Towards building a scalable graph stream library

Darpana Desai
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110085@iitgn.ac.in

Haarit Chavda
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110070@iitgn.ac.in

Aeshaa Shah
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110018@iitgn.ac.in

Siddharth Rajandekar
IIT Gandhinagar
Gandhinagar, Gujarat, India
23110310@iitgn.ac.in

1 INTRODUCTION

We will explore a range of innovative streaming algorithms designed for count-based data as well as for streams representing graph edges. This approach enables us to estimate key graph properties without the need to store every edge, opening exciting opportunities for efficient analysis. For example, we can accurately estimate degree distributions using frequency sketches like CM Sketch or Count Sketch, or develop sketches that can be queried to identify maximum density subgraphs and community structures. While theoretical work in graph streaming is extensive, our project will focus on practical, straightforward sketching algorithms, aiming to demonstrate their effectiveness on specific graph queries and drive positive advancements in the field.

2 RELATED WORK

In "On Sampling from Massive Graph Streams", the authors introduce Graph Priority Sampling (GPS), an innovative order-based reservoir sampling method that weights edges to enable accurate subgraph counts (triangles and wedges) with minimal storage and computational overhead. "Graph Stream Algorithms: A Survey" reviews a decade of research on streaming algorithms for massive graphs, summarizing key techniques and insights that have influenced data structures, approximation algorithms, and distributed computing. Finally, "Streaming Graph Partitioning for Large Distributed Graphs" presents lightweight, scalable heuristics for partitioning graphs during data ingestion, significantly reducing communication costs and improving the efficiency of distributed computations such as PageRank.

3 METHODOLOGY

We will be studying various graph streaming algorithms, implementing them on different datasets and visualising their performance through speed and accuracy. In particular, our project will focus on connectivity, k-connectivity, mincut, and sparsification algorithms. We plan to evaluate each method's computational efficiency and robustness across diverse graph structures, providing insights into their practical applications in large-scale data environments.

4 DATASETS

We use three datasets from the SNAP archive:

- **gemsec-Facebook**: The gemsec Facebook dataset contains data about Facebook pages (November 2017). This dataset represents blue verified Facebook page networks of different categories. Nodes represent the pages and edges are mutual likes among them.
- **musae-Twitch**: The musae-twitch dataset is used for node classification and transfer learning. These are Twitch user-user networks of gamers who stream in a certain language. Nodes are the users themselves and the links are mutual friendships between them.
- **feather-deezer-social**: This is a social network of Deezer users which was collected from the public API in March 2020. Nodes are Deezer users from European countries and edges are mutual follower relationships between them.

REFERENCES

- On Sampling from Massive Graph Streams
- Graph Stream Algorithms: A Survey
- Streaming Graph Partitioning for Large Distributed Graphs
- SNAP archive