Q1

Understand the working of Hyperlink Induced Topic Search (HITS). Give a pseudocode and illustrate the same over a sample dataset of your choice.

Introduction:

**Hyperlink Induced Topic Search**(HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.  
HITS uses hubs and authorities to define a recursive relationship between webpages. Before understanding the HITS Algorithm, we first need to know about Hubs and Authorities.

1. Given a query to a Search Engine, the set of highly relevant web pages are called **Roots**. They are potential **Authorities**.
2. Pages which are not very relevant but point to pages in the Root are called **Hubs**. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many authorities.

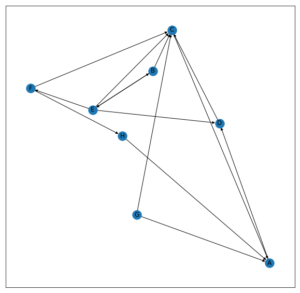
Pseudo Code:

Fix iterations: k

1. Each node is assigned a Hub score = 1 and an Authority score = 1.
2. Repeat k times:
   1. **Hub update:**
   2. **Authority update:**
   3. Normalize the scores by dividing each Hub score by square root of the sum of the squares of all Hub scores, and dividing each Authority score by square root of the sum of the squares of all Authority scores. (optional)

Trace:

Input Dataset



Fixing k = 3,

Initially,

Hub Scores: Authority Scores:

A -> 1 A -> 1

B -> 1 B -> 1

C -> 1 C -> 1

D -> 1 D -> 1

E -> 1 E -> 1

F -> 1 F -> 1

G -> 1 G -> 1

H -> 1 H -> 1

After 1st iteration,

Hub Scores: Authority Scores:

A -> 1 A -> 3

B -> 2 B -> 2

C -> 1 C -> 4

D -> 2 D -> 2

E -> 4 E -> 1

F -> 1 F -> 1

G -> 2 G -> 0

H -> 1 H -> 1

After 2nd iteration,

Hub Scores: Authority Scores:

A -> 2 A -> 4

B -> 5 B -> 6

C -> 3 C -> 7

D -> 6 D -> 5

E -> 9 E -> 2

F -> 1 F -> 4

G -> 7 G -> 0

H -> 3 H -> 1

After 3rd iteration,

Hub Scores: Authority Scores:

A -> 5 A -> 13

B -> 9 B -> 15

C -> 4 C -> 27

D -> 13 D -> 11

E -> 22 E -> 5

F -> 1 F -> 9

G -> 11 G -> 0

H -> 4 H -> 3