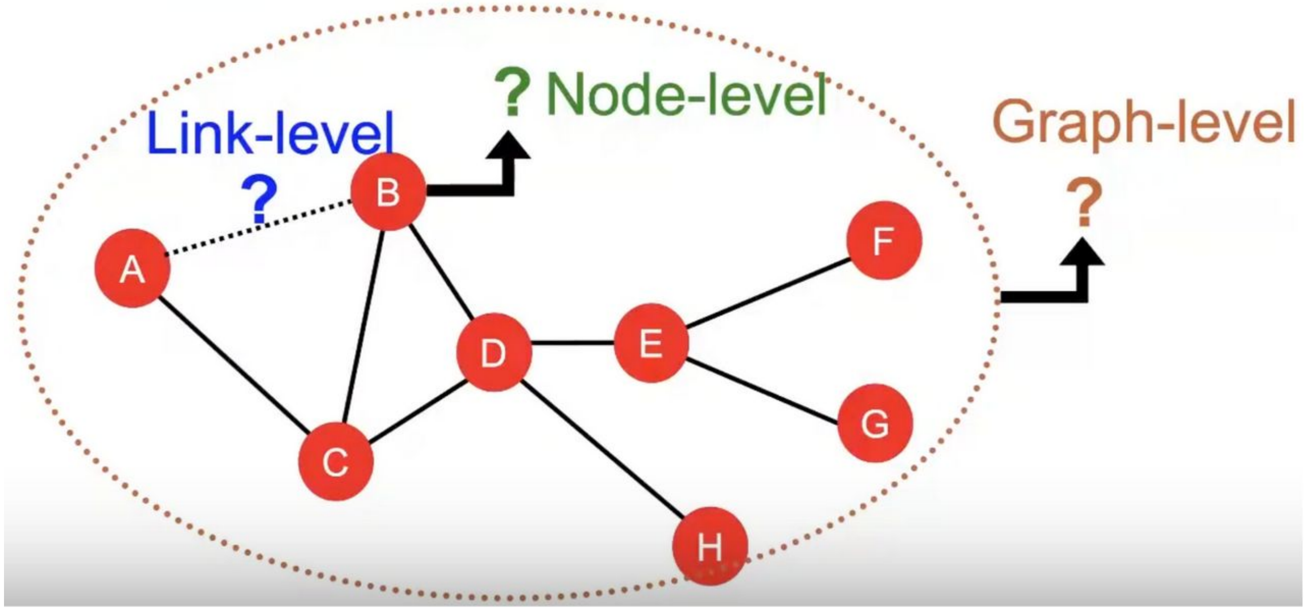
Traditional Feature-based methods: Link

4th April, Graphic Contents Reading Group

Praveen Selvaraj

Recap



Recap: Node-level features

- Importance-based features

- Structure-based features

Recap: Node-level features

- Importance-based features
  - Node degree
  - Centrality measures
    - Eigenvector centrality
    - Betweenness centrality
    - Closeness centrality

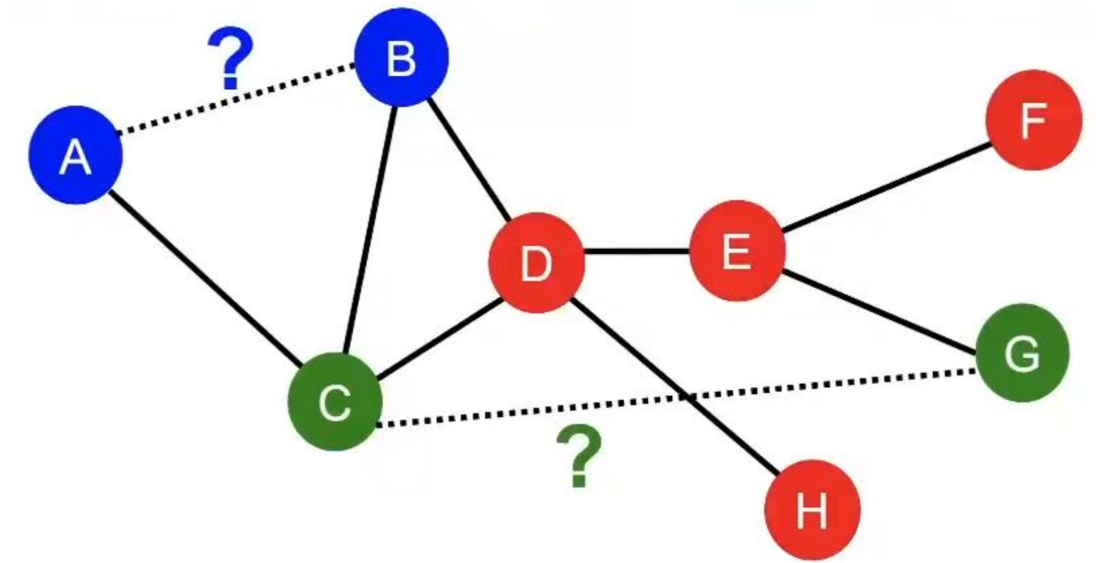
- Structure-based features

Recap: Node-level features

- Importance-based features
    - Node degree
    - Centrality measures
        - Eigenvector centrality
        - Betweenness centrality
        - Closeness centrality

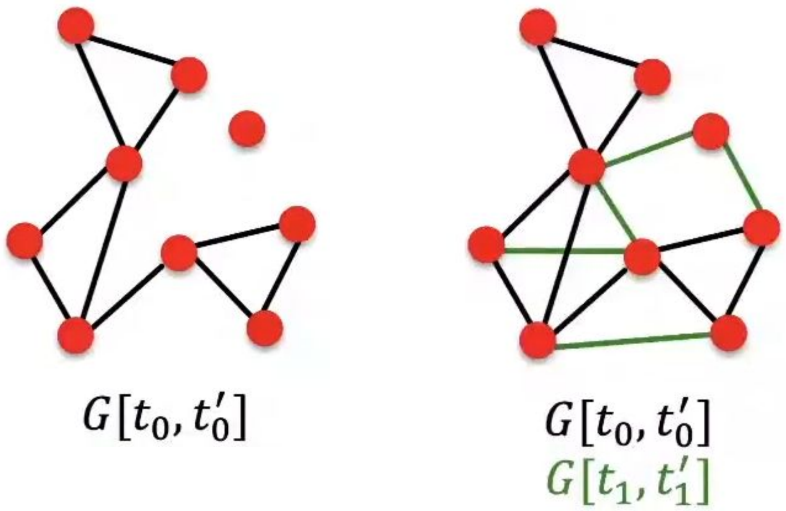
- Structure-based features
    - Node degree
    - Clustering coefficient
    - Graphlet degree vector

Link-level task



Link prediction formulation

- Links missing at random
- Links over time
  Given a graph $G[t_0, t'_0]$, output a ranked list L of edges predicted to appear in $G[t_1, t'_1]$.



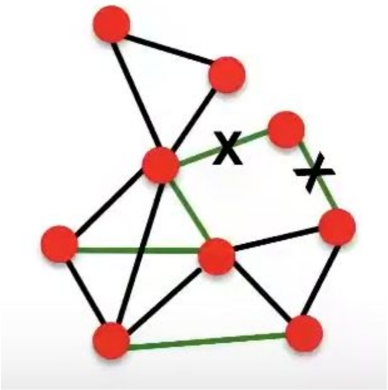
Links over time - algorithm

- For each pair of nodes (x,y), compute a score **c(x,y)**
  - Example: no of common neighbors between node x and y

- Sort the pairs of nodes (x,y) by their score **c(x,y)**

- Top n pairs would be the predicted links

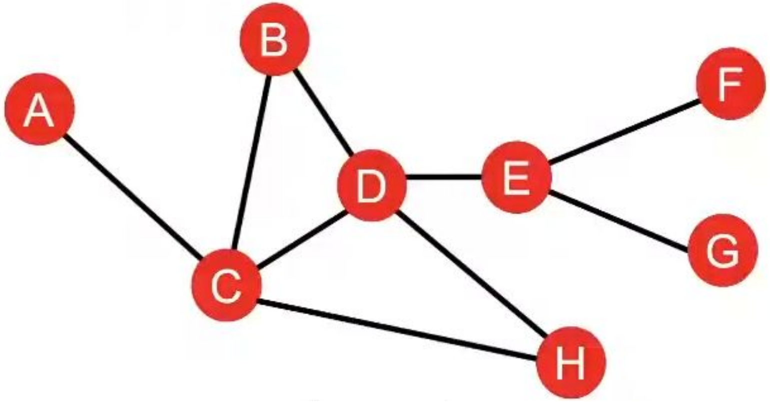
- Check which of these links actually appear in $G[t_1, t'_1]$



Link-level features

- Distance-based features
- Local neighborhood features
- Global neighborhood features

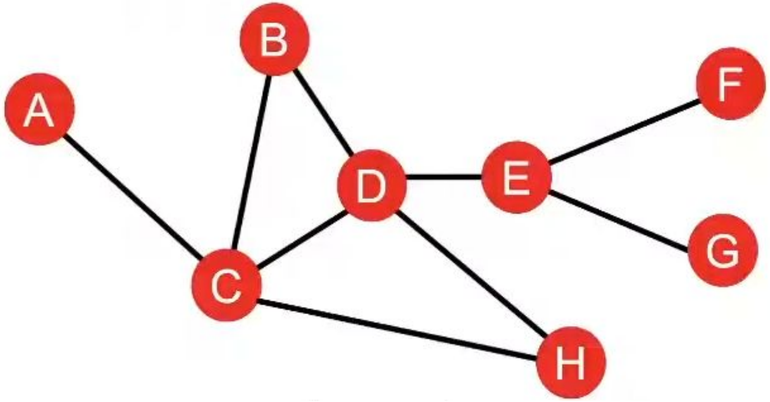
Shortest path distance



$$S_{BH} = S_{BE} = S_{AB} = 2$$

$$S_{BG} = S_{BF} = 3$$

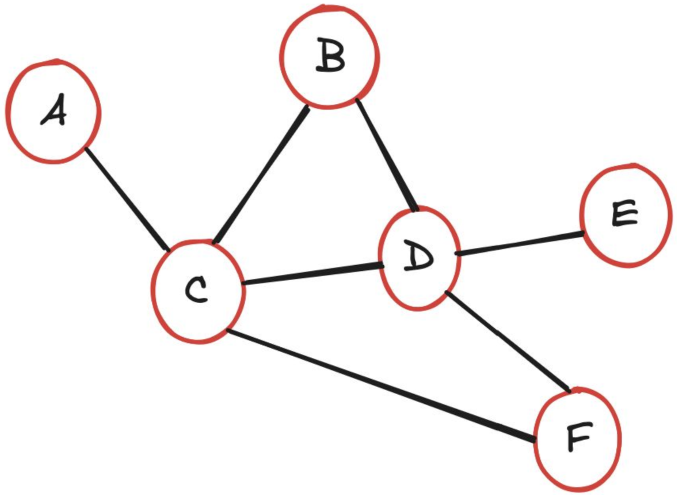
Shortest path distance



$$S_{BH} = S_{BE} = S_{AB} = 2$$
$$S_{BG} = S_{BF} = 3$$

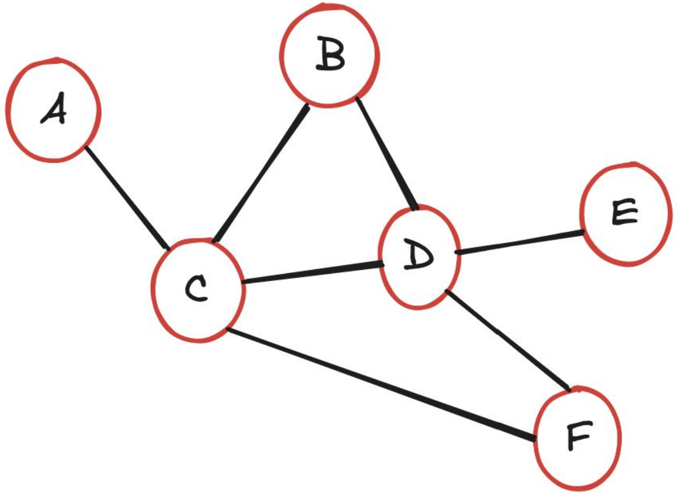
Treats BH and BE,AB as the same, but their ‘connectedness’ is different.

Local neighborhood overlap



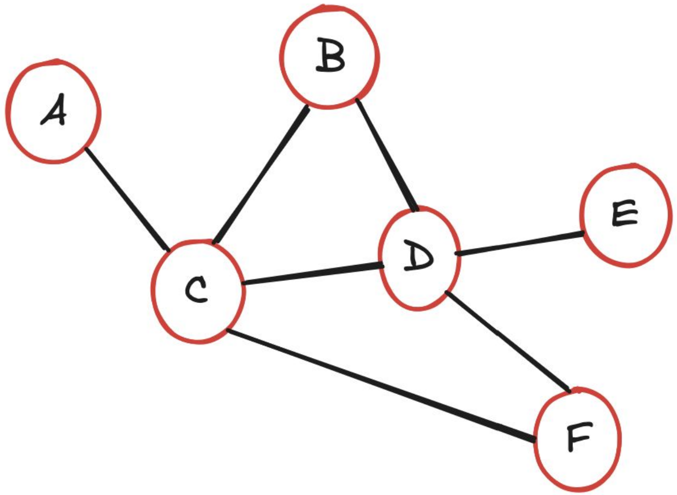
Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$



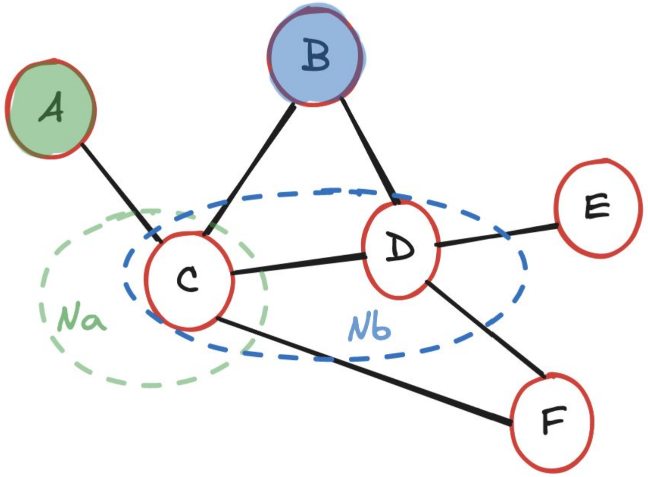
Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)|$



Local neighborhood overlap

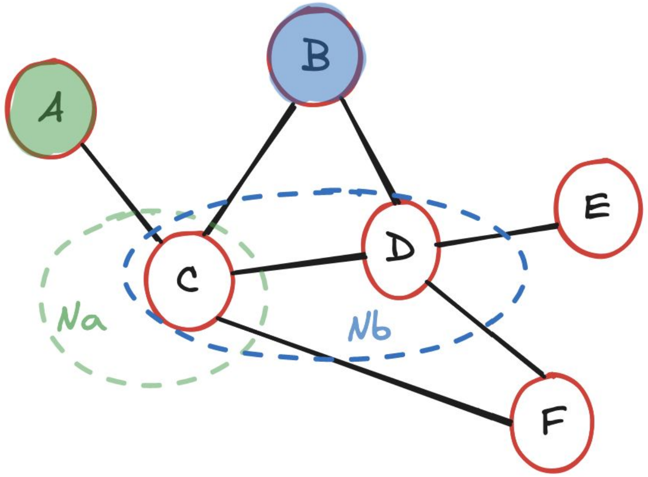
- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$



Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$

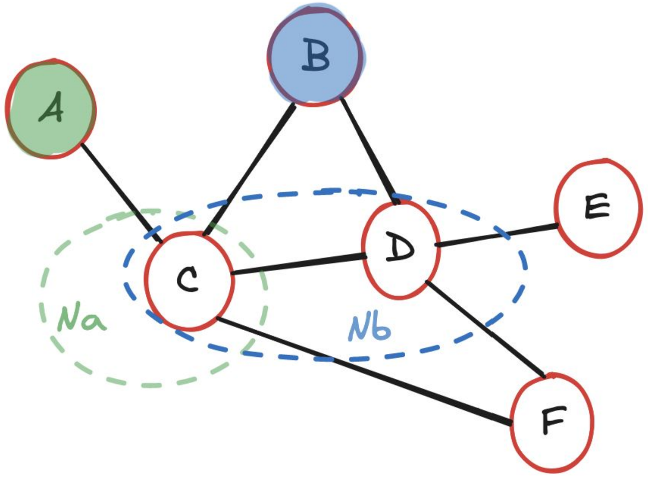
- Jaccard's coefficient: $|N(v_1) \cap N(v_2)| \,/\, |N(v_1) \cup N(v_2)|$



Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$

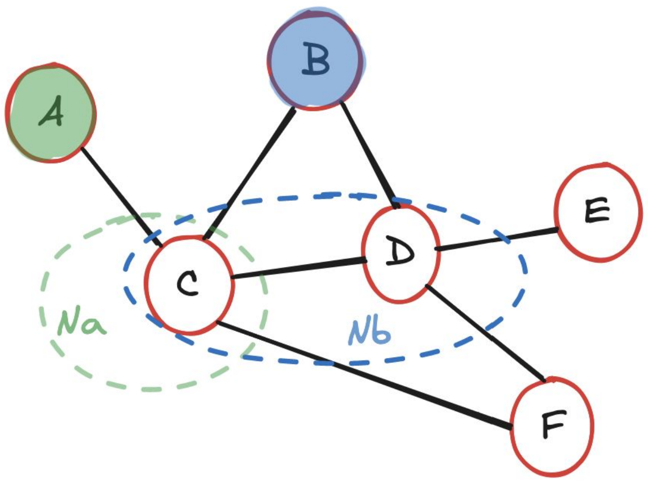
- Jaccard's coefficient: $|N(v_1) \cap N(v_2)| / |N(v_1) \cup N(v_2)|$
  e.g: $|N(A) \cap N(B)| / |N(A) \cup N(B)|$



Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$

- Jaccard's coefficient: $|N(v_1) \cap N(v_2)| \, / \, |N(v_1) \cup N(v_2)|$
  e.g: $|N(A) \cap N(B)| \, / \, |N(A) \cup N(B)|$
  $= |\{C\}| \, / \, |\{C,D\}| = 1/2$

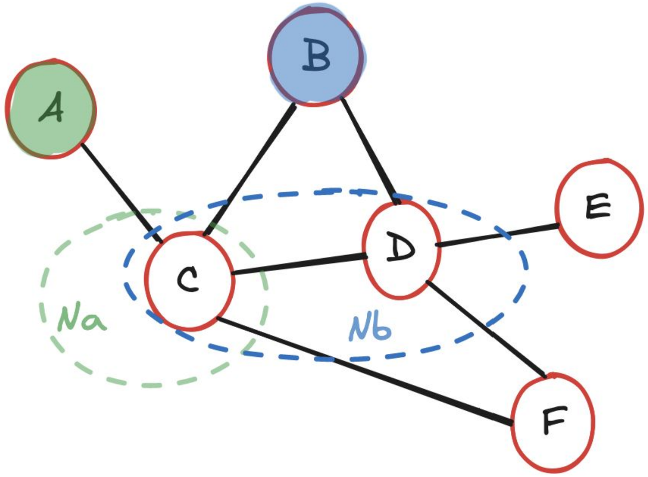


Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$

- Jaccard's coefficient: $|N(v_1) \cap N(v_2)| \,/\, |N(v_1) \cup N(v_2)|$
  e.g: $|N(A) \cap N(B)| \,/\, |N(A) \cup N(B)|$
  $= |\{C\}| \,/\, |\{C,D\}| = 1/2$

- Adamic-Adar index: $\sum_{u \in N(v_1) \cap N(v_2)} 1/\log(k_u)$

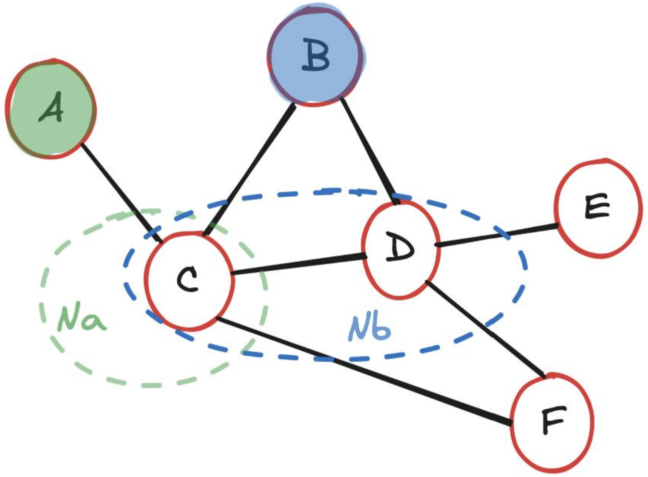


Local neighborhood overlap

- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$

- Jaccard's coefficient: $|N(v_1) \cap N(v_2)| \,/\, |N(v_1) \cup N(v_2)|$
  e.g: $|N(A) \cap N(B)| \,/\, |N(A) \cup N(B)|$
  $= |\{C\}| \,/\, |\{C,D\}| = 1/2$

- Adamic-Adar index: $\sum_{u \in \square(v1) \cap \square(v2)} 1/\log(k_u)$
  e.g: $1/\log(k_c)$

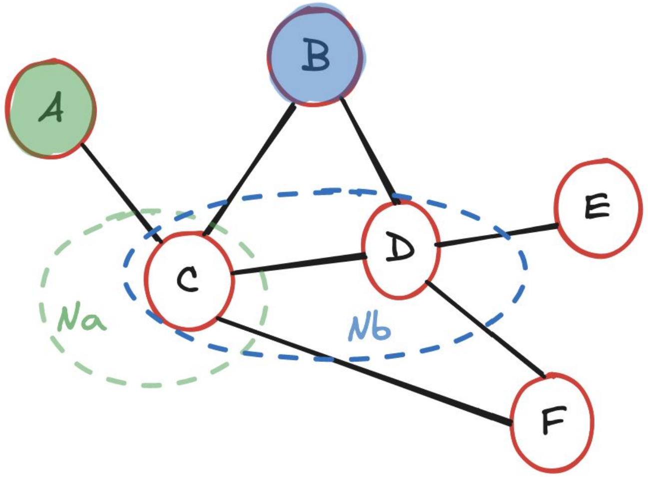


Local neighborhood overlap

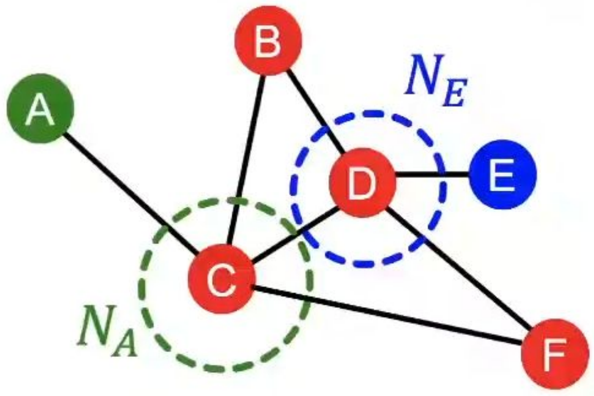
- Common neighbors: $|N(v_1) \cap N(v_2)|$
  e.g: $|N(A) \cap N(B)| = |\{C\}| = 1$

- Jaccard's coefficient: $|N(v_1) \cap N(v_2)| \,/\, |N(v_1) \cup N(v_2)|$
  e.g: $|N(A) \cap N(B)| \,/\, |N(A) \cup N(B)|$
  $= |\{C\}| \,/\, |\{C,D\}| = 1/2$

- Adamic-Adar index: $\sum_{u \in \square(v1) \cap \square(v2)} 1/\log(k_u)$
  e.g: $1/\log(k_c) = 1/\log 4 = 1.66$



Limitations



$$N_A \cap N_E = \phi$$
$$|N_A \cap N_E| = 0$$

Global neighborhood overlap

Global neighborhood overlap

Katz Index

Count the number of paths of all lengths between a pair of nodes.

Global neighborhood overlap

Katz Index

Count the number of paths of all lengths between a pair of nodes.

Q) how would you do this ?

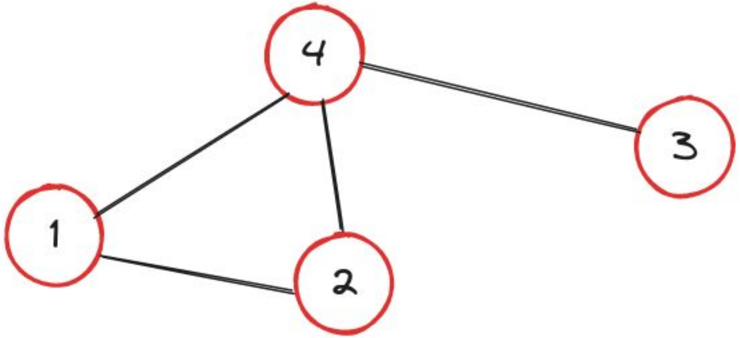
Global neighborhood overlap

Katz Index

Count the number of paths of all lengths between a pair of nodes.

Q) how would you do this ?

A) use powers of the adjacency matrix!

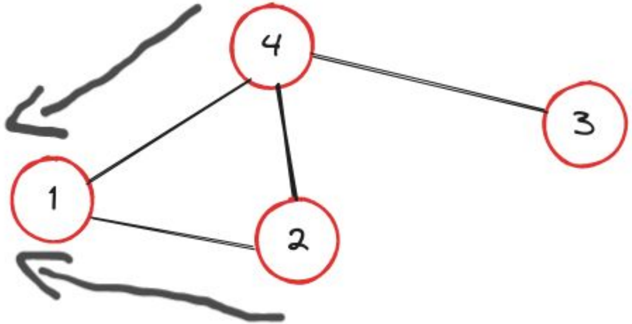


$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}_A$$

Node 1's neighbors

paths of len 1 between Node 1's neighbors itself

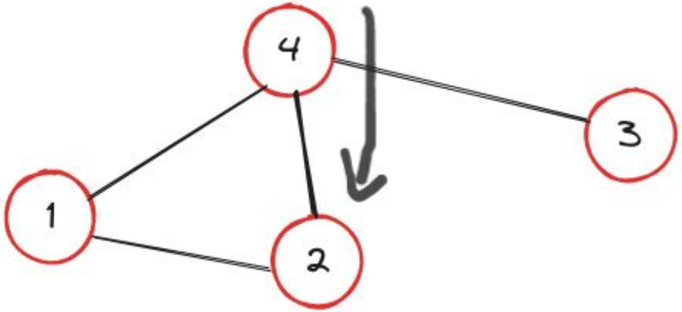
$$\begin{bmatrix} \boxed{0 \quad 1 \quad 0 \quad 1} \\ 1 \quad 0 \quad 0 \quad 1 \\ 0 \quad 0 \quad 0 \quad 1 \\ 1 \quad 1 \quad 1 \quad 0 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$



Node 1's neighbors

paths of len 1 between
Node 1's neighbors Node 2

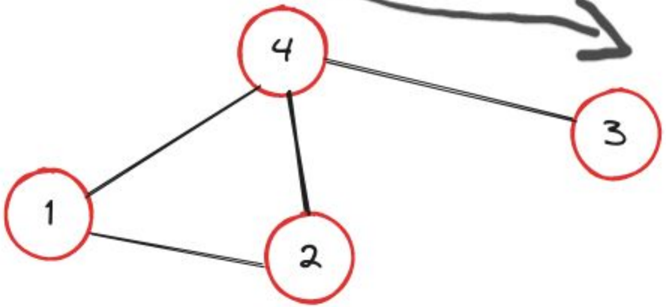
$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$



Node 1's neighbors

paths of len 1 between
Node 1's neighbors Node 3

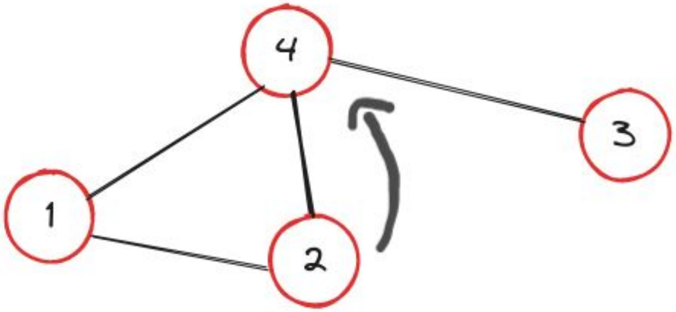
$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$



Node 1's neighbors

paths of len 1 between
Node 1's neighbors Node 4

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 1 \\ & & & \\ & & & \\ & & & \end{bmatrix}$$



Adjacency matrix powers

- A_{uv} specifies **#paths** of length **1** between u and v
- A^2_{uv} specifies **#paths** of length **2** between u and v
- A^l_{uv} specifies **#paths** of length **l** between u and v

Katz Index

Katz index between 2 nodes:

Sum over *all path lengths*

$$S_{v_1 v_2} = \sum_{l=1}^{\infty} \boxed{\beta^l} \boxed{A_{v_1 v_2}^l}$$

#paths of length l between v_1 and v_2

$0 < \beta < 1$: discount factor

Katz Index

Katz index between 2 nodes:

Sum over *all path lengths*

$$S_{v_1 v_2} = \sum_{l=1}^{\infty} \boxed{\beta^l} \boxed{A^l_{v_1 v_2}}$$

#paths of length l between v_1 and v_2

$0 < \beta < 1$: discount factor

Katz index matrix computed in closed form:

$$\boldsymbol{S} = \sum_{i=1}^{\infty} \beta^i \boldsymbol{A}^i = \underbrace{(\boldsymbol{I} - \beta \boldsymbol{A})^{-1}}_{= \sum_{i=0}^{\infty} \beta^i \boldsymbol{A}^i} - \boldsymbol{I},$$

Summary

- **Distance-based features**
  shortest path distance between 2 nodes.
  doesn’t capture neighborhood

- **Local neighborhood overlap**
  captures number of shared nodes between 2 nodes.
  null when no nodes are shared.

- **Global neighborhood overlap**
  counts number of paths of all lengths between 2 nodes.