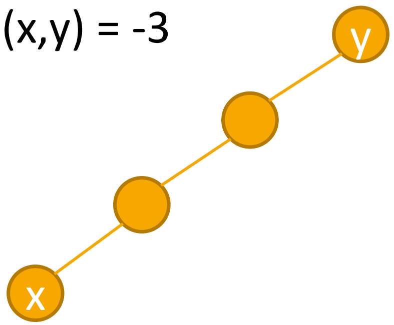
Methods for Link Prediction: Path based

Intuition
Use the (shortest) distance between two nodes as a link prediction measure

- For $(x, y) \in V \times V - E_{old}$

$score(x, y)$= (negated) length of shortest path between x and y

score(x,y) = -3



Very basic approach, it does not consider connections among (x,y) but only the distance

LP Methods: Path based

- Katz index

Element (x,y) in the Adjacency matrix

$$score(x,y) = \sum_{\ell=1}^{\infty} \beta^{\ell} \left| paths_{xy}^{(l)} \right| = \beta A_{xy} + \beta^2 A_{xy}^2 + \cdots$$

- Sum over ALL paths of length ℓ
- $0 < \beta < 1$ is a parameter of the predictor, exponentially damped to count short paths more heavily
- *Small β= predictions much like common neighbors*
- Two forms:
  - Unweighted: 1 if two authors collaborated, 0 otherwise
  - Weighted: strength of the collaboration

Closed form for the entire score matrix:
$(I - \beta A)^{-1} - I$

LP Methods: Path based

- Consider a random walk on G_{old} that starts at x and iteratively moves to a neighbor of x chosen uniformly random from $\Gamma(x)$
- The **Hitting Time** $H_{x,y}$ from x to y is the expected number of steps it takes for the random walk starting at x to reach y.

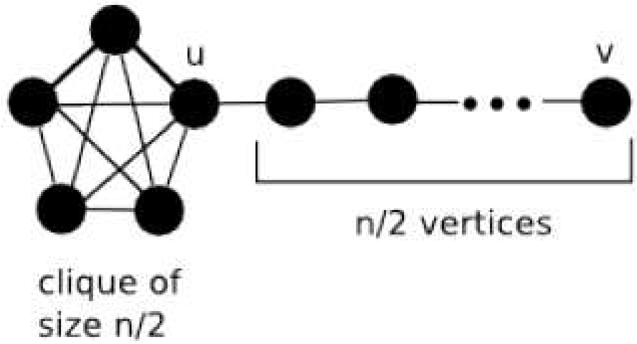
$$score(x,y) = -H_{x,y}$$

- The **Commute Time** from x to y is the expected number of steps to travel from x to y and from y to x

$$score(x,y) = -(H_{x,y} + H_{y,x})$$

Not symmetric, can be shown

$$h_{vu} = \Theta(n^2)$$
$$h_{uv} = \Theta(n^3)$$



LP Methods: Path based

- The hitting time and commute time measures are sensitive to parts of the graph far away from x and y -> periodically jump back to x
- Random walk on G_{old} that starts at *x* and has a probability *c* of returning to x at each step
- Random walk with restart: Starts from x, with probability $(1 - c)$ moves to a random neighbor and with probability c returns to x

$$s = (1 - c)(I - cD^{-1}A)^{-1}e_x$$

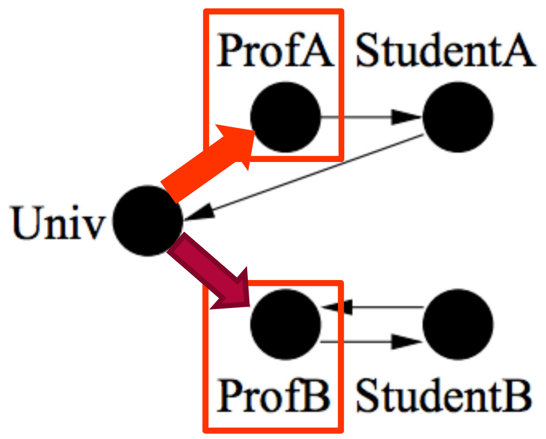
where s is a similarity vector between x and all the other nodes in the graph and e_x is the vector that has all 0, but a 1 in position x

$$score(x, y) = s_y$$

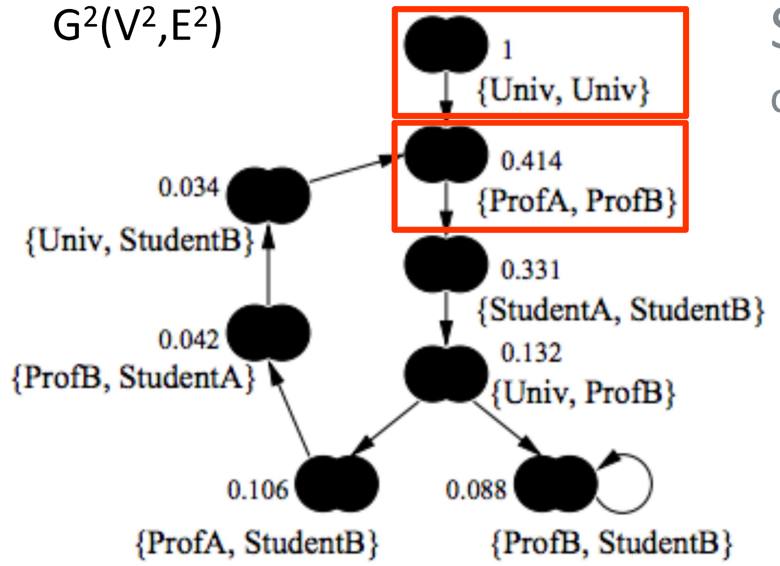
Path based: SimRank approaches

Intuition:
two objects are similar if they are referenced by similar objects

G(V,E)



$G^2(V^2,E^2)$

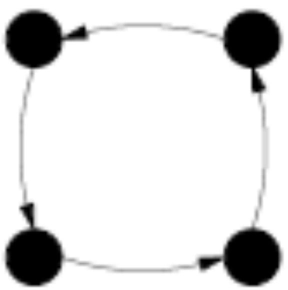


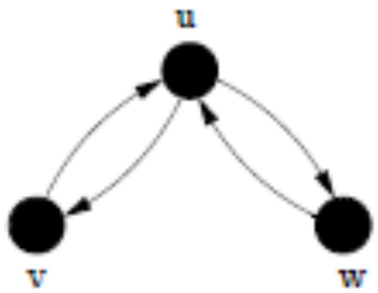
Structural
context

Glen Jeh and Jennifer Widom. **SimRank: a measure of structural-context similarity**. SIGKDD, 2002

Path based: SimRank approaches

Expected Meeting Distance (EMD): how soon two random surfers are expected to meet at the same node if they started at nodes x and y and randomly walked (in lock step) the graph backwards





- $score(\cdot,\cdot) = \infty$

=> no node will meet

- score(u, v) = score(u,w) = ∞
- score(v, w) = 1

=> v and w are much more similar than u is to v or w.

- $score(\cdot,\cdot) = 3$

=> any two node will meet in expectedly 3 steps, the similarity is lower than the previous for v,w

Path based: SimRank approaches

- Let us consider G^2
- A node (a, b) as a state of the tour in G: if a moves to c, b moves to d in G, then (a, b) moves to (c, d) in G^2

A tour in G^2 of length n represents a pair of tours in G where each has length n

- What are the states in G^2 that correspond to "meeting" points?

Singleton nodes (common neighbors)

- The EMD $m(a, b)$ is just the expected distance (hitting time) in G^2 between (a, b) and any singleton node
- The sum is taken over all walks that start from (a, b) and end at a singleton node