Social Network Analysis

LINK ANALYSIS

What are the Links?

- Model of interaction between entities defines types of entities being connected and types of links that connect these entities
- Diversities in connected entities
  - Homogeneous versus heterogeneous
- Diversities in connecting links
  - Directed versus undirected
  - Weighted versus unweighted
  - Signed versus unsigned, etc.
- Dynamics of link formation yields formation of substructures in the network
  - Communities emerges due to homophily
  - Strong ties and weak ties, etc.

Why Link Analysis?

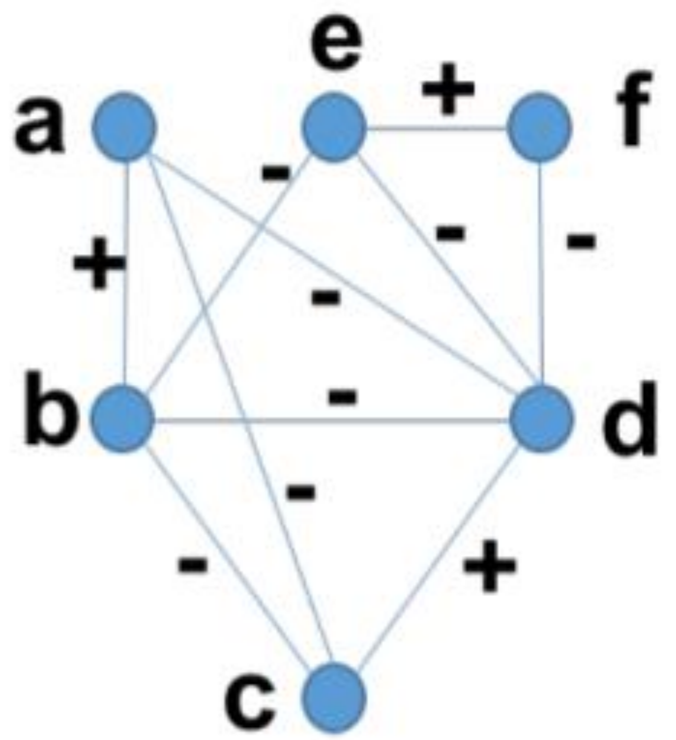
Fundamental output of link analysis task is to perform link-based object ranking, using global (network-wide) metric to measure the comparative importance of a node in the network.

- Entity Ranking
  - Search Engine Optimization
  - Scientific article Ranking
  - Scientific Author Ranking, etc.

Why Link Analysis?

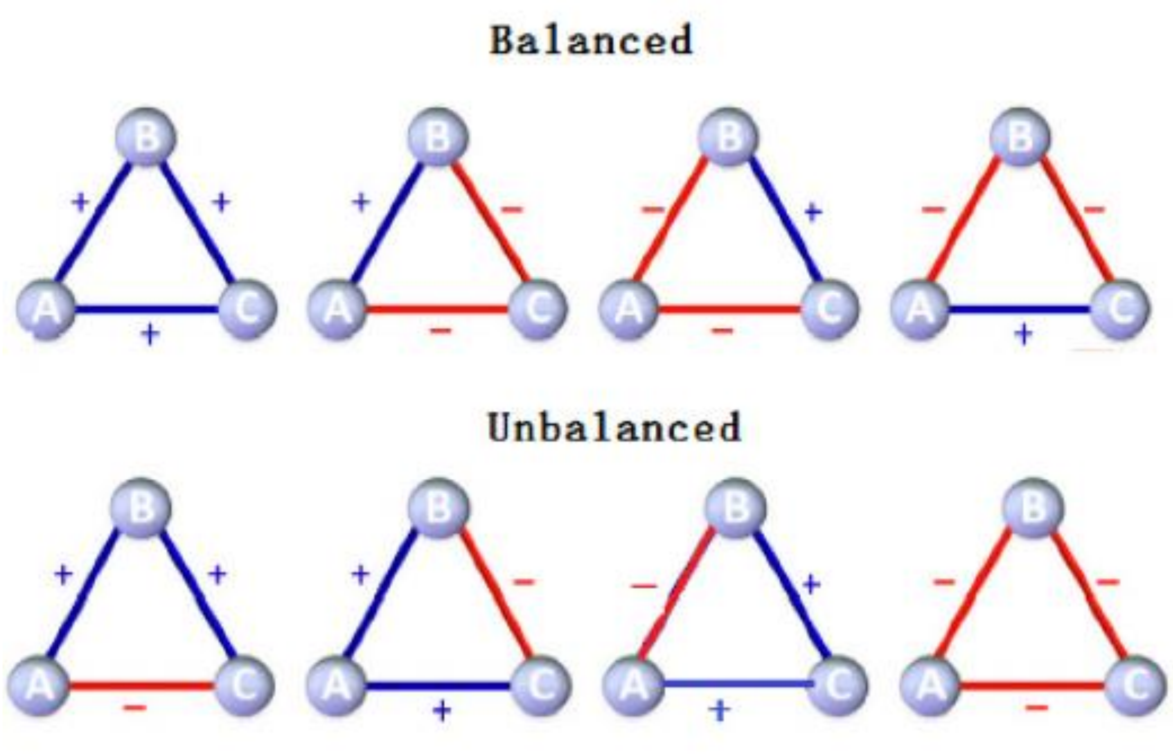
- Anomaly Detection
  - Online Fraud Detection
  - Counter Terrorism
  - Police/Military intelligence, etc.
- Adversarial Attacks
  - Adversarial attack on selected nodes to disrupt services
- Mining New Patterns
  - Crime Prevention
  - Future rank prediction
  - Link Prediction
  - Market Research, etc.

Signed Networks



- ❑Direction of a link in a network captures the direction of information flow across the link
- ❑Weight of a link in a network represents the strength of influence of information passing through that link
- ❑Neither of the above express how the information is perceived by the receiving node!
- ❑There often exist element pairs in perception/reaction towards information content –
  - ✓like/dislike (YouTube),
  - ✓agree/disagree (Reddit),
  - ✓Positive review/negative review (Amazon), etc.
- ❑Signed network captures the above opinion/relationship dynamics across entities

Balance Theory: Triads



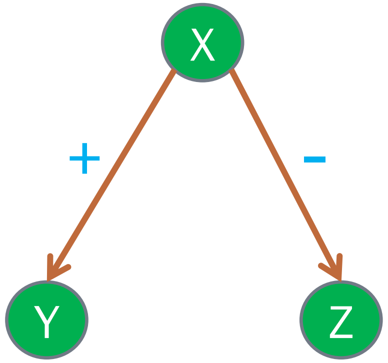
Positive = Friendship, Negative = Enmity

Li and Tang 2012

- Balance state occurs in triads when all sign multiplication of its sentiment relation charges positive

- Three Positive links
  - mutual trust and respect
  - Stable
- Two negative, one positive
  - trust between friends established based on distrust towards a common enemy
  - Stable
- Two positive, one negative
  - mutual friends would be under stress to take sides
  - Unstable
- Three negative links
  - No mutual trust
  - Unstable and likely to be disintegrated

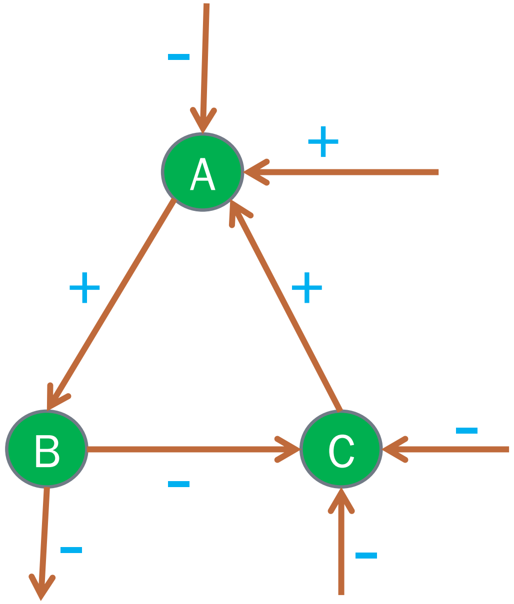
Signed Networks: Status Theory



$Y > X > Z$
Status relative to X

- Balance theory views signed links as model of likes and dislikes
- a signed link formed can have other possible interpretation!
  - Interpretation of link-sign as an indicator of relative status/prestige of a node with respect to the other
  - Status Theory
  - Assumes a signed, directed network of the entities
- A initiates a positive link to $B \implies A$ considers B to have a higher status than itself
- A initiates a negative link to $B \implies A$ considers B to have a lower status than itself

Signed Networks: Status Theory



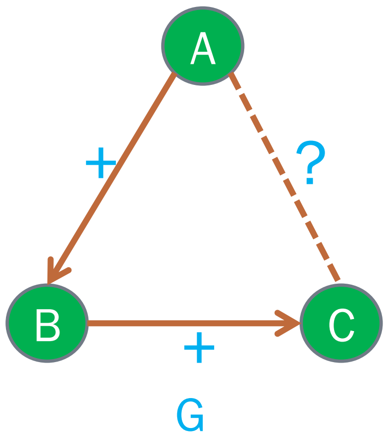
Snapshot of a signed graph

- Node-level metrics defined in this connection:
  - Generative Baseline (g): The fraction of positive signs generated by a node
  - Receptive Baseline (r): The fraction of positive signs received by a node

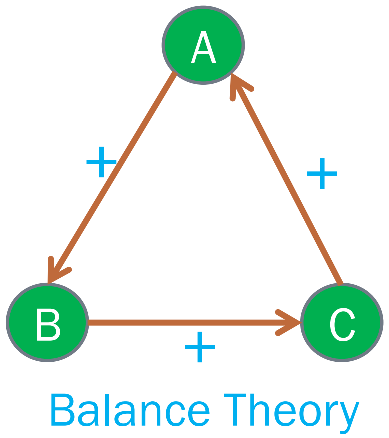
$$G = \frac{\text{number of positive signs generated by the node}}{\text{total number of signs generated by the node (outdegree)}} \qquad R = \frac{\text{number of positive signs received by node}}{\text{total number of signs receive by a node(indegree)}}$$

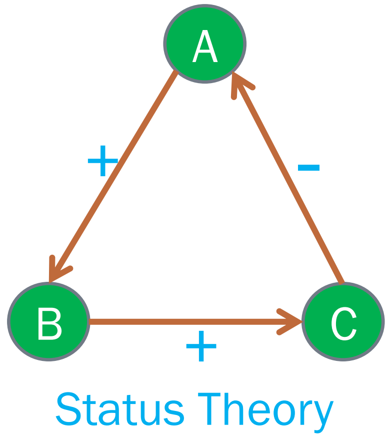
- Scores for generative baselines of the nodes of the signed graph beside are as follows:
  - $A_g = \frac{1}{1} = 1,$ $B_g = \frac{0}{2} = 0,$ $C_g = \frac{1}{1} =1$
- Scores for receptive baselines of the nodes of the signed graph beside are as follows:
  - $A_r = \frac{2}{3} = 0.67,$ $B_r = \frac{1}{1} = 1,$ $C_g = \frac{0}{3} = 0$

Comparison: Balance Theory and Status Theory

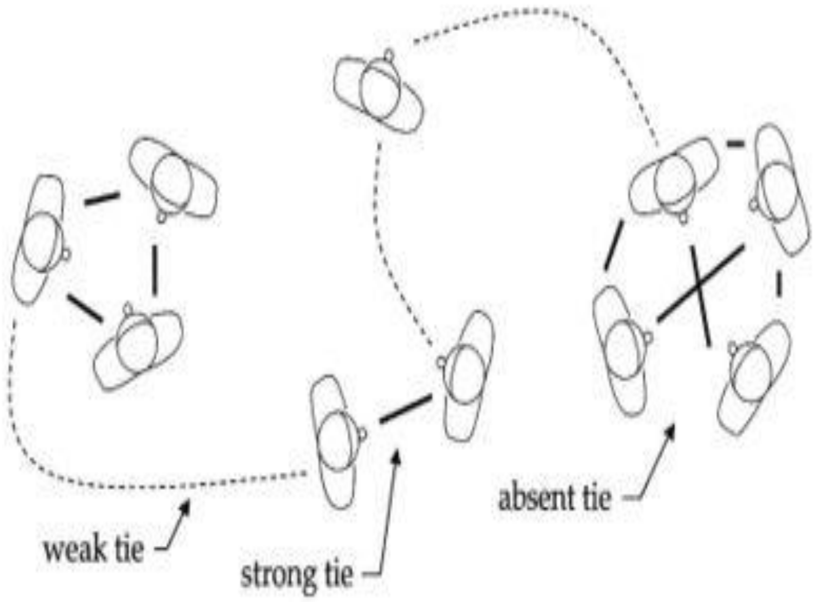


- Theory of status makes sense for directed networks only
- Theory of balance, though originated for undirected graphs, are also applicable for directed graphs
- In directed network G, if C forms a link to A, which link-sign is most likely to occur for that link?
  - According to theory of balance, link CA is predicted to be a positive link
  - According to theory of status, link CA is predicted to be a negative link!
- The two theories may infer conflicting predictions, as they have different interpretations altogether





Interpersonal ties



https://en.wikipedia.org/wiki/Interpersonal_ties

- Defined as information-carrying connections between entities/people
- Appear generally in three varieties: strong, weak or absent
- Strong ties
  - develop among entities that share interest and beliefs
  - thought of as source of confidence and emotional dependency
- Weak ties are mere acquaintances

Link Prediction approach 1: Strength of a Tie

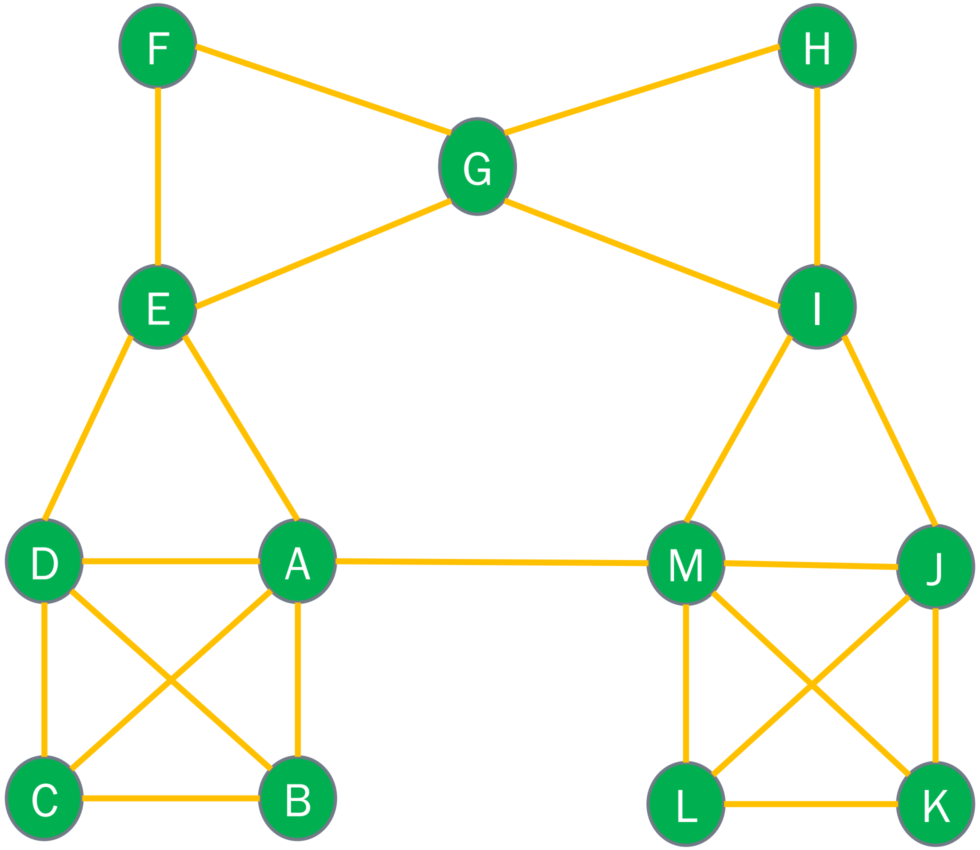
- Strength of ties captures a sense of closeness among entities/people
- Simplest metric to capture the same is via Jaccard score
- Corresponding metric, called Neighborhood Overlap (NO) is defined as:

$$NO(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}.$$

where $\Gamma(\cdot)$ denotes the neighbourhood of a node

- Higher the $NO(\cdot)$ score, higher the overlap between the nodes, and higher the chance forming a link in between

Link Prediction approach 1: Strength of a Tie ; Neighborhood Overlap: Example



$\Gamma(A) = \{B, C, D, E, M\}$,
$\Gamma(M) = \{A, I, J, K, L\}$,
$\Gamma(E) = \{A, D, F, G\}$

$|\Gamma(A) \cap \Gamma(M)| = |\phi| = 0$
$|\Gamma(A) \cap \Gamma(E)| = |\{D\}| = 1$

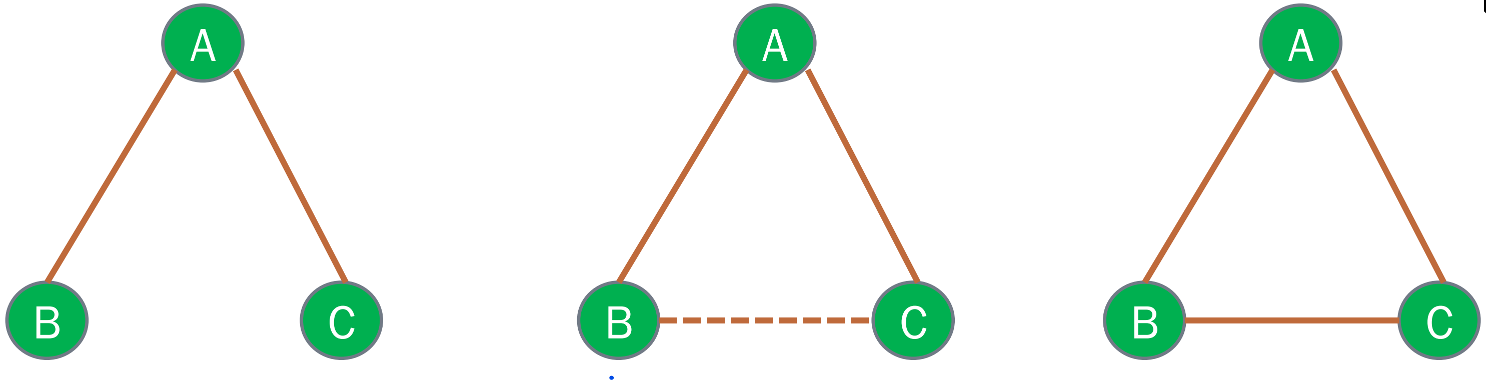
$|\Gamma(A) \cup \Gamma(M)| = |\{B, C, D, E, I, J, K, L\}| = 8$
$|\Gamma(A) \cup \Gamma(E)| = |\{B, C, D, F, G, M\}| = 6$

$NO(A, M) = \frac{0}{8} = 0$, $\quad NO(A, E) = \frac{1}{6}$

Link Prediction approach 2: Triadic Closure

toi

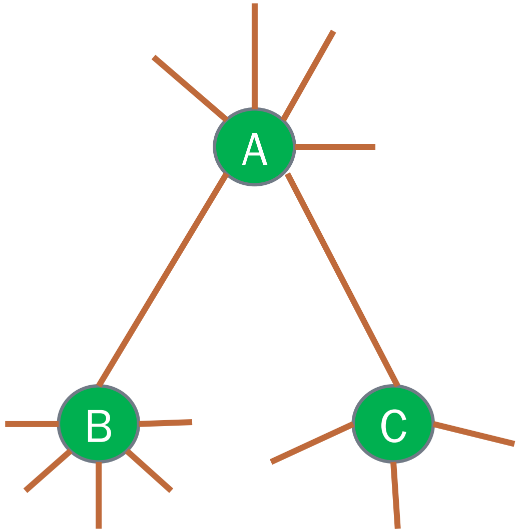
- A friend of a friend is also a friend – is the philosophy



$t-\Delta t$ B & C are not friends yet | t B & C gets introduced (via A) | $t+\Delta t$ B & C are friends now

- Reasons behind Triadic closure formation
  - Opportunity: of meeting via mutual connection
  - Trust: link formation based on mutual trust
  - Incentive: nodes may have incentives to bring their mutual friends together

Link Prediction approach 2: Triadic Closure…



❑Quantifying Strength of Triadic Closures

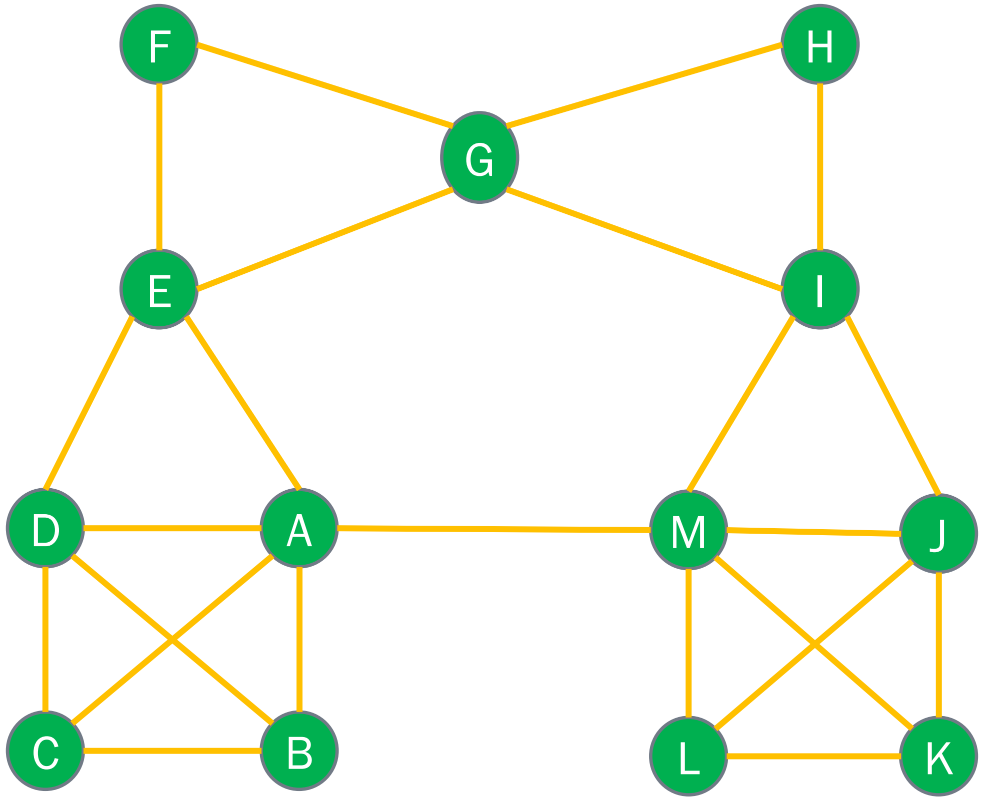
❑Strength of a triadic closure with respect to node A and the nodes B and C of which A is a mutual friend can be quantified using the clustering coefficient of node A

❑Clustering coefficient of a node (CC_A) measures the probability that the pair of friends (B and C) of the given node (A) are friends of each other

$$CC_A = \frac{2 \times \sum_{i,j \in \Gamma(A)} \boldsymbol{I}\big((i,j) \in E\big)}{k_A(k_A - 1)}$$

where $\boldsymbol{I}(\cdot)$ is the indicator function that returns 1 if condition is true, and 0, otherwise

Link Prediction approach 2: Triadic Closure - Example



- B and M are neighbours of node A. To find the how likely they form a link.

$\Gamma(A) = \{B, C, D, E, M\}$

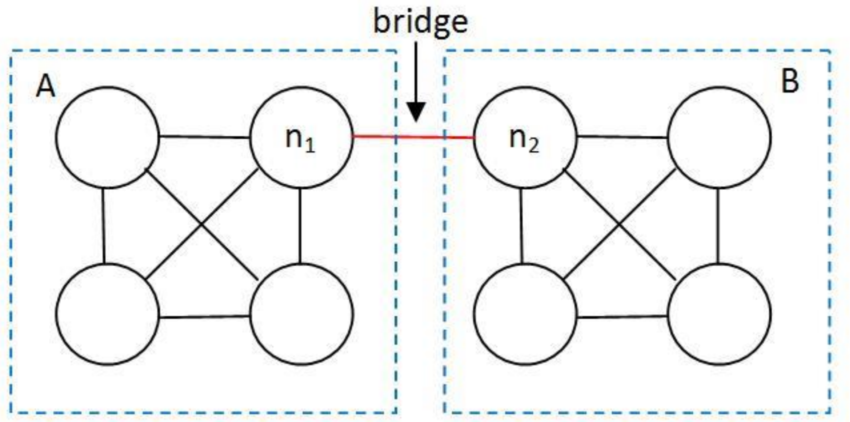
$k_A = 5$

Existing valid edges in $\Gamma(A)$ are $\{BC, BD, CD, DE\}$

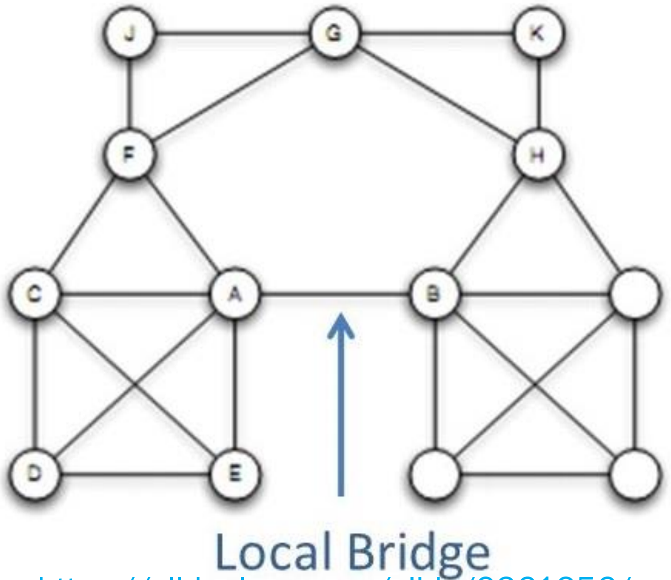
$$CC_A = \frac{2\times 4}{5\times 4} = 0.4$$

With 40% probability we may say that nodes B and M will form a link in the future.

Bridges and Local Bridges



https://en.wikipedia.org/wiki/Bridge_(interpersonal)

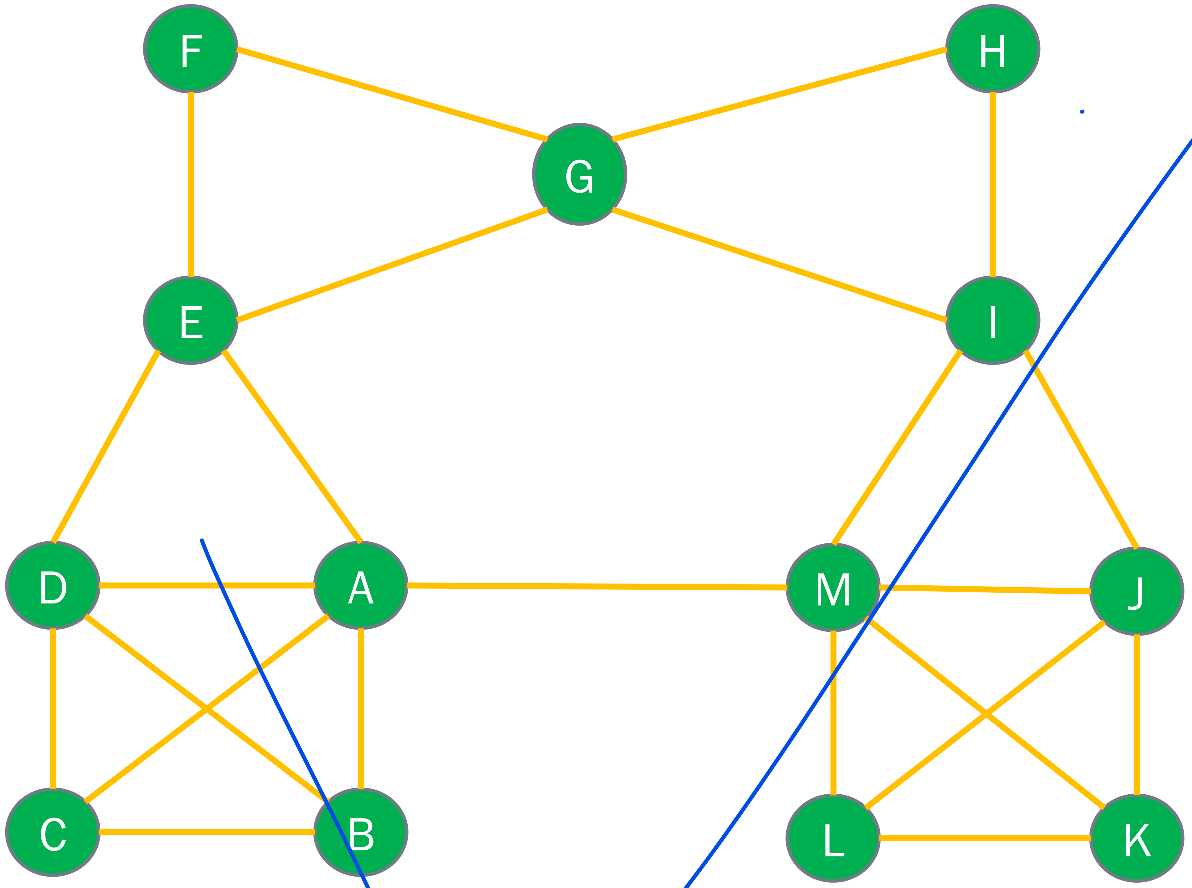


https://slideplayer.com/slide/9361256/

- A bridge is a direct tie between nodes that would otherwise be in disconnected components of the graph
- Removal of a bridge increases the number of disconnected components in a network

- Local bridges are ties between two nodes in a social graph that are the shortest route by which information might travel from those connected to one end to those connected to the other
- On removal of a local bridge the distance between these two nodes will be increased to a value strictly more than two

Local Bridges/Weak Ties



- An edge can be considered a local bridge if its Neighborhood Overlap Score (NO) is zero
- In other words, end-points of a local bridge have no mutual friends
- Local bridges are not a part of any triad in the network

atla mate to aene remove karva par ae nodes na vache no distance would be greater than 2

- (A, M) is a local bridge/weak tie

Local Bridges: Edge Embeddedness

- For an edge $\langle x, y \rangle$, its embeddedness can be defined as the number of mutual friends that the endpoints of the edge posses

$$Embeddedness(\langle x, y \rangle) = |\Gamma(x) \cap \Gamma(y)|$$

- A local bridge is an edge with embeddedness of zero

Local Bridges: Importance

- Close friends tend to move in the same circles that we do
  - Information close friends receive overlaps considerably
- Acquaintances, by contrast, know people that we do not,
  - People receive more novel information through acquaintances than from close friends
- Weaker ties act as a bridge and help a person gain access to newer and wider information (strength of weak ties)
- In case of stress/conflict between two groups, weak ties act as mediators
- In an adversarial setting, removing local bridges can lead to the formation of echo chambers
- During disease outbreaks, local bridges may cause the disease to transmit from one group to another

Revisiting PageRank…

Personalized PageRank

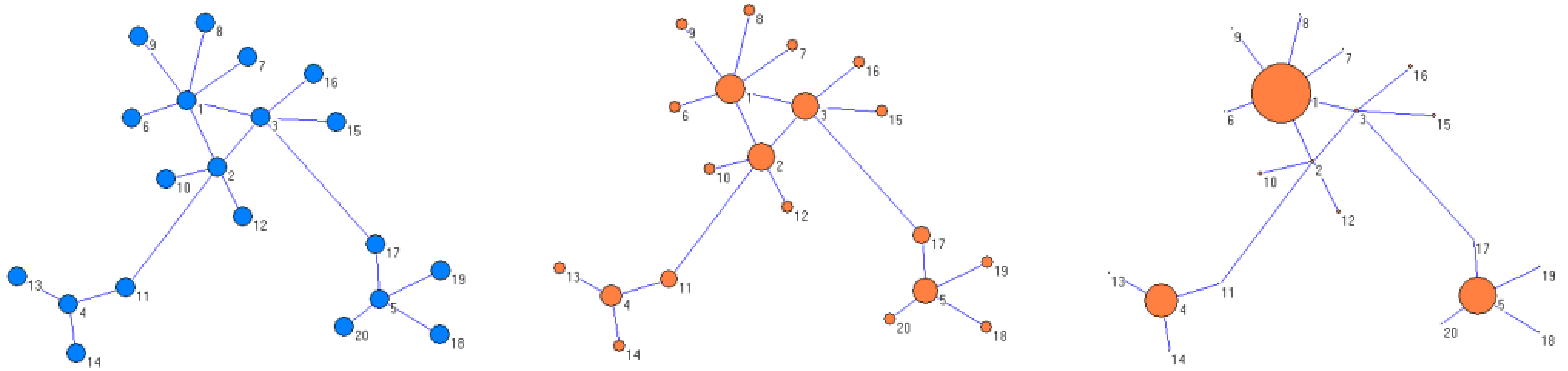
- The vector E characterizes the random jump after surfing hyperlinks from a page
- The landing page need not be equally-likely for all the pages of the graph
- The surfer may be biased to return to one or more selective pages based on the search
  - Surfer may land a specific page on return (say, index page)
  - Surfer may land one of a set S of pages
  - Surfer may land on one of a list S_w of pages based on her search pattern
- The distribution of E(S) or $E(S_w)$ will be different from being uniform distribution.
- The modified (Personalized) PageRank formula is as follows:

$$R(w) = (1-\alpha)\sum_{b \in B_w} \frac{R(b)}{N_b} + \alpha E(S_w)$$

Diverse PageRank : DivRank

- Lack of diversity in top-ranked nodes in PageRank
  - Suppose user is looking for a list of famous eateries in the city
  - If all the top-ranked places are non-veg eateries, and the user is vegetarian, the list is useless; and vice versa
- Output from PageRank often has redundant entities
- Redundancy is problematic in applications where space is a constraint
- A good combination of prestige and diversity is desirable
- DivRank (Diverse Rank) is a solution in the direction

DivRank: Prestige with Diversity



(a) An illustrative network. (b) Weighting with PageRank. (c) A diverse weighting.

- In example graph, Page may return entities 1, 2, and 3 as output
- However, these nodes, being part of a community, may be similar in nature
- Whereas choice 4 and 5 would have wiser, as they have information for different clusters

DivRank: Vertex-Reinforced Random Walks

ae to vertex par jetli pan vakht avso...atli j vakht tamare to farithi avani skyata 6e

Vertex-Reinforced Random Walks are random walks where the transition probability from one state to the next is reinforced by the number of previous visits to the state $N_T(v)$

SimRank: Measuring Similarity of Objects

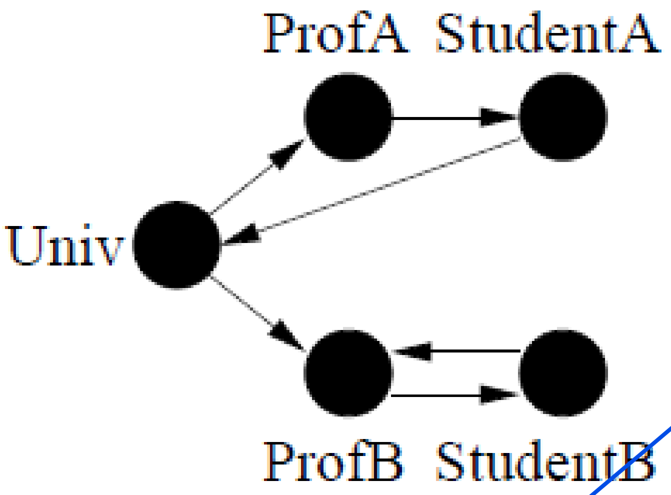
SimRank: Measuring Similarity of Objects

- SimRank measures similarity of the structural context in which objects occur, based on their relationships with other objects.
- Idea : Two objects are similar if they are related to similar objects
- For a given domain, SimRank can be combined with other domain-specific similarity measures.
- Example 1 : Citation graph
  - Two papers are similar if they are cited by similar papers
- Example 2 : E- Commerce graph
  - Two products are similar if they are bought by similar customers
  - Two customers are similar if they are buying similar products

SimRank: Measuring Similarity of Objects

- Idea : Two objects are similar if they are related to similar objects
- More precisely, objects a and b are similar if they are related to objects c and d, respectively, and c and d are themselves similar.
- The base case is that objects are similar to themselves.

SimRank: Measuring Similarity of Objects

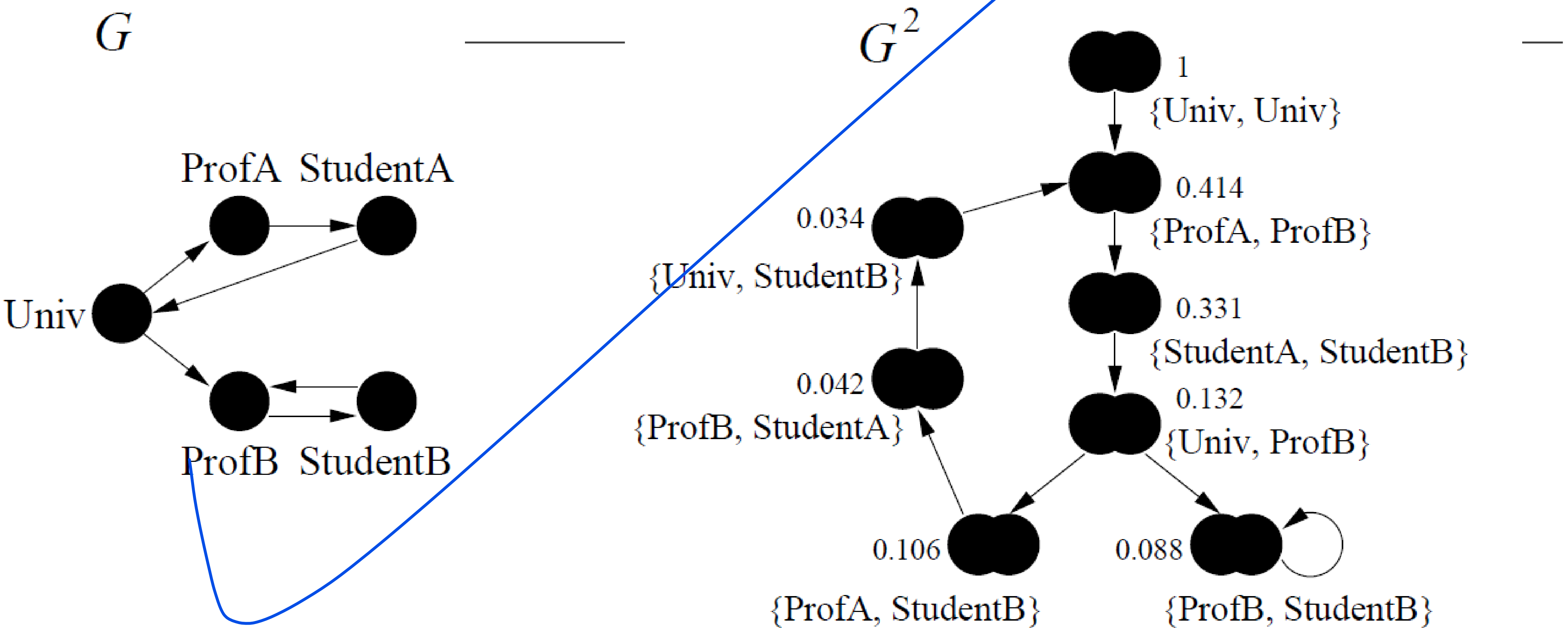


- Graph shows the Web pages of two professors ProfA and ProfB, their students StudentA and StudentB, and the home page of their university Univ.
- Edges between nodes represent hyperlinks from one page to another.
- From the fact that both are referenced (linked to) by Univ, we may infer that ProfA and ProfB are similar
- Can we infer that StudentA and StudentB are also similar based on the similarity of ProfA and ProfB ?
- Similar inference can be derived for other pairs of objects

SimRank: Measuring Similarity of Objects

- Logical representation of the SimRank computation by using a node-pair
- A new graph G^2 is formed in which each node represents an ordered pair of nodes of G.
- A node (a, b) of G^2 points to a node (c, d) if, in G, a points to c and b points to d.
- Each node-pair shows the similarity score between two nodes that they represent
- Scores are symmetric
- Draw (a, b) and (b, a) as a single node {a, b} (with the union of their associated edges).
- Iterative computation of SimRank scores for each node in G^2

SimRank: Measuring Similarity of Objects



SimRank: Basic Formulation

- For a node v in the network, $I(v) = \{I_i(v) | 1 \leq i \leq |I(v)|\}$ and $O(v) = \{O_i(v) | 1 \leq i \leq |O(v)|\}$ denotes the sets of indegree and outdegree neighbours, respectively.
- Formulate the similarity score $s(u, v) \in [0,1]$ as follows:

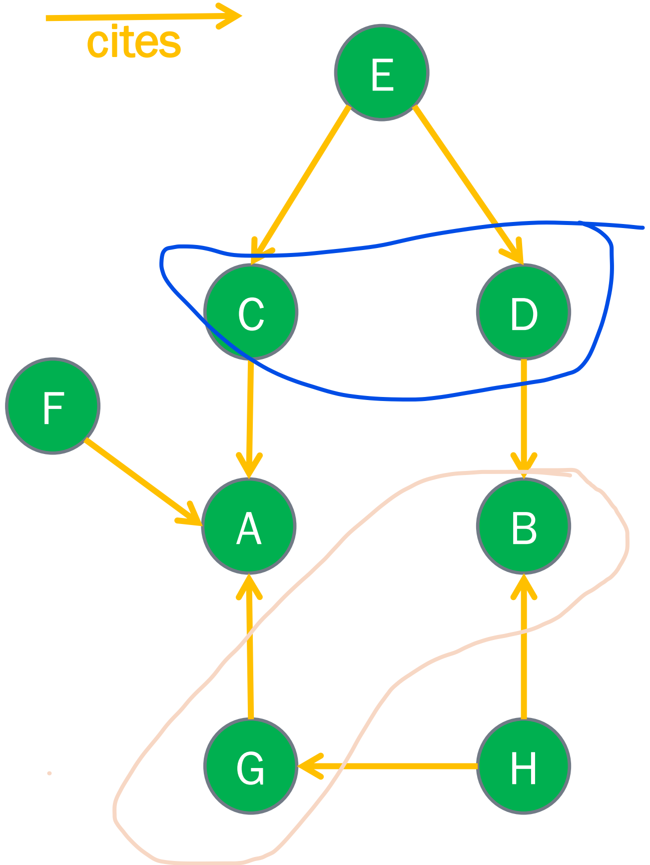
$$s(a,b) = \begin{cases} 1 & if\ a = b \\ 0 & if\ I(a) = \emptyset\ or\ I(b) = \emptyset \\ \dfrac{C}{|I(a)|.|I(b)|} \displaystyle\sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) & otherwise \end{cases}$$

  - A node is maximally similar to itself
  - No way of determining the score for a neighborhood that does not exist
  - Similarity between two randomly selected nodes is proportional to the average similarity between their neighbors

SimRank: Basic Formulation

- **Constant C** is considered as a confidence level or a decay factor
- Consider a simple scenario where page x references both c and d, so we conclude some similarity between c and d.
- The similarity of x with itself is 1, but we probably don't want to conclude that s(c, d) = s(x, x) = 1.
- Rather, we let s(c, d) = C x s(x, x), meaning that we are less confident about the similarity between c and d than we are between x and itself

SimRank: Example 2 Citation Network



- Paper E cites papers C and D
  - Papers C and D appears similar
- Paper H cites papers B and G
  - Papers B and G appears similar
- What about the similarity of papers A and B?

SimRank in Heterogeneous Bipartite Network

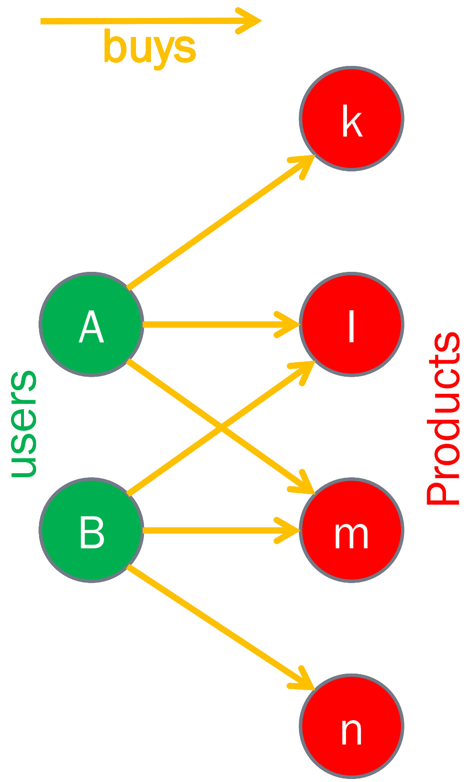
- In a heterogeneous network of users and products, the similarity of products and users are mutually-reinforced
  - two users can be considered similar if they buy similar products
  - two products can be considered similar if they are bought by similar users
- Similarity between two distinct users can be expressed as:

$$s(u_1, u_2) = \frac{C_1}{|O(u_1)|.|O(u_2)|} \sum_{i=1}^{|O(u_1)|} \sum_{j=1}^{|O(u_2)|} s(O_i(u_1), O_j(u_2))$$

- Similarity between two distinct products can be expressed as:

$$s(p_1, p_2) = \frac{C_2}{|I(p_1)|.|I(p_2)|} \sum_{i=1}^{|I(p_1)|} \sum_{j=1}^{|I(p_2)|} s(I_i(p_1), I_j(p_2))$$

Illustration: SimRank in Heterogeneous Bipartite Network



To calculate the similarity between users A and B

$O(A) = \{k, l, m\}$ and $O(B) = \{l, m, n\}$

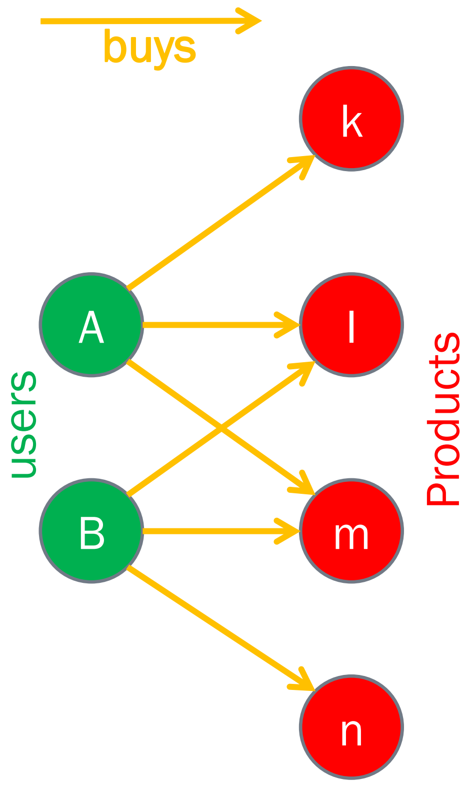
$I(k) = \{A\}$, $I(l) = \{A, B\}$, $I(m) = \{A, B\}$, and $I(n) = \{B\}$

$$s(A, B) = \frac{C_1}{3 \times 3}(s(k,l) + s(k,m) + s(k,n) + s(l,l) + s(l,m) + s(l,n) + s(m,l) + s(m,m) + s(m,n))$$

We have, $s(X, X) = 1$ and $s(X, Y) = s(Y, X)$

$$s(k,l) = \frac{C_2}{1 \times 2}[s(A,A) + s(A,B)] = \frac{C_2}{2} + \frac{C_2.s(A,B)}{2}$$

Illustration: SimRank in Heterogeneous Bipartite Network



Similarly, $s(k,m) = \frac{C_2}{2} + \frac{C_2.s(A,B)}{2}$, $s(k,n) = C_2.s(A,B)$

$s(l,l) = 1$, $s(l,m) = \frac{C_2}{2} + \frac{C_2.s(A,B)}{2}$, $s(l,n) = \frac{C_2}{2} + \frac{C_2.s(A,B)}{2}$

$s(m,l) = \frac{C_2}{2} + \frac{C_2.s(A,B)}{2}$, $s(m,m) = 1$, $s(m,n) = \frac{C_2}{2} + \frac{C_2.s(A,B)}{2}$

Solving, $s(A,B) = \frac{3C_1C_2+2C_1}{9-4C_1C_2}$

Further, setting $C_1 = C_2 = 0.8$,

$$\boldsymbol{s(A,B) = 0.547}$$

SimRank in Homogeneous Bipartite Network

Can you apply SimRank in following application ?

In a citation network, two scientific papers might be similar as survey papers if they cite similar result papers, while two papers might be similar as result papers if they are cited by similar survey papers.

PathSim: Measuring Similarity of Objects in Heterogeneous Network

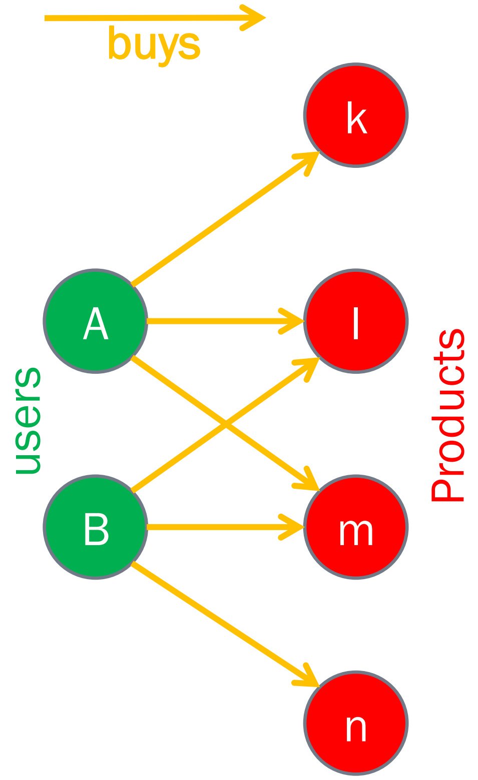
Heterogeneous Networks

- A tuple of the form $(V, E, \mathcal{A}, \mathcal{R}, \varphi, \psi)$ represents an information networking system if
  - V is the set of vertices
  - E is the set of edges
  - $\mathcal{A}$ is the set of different node types present in the network
  - $\mathcal{R}$ is the set of different link types present in the network
  - $\varphi(v): V \to \mathcal{A}$ maps each vertex to a node type
  - $\psi(e): E \to \mathcal{R}$ maps each edge to a link type
- If $|\mathcal{A}| = 1$ as well as $|\mathcal{R}| = 1$, then the system is termed as a homogeneous network
- On the contrary, if $|\mathcal{A}| > 1$ or $|\mathcal{R}| > 1$, or both, then the system is termed as a heterogeneous network

Heterogeneous Networks: Variants

- When $|\mathcal{A}| > 1$ and $|\mathcal{R}| = 1$, then we have a heterogeneous network consisting of vertices of more than one types, and only one types of links

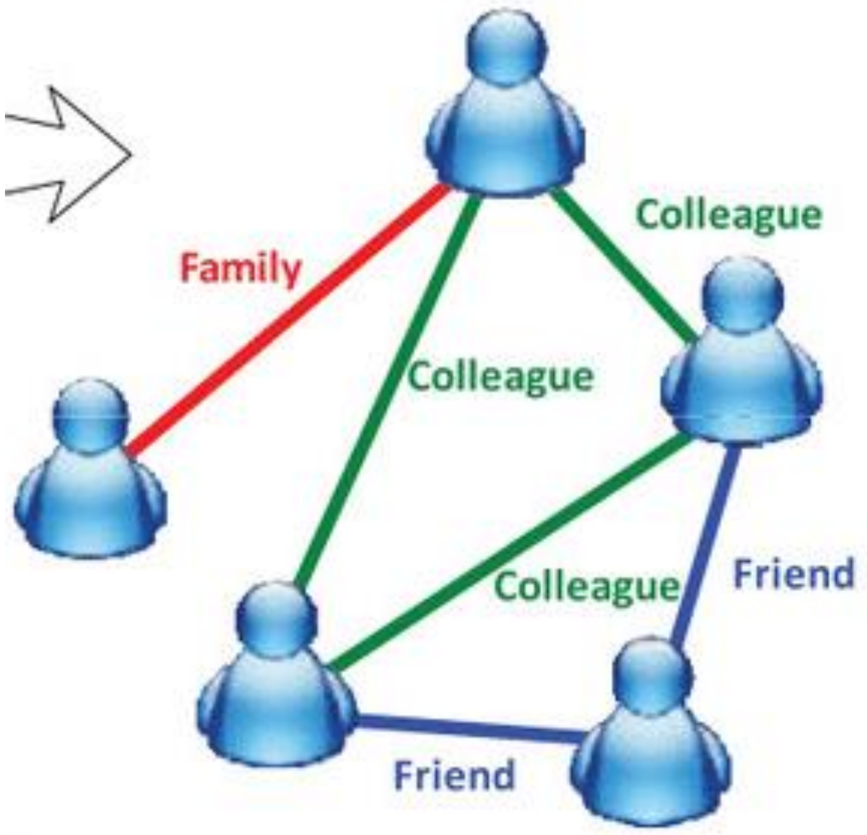
- A typical example is consumer-product purchase network, where
  - $\mathcal{A} = \{users, products\}$, and
  - $\mathcal{R} = \{user \longrightarrow products | user\ buys\ product\}$



Heterogeneous Networks: Variants

- When $|\mathcal{A}| = 1$ and $|\mathcal{R}| > 1$, then we have a heterogeneous network consisting of vertices of one type, but there are more than one type of links between these vertices

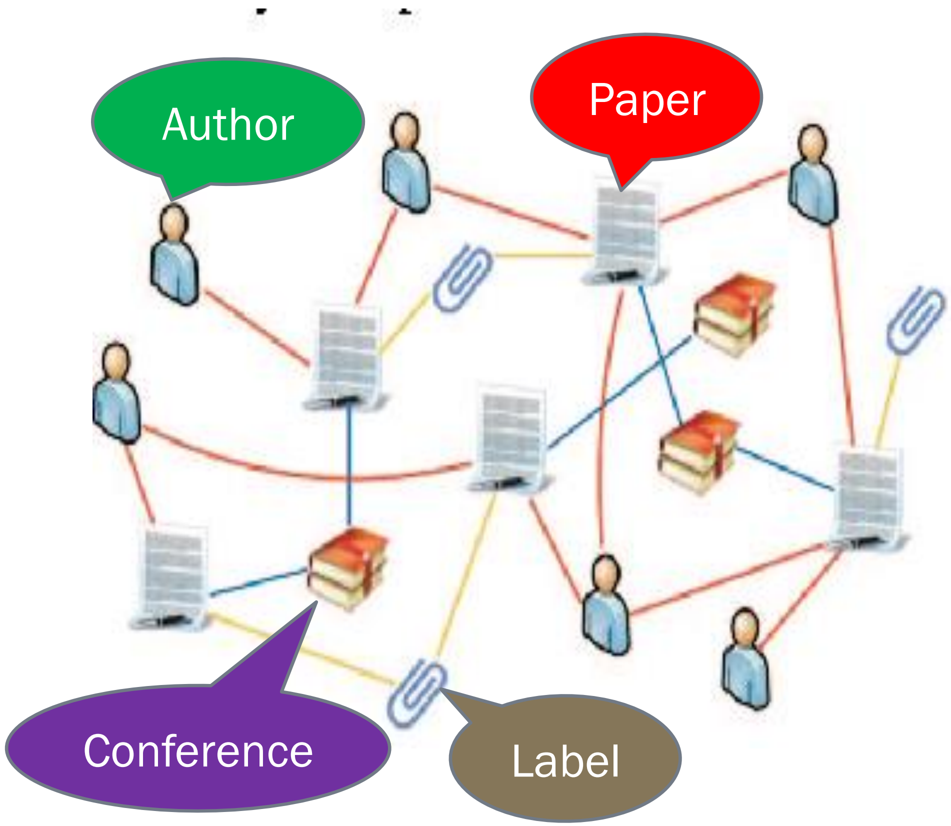
- A typical online social networking platform;
  - only one type of vertices, viz. users of the network;
  - There are more than one type of links: friends in real life, family members in real life, office colleague in real life, and so on.



Liu and Yu [2019]

Heterogeneous Networks: Variants

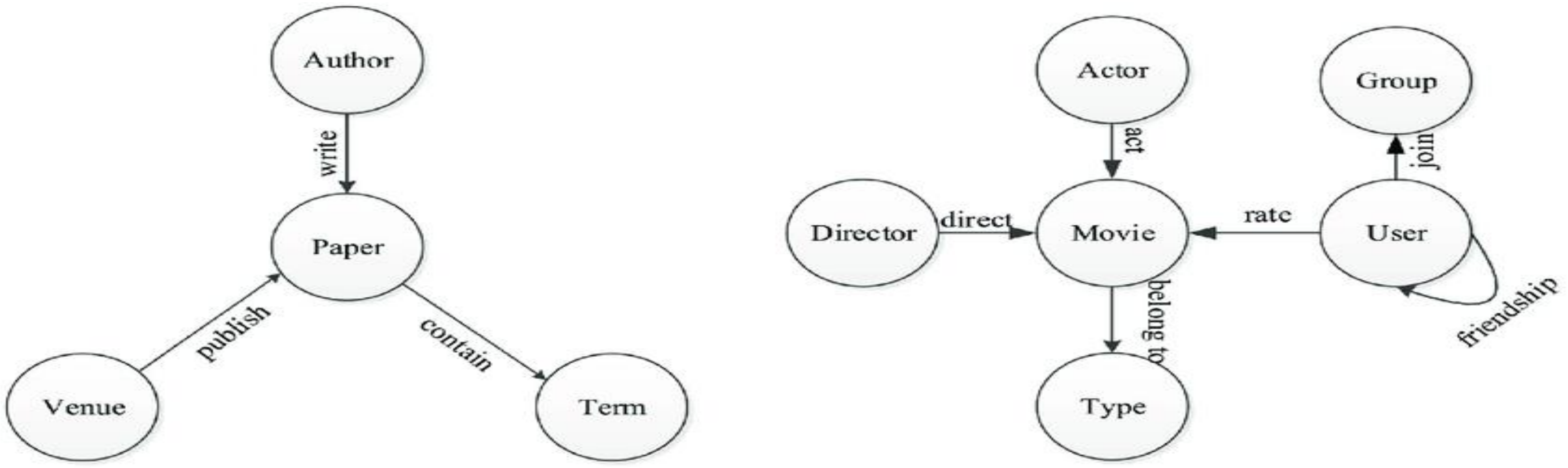
- When both $|\mathcal{A}| > 1$ and $|\mathcal{R}| > 1$, then we have a heterogeneous network consisting of vertices and links of more than one type
- A typical bibliographic network consisting of authors, papers, conference venues, etc., and various kinds of relationship between these entities



https://www.semanticscholar.org/paper/HRank%3A-A-Path-based-Ranking-Framework-in-Network-Li-Shi/186d8239daa10cedb7be946387a9326a0a3c9999

Heterogeneous Networks: Network Schema

- A meta-data level outline for a heterogeneous directed network $G(V, E)$ and the information tuple $(V, E, \mathcal{A}, \mathcal{R}, \varphi, \psi)$, where $\varphi: V \rightarrow \mathcal{A}$ is the object type mapping, and $\psi: E \rightarrow \mathcal{R}$ is the link type mapping. The corresponding network schema is given by $T_G = (\mathcal{A}, \mathcal{R})$



(A) DBLP network with a star network schema

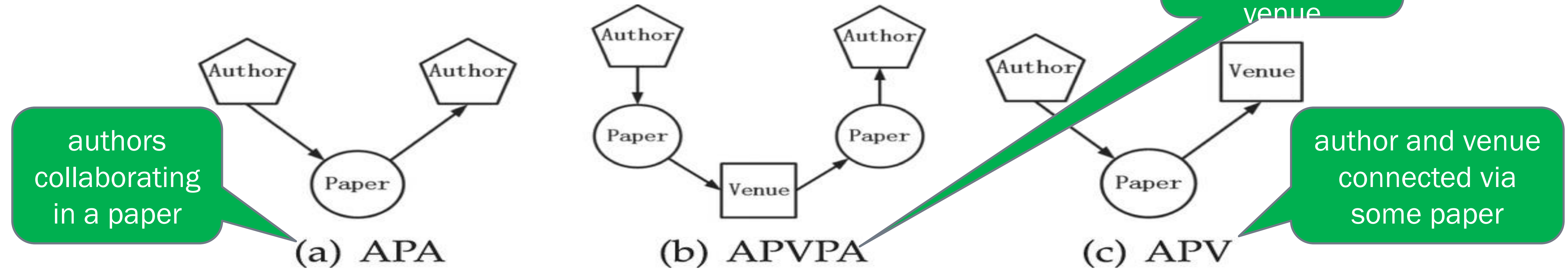
(B) Douban Movie network with a general network schema

https://www.researchgate.net/publication/314129795_Generic_network_schema_agnostic_sparse_tensor_factorization_for_single-pass_clustering_of_heterogeneous_information_networks

Heterogeneous Networks: Meta-Path

- A meta-path is a meta-level description of the structural connectivity between the entities
- Different paths deliver varying semantic similarity/differences or measure different topological connectivity
- A meta-path is a path $\mathcal{P}$ of length ℓ defining a composite relation over the ℓ links $\mathcal{R} = \mathcal{R}_1 \circ \mathcal{R}_2 \circ \mathcal{R}_3 \circ \cdots\cdots\cdots \circ \mathcal{R}_\ell$ and $\ell + 1$ objects $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \cdots\cdots\cdots, \mathcal{A}_{\ell+1}$ denoted in the form

$$\mathcal{A}_1 \xrightarrow{\mathcal{R}_1} \mathcal{A}_2 \xrightarrow{\mathcal{R}_2} \mathcal{A}_3 \cdots\cdots\cdots \mathcal{A}_\ell \xrightarrow{\mathcal{R}_\ell} \mathcal{A}_{\ell+1}$$



https://www.researchgate.net/figure/Example-for-Meta-path-in-HIN-on-the-bibliographic-network-2-Figure-3-defines-the-meta_fig1_339302745

PathSim: Formulation

- A meta-path based symmetric similarity measure, PathSim, between two objects x and y of the same type can be given as follows:

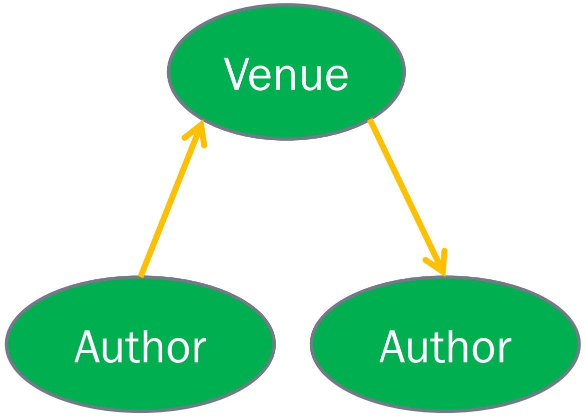
$$s(x, y) = \frac{2 \times |\{p_{x \rightsquigarrow y} | p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightsquigarrow x} | p_{x \rightsquigarrow x} \in \mathcal{P}\}| + |\{p_{y \rightsquigarrow y} | p_{y \rightsquigarrow y} \in \mathcal{P}\}|}$$

here $p_{x \rightsquigarrow y}$ is path instance between x and y, and $p_{x \rightsquigarrow x}$ and $p_{y \rightsquigarrow y}$ are roundtrip path instances

- The salient features of PathSim
  - Symmetric: $s(x, y) = s(y, x)$
  - Normalized: $s(x, y) \in [0,1]$
  - Self-Maximized: $s(x, x) = 1$

PathSim: Illustration

The table below depicts the venue based publication frequency of some authors. To find the author most similar to Mike



Author	MOD	VLDB	ICDE	KDD
Mike	2	1	0	0
Jim	50	20	0	0
Mary	2	0	1	0
Bob	2	1	0	0
Ann	0	0	1	1

PathSim: Illustration

The visibility V_p of individual authors:

$$V_p(Mike) = 2 \times 2 + 1 \times 1 + 0 \times 0 + 0 \times 0 = 5$$

$$V_p(Jim) = 50 \times 50 + 20 \times 20 + 0 \times 0 + 0 \times 0 = 2900$$

$$V_p(Mary) = 2 \times 2 + 0 \times 0 + 1 \times 1 + 0 \times 0 = 5$$

$$V_p(Bob) = 2 \times 2 + 1 \times 1 + 0 \times 0 + 0 \times 0 = 5$$

$$V_p(Ann) = 0 \times 0 + 0 \times 0 + 1 \times 1 + 1 \times 1 = 2$$

The overall connectivity C_p between Mike and other authors are as follows:

$$C_p(Mike, Jim) = 2 \times 50 + 1 \times 20 + 0 \times 0 + 0 \times 0 = 120$$

$$C_p(Mike, Mary) = 2 \times 2 + 1 \times 0 + 0 \times 1 + 0 \times 0 = 4$$

$$C_p(Mike, Bob) = 2 \times 2 + 1 \times 1 + 0 \times 0 + 0 \times 0 = 5$$

$$C_p(Mike, Ann) = 2 \times 0 + 1 \times 0 + 0 \times 1 + 0 \times 1 = 0$$

PathSim: Illustration

Similarity scores in terms of V_p and C_p are as follows

$$s(Mike, Jim) = \frac{2 \times 120}{5 + 2900} = 0.0826$$

$$s(Mike, Mary) = \frac{2 \times 4}{5 + 5} = 0.8$$

$$\boldsymbol{s(Mike, Bob) = \frac{2 \times 5}{5 + 5} = 1.0}$$

$$s(Mike, Ann) = \frac{2 \times 0}{5 + 5} = 0.0$$

PathSim: Exercise

Consider a restaurant review network containing objects of two types : restaurant (R) and user (U). There exists a review (V) relationship between U and R as shown in below table, where each cell shows the number of reviews given by a user to a restaurant. The task is to find the peer restaurant for Mint.

Author	Michelle	Alice	Bob	Eve
Mint	2	4	0	0
Pavilion	4	0	2	1
Symposium	2	4	0	0
Sky Route	0	0	1	3