

# Crawling the Community Structure of Multiplex Networks

Ricky Laishram  
Syracuse University  
rlaishra@syr.edu

Jeremy D. Wendt  
Sandia National Laboratories  
jdwendt@sandia.gov

Sucheta Soundarajan  
Syracuse University  
susounda@syr.edu

## Multiplex Networks

- **Multiplex Networks:** A type of multilayer network in which all nodes can participate in all layers.
- Challenges of data collection in multiplex networks:
  - Different layers have different data collection costs.
  - Data collected from different layers have different reliabilities.

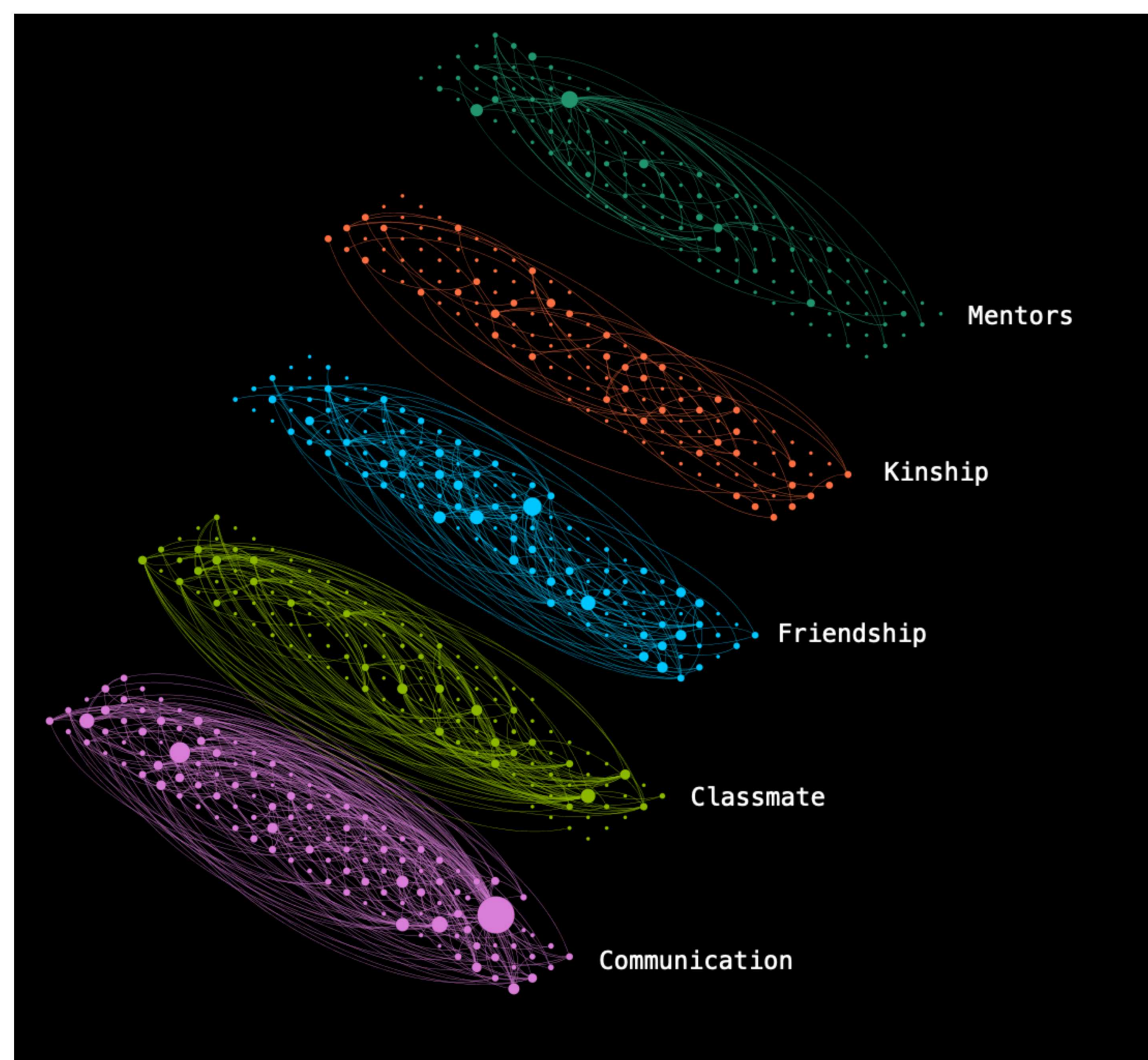


Figure: NoordinTop Multiplex Network

## Problem

- Let  $M$  be a multiplex network with layers  $L_0, L_1, \dots$  as the different layers.
- Query costs of the layers:  $c_0, c_1, \dots$
- Given an initial set of nodes  $V'$ , query budget  $B$ , and layer of interest  $L_0$ ; how can we sample  $M$  through crawling so that the sample of  $L_0$  found is community representative of  $L_0$  without exceeding the query budget?

## Query Response Models

- **Reliable Query Response (RQR):** A query for neighbors of a node returns all the neighbors.
- **Unreliable Query Response (UQR):** A query for the neighbors of a node may not return all the neighbors.

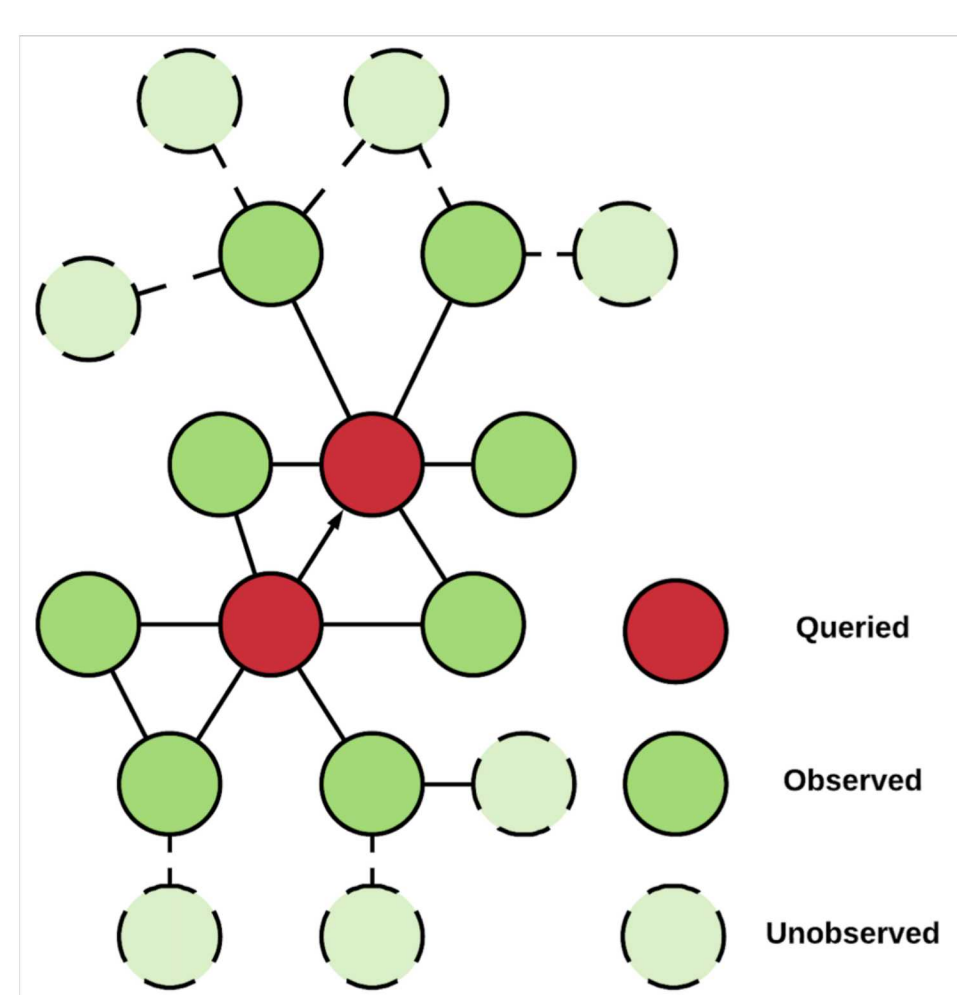


Figure: Reliable Query Response

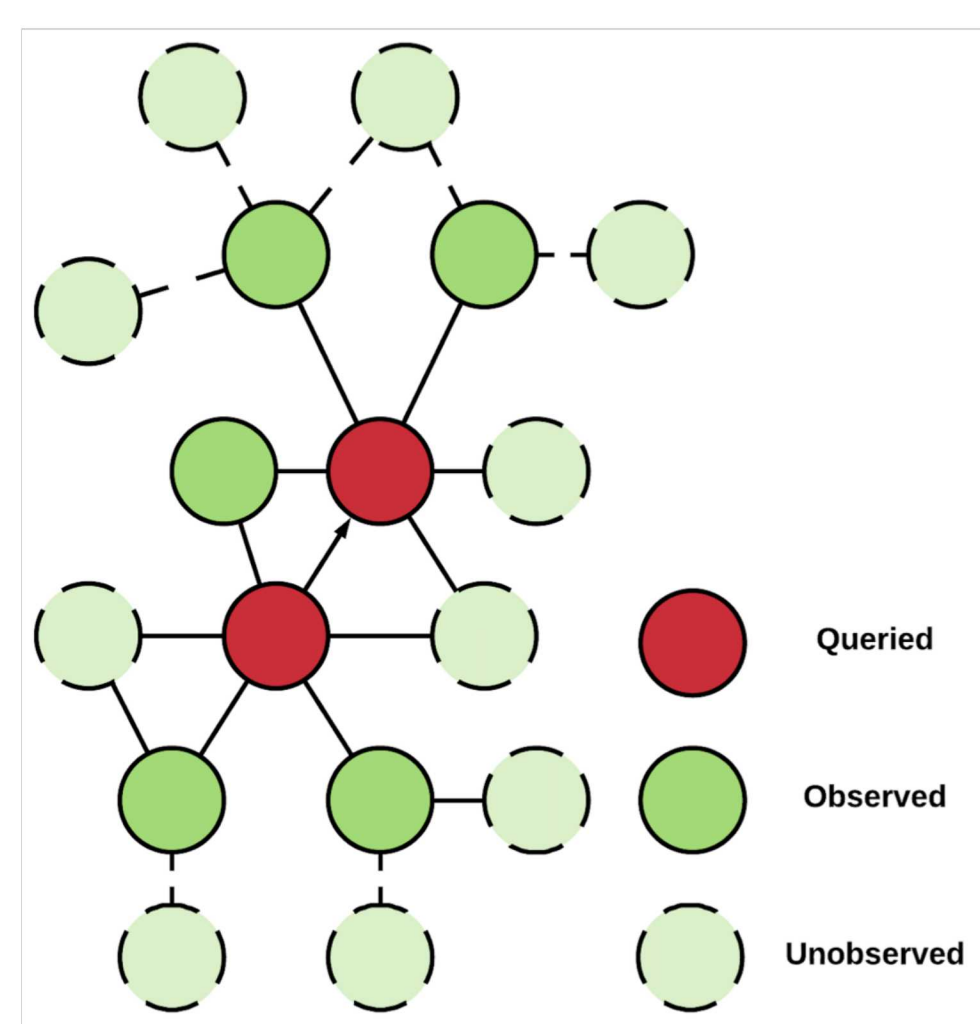


Figure: Unreliable Query Response

## MCS

MCS consists of two steps:

- **RNDSample:** Sample the 'cheaper' layers.
- **MABSsample:** Sample the 'layer of interest' using information from RNDSample.

**RNDSample:**

- Each layer is allocated some fraction of the budget.
- Random walk (with jump) on layers with the allocated budget.

**MABSsample:**

- Operates on the 'layer of interest'.
- Consists of three multi-armed bandits:
  - **LBandit:** Selects layer that is more likely to have high edge overlap with  $L_0$ .
  - **CBandit:** Selects community in the layer selected by LBandit.
  - **RBandit:** Selects node in the community selected by CBandit.

## MABSsample

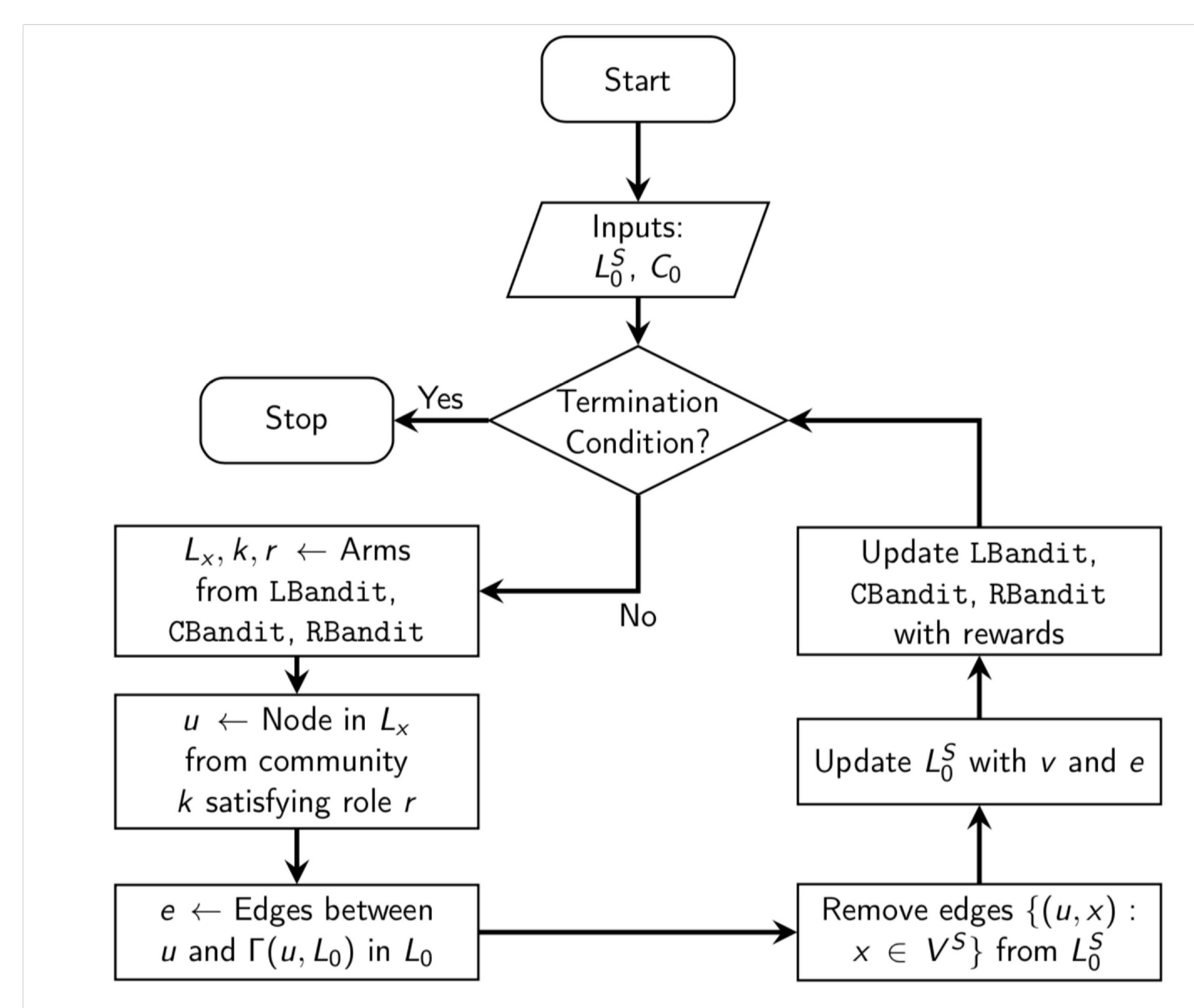


Figure: Flow chart of MABSsample.

## MABSsample Rewards

- **LBandit:** Edge Overlap.
- **CBandit** and **RBandit:** Community update distance. (Edge overlap and Community update distance are calculated only from the observed sample.)

## RQR vs UQR

- **RQR:** If a node is queried, it is never queried in that layer again.
- **UQR:** Estimate the uncertainty of the layers. Queried nodes have a chance of being queried again.

## Performance Comparison

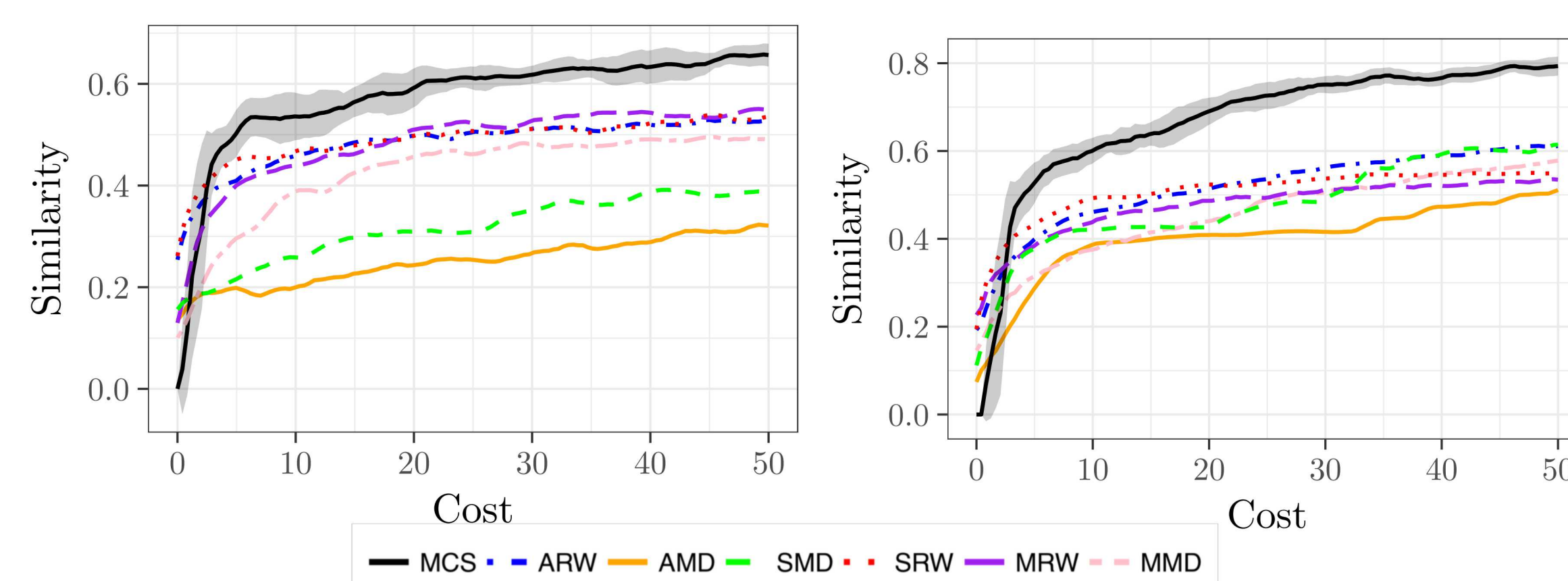


Figure: Results for UQR.

Figure: Results for RQR.

- MCS outperforms all the baselines in finding samples whose community structure is more similar to the original network.

## Regret Analysis

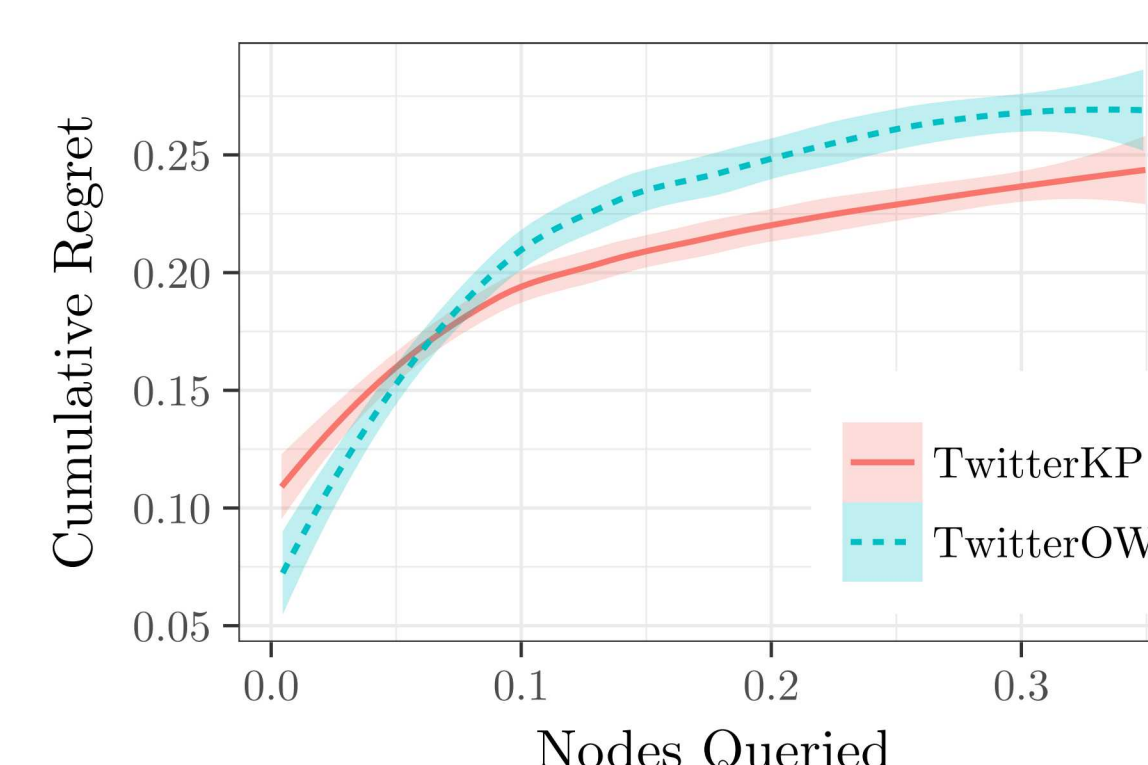


Figure: Cumulative regret of MCS for TwitterKP and TwitterOW networks.

- MCS gets close to the oracle after around 10%-20% of the nodes has been queried.

Laishram and Soundarajan are supported by the U. S. Army Research Office under grant number #W911NF1810047. Wendt's work is supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This research was supported in part through computational resources provided by Syracuse University.