Summary (Tight Sampling in Unbounded Networks):

The basic idea is that to take a smaller subnetwork of the entire social networking graph to try and represent the entire graph (which is a conceptually unbounded network meaning that the graph is not bounded in its scope and is not fully understood)

The point is to take a smaller group as to say, “a friend's friend is also my friend”, they approach a variant of “snowball sampling” that is designed to include the strongly connected subgraphs.

Twitter is a valuable source of information on social science however, getting the entire network can't be given due to the large data size.

Snowball sampling is like the topology-based method, the idea being explored via the nodes that are adjacent.

The paper is a novel snowball-type sampling designed to prioritize the cohesive subgroups, this works like community detection, but the main difference is that only partial part of the graph is given.

The other clustering methods don’t work as well because they depend on page rank or random walk that need a large portion of the graph to conduct their community detection.

The basic idea is that:

Build the maximum adjacency search that finds the minimum graph cuts (The idea being that you cut into the space with two disjoint endpoints) from the seed nodes.

Add the various attributes needed (like retweets, replies likes and so on), you also prioritize the new profiles to show the maximum levels of engagement with profiles inside.

Sampling Networks

Node-Based Sampling Methods = This method talks about sampling the nodes at random, this will allow to have the properties of nodes, but can't preserve the connectivity.

Traversal-Based Sampling Methods = They add nodes as the traverse the graph

The snowball sampling is a Traversal based method, this uses the BFS search method with but has the issue with a boundary bias as the neighbor connections are not captured.

Snowball sampling is also based on random walk to remove some of the outgoing links, and this is done to avoid repeated sampling.

Tight Sampling, this is the process mentioned in the paper to prevent parts of the graph from expanding into the entire network.

When this sampling is done, the new vertex must be a Neighbour of the already sampled node, this is essentially a back traversal of the edges.

We select the node which is a backward neighbor to the sampled node and out of all the ones present we select the node whose in-neighbor is the highest number of edges already present in the graph using maximum-adjacency search

The aggregation formula used to calculate the weighted relation is that we just add the edge weight

The paper uses 4 types of interactions = likes, retweets, replies, and quotes

Distinct interaction patterns = patterns between tweets and users

Nested Interaction Patterns = frequency of interaction patterns

Audience-Facing Interactions (A-F): distinction between likes and replies (more personal forms) and retweets and quotes (aimed at visibility).

Note: retweets and quotes are as interchangeable as retweets and quotes often just telling the same thing.

Now to scale the importance of liking and quoting they assign a weight so that all the parameters have the same amount of weightage and there is no bias towards them,

Global normalization = Ignoring the users involved and take the entire tuple, and normalize

Source Normalization = Take the average of the engagement and then normalize the tuple.

Target Normalization = The interaction of the users from various demographics.

The paper takes the minimum of all three squared errors and finds the correspondng vector.

Synthetic Data:

The main difficulty is to create a clear community boundary, so they perform an iterative ratio change between the inter and intra nodes and gradually change the performance.

We make the nodes based on the node centrality.

They are distributed into various blocks with each block having various seeds and this allows to show how all the various possibilities that can exist to capture the best cohesive graph.

Sampling:

1.Maximum Adjacency Search (MAS): Selects node with the highest number of edges to the required node.

2. Random Insider and MAS (RI MAS): Random sampled node, has a random neighhour sampled

3. Random Outsider (RO): chooses a random node outside set, with equal probability

4. Random Insider and Random Outsider (RI RO): as the name says random inside and outside the seed nodes respectively.

Evaluation:

The focus was the boundary vs timestep plot, it was noticed that the graph was growing properly and fully in a connected manner and did not exhibit a preference for obtaining one community.

Inferences on the synthetic data:

This said that the starting sample node dint have much effect, the main factor was the degree size and the random selections.

Observations:

The plot between the boundary vs timestep is the most uniform as once a community has been explored and now looking for the next node in the community this duration depends on the cluster ratio value, the larger values of r have better boundaries and the ones with smaller r tend to mix with other communities.

Empirical data:

Now when working with the twitter data, they wish to look at highly networked profiles and focus on the subset of seed users with the working domain as “civil society”, they even went ahead to collect the tweets posted by these users recently to show that the graph can grow dynamically and for a more cohesive network, they ensured that the dataset is not skewed by they employed a ranking approach, focusing on the lower interacted tweets.

Sampling:

The random sampling was done with two key parameters; selection probability (can be uniform or weighted) and selection strategy (either direct, random outsider or staged, random insider, to a random outsider)

Evaluation:

Despite not having the labelled “community data” as it was present in the synthetic data as “blocks” here they applied the Louvain community method and observed that there was a similar graph along the same “boundary vs timestep plot”, occurring when they sampled two communities at the same time,

So, when there is a sudden spike, this means that there is another community that has been found and potentially mixed into.

They even performed statistical metrics like the shortest path, clustering coefficient,

To ensure no bias in the metrics above they sample on smallest common networks for equal size and comparability

Naturally priority-based operation performs better, the main task being all the combinations of the variants outperform any individual random sampling scheme.

Strengths Of the Paper:

1. Taking into account the interactions between the various users normalizing it as the count of the number of likes is not the same as someone adding a quote (Importance scaling), so that normalization along with them will be weighted the same.
2. The effect of taking the new samples/tweets from the users was an added benefit of the paper as it also shows that the paper implementation works strongly on a dynamic based graph.
3. The statistical visualization of the graphs such as the “boundary VS timestep” is a uniform plot and considering all the possibilities of sampling and seed selection and taking their permutations and combinations show that the algorithm still outperforms, and even outperforms a basic random walk as well.

Weakness of the Paper:

1. Maybe while taking only 90% of the least interacted tweets, to show the dynamic change in the graph might be insignificant, maybe take a larger percentage is needed.

Suggestions of the Paper:

1. While dealing with the interactions between the likes and comments, a tweet can have 100 thousand likes but only 5 comments and such equal normalization won't work evenly.
2. Test and see if the seed node don’t contain any influential profiles in the network as a whole and identify what happens (this is in the empirical data)