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.

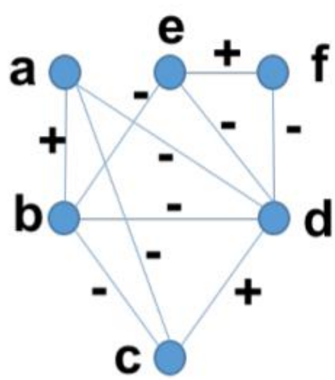
- Entity Ranking
  - Search Engine Optimization
  - Scientific article Ranking
  - Scientific Author Ranking, etc.

Why Link Analysis?

- Anomaly Detection
  - Online Fraud Detection
  - Counter Terrorism
  - Police/Military intelligence, etc.

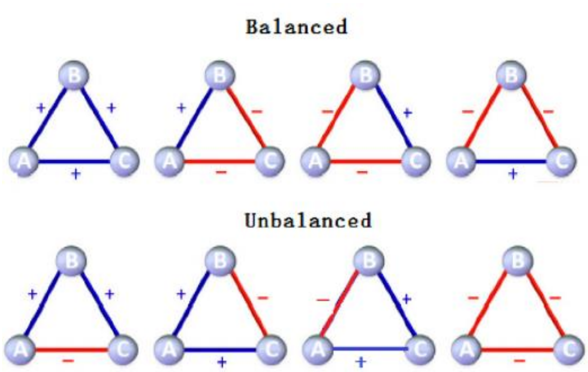
- Mining New Patterns
  - Crime Prevention
  - Future rank prediction
  - Link Prediction
  - Market Research, etc.

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

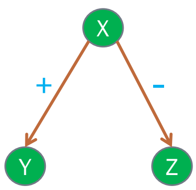
Balance Theory: Triads



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

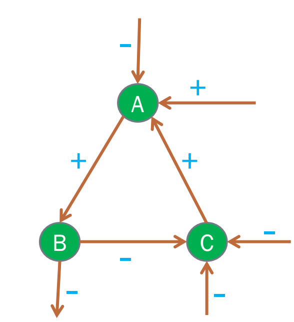
Signed Networks: Status Theory



$Y > X > Z$
Status relative to X

- Balance theory views signed links as model of likes and dislikes
- a signed link from 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 $B \Longrightarrow A$ considers B to have a higher status than itself
- A initiates a negative link to $B \Longrightarrow A$ considers B to have a lower status than itself

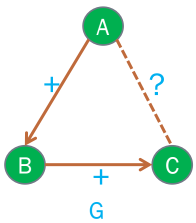
Signed Networks: Status Theory



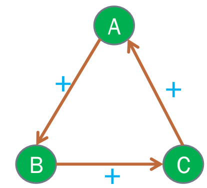
Snapshot of a signed graph

- 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
- 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:
  - $A_r = \frac{2}{3} = 0.67,$ $B_r = \frac{1}{1} = 1,$ $C_g = \frac{0}{3} = 0$

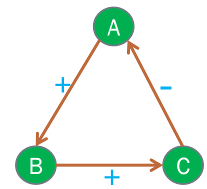
Comparison: Balance Theory and Status Theory



- ❑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

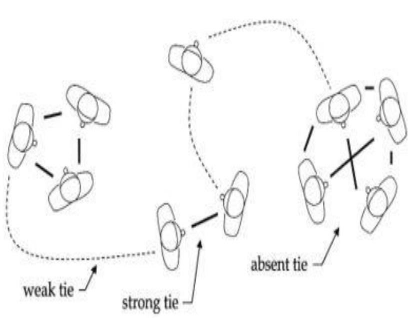


Balance Theory



Status Theory

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
- ❑ Granovetter studied the notion of strength and the impact of these ties on a network in 1973

Strength of a Tie

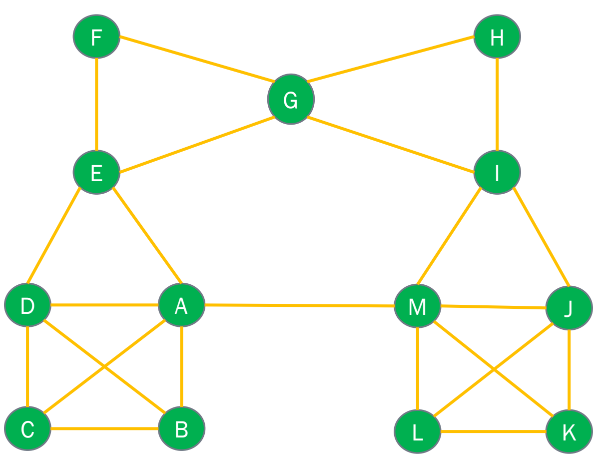
- strength of ties captures a sense of closeness among entities/people
- Simplest metric to capture the same is via Jaccard score
- Corresponding metric, called Neighborhood Overlap (NO) is defined as:

$$N(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

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

Neighborhood Overlap: Example



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

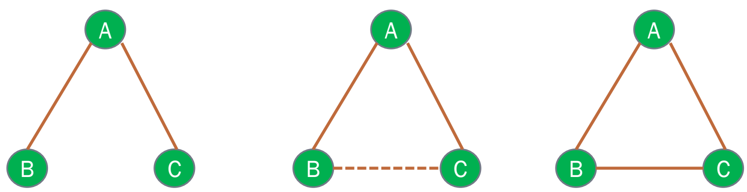
$|\Gamma(A) \cap \Gamma(M)| = |\phi| = 0$
$|\Gamma(A) \cap \Gamma(E)| = |\{D\}| = 1$

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

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

Triadic Closure

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

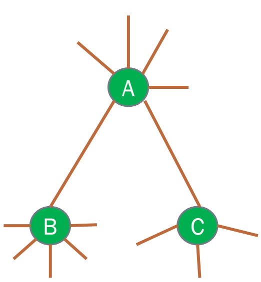


$t - \Delta t$ B & C are not friends yet | t B & C gets introduced (via A) | $t + \Delta t$ B & C are friends now

❑ 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

Quantifying Strength of Triadic Closures



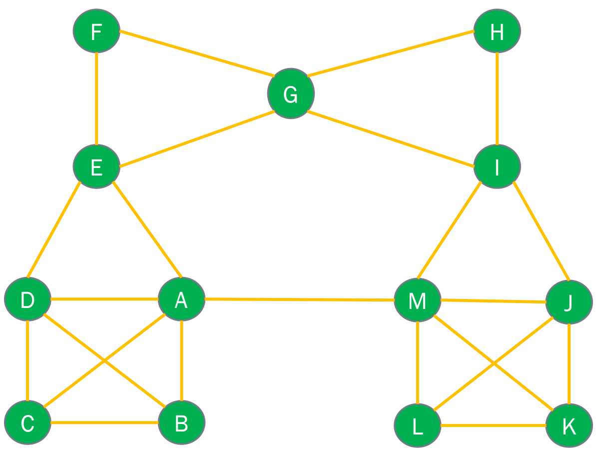
❑ Strength of a triadic closure with respect to node A and the nodes B and C of which A is a mutual friend can be quantified using the clustering coefficient of node A

❑ 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

Clustering Coefficient: Application



- 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 $\{BC, BD, CD, DE\}$

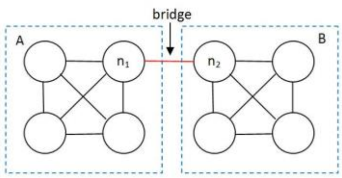
$CC_A = \frac{2\times4}{5\times4} = 0.4$

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

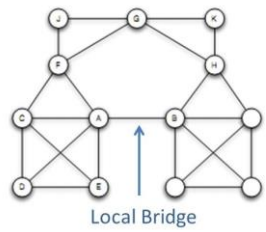
Dunbar Number

- An empirical study that supports the presence of strong and weak ties in real world
- Refers to a suggested cognitive limit to the number of people with whom one can maintain stable social relationships
- First proposed in 1990 by British anthropologist Robin Dunbar
- Observed a correlation between primate brain size and average social group size
- Which comes out to be 150
- The number informally represents the set of people one can be in close contact with (strong ties)
- Rest of the social contacts are likely to be acquaintances (weak ties)

Bridges and Local Bridges



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

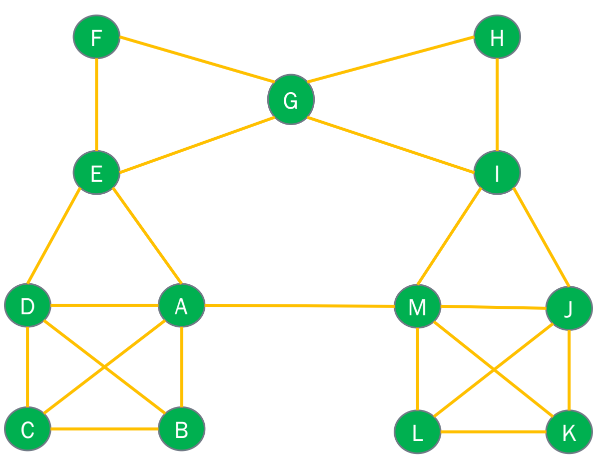


https://slideplayer.com/slide/9361256/

- 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
- (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

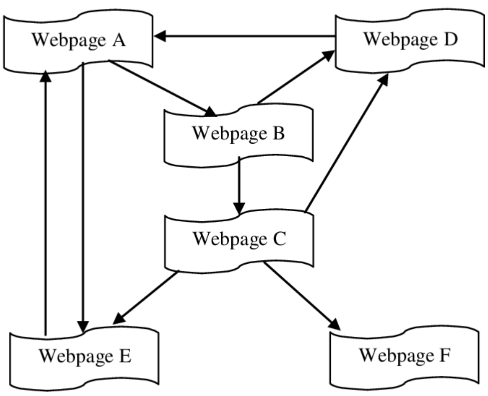
$$Embeddedness(\langle x, y \rangle) = |\Gamma(x) \cap \Gamma(y)|$$

- 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

PageRank: Intuition



Navadiya and Garg [2011]

- Outgoing hyperlink from a page is termed as out-edge or forward link
- Incoming hyperlink to a page from the second one is termed as an in-edge or backward link
- With every forward link a page establishes,
  - it transfers some of its importance/rank influence to the forward page
- If a highly important node points to a lesser important one,
  - there is an enhancement in the status of the latter node
- Importance of each node is determined by its in-edges/backward links

PageRank: Simple Ranking

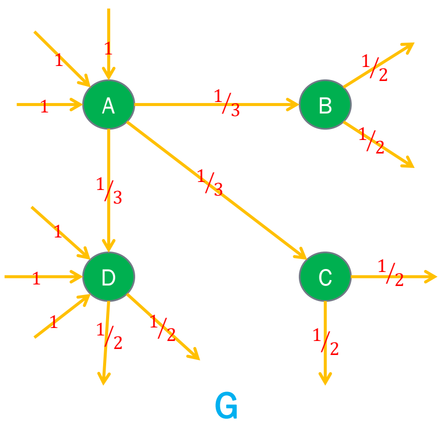
For a node w, let F_w be the set of nodes that w points to (Forward links) and B_w be the set of node that points to w (Backward links). Further, let $N_w = |F_w|$, the number of forward links from w. Then, the simple ranking of w, denoted $R(w)$, is given by

$$R(w) = \sum_{b \in B_w} \frac{R(b)}{N_b}$$

- The underlying web graph is assumed to be a connected component
- There could be pages that neither refer to any other page nor are referred to by any other page
- In a scenario where no hyperlinks exist in the network
  - Each page is assumed to be equally (un)important with a uniform rank given by

$$R(p) = \frac{1}{\#Webpages}$$

Simple PageRank: Illustration



Let us compute simple PageRank for the nodes in network G

$R(A) = 1 + 1 + 1 = 3$

$R(B) = \frac{1}{3}$

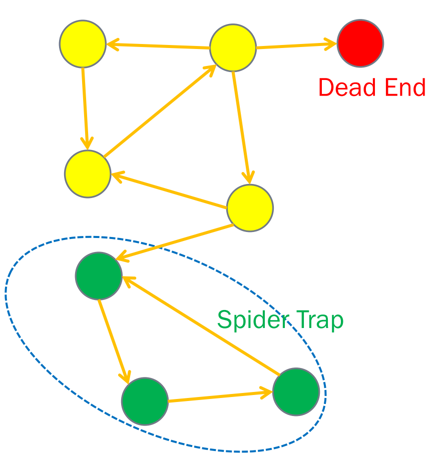
$R(C) = \frac{1}{3}$

$R(D) = \frac{1}{3} + 1 + 1 + 1 = \frac{10}{3}$

So, $PageRank_{raw} = [3, \frac{1}{3}, \frac{1}{3}, \frac{10}{3}]$

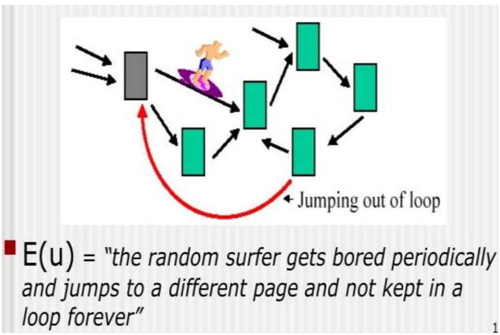
$PageRank_{normalized} = \frac{PageRank_{raw}}{Sum(PageRank_{raw})} = [0.4, 0.05, 0.05, 0.5]$

Simple PageRank: Drawbacks



- PageRank method follows a recursive approach
- Scores obtained at $(n-1)^{th}$ iteration is used as input scores at n^{th} iteration
- The above process stumbles at two extra-ordinary situations shown in the diagram
- Scores of nodes at dead ends does not impact rest of the nodes in the network
- The nodes forming a spider trap can revise their scores indefinitely without having any impact on the rest of the nodes in the network

PageRank: Random Surfer Model



https://www.slideserve.com/leroy-wright/the-pagerank-citation-ranking-bringing-order-to-the-web

1) Surfer starts a random page P_1 and moves to subsequent pages $P_2, P_3, \cdots\cdots\cdots, P_m$ in random order

2) Upon landing at a page P_i, the surfer choose either of the following
   a) With probability α, jump to random page P_j and repeat step 2. This random jump action is denoted $E = \frac{1}{N}$, where N is the number of pages in the network
   b) With probability $1 - \alpha$, it continues in its course of following hyperlinks

❖ the more number of times the surfer visits a node during the above random surfing, the higher the importance of the node

PageRank: Random Surfer Model

With the help of model and the analogy discussed here, the PageRank formulation is revised as:

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

$$\sum_{i=1}^{N} R(i) = 1$$

❑ The parameter α is a parameter that controls the balance between the importance of two components of the formulation above

❑ The random jump action is introduced in the revised PageRank method to deal with Dead Ends and Spider Traps in the network

PageRank: Matrix Representation

- A web graph of N webpages; A denotes the adjacency matrix for the web graph
- PageRank vector: $R = \langle r_1, r_2, r_3, \cdots, r_N \rangle$ with $0 \leq r_i \leq 1$
- Initial PageRank scores, R_0: $r_1 = r_2 = r_3 = \cdots = r_N = \frac{1}{N}$
- Normalize A to a stochastic matrix by setting: $A_{ij} = \frac{1}{N_i}$, where $N_i = |F_i|$
- For the first iteration, the initial PageRank score is R_0
- Then, the PageRank scores after the first iteration: $R_1 = R_0 A$
- Generalizing, the PageRank score can be obtained as: $R_{i+1} = R_i A$

PageRank: Matrix Formulation

- In order to include random jump in the above equation, we set $E = \langle e_1, e_2, e_3, \cdots, e_N \rangle$
- Since every page is equally probable to reach during random jump by the random surfer,

$$e_1 = e_2 = e_3 = \cdots = e_N = \frac{1}{N}$$

- Then, the updated PageRank equation is as follows:

$$R_{i+1} = (1 - \alpha) R_i A + \alpha E$$

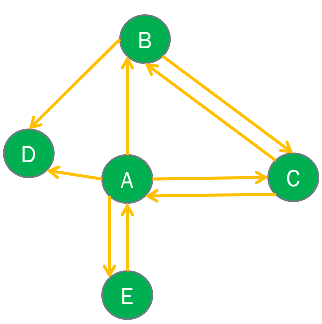
PageRank: Damping Factor

- PageRank theory centers around a random surfer who
  - randomly clicks on hyperlinks,
  - will eventually stop clicking, move to another random page, and
  - repeat the above sequence
- The damping factor d refers to the probability that the surfer continue random clicking the current chain of hyperlinks
- We usually set $d = 1 - \alpha$
- Then the revised PageRank formula:

$$R_{i+1} = \frac{1-d}{N} + dR_iA$$

- We may any value as damping factor; however, historically, it is often set as $\mathrm{d} = 0.85$

Revised PageRank: Illustration

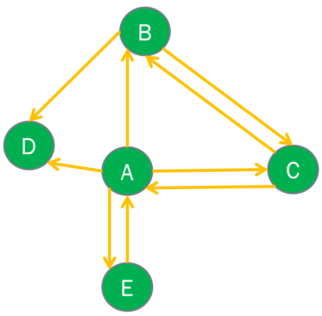


Let us find PageRank for the nodes in the graph

We set, $R_0 = \left[\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right]$, $\mathrm{E} = \left[\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right]$, and $d = 0.8$

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$, Normalized matrix, $$A = \begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Revised PageRank: Illustration



$R_0 A = [0.1, 0.1666, 0.1666, 0.3666, 0.2]$

$dR_0 A = [0.08, 0.1336, 0.1336, 0.2936, 0.16]$

$(1-d)E = [0.04, 0.04, 0.04, 0.04, 0.04]$

Thus, we have the updated PageRank scores after the first iteration as,

$R_1 = (1-d)E + dR_0 A = [0.12, 0.1736, 0.1736, 0.3336, 0.2]$

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:

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

Random Walks: Stationary Distribution

- Random walks over a network can be represented as a Markov Chain
  - Each page is state
  - Random walk defines a series of transitions from one state to another
- For a network of N webpages, the precomputed transition probabilities, $p_0 : N \times N \rightarrow [0,1]$, of the induced Markov chain above can be estimated as:

$$P_0(u, v) = (1 - d)p^*(v) + d \frac{w(u, v)}{\sum_{b \in F_u} w(u, b)}$$

$w(u, v)$ is the weight of the out-edge $\langle u, v \rangle$

$p^*(v)$ is the prior distribution of the vector E

Random Walks: Stationary Distribution

If $p_T(u)$ is the probability that the random surfer is at page u at iteration T, then the probability of its reaching node v at iteration $T + 1$ is obtained using the results of Markov chains as,

$$p_{T+1}(v) = \sum_{u \in B_v} (1 - d)p^*(v) + d.\, p_T(u) \frac{w(u, v)}{\sum_{b \in F_u} w(u, b)}$$

From the principle of time-homogeneous Markov chains, we have $p_t(u, v) = p_0(u, v)$, which yields

$$p_{T+1}(v) = \sum_{u \in B_v} p_0(u, v) p_T(u)$$

The above would converge when $p_{T+1}(v) \approx p_T(v)\ \forall v \in V$

In that case, it converges to a score $\pi(v)$, that provides the prestige of all the nodes.

PageRank: Advantages

- ❑Vectorized system of equations fast to compute
- ❑Guaranteed to converge to a unique solution
- ❑ranks can be pre-computed during indexing and re-used during query time
- ❑Ranks are robust and stable as in-edges to a page are harder to manipulate than out-edges
- ❑Conforms with the intuitive notion of importance of entities from the real world

PageRank: Disadvantages

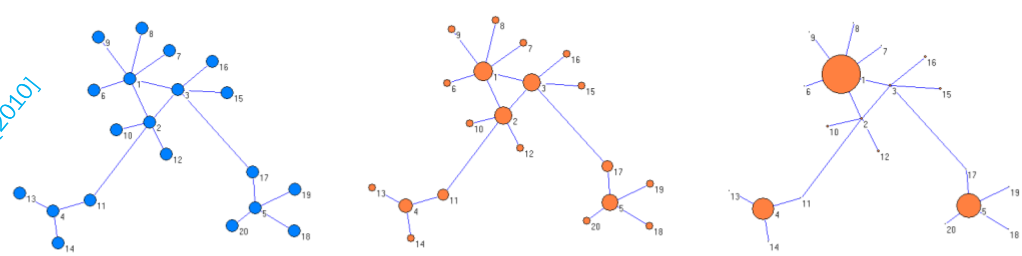
- ❑Prone to spamming, as it considers only the connections of node rather than its content
- ❑A page can get high rank by connecting a lot of trivial (possibly dummy) pages
- ❑Possibility of manipulation cannot be avoided completely, a malicious node could make hyperlinks with important pages and elevates its rank
- ❑The basic PageRank system assumes a static system; no modification in adjacency matrix allowed during computation
- ❑For dynamic systems, any modification requires all ranks to be re-computed
- ❑Formulation has been extended for dynamic networks

DivRank

- Top-ranked nodes in PageRank are often not diverse
  - 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

Mei et al. [2010]



(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

Vertex-Reinforced Random Walks are random walks where the transition probability from one state to the next $p_T(u,v) \longrightarrow p_{T+1}(u,v)$ is reinforced by the number of previous visits to the state $N_T(v)$; i.e., $p_T(u,v) \propto p_0(u,v).N_T(v)$

DivRank: Random Walk Formulation

- The organic and precomputed transition probabilities

$$p_0(u,v) = \begin{cases} \alpha \dfrac{w(u,v)}{\sum_{b \in F_u} w(u,b)} & u \neq v \\ 1-\alpha & u = v \end{cases}$$

Here α would capture whether the random walk will follow one of the neighbors or choose to stay at the current state/node

- At a given timestamp, there is a chance that the surfer stays at the node
- The probability of the above is reinforced by the number of previous visits at the current node

DivRank: Random Walk Formulation

- Then the overall transition probability:

$$p_T(u,v) = (1-d)p^*(v) + d.\frac{p_0(u,v).N_T(v)}{\sum_{b\in F_u} p_0(u,b).N_T(b)}$$

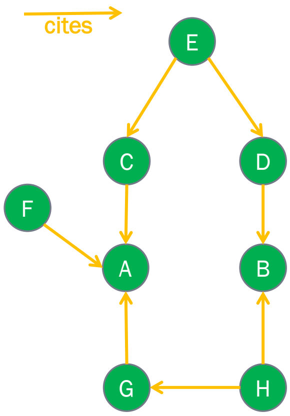
Then the overall probability of the random surfer to move to node v at time $T+1$, from one of its neighbors B_v, can be obtained as

$$p_{T+1}(v) = (1-d)p^*(v) + \sum_{u\in B_v} d.p_T(u)\frac{p_0(u,v).N_T(v)}{\sum_{b\in F_u} p_0(u,b).N_T(b)}$$

Measuring Similarity of Objects

- Metadata used to measure similarity between objects are often hard to determine and quantify in practice
- Contextual information may be used for the purpose
  - Two objects are similar if they are related to similar objects
  - Easier to determine in practice
- SimRank follows the above paradigm to measure similarity between entities
  - For a network of size N, we require N^2 similarity score, one per each pair of objects
  - For the same network, a score like PageRank or DivRank would form a list of length N.

SimRank: Measuring Similarity of Objects



- 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?
  - $\Gamma(A) = \{C, F, G\}$ and $\Gamma(B) = \{D, H\}$
  - SimRank can answer such question

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:

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

- 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: Naïve Solution

- An iterative solution for SimRank is as follows:

$$R_0(a,b) = \begin{cases} 0 & if\ a \neq b \\ 1 & if\ a = b \end{cases}$$

and

$$R_{k+1}(a,b) = \begin{cases} 1 & if\ a = b \\ \dfrac{C}{|I(a)|.|I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k(I_i(a), I_j(b)\} & if\ a \neq b \end{cases}$$

with

$$\lim_{k\to\infty} R_k(a,b) = s(a,b)$$

SimRank in Heterogeneous Bipartite Network

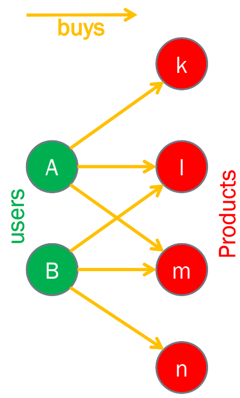
- In a heterogeneous network of users and products, the similarity of products and users are mutually-reinforced
  - two users can be considered similar if they buy similar products
  - two products can be considered similar if they are bought by similar users
- Similarity between two distinct users 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 $O(B) = \{l, m, n\}$

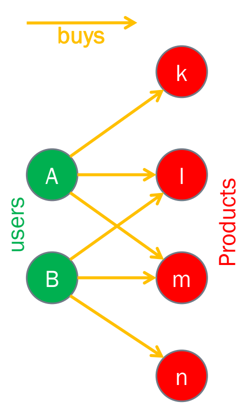
$I(k) = \{A\}$, $I(l) = \{A, B\}$, $I(m) = \{A, B\}$, and $I(n) = \{B\}$

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

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 $C_1 = C_2 = 0.8$,

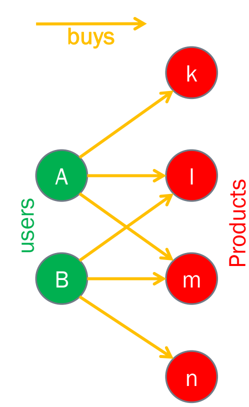
$$\boldsymbol{s(A,B) = 0.547}$$

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
  - $\varphi(v): V \rightarrow \mathcal{A}$ maps each vertex to a node type
  - $\psi(e): E \rightarrow \mathcal{R}$ 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

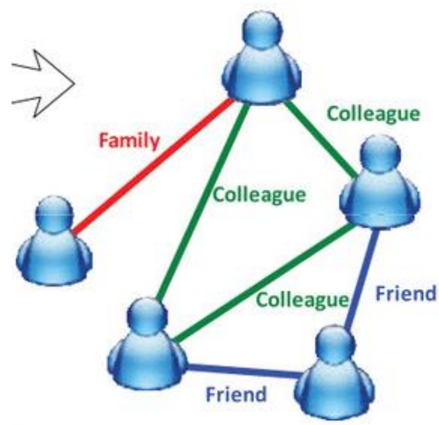
- When $|\mathcal{A}| > 1$ and $|\mathcal{R}| = 1$, then we have a heterogeneous network consisting of vertices of more than one types, and only one types of links
- A typical example is consumer-product purchase network, where
  - $\mathcal{A} = \{users, products\}$, and
  - $\mathcal{R} = \{user \longrightarrow products | user\ buys\ product\}$



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.

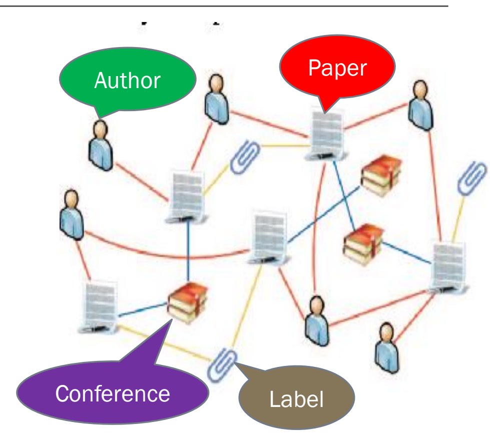


Liu and Yu [2019]

Heterogeneous Networks: Variants

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

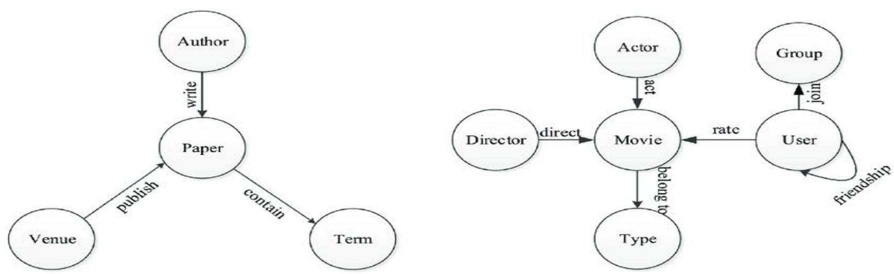
❑A typical bibliographic network consisting of authors, papers, conference venues, etc., and various kinds of relationship between these entities



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

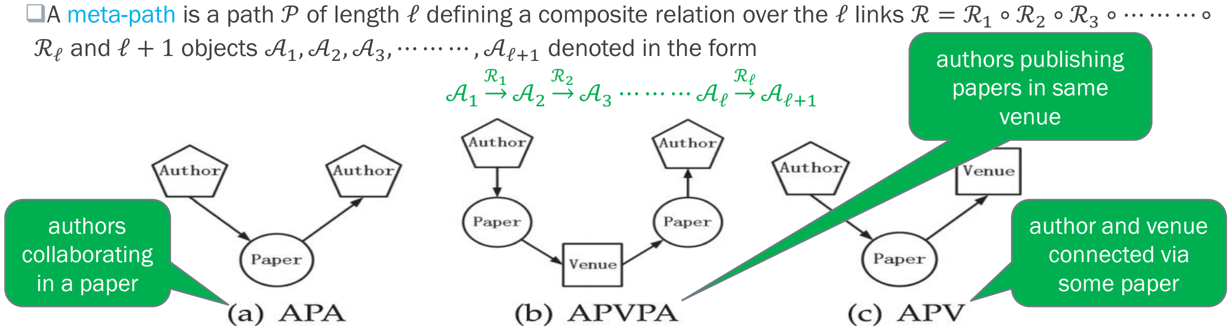
Heterogeneous Networks: Meta-Path

❑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

Object Similarity via Meta-Path

- Path Count: It indicates the number of path instances p of $\mathcal{P}_\ell$, which begin at x and end at y. The similarity score is

$$s(x, y) = |\{p \in \mathcal{P}_\ell | x \in \mathcal{A}_1, y \in \mathcal{A}_{\ell+1}\}|$$

- Random Walk: For a random surfer starting at x and following the path $\mathcal{P}_\ell$, what is the probability of it ending at y

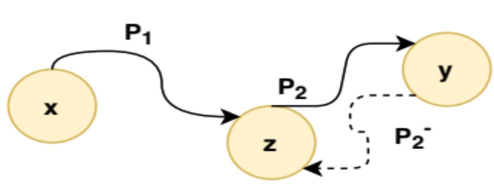
$$s(x, y) = \sum_{p \in \mathcal{P}_\ell} Prob(p)$$

Object Similarity via Meta-Path

- Pairwise Random Walk: For a concatenated meta-path $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2)$ with instances starting at x and y, if we reverse the second sub-path to have two sets of random walkers starting at x and y and reaching a mid-point z, it forms a valid instance as $(x \rightarrow z \leftarrow y)$. Here, the similarity score is given by

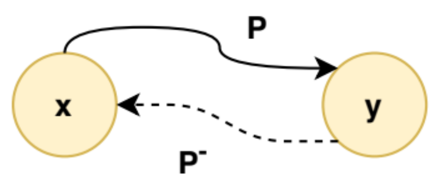
$$s(x, y|z) = \sum_{p_1 \in \mathcal{P}_1, p_2^- \in \mathcal{P}_2^-} Prob(p_1). Prob(p_2^-)$$

here p^- is the reverse path instance of the path p



PathSim: Formulation

- A measure of similarity search scoring and ranking in heterogeneous information networks
- Use the notion of meta-paths for the formulation
- A meta-path of the form $\mathcal{P} = (\mathcal{P}_\ell, \mathcal{P}_\ell^-)$ where the starting and ending object is the same, is termed as a round-trip meta-path. By default, it is always symmetric.



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:

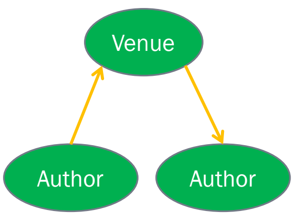
$$s(x,y) = \frac{2 \times |\{p_{x \rightsquigarrow y} | p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightsquigarrow x} | p_{x \rightsquigarrow x} \in \mathcal{P}\}| + |\{p_{y \rightsquigarrow y} | p_{y \rightsquigarrow y} \in \mathcal{P}\}|}$$

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

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

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

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

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

END