Graph Summarization

[Summer Research Internship]

Jemish Variya | 201901112
Raj Patel | 201901306

DA-IICT

Guided by : Prof. Minal Bhise

Problem Definition

- Given a graph G, find a compact representation, so called summarized graph, of G, $\overline{G}$.

Problem Definition

- Given a graph G, find a compact representation, so called summarized graph, of G, $\overline{G}$.
- In the most popular group based approach, summarized graph $\overline{G}$ encoded as **supernodes** and **superedges**.

Problem Definition

- Given a graph G, find a compact representation, so called summarized graph, of G, $\overline{G}$.
- In the most popular group based approach, summarized graph $\overline{G}$ encoded as **supernodes** and **superedges**.
- Graph G is **reconstructed**, as per a fixed scheme, from its summarized graph $\overline{G}$ either perfectly **(lossless summarization)** or with some loss in information **(lossy summarization)**.

Current State of The Art...

- While reconstructing the original graph from its summarized graph superedges decoded uniformly among possible edges between two supernodes.

Current State of The Art...

- While reconstructing the original graph from its summarized graph superedges decoded uniformly among possible edges between two supernodes.
- It's obvious that these uniform scheme may lead to worst accuracy scenario in the **highly degree - skewed** graph.

Current State of The Art...

- While reconstructing the original graph from its summarized graph superedges decoded uniformly among possible edges between two supernodes.
- It's obvious that these uniform scheme may lead to worst accuracy scenario in the **highly degree - skewed** graph.
- And mostly real - world graphs consist of highly degree skewed nodes, i.e., linkage graph of social media platform. Thus this uniform reconstruction scheme is not suitable for real - world graphs.

Current State of The Art...

- While reconstructing the original graph from its summarized graph superedges decoded uniformly among possible edges between two supernodes.
- It's obvious that these uniform scheme may lead to worst accuracy scenario in the **highly degree - skewed** graph.
- And mostly real - world graphs consist of highly degree skewed nodes, i.e., linkage graph of social media platform. Thus this uniform reconstruction scheme is not suitable for real - world graphs.
- So the solution is **Degree Preserving Graph Summarization (DPGS)** which considers degrees of nodes while reconstructing the original graph.

Solution - DPGS

- While relaxing a superedge, assign weights of reconstructed edges proportional to multiplication of degrees of endpoint nodes.
- Uses **Minimum Description Length (MDL)** principle to minimize the **cost of summary graph** and **reconstruction error**.
- Uses **Locality Sensitive Hashing (LSH)** to group candidate nodes and it merges nodes greedily within the groups.
- It's lossy summarization method.

How to find candidate pair of nodes to be merged?

- If we naively select two best pair of nodes to be merged in the large dataset, then it takes $O(n^2)$ time.
- But it would slow down the performance. So how to find candidate pair of nodes efficiently?

How to find candidate pair of nodes to be merged?

- If we naively select two best pair of nodes to be merged in the large dataset, then it takes $O(n^2)$ time.
- But it would slow down the performance. So how to find candidate pair of nodes efficiently?
- The solution to this is **Locality Sensitive Hashing (LSH)**. With this LSH time reduces to $O(n)$.

Minimizing The Description Length

- For minimizing the total description (encoding) length, DPGS uses **Minimum Description Length (MDL)** principle.
- So formulating our objective in mathematical terms,

Minimum Description Length (MDL)

$$\text{Minimize } L(G, \overline{G}) = L(\overline{G}) + L(G|\overline{G})$$

$$\text{where, } L(\overline{G}) = \text{Description length of summarized graph.}$$

$$L(G|\overline{G}) = \text{Description length of errors.}$$

DPGS Algorithm

Input: $G = (V, E)$, iteration T
Output: $\overline{G} = (\overline{V}, \overline{E}, \overline{A})$

$\overline{G} \leftarrow G, \overline{V} \leftarrow V, \overline{E} \leftarrow E$
$t \leftarrow 0$
while $t < T$ do
    $t \leftarrow t + 1$
    Update LSH
    Divide supernodes into disjoint groups by LSH
    for each group g do
        MergeGroup(g)
    end for
end while
return $\overline{G}$

Merge Group Algorithm

Input: $g \subset \overline{V}$

$times \leftarrow log_2|g|$
$nskip \leftarrow 0$
while $nskip < times$ and $|g| >= 1$ do
    $pairs \leftarrow$ Sample $log_2|g|$ node pairs from g
    $(u, v) \leftarrow argmax_{(i,j) \in pairs}\ gain(i, j)$
    if $gain(u, v) > 0$ then
        Merge u and v
        $nskip \leftarrow 0$
    else
        $nskip \leftarrow nskip + 1$
    end if
end while

Novel Reconstruction Scheme(CR Scheme)

- So DPGS uses new scheme for reconstruction, **configuration based reconstruction scheme (CR Scheme)**.

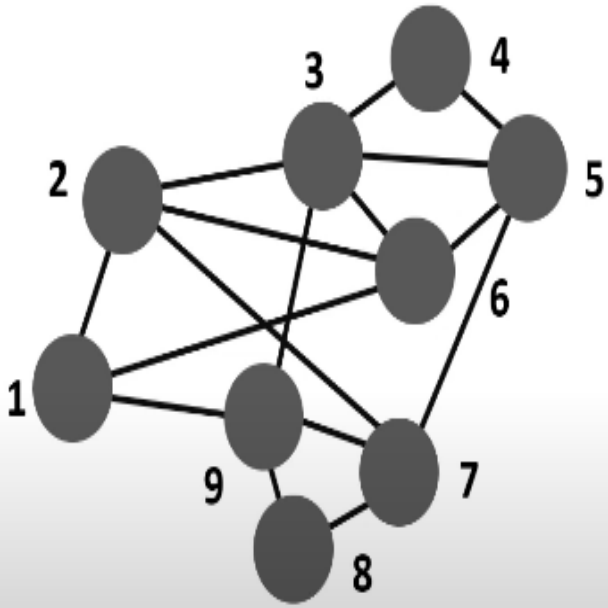
  Let $\overline{A}$ and A' be adjacency matrices of summarized and reconstructed graph respectively,

$$A'(i,j) = \frac{d_i}{D_k}\overline{A}(k,l)\frac{d_j}{D_l}$$

  where, d_i & d_j are degrees of nodes i and j;

  D_k & D_l are degrees of supernodes k and l.

Input Graph

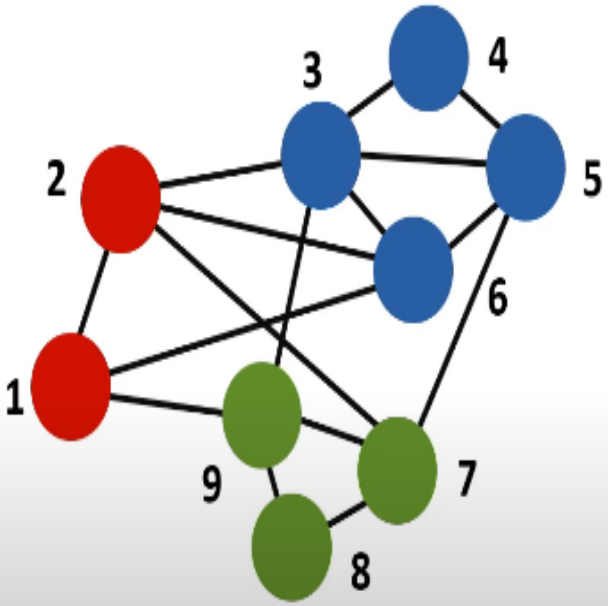


Input Graph

	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	1	0	0	1
2	1	0	1	0	0	1	1	0	0
3	0	1	0	1	1	1	0	0	1
4	0	0	1	0	1	0	0	0	0
5	0	0	1	1	0	1	1	0	0
6	1	1	1	0	1	0	0	0	0
7	0	1	0	0	1	0	0	1	1
8	0	0	0	0	0	0	1	0	1
9	1	0	1	0	0	0	1	1	0

Adjacency Matrix

Input Candidate Grouped Graph

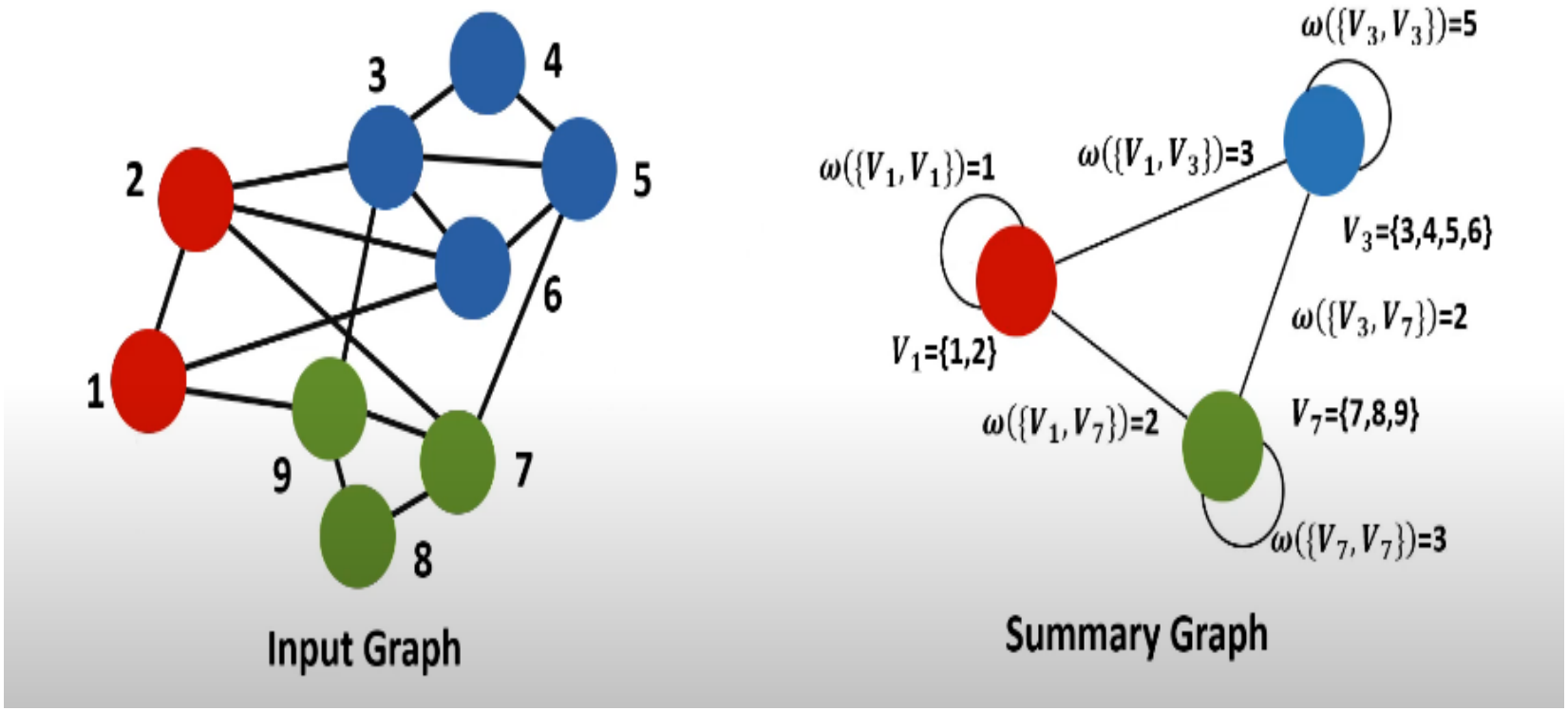


Input Graph

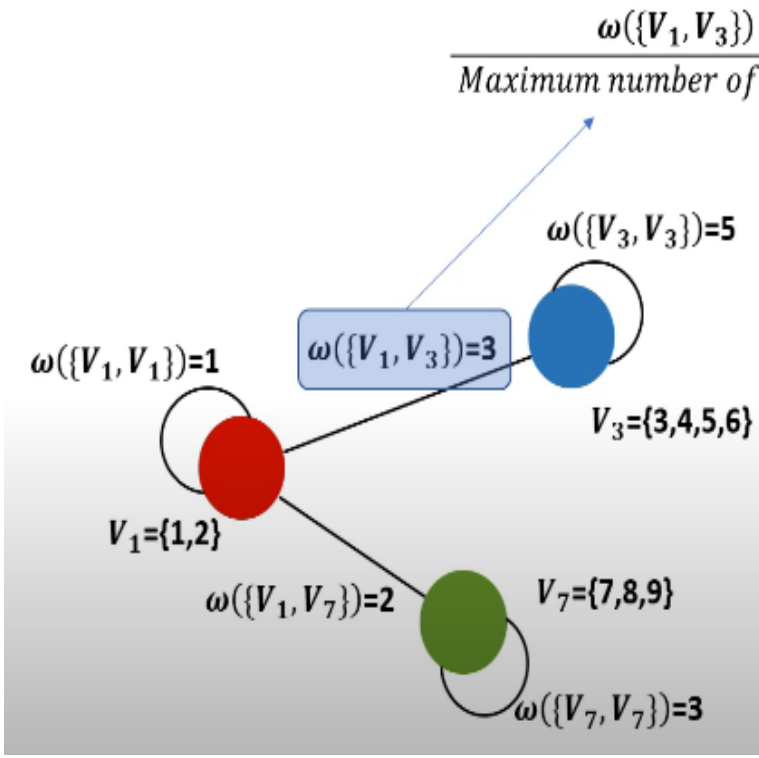
	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	1	0	0	1
2	1	0	1	0	0	1	1	0	0
3	0	1	0	1	1	1	0	0	1
4	0	0	1	0	1	0	0	0	0
5	0	0	1	1	0	1	1	0	0
6	1	1	1	0	1	0	0	0	0
7	0	1	0	0	1	0	0	1	1
8	0	0	0	0	0	0	1	0	1
9	1	0	1	0	0	0	1	1	0

Adjacency Matrix

Visualizing Summarized Graph from Input Graph



Summarized Graph



	1	2	3	4	5	6	7	8	9
1	0	1	3/8	3/8	3/8	3/8	1/3	1/3	1/3
2	1	0	3/8	3/8	3/8	3/8	1/3	1/3	1/3
3	3/8	3/8	0	5/6	5/6	5/6	0	0	0
4	3/8	3/8	5/6	0	5/6	5/6	0	0	0
5	3/8	3/8	5/6	5/6	0	5/6	0	0	0
6	3/8	3/8	5/6	5/6	5/6	0	0	0	0
7	1/3	1/3	0	0	0	0	0	1	1
8	1/3	1/3	0	0	0	0	1	0	1
9	1/3	1/3	0	0	0	0	1	1	0

What if CR Scheme is used instead of uniform scheme for reconstruction?

- As mentioned in the formulation, it incorporates degrees of nodes.

What if CR Scheme is used instead of uniform scheme for reconstruction?

- As mentioned in the formulation, it incorporates degrees of nodes.
- For e.g., let's assign weight of edge $(1, 4)$ using CR scheme. Node 1 and 4 belongs to supernodes *Red* and *Blue* respectively.

$$\begin{aligned} A'(1,4) &= \frac{d(1)}{D(Red)} \overline{A}(Red, Blue) \frac{d(4)}{D(Blue)} \\ &= \frac{3}{7} 3 \frac{2}{14} \\ &= 0.1836 \end{aligned}$$

But previously with uniform scheme, value of $A'(1,4)$ was $\frac{3}{8} = 0.375$ which is unnecessarily high as 4 have less degree.

Correction Set Based Approach

- There are also current state of the art summarization methods in the family of popular group based approach.

Correction Set Based Approach

- There are also current state of the art summarization methods in the family of popular group based approach.
- Out of which **Correction Set - Based Approach** is commonly used.

Correction Set Based Approach

- There are also current state of the art summarization methods in the family of popular group based approach.
- Out of which **Correction Set - Based Approach** is commonly used.
- Here as output, summarization method gives correction set in addition to summarized graph.

Correction Set Based Approach

- There are also current state of the art summarization methods in the family of popular group based approach.
- Out of which **Correction Set - Based Approach** is commonly used.
- Here as output, summarization method gives correction set in addition to summarized graph.
- **Correction set:** Set of edges we need to add or remove to reconstruct the graph perfectly or with less error.

Correction Set Based Approach

- There are also current state of the art summarization methods in the family of popular group based approach.
- Out of which **Correction Set - Based Approach** is commonly used.
- Here as output, summarization method gives correction set in addition to summarized graph.
- **Correction set:** Set of edges we need to add or remove to reconstruct the graph perfectly or with less error.
- The current state of the art correction set based graph summarization algorithm is **SWeG**.

Correction Set Based Approach...

- This method consists of 3 steps:

Correction Set Based Approach...

- This method consists of 3 steps:
  1. Merge the nodes into supernodes.

Correction Set Based Approach...

- This method consists of 3 steps:
  1. Merge the nodes into supernodes.
  2. Encode the edges into superedges and correction set.

Correction Set Based Approach...

- This method consists of 3 steps:
  1. Merge the nodes into supernodes.
  2. Encode the edges into superedges and correction set.
  3. Drop some edges from superedges and correction set to make reconstruction more compact.

Correction Set Based Approach...

- This method consists of 3 steps:
  1. Merge the nodes into supernodes.
  2. Encode the edges into superedges and correction set.
  3. Drop some edges from superedges and correction set to make reconstruction more compact.
- SWeG is faster and elegant than all of its competitors, yields better compression than other methods, and can also run in a distributed setting.

Correction Set Based Approach...

- This method consists of 3 steps:
  1. Merge the nodes into supernodes.
  2. Encode the edges into superedges and correction set.
  3. Drop some edges from superedges and correction set to make reconstruction more compact.
- SWeG is faster and elegant than all of its competitors, yields better compression than other methods, and can also run in a distributed setting.
- Despite the impressive performance of SWeG compared to other algorithms, there are several steps in the algorithm which bottleneck its performance.

Correction Set Based Approach...

- This method consists of 3 steps:
  1. Merge the nodes into supernodes.
  2. Encode the edges into superedges and correction set.
  3. Drop some edges from superedges and correction set to make reconstruction more compact.
- SWeG is faster and elegant than all of its competitors, yields better compression than other methods, and can also run in a distributed setting.
- Despite the impressive performance of SWeG compared to other algorithms, there are several steps in the algorithm which bottleneck its performance.
- Still there are major bottlenecks in SWeG, i.e., finding the candidate pair to be merged and the merging algorithm itself.

Correction Set Based Approach...

So the solution to this problem is **Locality Sensitive Hashing Divide Merge Encode [LDME]**.

Learning

- First of all we have researched the field of graph summarization.
- Then we have focused on impact of degree skewed nodes, which is our main area of research.
- Then we have compared analytical results provided by DPGS method.
- We also have looked upon the vast topic of ML for grouping similar objects, LSH.

References

[1] Efficient graph summarization using weighted lsh at billion-scale.

[2] Incremental lossless graph summarization.

[3] Locality sensitive hashing: How to find similar items in a large set, with precision.

[4] L Durbeck and Peter Athanas.
Dpgs graph summarization preserves community structure.
In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9, 2021.

[5] Houquan Zhou, Shenghua Liu, Kyuhan Lee, Kijung Shin, Huawei Shen, and Xueqi Cheng.
DPGS: Degree-Preserving Graph Summarization, pages 280–288.