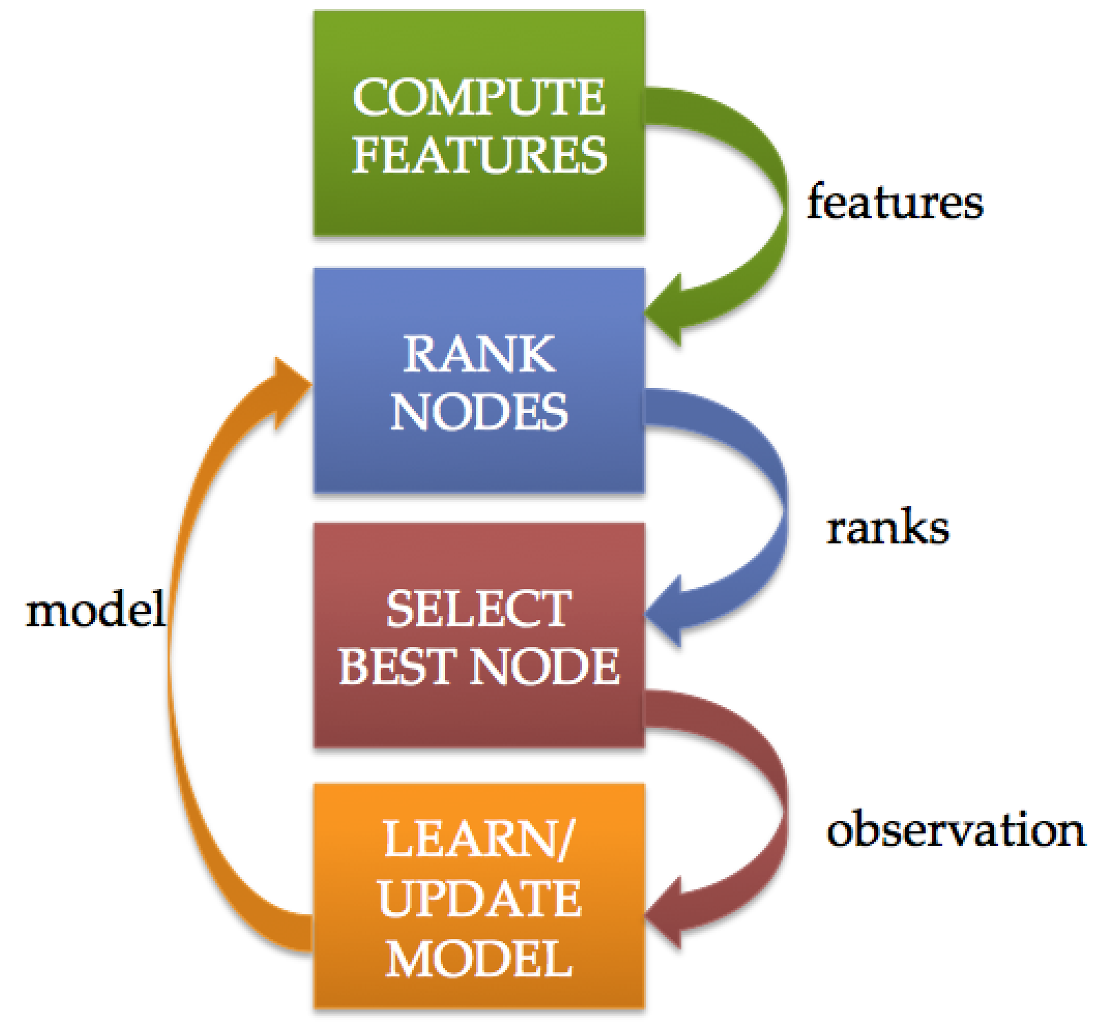

Figure 2: Algorithm steps, inputs and outputs.

Results

Baseline assumes strong *homophily* w.r.t. node types; i.e., the next node to be recruited is the one that has the largest difference between the number of recruited and not recruited neighbors.

Reference is based on **weighted averaging** of two-hop neighbors and **collective classification** via MCMC [2].

DBLP: main findings.

- Baseline does very well, even outperforming Reference (by 20% to 38%).
- Reference takes 1 week to run, whereas our method takes 10h in a Intel Xeon @ 2.6Ghz.
- Proposed was a bit better than Baseline (5 to 11%).

YouTube: main findings.

- Reference could not handle a larger network.
- Proposed improved number of successful recruitments four fold over the baseline.

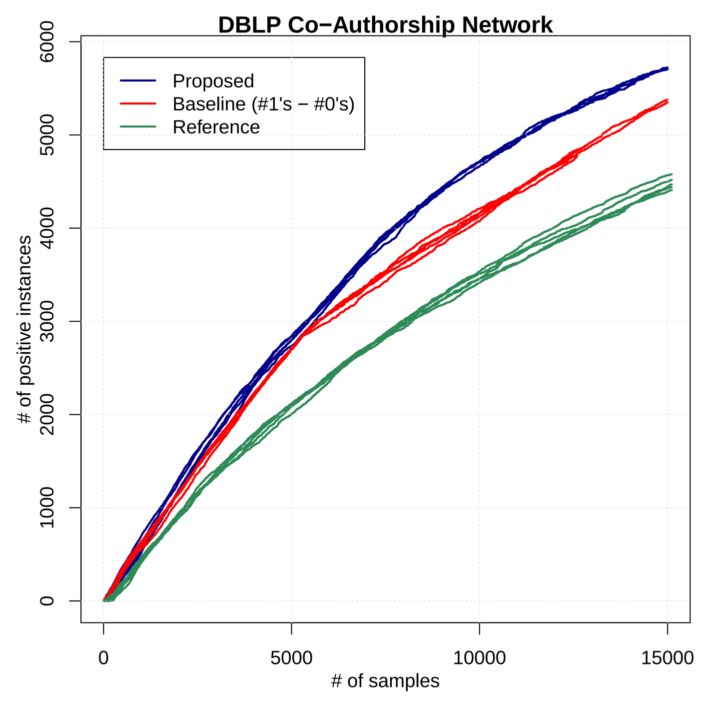

Figure 3: Sample paths for DBLP.

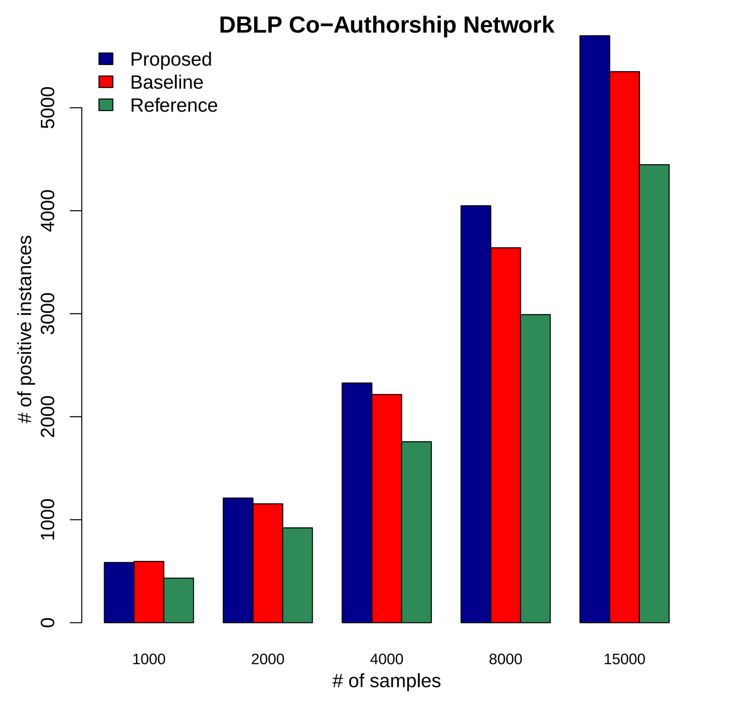

Figure 4: Average performances for DBLP (32 runs).

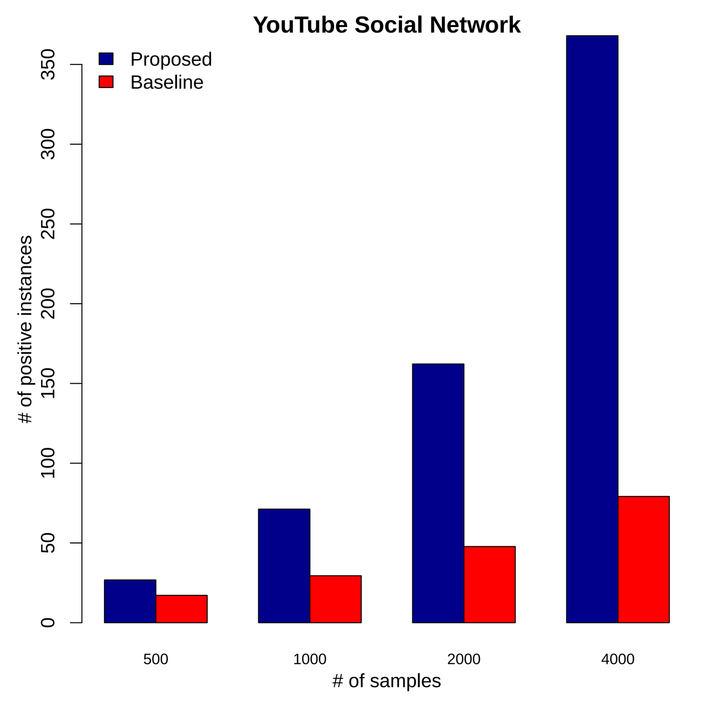

Figure 5: Average performances for YouTube (20 runs).

Conclusion

The proposed method is able to learn from node attributes and structural features to greatly improve the recruitment, even when compared to more costly methods. In homophily-based networks, it performs at least as well as the baseline.

References

[1] T. H. McCormick, A. E. Raftery, D. Madigan, and R. S. Burd. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics*, 68(1):23–30, 2012.

[2] J. J. Pfeiffer III, J. Neville, and P. N. Bennett. Active sampling of networks. In *10th International Workshop on Mining and Learning with Graphs*, 2012.

Contact Information

Web cs.umass.edu/~fabricio

Email fabricio@cs.umass.edu