# COMP631 Introduction to IR

Homework 2

**1 HITS and PageRank (35 pts)**

Given the figure including a graph as below:

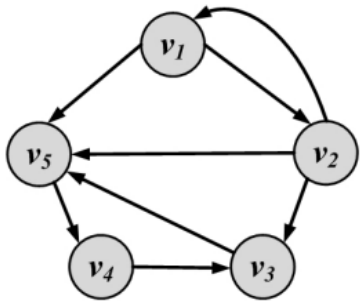



Figure 1: A toy example of a graph

**1) (8 pts) For each graph node in Figure 1, calculate the degree centrality normalized by the degree sum, as well as the rank with respect to the centrality.**

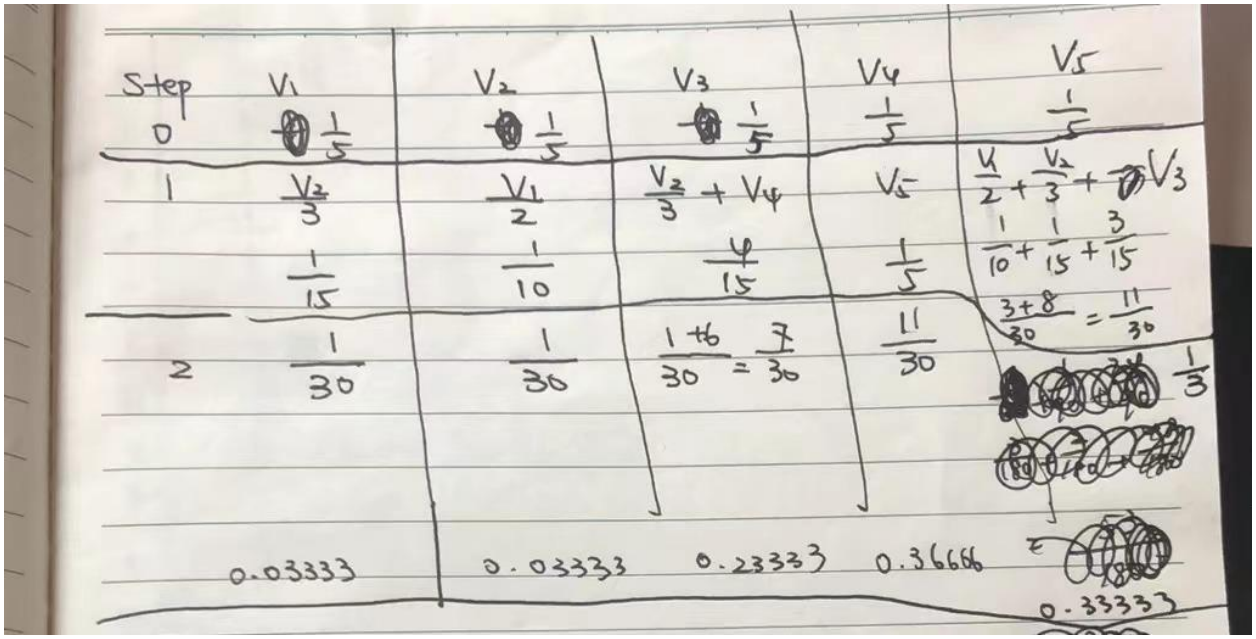
| Node | Indegree | Out Degree | Centrality | Rank |
|---|---|---|---|---|
| v1 | 1 | 2 | 0.25 | 2 |
| v2 | 1 | 3 | 0.375 | 1 |
| v3 | 2 | 1 | 0.125 | 3 |
| v4 | 1 | 1 | 0.125 | 3 |
| v5 | 3 | 1 | 0.125 | 3 |
| sum | | 8 | | |

**2) (8 pts) In the context of world wide web or social media, given a concrete real-world example for entities that can be recognized as “authorities”.**

An example of the “authorities” on the world wide web would be CDC covid guidelines website, many local news websites, university has websites on covid that would link to the the cdc covid

guideline website, thus cdc covid guideline website would have a very huge indegree from other websites which means it would have a very large authority scores.

**3) (12 pts) Calculate the PageRank values for the graph in Figure 1 by applying the iterative algorithm. (Please show 2 iteration steps.)**




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | step | v1 | v2 | v3 | v4 | v5 | |
| 2 | 0 | 0.200000 | 0.200000 | 0.200000 | 0.200000 | 0.200000 | |
| 3 | 1 | 0.066667 | 0.100000 | 0.266667 | 0.200000 | 0.366667 | |
| 4 | 2 | 0.033333 | 0.033333 | 0.233333 | 0.366667 | 0.333333 | |
| 5 | | | | | | | |

**4) (7 pts) Explain the limitations of Katz Centrality, and how PageRank overcome these limitations.**

**ANSWER**

In directed graphs, once a node has a high Katz centrality, it passes all its centrality along all of its out-links

This is less desirable: not everyone known by a well-known person is well-known

To mitigate this problem pagerank divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node such that each connected neighbor gets a fraction of the source node's centrality

**2 Graph Analysis (35 pts)**

Given the graph as below:

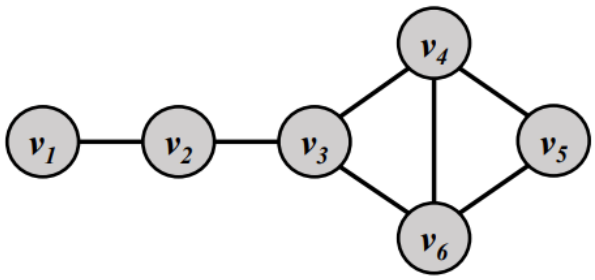


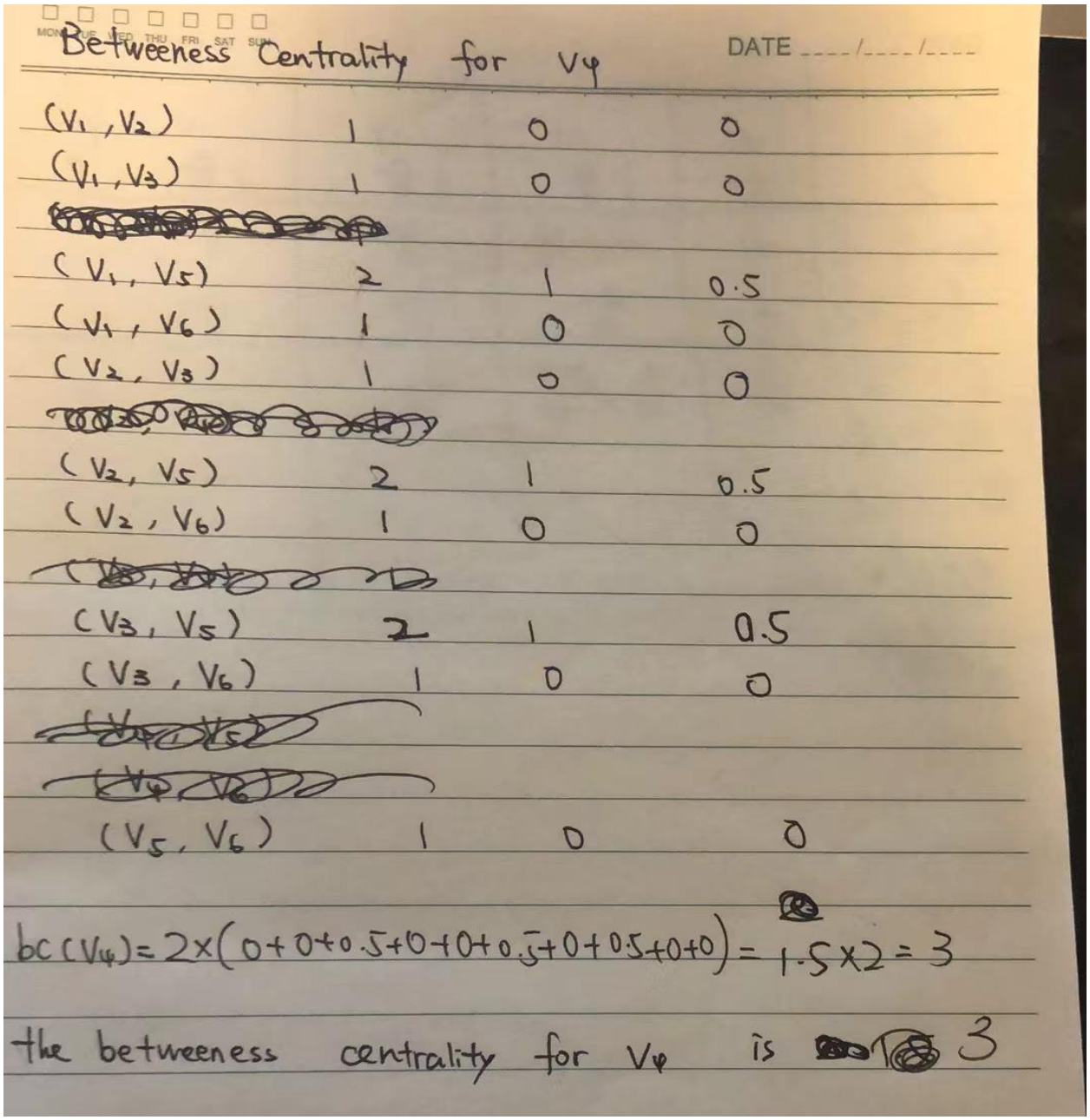

Figure 2: Another toy example of a graph

**1) (7 pts) Calculate the Closeness centrality of node v5.**

| Node | v1 | v2 | v3 | v4 | v5 | v6 | sum | closeness |
|---|---|---|---|---|---|---|---|---|
| v1 | | | | | | | | |
| v2 | | | | | | | | |
| v3 | | | | | | | | |
| v4 | | | | | | | | |
| v5 | 4 | 3 | 2 | 1 | 0 | 1 | 11 | 5/11 |
| v6 | | | | | | | | |

Closeness centrality = 1/(11/5) = 5/11

2) (7 pts) Calculate the Betweenness centrality of node v4.




3) (7 pts) Calculate the Jaccard similarity between v3 and v5.

Jaccard similarity (v3,v5) = |N(v3) n N(v5)|/ |N(v3) u  N(v5)|

=| {v2,v4,v6} n {v4,v6}|/| {v2,v4,v6} u {v4,v6}|
= 2/3

4) (7 pts) Calculate the local clustering coefficient of v4.

cc(v4)= 2/3

**5) (7 pts) Explain how we can use similarity measure or clustering coefficient for friend recommendation in social network.**

Based on the mutual friend list like in facebook various features can be applied similarity measures between existing and new friends. Based on their schools, works  etc similarity measures can be used to make recommendations based on common workplace or place of study.

Further clustering can take people to pages they like.  People with common interests can chat together and make better friend suggestions.

**3 Results Assembly (30 pts)**

**1) (12 pts) For a multi-term query q, explain why we can only consider high-id f query terms for index eliminati**

First, we only consider documents containing terms whose idf exceeds a preset threshold.

Thus, in the postings traversal, we only traverse the postings for terms with high idf. This has a fairly significant benefit: the postings lists of low-idf terms are generally long; with these removed from contention, the set of documents for which we compute cosines is greatly reduced.

low-idf terms are treated as stop words and do not contribute to scoring. For example, on the query catcher in the rye, we only traverse the postings for catcher and rye. The cutoff threshold can of course be adapted in a query-dependent manner.

Second, we can only consider documents that contain many of the query terms. This can be accomplished during the postings traversal, we can only compute scores for documents containing all of the query terms. The problem is that by requiring all query terms to be present in a document before considering it for cosine computation, we may end with fewer than K candidate documents in the output.

**2. (18 pts) Consider a more general form of net score, netscore(q, d) = α · g(d) + β · cosine(q, d), (1) where α and β are weights, 0 ≥ α ≥ 1, 0 ≥ β ≥ 1, α + β = 1, explain how the IR system would behave with different relative weighting.**

A small **α** means less smoothing and more emphasis on relative term weighting. When **α** approaches zero, the weight of each term will be dominated by the term that is term-independent, so the scoring formula will be dominated by the coordination level matching, which is simply the count of matched terms.

This means that documents that match more query terms will be ranked higher than those that match fewer terms, implying a conjunctive interpretation of the query terms.