Sampling cohesive communities in unbounded networks

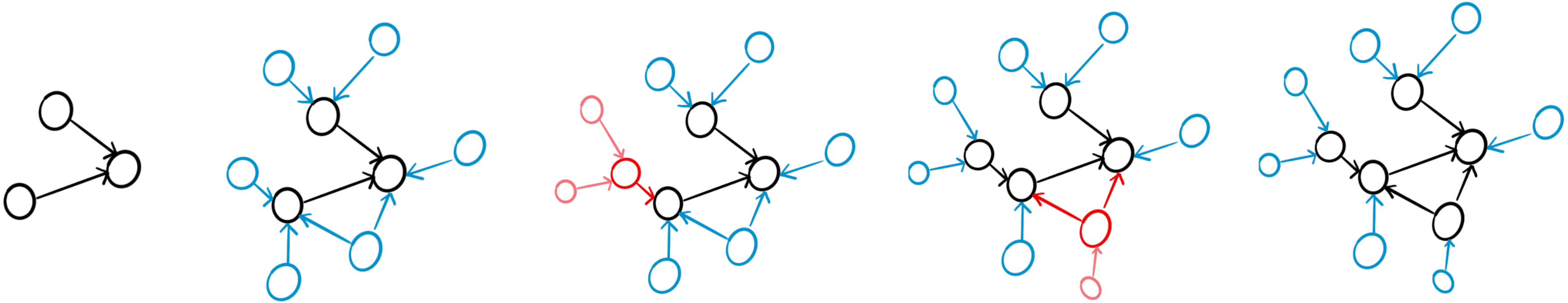
Kshitijaa Jaglan

Advisor: **Sushmita Banerji**
Co-advisor: **Ponnurangam Kumaraguru (PK)**

ETH zürich
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD
Precog IIIT HYDERABAD

Sampling cohesive communities in unbounded networks

Specifically, **Snowball sampling**

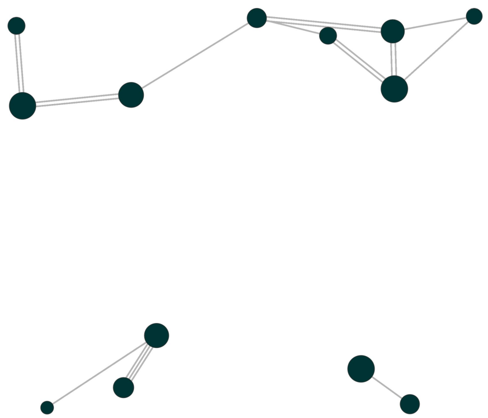


Seed network | *Seed network* + their connections | Sampling iteration 1 | Sampling iteration 2 | Network after sampling 2 nodes

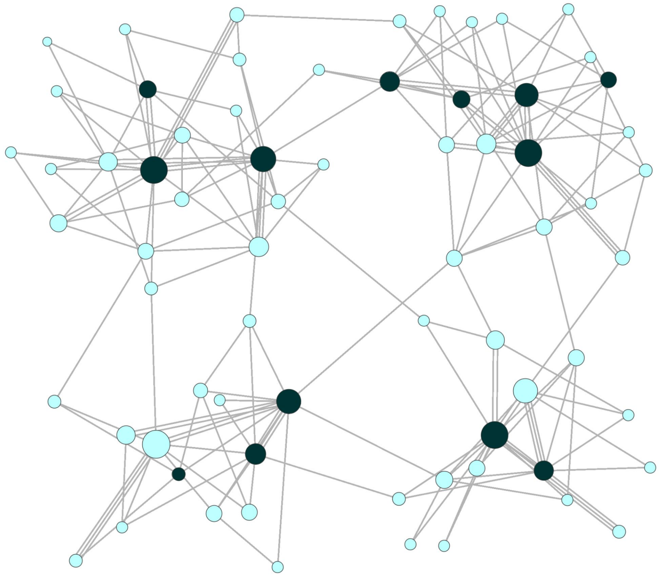
Network growing in each iteration

Problem Statement

Sampling the cohesive community around seed nodes



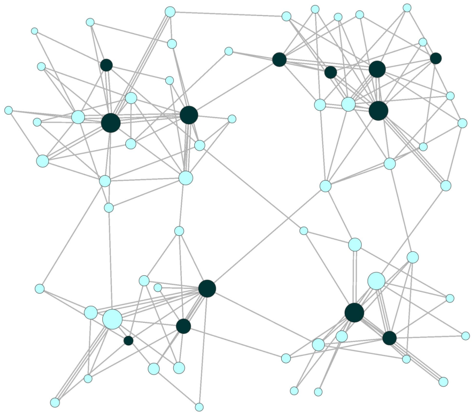
Seed network



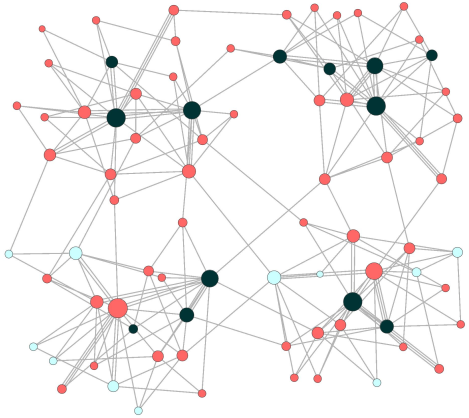
Seed network + their connections

Problem Statement - other methods

Using breadth first search (BFS)



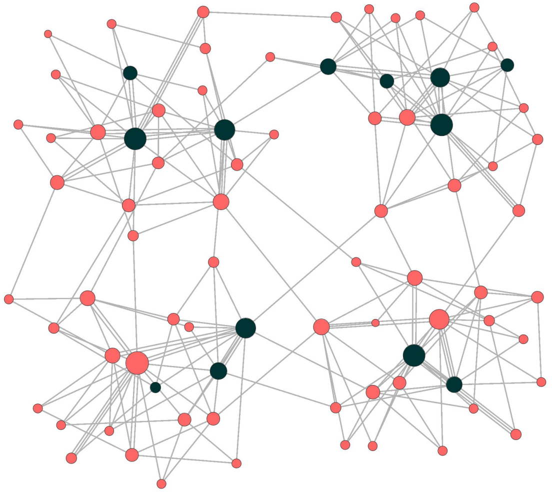
Seeds + connections



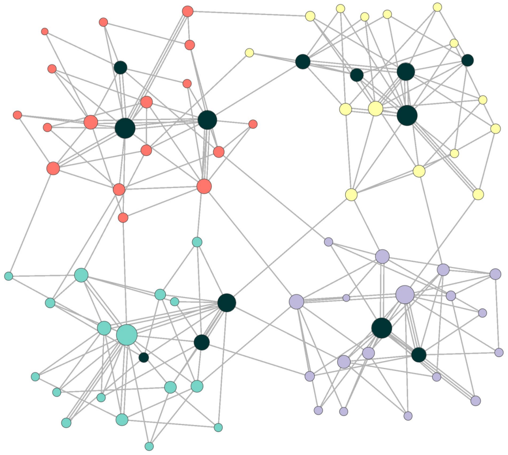
Getting the the first layer

Problem Statement - other methods

Using breadth first search (BFS)



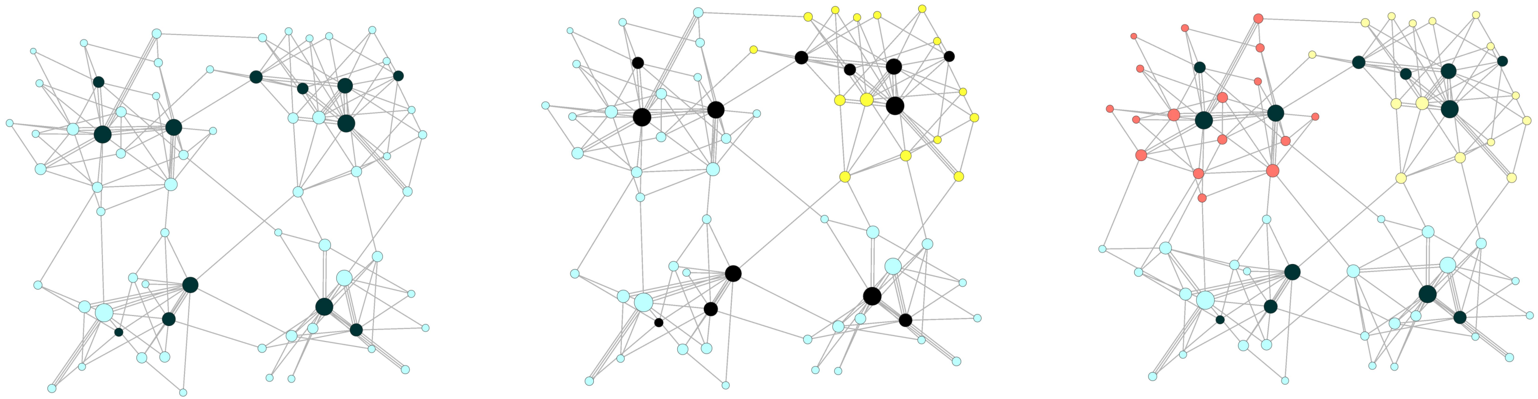
Getting the the second layer (entire network)



Run community detection and get the required cluster

Problem Statement - proposed method

We argue that assigning priorities using **Maximum Adjacency** addresses the problem statement

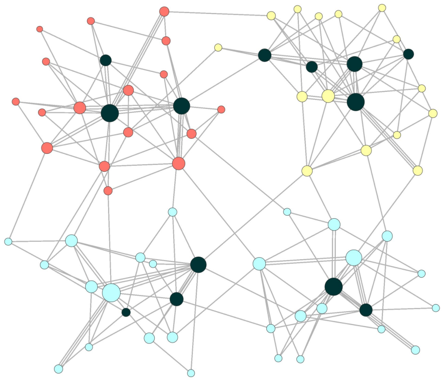


Seeds + connections

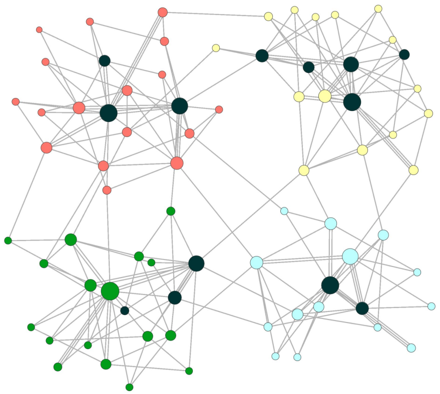
Sampling the first cluster

Sampling the second cluster

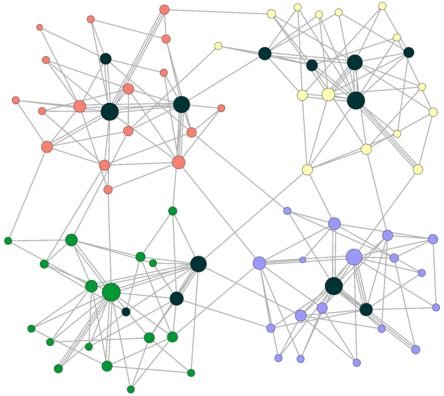
Problem Statement - proposed method



Sampling the second cluster



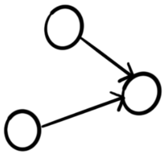
Sampling the third cluster



Sampling the fourth cluster

Problem Statement - How to check the quality?

Using boundary to understand sampling



Seed network

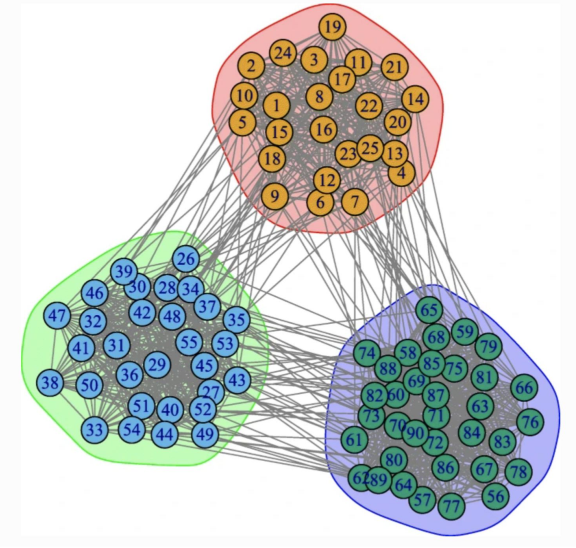
Experiments - Using synthetic networks

Stochastic Block Model

Number of nodes: 90
Group sizes: {25, 30, 35}

Block matrix:

$$\begin{pmatrix} 0.8 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.8 \end{pmatrix}.$$



Source: https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0232-2

Experiments - Using synthetic networks

Group sizes

1. {400, 800, 1200, 1600}
2. {800, 1200, 1600, 2000}
3. {1000} * 8

Selection of seed nodes

1. {20, 50} per block, two blocks at a time
2. {20, 50} per block, two blocks at a time
3. For eight blocks of size 1000 each
   a. [1] * 8
   b. [10] * 8
   c. [20] * {2, 3}

To get block probabilities

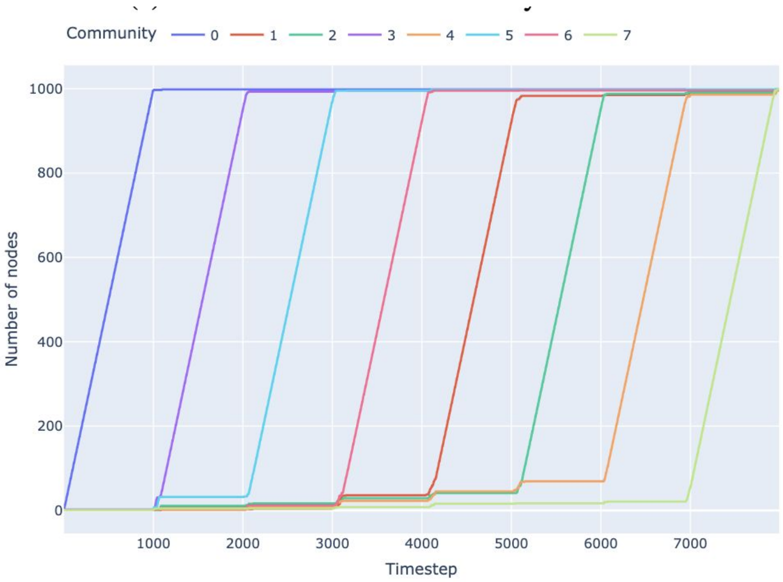
1. **Uniform average degree** (<*k*> = 10)

2. **Ratio of intra-block to inter-block edges** *(r)* : {1/(num_blocks-1), 0.5, 1, 2, 4, 8}

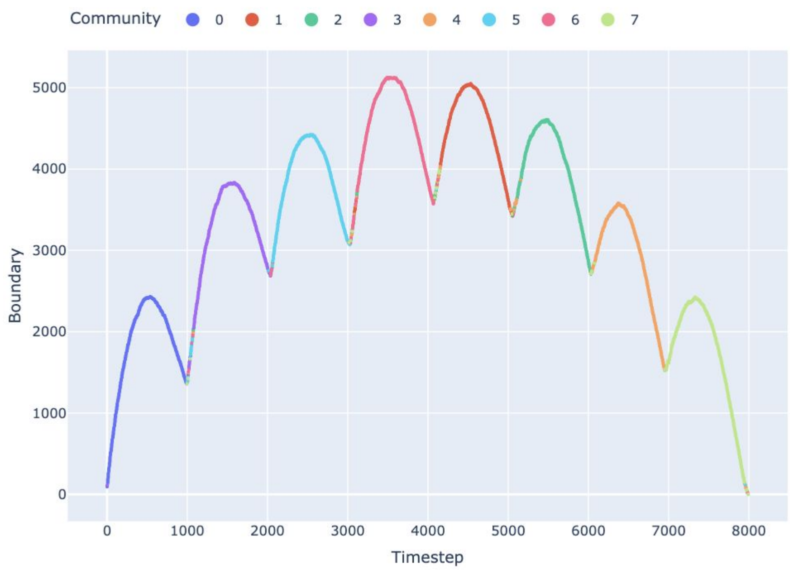
   For every edge within the cluster, how many are going outside it?

Configuration: 8 blocks of size 1000 each (1 seed per block) ; r = 4

Looks very close to an ‘Ideal case’



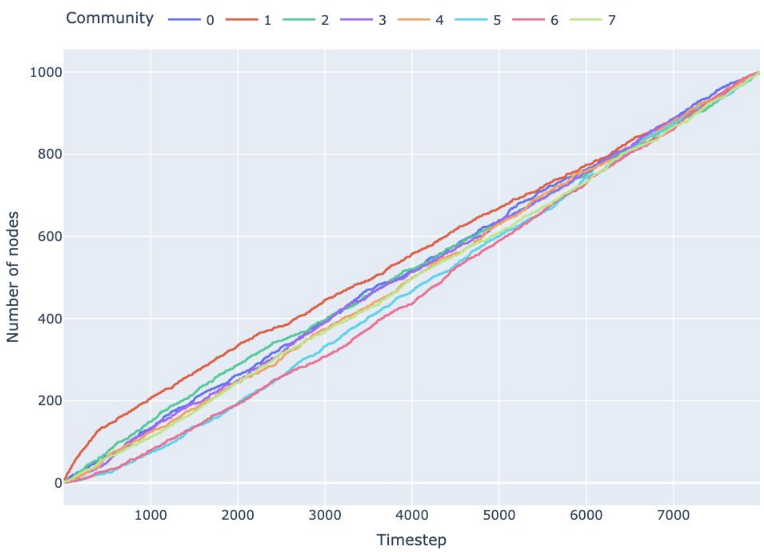
Evolution of community sizes



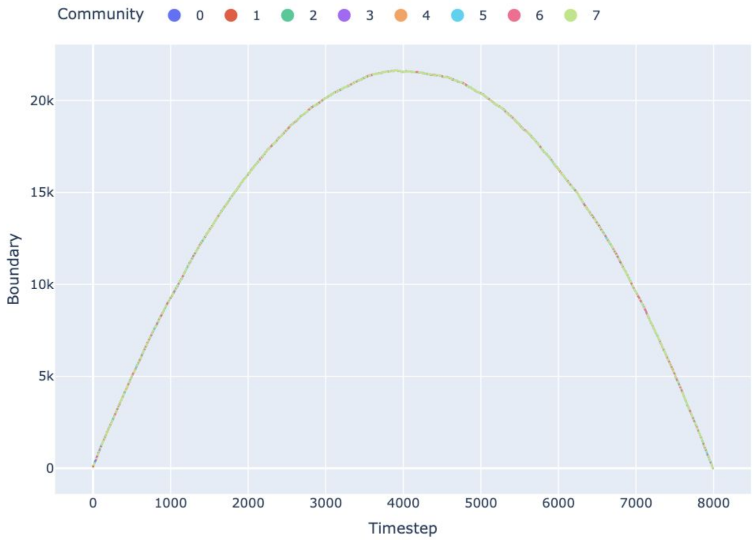
Evolution of boundary

Configuration: 8 blocks of size 1000 each (1 seed per block) ; r = 4

No community distinction when we randomly sample



Evolution of community sizes

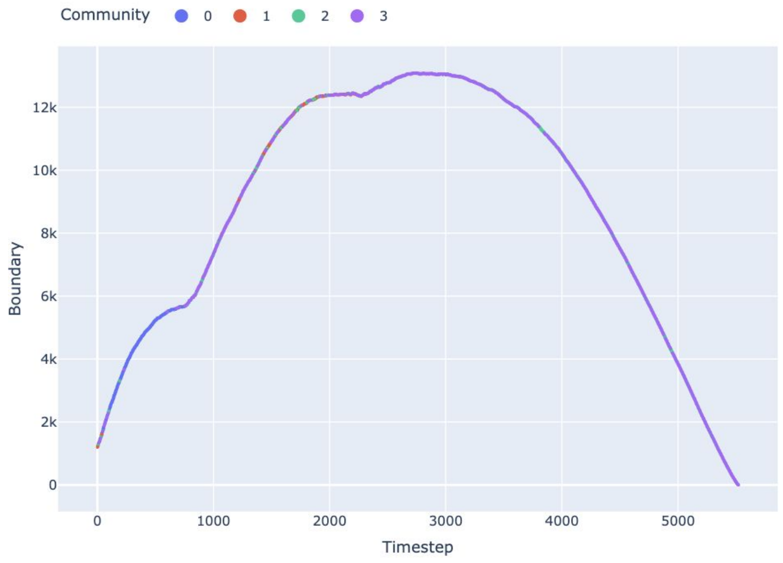


Evolution of boundary

Configuration: Block sizes: {800, 1200, 1600, 2000} (20 seed per block) ; r = 1

Can identify only the first community

Evolution of community sizes



Evolution of boundary

Experiments - Using real-world networks

Twitter networks

To get a group of politically active users (tweeting, or interacting), to study properties like:

1. Spread of influence
2. Structural regularities vs Linguistic regularities

Seed set: DISMISS dataset of Indian Social Media Influencers on Twitter

Arya, A., et al. (2022). DISMISS: Database of Indian Social Media Influencers on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 1201-1207. https://doi.org/10.1609/icwsm.v16i1.19370

Experiments - Using real-world networks

Twitter networks

Preliminary task: How to build a network?

- Follower-followee network
- Retweet network
- Likes network
- …

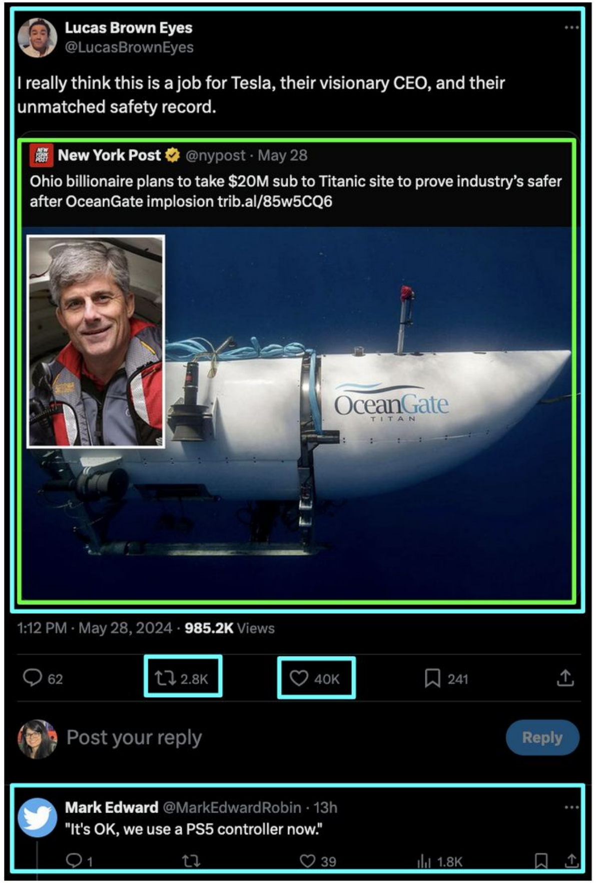
Which network to use?

All?

Building an ‘interaction’ network

Types of interactions:

1. Like
2. Retweet
3. Reply
4. Quote

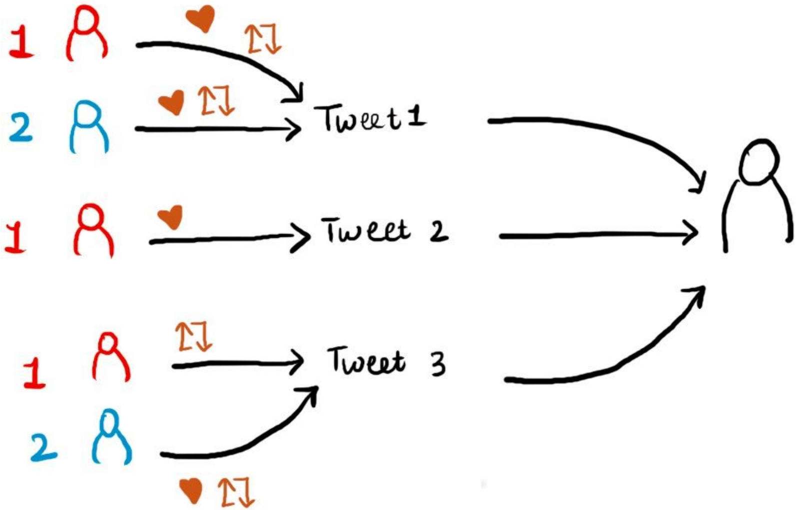


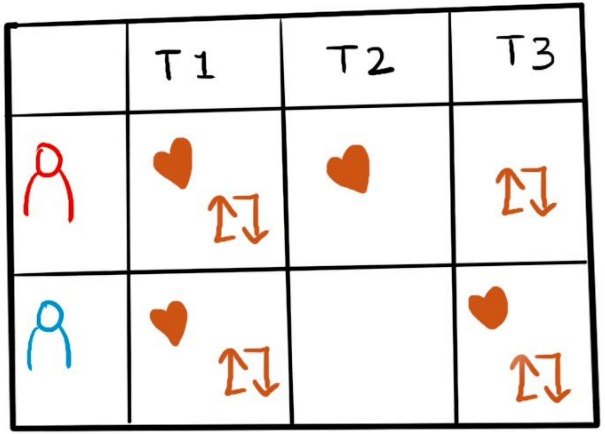
Should we simply combine the *four* interaction networks by assigning them four weights?

We can’t!

Inherent assumption: All interactions are independent of each other

Consider the following situation:





Are red and blue users interacting in the same way?

Not necessarily!

Modelling interactions:

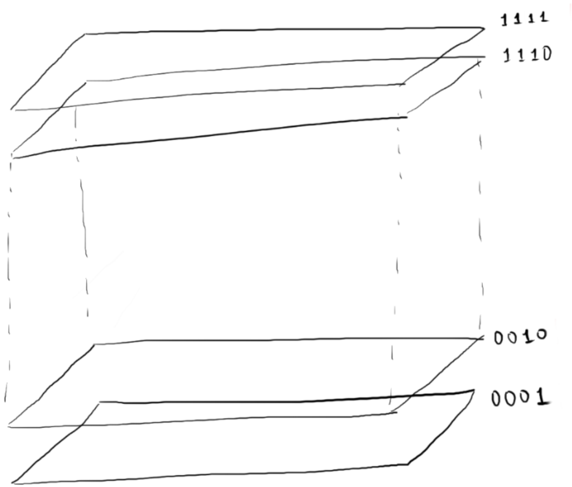
Using a four length vector

[{0,1}, {0,1}, {0,1}, {0,1}] representing [likes, retweet, replies, quotes]

Resulting in $2^4 - 1 = 15$ *‘networks’*

- [0, 0, 1, 1] : Quotes and replies
- [1, 1, 1, 0] : Like, retweet and quotes
- …

Combining multiplex network into monoplex one



$$W_1 \{0001\} + W_2 \{0010\} + \ldots\ldots + W_{15} \{1111\}$$

Combining multiplex network into monoplex one

Want to capture user behaviour through weights

Intuition: More common an interaction type is, lower it should be weighed
(Horvitz-Thompson principle)

Capturing behaviour at **global and local scale**

Introducing types of frequencies (and normalisations)

Combining multiplex network into monoplex one

Global behaviour through **global normalisation**

$$\eta(x) = \frac{n(x)}{N}$$

If I choose an interaction between a user and a tweet, what is the probability of that interaction being *x* ?

Local behaviour through **source and target normalisation**

$$\overleftarrow{\eta}(x) = \frac{1}{|S|} \sum_{i \in S} \frac{n(i,x)}{\overleftarrow{N}(i)}$$

For an average source of interactions, what is the frequency distribution between all the interaction types?

Similarly for target normalisation too..

$$\overrightarrow{\eta}(x) = \frac{1}{|T|} \sum_{j \in T} \frac{n(x,j)}{\overrightarrow{N}(j)}$$

Combining multiplex network into monoplex one

Combining the three ‘frequencies’ to get the weights:

Finding a fourth distribution which is at a minimum distance from the three distributions (global, target, source)

$$\sum_{x \in X} (\eta^*(x) - \eta(x))^2 + (\eta^*(x) - \overleftarrow{\eta}(x))^2 + (\eta^*(x) - \overrightarrow{\eta}(x))^2$$

= mean of the three distributions

Followed by taking the reciprocal to get the weights

Experiments - Using real-world networks

Twitter networks

Preliminary task: How to build a network? - Done!

Next task: Sampling

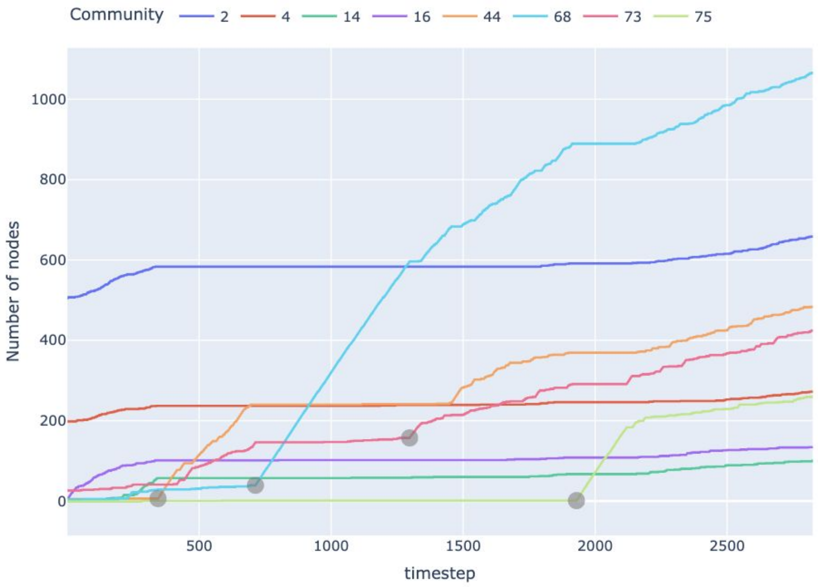
Experiments - Using real-world networks

How is it different from sampling on synthetic network?

- Directed (we consider *incoming edges)*
- Weighted
- Unbounded
  - We do not have ‘ground-truth’ communities
  - Do community detection on *sampled network*

Executing the sampling scheme

Sharp increase in boundary when a community is being sampled



Evolution of community sizes

Evolution of boundary

A few other checks..

Sampling scheme		CC_{local}	CC_{global}	$\langle L \rangle$	$\langle k \rangle$
Priority	Distinct	0.2566	**0.4239**	5.34	12.97
	Nested	0.3747	0.4145	4.62	21.65
	A-F	**0.4004**	0.4035	4.40	**26.49**
Random	RS_DU	0.0646	0.0698	5.25	3.40
	RS_DW	0.1360	0.0608	4.87	5.32
	RS_SU	0.1179	0.0559	4.95	4.81
	RS_SW	0.1237	0.0562	**4.33**	9.11

Priority based sampling gives a more 'cohesive' network than any other random-based scheme

A-F (Audience facing) interactions seem to give the best results, especially in getting higher degree nodes

Surprisingly, RS_SW (Random, staged, weighted) scheme gives better results for average shortest path length

Conclusion

We introduce a sampling mechanism to get cohesive groups around given seed nodes

Also applicable for seeded community detection

Propose a method of integrating different modes of interactions - especially useful in social networks

Providing Twitter datasets containing the sampled groups

Interesting application:

- Getting terrorist networks

Key limitations:

- Baseline comparison (shut down of API)
- Usage of *detected communities* for Twitter network analysis

Thesis reviews

Overall Evaluation : Suitable for MS
As a cultural anthropologist teaching HCI I am interested in computational approaches pushing research boundaries to include the messiness of 'context' , 'User attributes', capturing rich interactions of users and 'expanding attributes' to annotate tweets
I was intrigued and impressed by the ambition of the paper to explore methodologies overcoming the pitfalls of diversity/uncertainty of unboundedness of networks especially in political Twitter communities and render these into 'cohesive' groups' in order to fetch effective sampling techniques.

User attributes

- Context annotations are provided by Twitter from an exhaustive list spanning ~80 categories
- Studies about disparities in context annotations with respect to language etc.
- Using tweet attributes to craft user attributes vs vice versa

Thesis reviews

Overall Evaluation : Exceeds Expectations
Overall, it is a good thesis. Very nice work. I myself learned something from it. I liked the survey of node-based, edge-based and traversal-based sampling methods, and the explanations of different normalization techniques. Also, the writing is crisp and easy to follow. Here are a few remarks:

(1) SBMs should be described at the beginning of Chapter 5.
(2) The plots in Figure 5.1 require some more explanation.
(3) Datasets other than DISMISS could also be considered.

The revised version of thesis will have information about SBMs, and an in-depth description for Figure 5.1

Using DISMISS, we were able to get a group of users who are politically active (tweeting or interacting). By changing the seed set, and applying the sampling scheme in real world, we can even have applications like identification of terrorist networks

Acknowledgements

A big thanks goes to:

- Advisors (Sushmita Banerji, PK)
- Examiners (Nimmi Rangaswamy and Kshitij Gajjar)
- Prof. Ulrik Brandes from ETH Zurich
- Co-authors (Meher, Abhi, Triansh, Nidhi)
- Family
- Friends
- You all, for attending this talk!

Thank you!