APPROVAL EMAIL

PES UNIVERSITY

PES UNIVERSITY

PROJECT TITLE:

DETECTING ANOMALOUS SELF-CITATIONS USING RELATION NETWORK

PES2UG21CS170 : Farhaan Ebadulla
PES2UG21CS175 : Gaurav BV
PES2UG21CS182 : H Manoj
PES2UG21CS243 : Kruthik K

Project ID: 40

Project Guide: Dr. Nazmin Begum

PES UNIVERSITY

ABSTRACT

Anomalous self-citations distort research impact by inflating an author's perceived importance. While some self-citations provide context or build upon prior findings, their disproportionate use threatens the integrity of research metrics.

Detecting these anomalous patterns is crucial for ensuring fair and accurate scholarly evaluation. This study develops a robust methodology to identify and address the issue of anomalous self-citation behavior.

SUGGESTIONS FROM PREVIOUS REVIEW

- Check if the summarization module works on all papers
- Check if the citation contexts can be extracted for different types of citations

PROJECT PIPELINE

PES UNIVERSITY

Dataset Preprocessing: ✅
Construction of Citation Network: ✅
Get Self Citation Loops: ✅
Get TSR For the papers in the Self Citation Loop ✅
Use Bart for summarization of the abstracts to get a paper summary ✅
Use ParsCit to get the citation context ✅
Fine-tune prompt using Gemini: ✅
Fine-tune GPT-4-o1 for Rule-based Classification: ✅

DATA PREPROCESSING

PES UNIVERSITY

Data Collection:

The dataset is taken from a website called Aminer. The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. We are using version 14 with 5 million papers and 36 million citations.

Data Preparation:

Removing the data attributes like artworks that are irreverent to our study, and removing the fields that are not required to our study like
useless_attributes=
['lang','volume','v12_id','v12_authors','indexed_abstract','page_start','page_end','isbn','issn','doc_type','url','issue','doi','keywords','year']

DATA PREPROCESSING

PES UNIVERSITY

Data Input:

After Data Preparation, the data input is in this schema:

- Author
- Source
- Destination
- Contents of source
- Contents of destinations

DATA PREPROCESSING

PES UNIVERSITY

Data Preprocessing:

- We convert the data points into a graph structure using the paper ID and references field from the cleaned dataset. This allows us to construct the citation network, where all edges are undirected. We then select random nodes and their K-hop neighborhoods.

- To find self-citations, we sample and construct a heterogeneous graph with two types of nodes—authors and papers—and two types of edges—cited and published

- We then convert this graph into a homogeneous graph where all nodes are of the same type. This is achieved by converting 'cited' edges into directed edges, similar to those found in natural networks, and converting 'published' edges into undirected or bidirectional edges.

SELF CITATION DETECTION

PES UNIVERSITY

- The concept used to detect self-citations is the detection of triangles in a graph, which is a specific type of cycle, and in this case, it's used to identify potential self-citations through 3-cycles.

- If an author does not have a closed walk involving their papers (i.e., no self-citation through intermediaries), then Aij = 0.

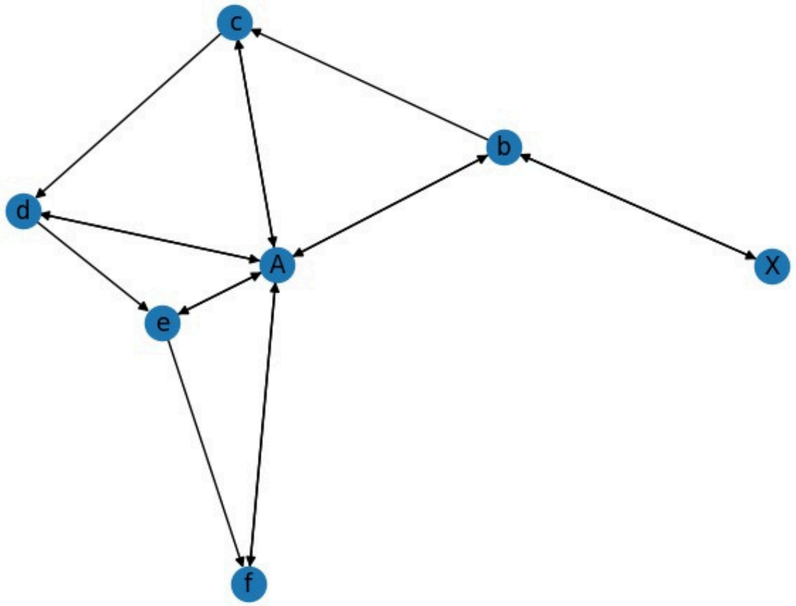
- We compute A^3 for each author, where A is the adjacency matrix that includes the author and their papers. This identifies instances where the author is self-citing. We then focus on these authors, examine their one-hop neighborhoods, and apply our cycle-finding algorithm. This algorithm first considers the one-hop connections in A, then uses A^2 to re-examine their rows. This approach allows us to identify self-citing papers that other cycle-finding algorithms might miss.

SELF CITATION DETECTION

PES UNIVERSITY

```
A = nx.adjacency_matrix(G)
A_squared = A.dot(A)
A_cube = A_squared.dot(A)
self_citing_datas=[]
for i in author:
    cycles = find_cycles_for_node(G, i)
    flag=[]
    for j in cycles:
        j.index(i)
        rr=rotate_list(j,j.index(i))
        flag.append(rr)
    print(flag)
    df1=pd.DataFrame(flag,columns=['author','source','destination'])
    self_citing_datas.append(df1)
self_citing_datas=pd.concat(self_citing_datas).reset_index(drop=True)
```

SELF CITATION DETECTION



self_citing_datas

✓ 0.0s

	author	source	destination
0	A	b	c
1	A	c	d
2	A	e	f
3	A	d	e

TOTAL SELF CITATION RATE

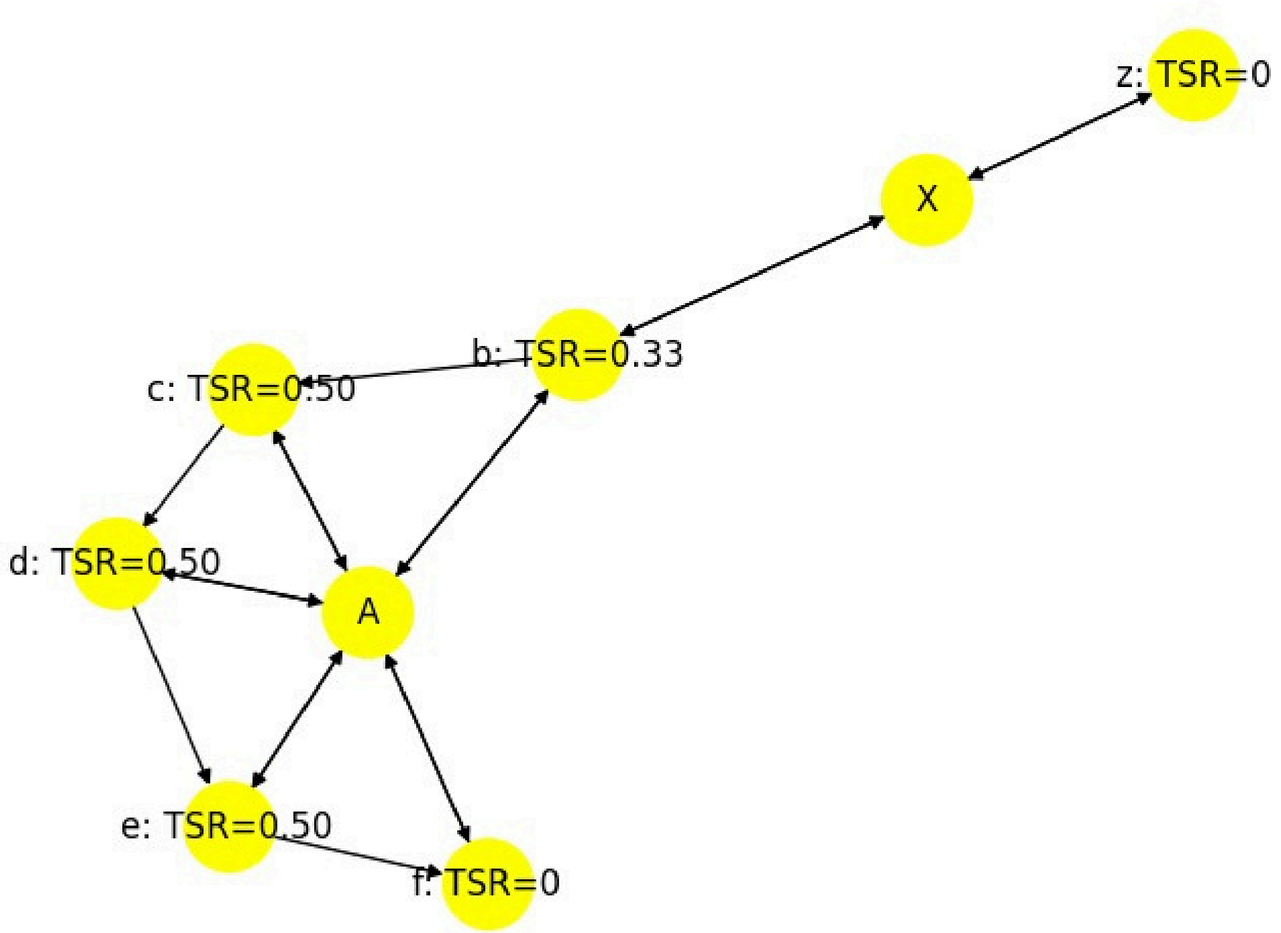
PES UNIVERSITY

For each paper in a Self Citation Loop, the TSR value is calculated

$$TSR = \frac{TotalNumberOfSelfCitations}{TotalNumberOfCitations}$$

The Mean TSR for all the papers is calculated, and only the papers with TSR greater than the mean are retained to further reduce the size of the dataset.

TOTAL SELF CITATION RATE



TOTAL SELF CITATION RATE

PES UNIVERSITY

```
tsr=[]
for i in df['id']:
    if i in list(self_citing_datas['source']):
        condition1=self_citing_datas['source'].isin([i])
        a=self_citing_datas[condition1].shape[0]/(G.degree(i)-2)
        a = "{:.2f}".format(a)
        tsr.append(a)
    else:
        tsr.append(0);
df['TSR']=tsr
```

SUMMARIZATION MODEL

To summarize the paper, we perform a multi-level summarization (two levels) using the BART model, specifically Facebook's bart-large-cnn.

First Level of Summarization: We start by summarizing the paper page by page to ensure that no important information is lost, using the default parameters: num_beams=4 and early_stopping=True. Beam search, a heuristic search algorithm, explores a graph by expanding the most promising nodes within a limited set. In text generation, it tracks multiple candidate sequences (beams) at each step, selecting the top num_beams sequences based on their probabilities. With early_stopping=True, the generation process stops as soon as all beams have generated the end-of-sequence token, speeding up the process and preventing the model from generating unnecessarily long sequences.

SUMMARIZATION MODEL

Second Level of Summarization: We then take the summaries from each page, concatenate them, and feed them to Google's Gemini. This helps us to get a cleaner, better and more comprehensive summary of the paper

SUMMARIZATION MODEL

PES UNIVERSITY

```python
from transformers import BartForConditionalGeneration, BartTokenizer

model = BartForConditionalGeneration.from_pretrained("facebook/bart-large-cnn")
tokenizer = BartTokenizer.from_pretrained("facebook/bart-large-cnn")


import time
def summaryies(text):
    prompt = "Summarize the following paragraph in 500 words and return only the summary, without any explanation: \"" + text + "\""
    model = genai.GenerativeModel(model_name="gemini-1.5-flash")
    response = model.generate_content(prompt)
    time.sleep(13)
    return response.text

def level1(path):
    reader = PdfReader(path)
    summary=""
    for i in reader.pages:
        inputs = tokenizer([i.extract_text().replace("\n","")], max_length=1024, return_tensors="pt", truncation=True).to(device)
        summary_ids = model.generate(inputs["input_ids"], num_beams=32, max_length=200, early_stopping=False)
        summary =summary +"\n"+ tokenizer.decode(summary_ids[0], skip_special_tokens=True)
    return summary

from tqdm.auto import tqdm
tqdm.pandas()
noice['l1_summary']=noice['summary'].progress_apply(summaryies)
noice['final_summary']=noice['l1_summary'].progress_apply(summaryies)
```

CITATION CONTEXT

- To assess the relevance of self-citations, we plan to extract the citation context using the citation parsing tool available in the ParsCit repository.

- The process involves extracting the text surrounding a citation reference, known as the citation context. This context provides insight into how and why the authors referenced their own work, which is essential for determining whether the self-citation is relevant and meaningful or merely an attempt to inflate citation metrics.

CITATION CONTEXT

PES UNIVERSITY

```
# Loop over all .txt files in the input directory
for filename in os.listdir(input_dir):
    if filename.endswith('.txt'):
        input_file = os.path.join(filename)
        output_file = os.path.join(output_dir, f'{os.path.splitext(filename)[0]}.xml')

        # Construct the command to run the Perl script
        #command = f'perl citeExtract.pl -m extract_all out/{input_file}> {output_file}'
        command = f'perl citeExtract.pl -m extract_all out/{input_file}> {output_file}'
        # Run the command using subprocess
        subprocess.run(command, shell=True)

print("Processing completed.")
```

✓ 29m 4.7s

CITATION CONTEXT

PES UNIVERSITY

```python
def extract_citations(xml_path):
    tree = ET.parse(xml_path)
    root = tree.getroot()

    citations = []
    for citation in root.findall('.//citation'):
        # Try to extract the title element
        title = citation.find('title').text if citation.find('title') is not None else 'No Title'
        raw_string = citation.find('rawString').text if citation.find('rawString') is not None else None

        # Use rawString to extract the title if the title is too short or absent
        if is_author_name(title) and raw_string:
            title = extract_title_from_raw_string(raw_string)

        # Collect context from context tags
        contexts = [context.text.strip() for context in citation.findall('.//context')]
        context_text = '\n'.join(contexts)

        citations.append({'title': title, 'citation context': context_text})

    return citations

def save_to_csv(citations, output_path):
    headers = ['title', 'citation context']
    with open(output_path, 'w', newline='', encoding='utf-8') as csvfile:
        writer = csv.DictWriter(csvfile, fieldnames=headers)
        writer.writeheader()
        for citation in citations:
```

CITATION CONTEXT

- To streamline our analysis, we will focus on extracting the citation contexts only from those papers that contain self-citations. These extracted contexts will be 3 lines before the citation and 5 lines after the citation. This structured data will allow us to systematically evaluate the relevance of self-citations across multiple papers.

- However, while extracting the citation context using ParsCit, we have encountered an issue where formulas and mathematical expressions within the citation context are being converted into special characters. This issue requires further processing of the extracted context to ensure that the original content, including formulas, is accurately represented

FINE-TUNING OF PROMPT

1. **Initial Data Flagging**
   - **Selected a sample of 30 data points for manual review.**
   - **Three reviewers manually flagged each citation as “essential” or “non-essential” to establish a baseline.**
2. **Developing Initial Prompt**
   - **Created an initial prompt defining the criteria for “essential” (critical to the paper’s main argument) and “non-essential” (general background acknowledgment) citations.**
   - **This prompt served as a foundation for distinguishing citation types.**
3. **Fine-Tuning with Gemini**
   - **Used Gemini to refine and optimize the initial prompt based on the manually flagged data.**
   - **Developed a set of nuanced rules that accurately capture citation essentiality.**
4. **Finalized Rules for Classification**
   - **The final set of rules now provides a consistent framework to classify citations across larger datasets as “essential” or “non-essential.”**

FINE-TUNING OF PROMPT

PES UNIVERSITY

INITIAL PROMPT:

Identify the contexts where the specified reference is used in the given passage. For each context, determine if the citation is essential or non-essential for comparison purposes.

Essential: Consider the citation essential if it is used to critically evaluate, compare performance, or highlight methodological differences directly related to the main contributions or findings of the passage. Essential comparisons should influence the core analysis, conclusions, or novelty of the work.

Non-Essential: Consider the citation non-essential if it merely acknowledges the existence of similar work without contributing to the critical evaluation or influencing the core analysis or methodology.

FINE-TUNING OF PROMPT

FINAL PROMPT:

Analyze the specified reference in the passage to identify all instances where it is cited. For each instance, determine if the citation is essential or non-essential to the passage's main contributions or findings, based on its role in the comparison between the reference and the source work.

Essential: Consider the citation essential if it is used to critically evaluate, compare performance, or highlight methodological differences directly related to the main contributions or findings of the passage. Essential comparisons should influence the core analysis, conclusions, or novelty of the work or provide a important information about a process being used.

Non-Essential: Consider the citation non-essential if it merely acknowledges the existence of similar work without contributing to the critical evaluation or influencing the core analysis or methodology or help the user in understanding the concepts or provide any good insites for the papers aim.

RULE BASED CLASSIFICATION

PES UNIVERSITY

Few-Shot Dataset Preparation

- Used the manually flagged sample of 30 citations, each labeled as “essential” or “non-essential” based on the finalized rules from Gemini.
- This dataset provided specific examples illustrating the nuances of citation essentiality.

Prompt-Based Fine-Tuning

- Fed the initial prompt, along with the 30 labeled examples, into GPT-4-o1 for fine-tuning.
- The prompt was designed to help GPT-4-o1 recognize essentiality criteria, such as relevance to the main argument vs. general background acknowledgment.

RULE BASED CLASSIFICATION

Iterative Adjustment

- Conducted iterative runs, refining the model's responses based on accuracy in predicting the correct label.
- Adjusted the prompt to improve clarity, focusing on examples where GPT-4-o1 had difficulty discerning essentiality.

Model Validation and Testing

- Tested the fine-tuned GPT-4-o1 model on the 30-item dataset to confirm it could reliably predict essential vs. non-essential labels.
- Verified that the model achieved a high level of accuracy, indicating it had effectively learned the essentiality rules.

Deployment for Larger Dataset Classification

- With fine-tuning complete, GPT-4-o1 was deployed to classify essentiality across the broader citation dataset, maintaining the criteria established in the few-shot training.

RESULTS

Reference [8]: Non-Essential Citation

- Citation Context: Discussed in relation to bounds on Price of Anarchy (PoA) alongside similar studies.
- Paper Summary: Examines routing games, analyzing mixed/pure Nash equilibria and PoA bounds.
- Results: Classified as non-essential; provides background on PoA bounds.
- Explanation: Serves as supportive context without impacting main analysis or conclusions.

RESULTS

Reference [9]: Non-Essential Citation

- Citation Context: Part of a discussion on Fully Mixed Nash Equilibrium Conjecture with other studies.
- Paper Summary: Analyzes routing games, focusing on PoA, equilibria types, and computational challenges.
- Results: Classified as non-essential; acknowledges related work without direct contribution.
- Explanation: Acknowledges similar studies, contextualizing but not impacting core findings.

RESULTS

Reference [26]: Essential Citation

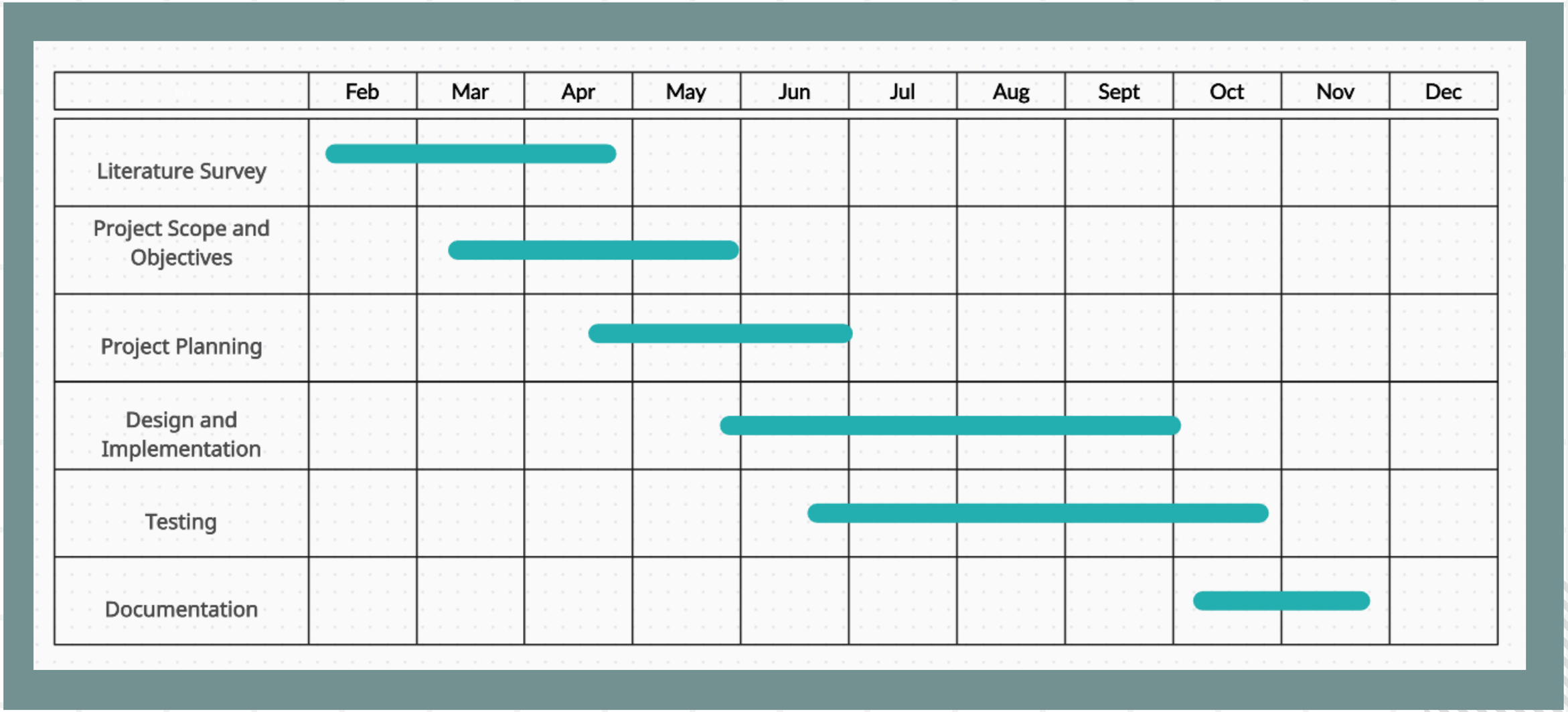
- Citation Context: Used to explain SPT performance in the LMCF game in relation to algorithmic behavior.
- Paper Summary: Examines LMCF in wireless networks, focusing on Nash equilibria, PoA, and DeltaHeur algorithm.
- Results: Classified as essential; supports key analysis of LocalHeur's behavior.
- Explanation: Central to understanding LocalHeur effectiveness, directly supporting main findings.

CONCLUSION

- We developed a method to classify citations as Essential or Non-Essential by evaluating their role relative to a paper's key findings or contributions.
- Our approach interprets citation intent, distinguishing citations essential for supporting arguments from those that simply acknowledge prior work.
- The model identifies citation importance, helping highlight patterns of repeated or unreferenced self-citations that may affect scholarly integrity.
- This classification framework provides a structured way to assess citation relevance, offering insights into citation behaviors across academic networks.

FUTURE WORK

- Integrate temporal filtering to prioritize the most recent, relevant citations, reducing redundancy by avoiding outdated sources.
- Enhance recognition of specific citation formats (e.g., "[Suri et al.]") through customized regex patterns, improving accuracy across various academic styles.


Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
Literature Survey
Project Scope and Objectives
Project Planning
Design and Implementation
Testing
Documentation

PES UNIVERSITY

THANK YOU