Crawling the Community Structure of Multiplex Networks

Ricky Laishram[1] Jeremy D. Wendt[2] Sucheta Soundarajan[1]

[1]Syracuse University, Syracuse NY, USA

[2]Sandia National Laboratories, Albuquerque NM, USA

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.

Multiplex Networks

- Nodes have multiple types of edges between them[1].
- Edges of the same type can be considered as belonging to the same 'layer'.
- A special type of multilayer network in which nodes can participate in all layers.
- Example: Terrorist Network. Layers: Face-to-face communication, kinship, classmates, mentors.

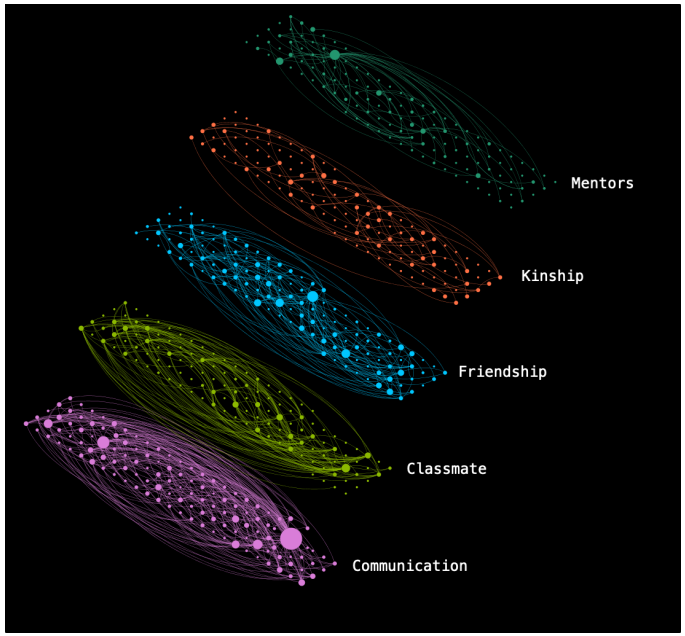


Figure: *NoordinTop* Multiplex Network.

[1]Mucha, Peter J., et al. "Community structure in time-dependent, multiscale, and multiplex networks." science 328.5980 (2010): 876-878.

Data Collection in Multiplex Networks

Before a multiplex network can be analyzed, we need data!

Challenges of data collection in multiplex networks:

1. Different layers have different data collection costs.
2. Data collected from different layers have different reliabilities.

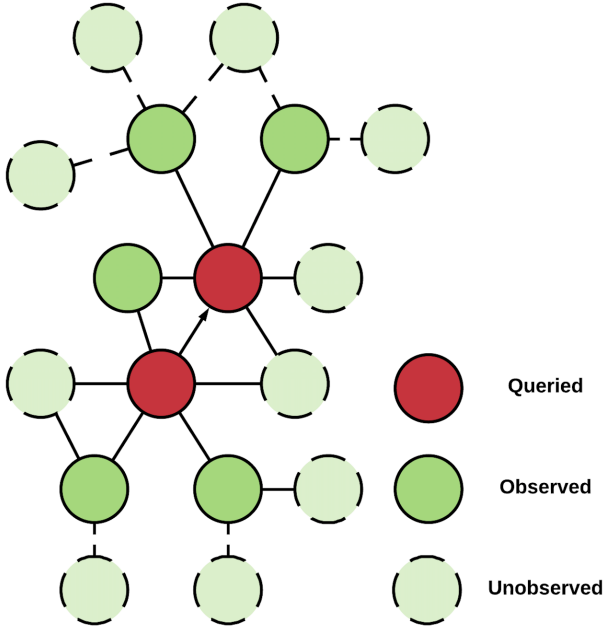
Layer	Cost of query	Reliability of response
Kinship	Low	High
Communication	High	Low

Problem Definition

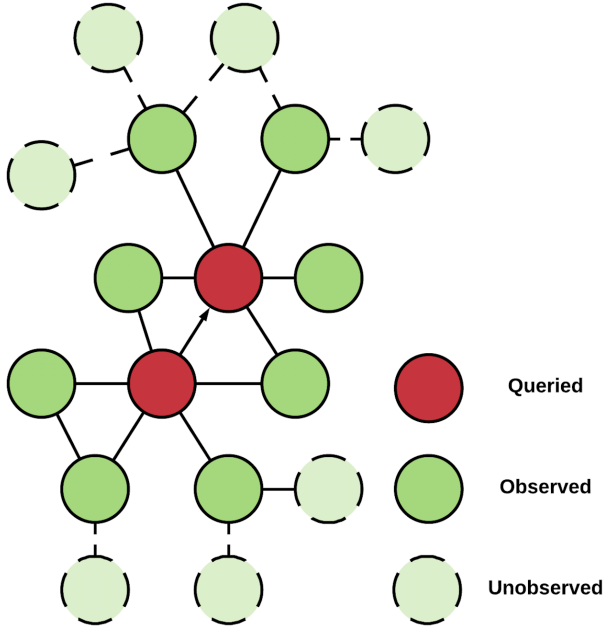
- Let M be a multiplex network, with $L_0, L_1, \ldots$ as the different layers.
- Query costs of the layers: $c_0, c_1, \ldots$.
- *Given the initial set of nodes V', query budget B, and layer of interest L_0, how can we sample M through crawling so that the sample of L_0 found is community representative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

Query Response Models

1. **Reliable Query Response (RQR)**: A query for the neighbors of a node returns all the neighbors.
2. **Unreliable Query Response (UQR)**: A query for the neighbors of a node may not return all the neighbors.
   - Every node has an uncertainty factor that determines the probability of including a neighbor in the response.



(a) Example of UQR



(b) Example of RQR

Challenges

1. The layer of interest is costly to explore.
2. Need to balance trade-off between exploring the layer of interest and the other layers.
3. The true properties of many nodes are not known initially[1].
4. In UQR, a queried node may still have unobserved neighbors.

[1] This is a challenge related with data collection with crawling in general; not just in multiplex networks.

Contributions

1. We are the first to consider the problem of sampling a multiplex network to generate a sample that is representative of the community structure of the layer of interest.
2. We propose *MultiComSample*(`MCS`), a novel sampling algorithm for crawling the community structure of the layer of interest.
3. We perform extensive experimental evaluations, and demonstrate thet `MCS` outperforms all the baseline algorithms.

Methodology

`MCS` consist of two steps:

1. `RNDSample`: Sample the 'cheaper' layers.
2. `MABSample`: Sample the 'layer of interest' using the information from `RNDSample`

RNDSample

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

MABSample: Overview

MABSample has three multi-armed bandits.

1. LBandit: Selects the layer that is more likely to have high *edge overlap* with L_0.
2. CBandit: Selects a community in the layer selected by LBandit.
3. RBandit: Selects a node in the community selected by Cbandit.

Each layer has its own CBandit and RBandit.

MABSample: Details

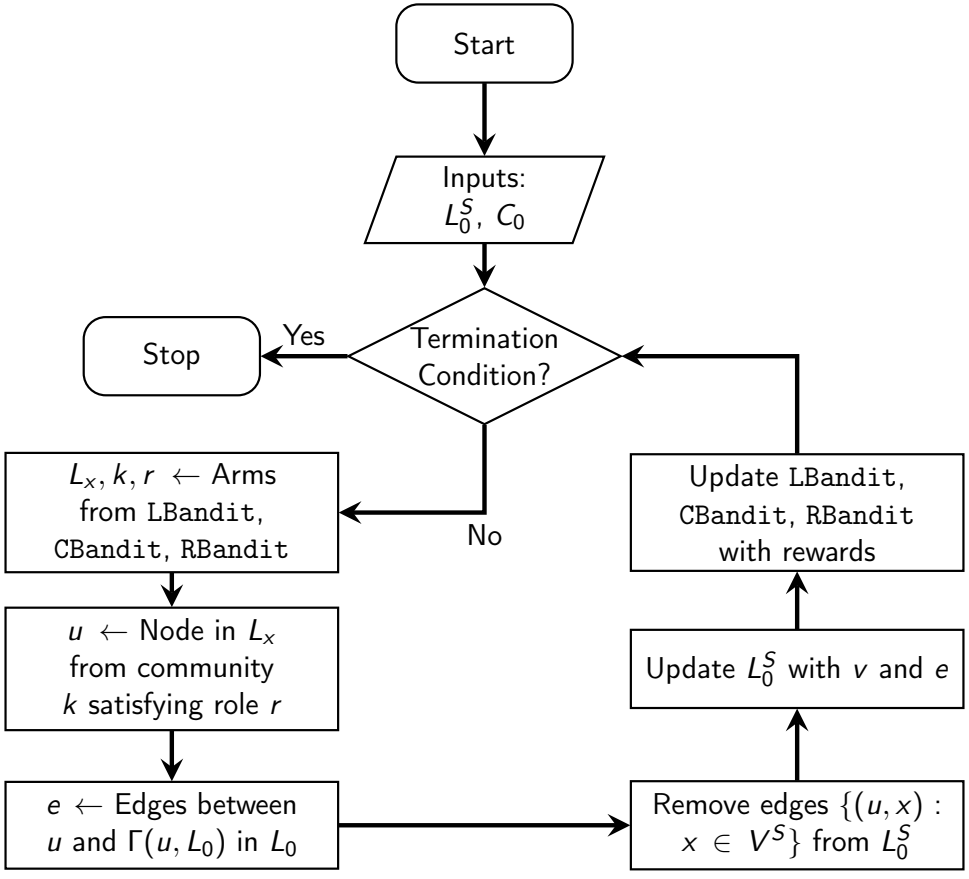


Figure: The flowchart for MABSample.

`MABSample`: Rewards

- Edge Overlap: Measures how similar a layer L_x is to L_0 based on observed edges.
- Community Update Distance: Normalized partition distance before and after querying some nodes.

	Reward
`LBandit`	Edge Overlap
`CBandit`	Community Update Distance
`RBandit`	Community Update Distance

MultiComSample (MCS)

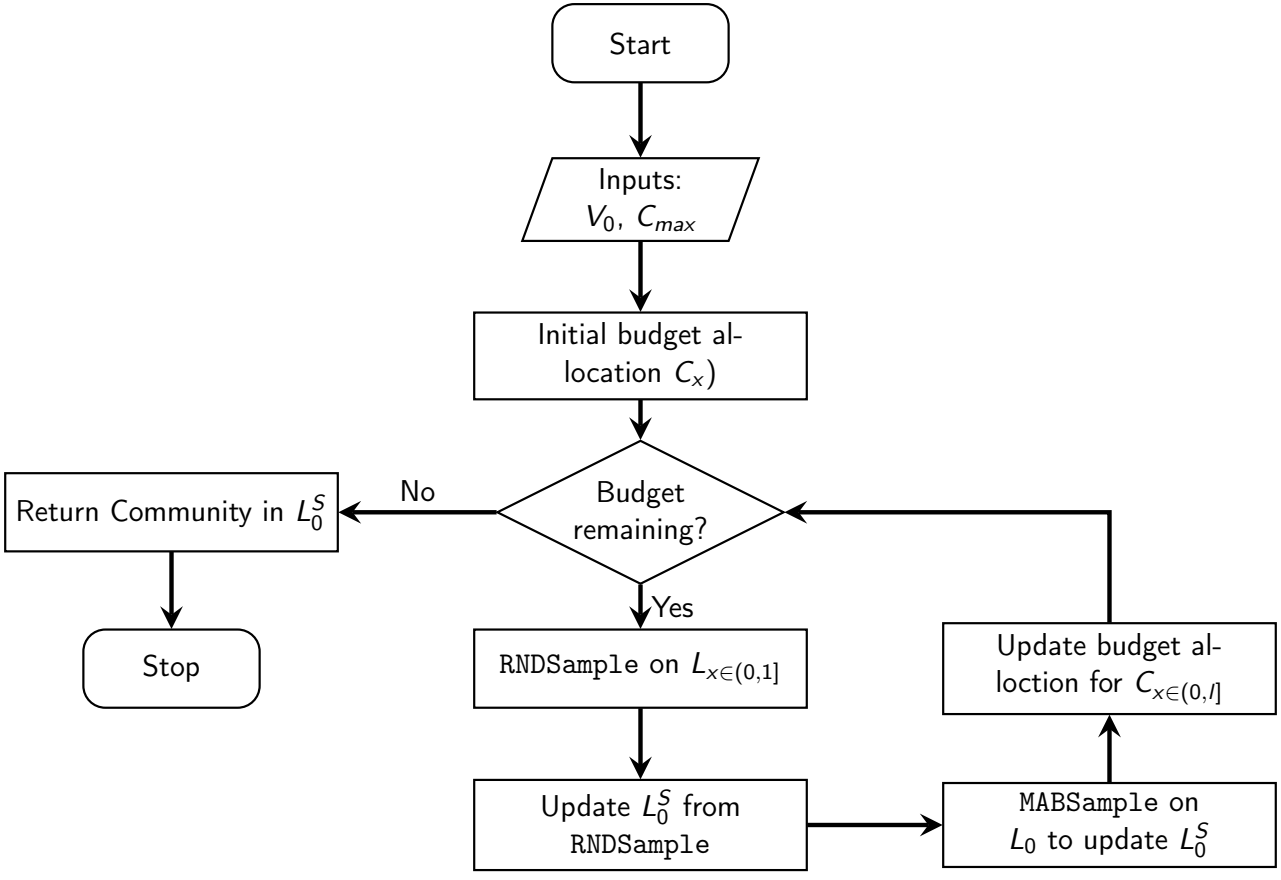
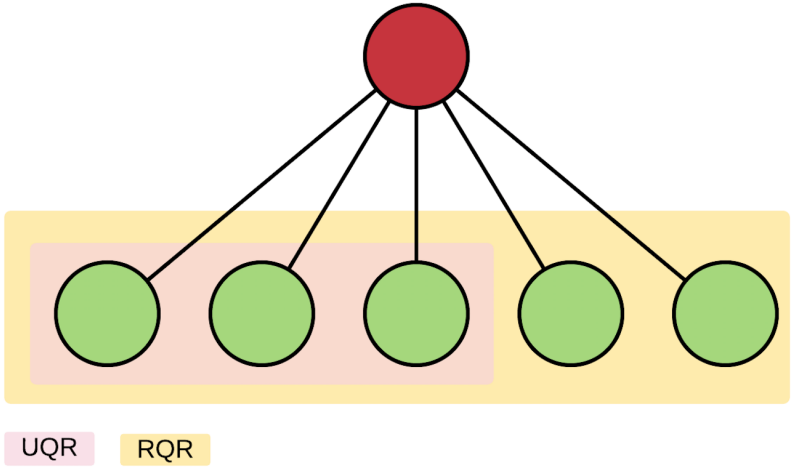


Figure: The flowchart of the MCS algorithm.

RQR vs UQR



- RQR: Once queried, a node is never queried in that layer again.
- UQR:
  - Estimate the uncertainty of the queried nodes.
  - Already queried nodes have some chance of being queried again in that layer.

Datasets

Network	Number of Nodes	Number of Layers	Max Budget
TwitterKP	2420	3	50%
TwitterOW	2182	3	50%
TwitterSC	2116	3	50%
TwitterTR	3036	3	50%
CaHepPhTh	1324	2	50%
NoordinTop	120	5	50%
DBLP	6×10^5	2	5%

Table: Statistics of datasets used for experiments.

Baseline Algorithm

Operates on	**Name**	**Next node to query**
Layer of interest, L_0	SMD	Node with most neighbors in L_0^S.
	SRW	Random node in L_0^S
Aggregate of all layers	AMD	Node with most neighbors in aggregated sample
	ARW	Random node in aggregated sample
Multiplex Network	MMD	Layer with highest edge overlay is selected Node with highest neighbors in selected layer
	MRW	Random node in selected layer Node is queried in both L_0 and selected layer

Appropriate modifications are made to the set of candidate node in the case of UQR.

Performance Comparison

(a) RQR

(b) UQR

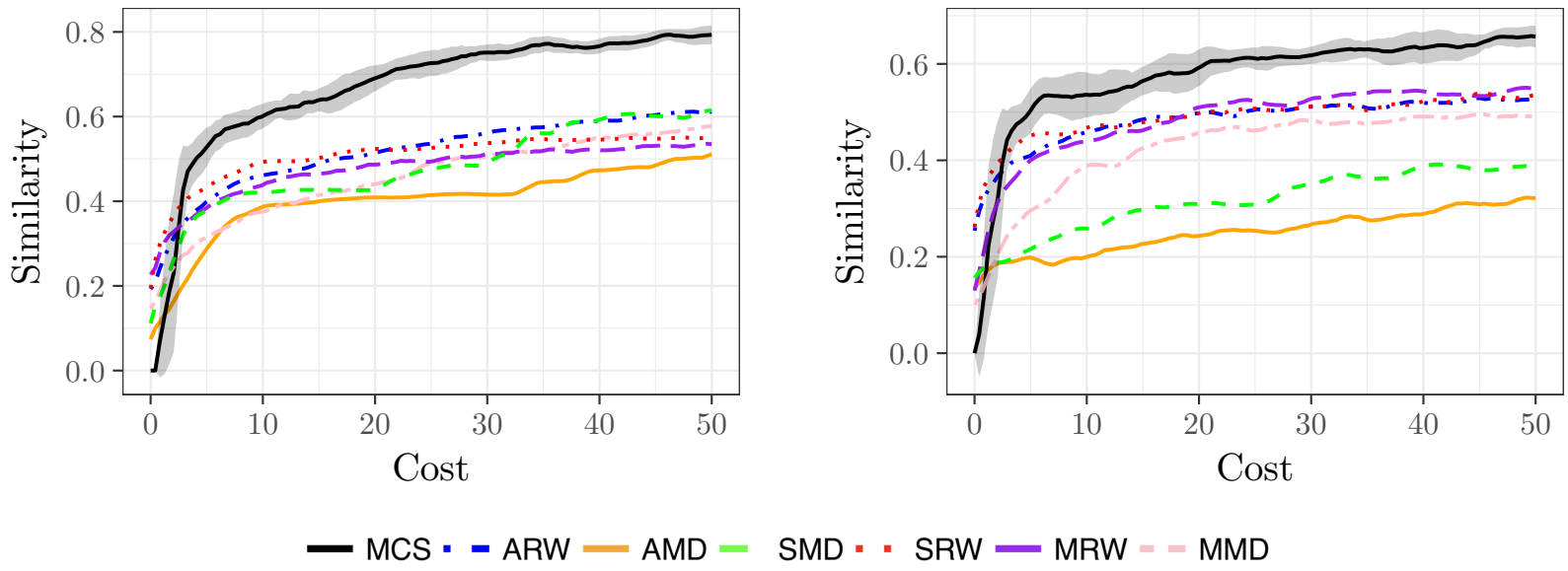


Figure: Comparison between MCS and baselines on TwitterKP dataset.

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

Regret Analysis

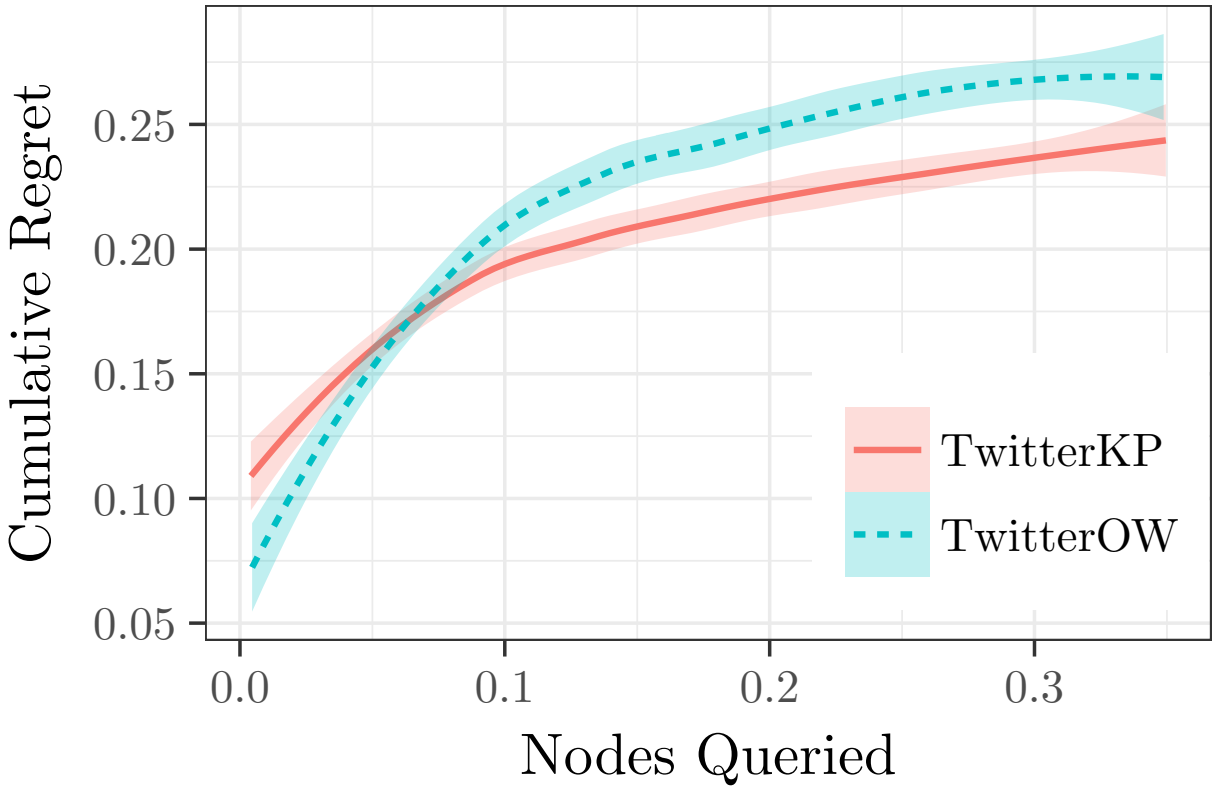


Figure: Cumulative regret for MCS for TwitterKP and TwitterOW.

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

Conclusion

- Addressed the problem of sampling community structure of a layer of interest in multiplex network.
- Proposed a novel algorithm called *MultiComSample* (`MCS`).
- Showed that `MCS` outperforms baseline on multiple real-world networks.

Thank You.

Questions?

rlaishra@syr.edu