Targeted Network Recruitment on a Budget

(preferences: ignite talk, talk, poster)

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.*

Recruiting network nodes of a given type – say, nodes with a specific political affiliation in an online social network – is one of the most fundamental problems in the computational social sciences. In real world applications this recruitment takes place in scenarios where only partial data (node attributes and topology) is available, more specifically, the data collected **from the already recruited nodes**. Moreover, in these scenarios there is often a steep penalty for trying to recruit the "wrong" nodes; for instance, a "get out the vote" campaign should not reach voters that will likely vote against the intent of the campaign. A key question, then, is whether it is possible to learn on-the-fly from the observed node attributes and local topology to better recruit the "right" nodes, even when these nodes are embedded in, say, a dissortative network. Moreover, the requirement of learning on-the-fly also implies that the algorithm must be **scalable**, that it be able to recruit the appropriate nodes regardless of the network size.

In our work, we propose *a highly-scalable method to identify node types in partially observed networks.* Our method aims at maximizing the number of correctly recruited nodes given a campaign's budget (specified in terms of the total number of recruitments). In particular, we generalize the Bayesian sequential analysis of the logistic regression equation (McCormick et. al 2011) to consider network structural features. The resulting method can model the relationship of a node's type with the types and attributes of other nodes taking into consideration their relative positions in the network structure; our never-ending learner updates the logistic model with new observations on-the-fly, without ever needing to rerun the costly logistic regression over the ever-growing observed data. We also propose a way to select the relevant attributes at each point in time, which is essential to achieve good performance as the amount of information conveyed by each attribute continually increases as we collect more observations.

We evaluate the performance of our search method w.r.t. the number of correctly recovered individuals in annotated datasets, finding that – for instance in the DBLP co-authorship network – we can improve the probability of finding authors that belong the largest scientific community in the dataset four fold over the state-of-the-art method (56% v.s. 14.7%) and twelve fold over random targeting (56% v.s. 4.7%). Figure 1 shows, for different budgets, the performance of: (i) an "oracle", (ii) the prior state-of-the art (Pfeiffer III et. al), (iii) random recruitment and (iv) our method. Curves represent averages over 50 runs and shaded areas indicate confidence intervals. Although the oracle was always able to find target nodes, this is unachievable in practice, as correlations between types, attributes and topological structure of nodes impose a fundamental bound to the recruitment performance. The prior state-of-the-art method performs significantly better than random recruitment, except for small budgets (where the performance is similar). Yet, our method outperforms the prior state-of-the-art algorithm by a large margin, which widens as the campaign budget increases.

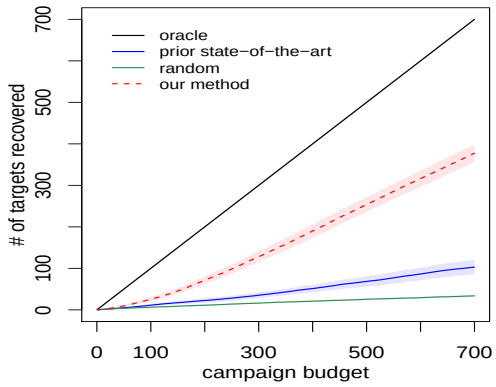

Figure 1: Average performance of each method as a function of the budget (50 runs). Shaded areas represent confidence intervals.

McCormick, T. H., Raftery, A. E., Madigan, D. and Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. Biometrics 68, (2011).
Pfeiffer III, J., Neville, J. and Bennett, P. N. Active sampling of networks, International Workshop on Mining and Learning with Graphs (2012).

This research was sponsored by the NSF under CNS-1065133, ARO under MURI W911NF-08-1-0233, and the U.S. Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053, the CNPq, National Council for Scientific and Technological Development – Brazil, NSF under SES-1230081 and the National Agricultural Statistics Service. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the NSF, ARO, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.