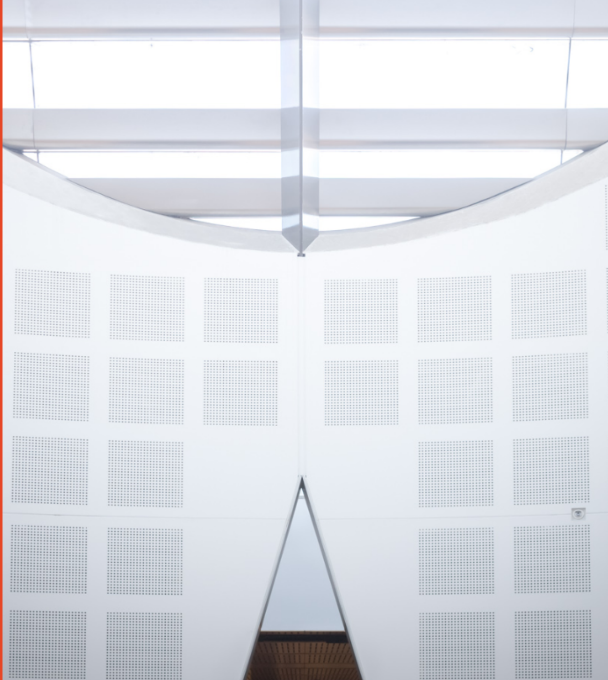
COMP5313/COMP4313 - Large Scale Networks

Week 6: PageRank and Its Generalizations

Lijun Chang

April 3, 2025

THE UNIVERSITY OF SYDNEY



Introduction

- In the previous lecture, we have talked about:
  - Hubs whose scores depend on the authority scores of the nodes they point to
  - Authorities whose scores depend on the hub scores of the nodes pointing to them
  - In the *optional* advanced material, we show that the vector of limiting hub scores $\mathbf{h}^*$ is the eigenvector, corresponding to the largest eigenvalue, of the matrix $\mathbf{A}\mathbf{A}^\top$
  - Nodes play multiple roles (i.e., hub and authority) in the network
    - Nodes (e.g., hubs) can play an important endorsement role without being heavily endorsed

- Now, we will see that
  - There is only one role
  - Endorsement is viewed as passing directly from one prominent node to another

Outline

PageRank

Scaled PageRank

Personalized PageRank

Generalized PageRank (Optional)

Hubs-and-authorities vs. PageRank

- The hubs-and-authorities analysis are used in settings where prominent (i.e., authority) pages do not directly link to each other
  - E.g., competitor companies may not endorse each other
  - The only way to conceptually pull them together is through a set of hub pages that link to all of them at once
- But most of the time, prominent pages are directly endorsed by other prominent pages:
  - Academic pages
  - Government pages
  - Bloggers
  - Personal pages
  - Scientific literature

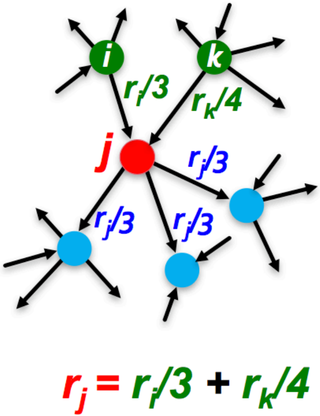
PageRank: Links as Votes

- This form of endorsement is at the heart of the PageRank. [1]
  - Votes and repeated improvements are used to determine the PageRank of a page
  - Page is more important if it has more links
    - In-coming links or out-going links?
    - Think of in-links as votes (endorsements)
  - Are all in-links equal?
    - Links from important pages count more
    - Recursive definition!
  - That is, endorsement is passed through outgoing links with a weight that is proportional to the current PageRank of the source page

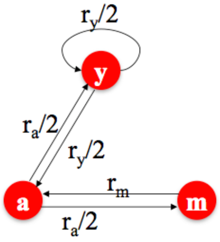
[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, vol. 30, no. 1–7, 1998.

PageRank: the "flow" model

- Define the importance of page i as r_i
- A page is important if it is pointed to by other important pages
  - Each link's vote is proportional to the importance of its source page
  - If page i with importance r_i has d_i out-links, each link gets $\frac{r_i}{d_i}$ votes
  - Page j's own importance r_j is the sum of the votes on its in-links $r_j = \sum_{i \in V: i \to j} \frac{r_i}{d_i}$



PageRank: the "flow" model



"Flow" equations:

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

PageRank Analogy

The PageRank can be viewed as a fluid circulating through the network and pooling at the nodes that are the most important

You might wonder: Let's just use Gaussian elimination to solve this system of linear equations. Bad idea due to high time complexity!

PageRank: Basic Update Rule

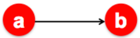
- PageRank is computed as follows:
  - In a network with n nodes, we assign all nodes the same initial PageRank, $\frac{1}{n}$.
  - We choose a number of steps k.
  - We then perform a sequence of k updates to the PageRank values, using the following rule for each update:

    Basic PageRank Update Rule: Each page divides its current PageRank equally across its outgoing links, and passes these equal shares to the pages it points to. (If a page has no outgoing links, it passes all its current PageRank to itself.) Each page updates its new PageRank to be the sum of the shares it receives.

- Since each page divides its PageRank among its outgoing link, there is no need to normalize it, the total PageRank in the network is constant.

The "Dead End" Problem

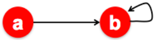
- In the basic PageRank update rule, why we need the rule "If a page has no outgoing links, it passes all its current PageRank to itself"?
- Consider the graph



  – The total PageRank value in the network would decrease

Step	0	1	2	3
r_a	$1/2$	0	0	0
r_b	$1/2$	$1/2$	0	0

- The so-called remedy

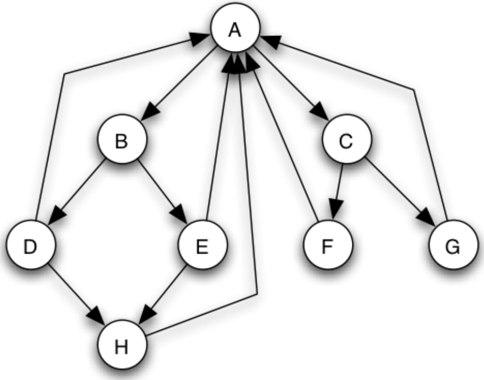


  – Ensures the total PageRank value in the network is constant

Example

- A collection of eight pages: A has the largest PageRank, followed by B and C (which collect endorsements from A).

- Let's consider how this computation works on this collection of eight pages.



Example

- All pages start out with a PageRank of $\frac{1}{8}$ and their PageRank values after the first two updates are given by the following table:

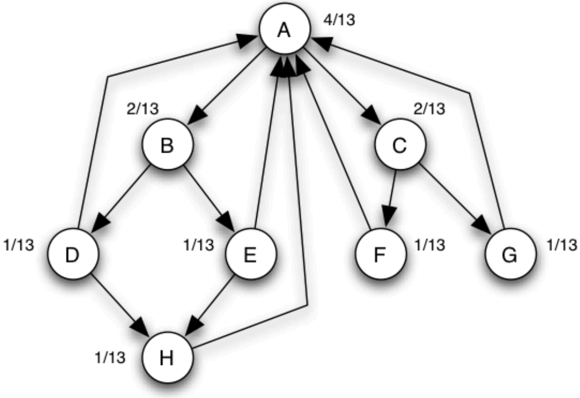
Step	r_A	r_B	r_C	r_D	r_E	r_F	r_G	r_H
1	$1/2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/8$
2	$5/16$	$1/4$	$1/4$	$1/32$	$1/32$	$1/32$	$1/32$	$1/16$

- For example, A gets PageRank of $\frac{1}{2}$ after the first update because it gets all of F's, G's, and H's PageRank, and half each of D's and E's. On the other hand, B and C each get half of A's PageRank, so they only get $\frac{1}{16}$ each in the first step.
- This is in keeping with the principle of repeated improvement: the first update causes us to estimate that A is an important page, we weigh its endorsement more highly in the next update.

Equilibrium Values of PageRank

- A state is an equilibrium if applying the Basic PageRank Update Rule does not update anything

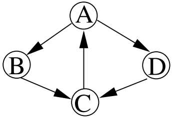
- Equilibrium PageRank values for the network of eight Web pages
  - Assigning a PageRank of $\frac{4}{13}$ to page A, $\frac{2}{13}$ to each B and C, and $\frac{1}{13}$ to the five other pages achieves the equilibrium



Equilibrium Values of PageRank

- If the network is strongly connected, then there is a unique set of equilibrium values

- But the iterative algorithm may not be able to obtain the equilibrim PageRank values (e.g., consider the following graph)



- Under reasonable assumptions, the PageRank values of all nodes, as computed by the iterative algorithm, converge to limiting values when the number of update steps, k, goes to infinity

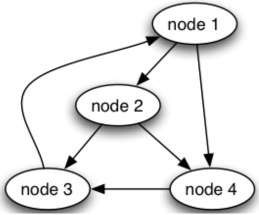
Equilibrium Values of PageRank in Undirected Graphs

- For undirected graphs, there is a natural equilibrium $r_i = \frac{d_i}{2m}$.
  - m is the number of undirected edges.
- For connected and non-bipartite undirected graph, iteratively applying the Basic PageRank Update Rule will converge to limiting value (i.e., equilibrium), regardless of initialization.
  - As long as the initial values are non-negative and sum to one.

Spectral Analysis

Let's analyze PageRank using matrix-vector multiplication and eigenvectors

- Under the Basic Update Rule, each node takes its PageRank and divides it equally over all the nodes to which it points
- The flow of PageRank can be represented using the row-stochastic matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$: $W_{i,j}$ is the share of i's PageRank that j should get in one update step. $\mathbf{W} = \mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix with $D_{i,i} = d_i$.
  - $W_{i,j} = 0$ if i has no link to j
  - $W_{i,j}$ is reciprocal of the number of nodes i points to, otherwise
  - $W_{i,i} = 1$ if i has no outgoing link (it passes its PageRank to itself)



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Spectral Analysis

Let's represent the PageRank of all nodes using a row vector $\mathbf{r} \in \mathbb{R}^{1\times n}$

- The coordinate r_i is the PageRank of node i
- The Basic PageRank Update Rule becomes:

$$r_i^{(t+1)} = r_1^{(t)} \times W_{1,i} + r_2^{(t)} \times W_{2,i} + \cdots + r_n^{(t)} \times W_{n,i}$$

  This corresponds to multiplication on the right by the matrix $\mathbf{W}$, just as we saw for the Authority Update Rule. Thus, this equation can be rewritten as:

$$\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)}\mathbf{W}$$

- $\mathbf{r}^{(k)} = \mathbf{r}^{(0)}\mathbf{W}^k = \frac{1}{n}\mathbf{e}\mathbf{W}^k$. Here $\mathbf{e} = (1, 1, \ldots, 1) \in \mathbb{R}^{1\times n}$.

- The equilibrium PageRank should satisfy

$$\mathbf{r}^* = \mathbf{r}^*\mathbf{W}$$

Random Walks

- The PageRank can be equivalently formulated using a *random walk*
- Consider someone who is randomly browsing a network of Web pages
  - She starts by choosing a page at random, picking each page with the same probability.
  - Then, she follows links for a sequence of k steps
  - In each step, she picks a random outgoing link from the current page and follows it to where it leads
    - If the current page has no outgoing link, she stays where she is.
- This exploration of the network is called a random walk on the network

Claim

The probability of being at page X after k steps of this random walk is precisely the PageRank of X after k applications of the Basic PageRank Update Rule.

Outline

PageRank

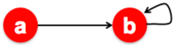
Scaled PageRank

Personalized PageRank

Generalized PageRank (Optional)

The "Spider Trap" Problem

- In many networks the wrong nodes can end up with all the PageRank

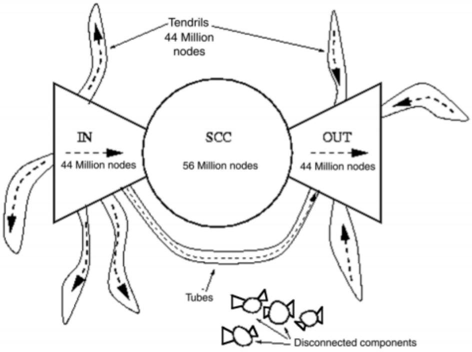


Step	0	1	2	3
r_a	$1/2$	0	0	0
r_b	$1/2$	1	1	1

- As long as there are a small set of nodes that can be reached from the network but do not have any path back to the network, then PageRank will accumulate there

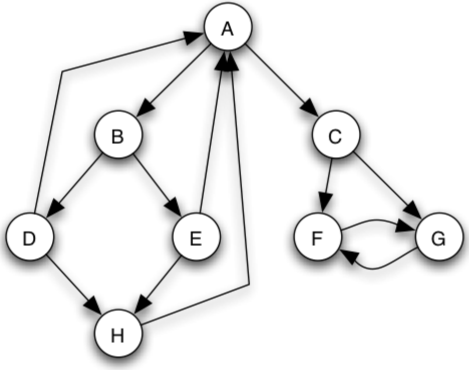
The "Spider Trap" Problem

- This problem actually exists, given the bow-tie structure of the Web
  - There is one giant strongly connected component (SCC)
  - All the PageRank would accumulate at the end of the downstream (OUT) nodes



The "Spider Trap" Problem

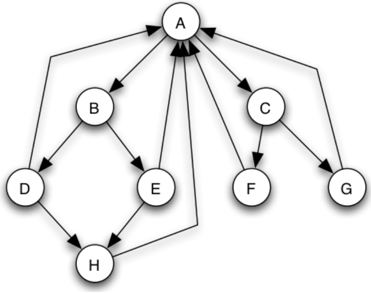
- The same collection of eight pages, but F and G have changed their links to point to each other instead of A. Then, all the PageRank would go to F and G.
- PageRank that flows from C to F and G can never circulate back into the rest of the network
- There is a kind of slow leak that causes all the PageRank to end up at F and G. We converge to $\frac{1}{2}$ for F and G and 0 for others.



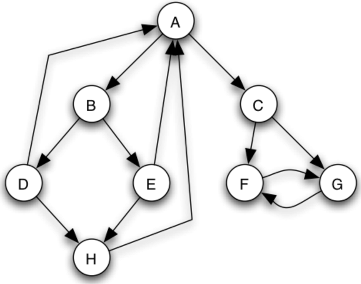
The "Spider Trap" Problem

Claim

The spider trap problem occurs if and only if the graph is not strongly connected.



Strongly connected



Not strongly connected

Scaled PageRank Update Rule (Power Iteration/Method)

- We can solve this problem similarly to the observation that rain water does not converge to the same lowest points due to a counter-balancing process of evaporation and rains on the highest points
- The idea is to pick a scaling factor α strictly between 0 and 1 and replace the Basic PageRank Update Rule by the following:

  Scaled PageRank Update Rule: First apply the Basic PageRank Update Rule. Then scale down all PageRank values by a factor of α. This means that the total PageRank in the network has shrunk from 1 to α. We divide the residual $1 - \alpha$ units of PageRank equally over all nodes, giving $\frac{1-\alpha}{n}$ to each node.
- The scaled PageRank update rule also preserves the PageRank of the network since it is based on redistribution according to a water cycle that evaporates $1 - \alpha$ units of PageRank in each step and rains it down uniformly across all nodes.

Spectral Analysis

The scaled update rule can also be represented using matrix-vector multiplication

- In the scaled PageRank update rule, the updated PageRank is scaled down by a factor of α and the residual $1-\alpha$ units are divided equally over all nodes.
  - $r_i^{(t+1)} = \alpha \times \left(r_1^{(t)} \times W_{1,i} + r_2^{(t)} \times W_{2,i} + \cdots + r_n^{(t)} \times W_{n,i}\right) + \frac{1-\alpha}{n}$

$$r_i^{(t+1)} = \left(r_1^{(t)} \times \alpha W_{1,i} + r_2^{(t)} \times \alpha W_{2,i} + \cdots + r_n^{(t)} \times \alpha W_{n,i}\right) + \left(r_1^{(t)} \times \frac{1-\alpha}{n} + r_2^{(t)} \times \frac{1-\alpha}{n} + \cdots + r_n^{(t)} \times \frac{1-\alpha}{n}\right)$$

Note that, this works only when the PageRank values sum to 1.

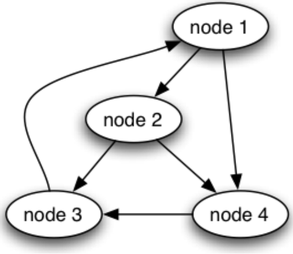
- We can define $\tilde{W}_{i,j}$ to be $\alpha W_{i,j} + \frac{1-\alpha}{n}$ and then the scaled PageRank update rule can be written:

$$r_i^{(t+1)} = r_1^{(t)} \times \tilde{W}_{1,i} + r_2^{(t)} \times \tilde{W}_{2,i} + \cdots + r_n^{(t)} \times \tilde{W}_{n,i}$$

Or equivalently:

$$\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)}\tilde{\mathbf{W}}$$

Spectral Analysis



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

- The flow of PageRank under the scaled PageRank update rule represented with matrix $\tilde{\mathbf{W}}$ with scaling factor $\alpha = 0.8$
- The entry $\tilde{W}_{i,j}$ specifies the portion of i's PageRank that should be passed to j in one update

Spectral Analysis

- *The scaled PageRank update rule virtually makes an input graph a complete graph by adding a link from each node to every other node. Therefore, the resulting graph is strongly connected.*

Spectral Analysis

Recall that the scaled PageRank update rule can be written as $\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)}\tilde{\mathbf{W}}$ where $\tilde{W}_{i,j} = \alpha W_{i,j} + \frac{1-\alpha}{n}$

- This is equivalent to $\mathbf{r}^{(t+1)} = \alpha \mathbf{r}^{(t)}\mathbf{W} + \frac{1-\alpha}{n}\mathbf{e}$ as $\tilde{\mathbf{W}} = \alpha \mathbf{W} + \frac{1-\alpha}{n}\mathbf{e}^\top \mathbf{e}$.
  - In the scaled PageRank update rule, the updated PageRank is scaled down by a factor of α and the residual $1-\alpha$ units are divided equally over all nodes.

- The equilibrium PageRank should satisfy

$$\mathbf{r}^* = \alpha \mathbf{r}^* \mathbf{W} + \frac{1-\alpha}{n}\mathbf{e}$$

Spectral Analysis

$\mathbf{r}^* = \alpha \mathbf{r}^* \mathbf{W} + \frac{1-\alpha}{n}\mathbf{e}$ has a unique solution $\mathbf{r}^* = \frac{1-\alpha}{n}\mathbf{e}(\mathbf{I} - \alpha\mathbf{W})^{-1}$ when $0 < \alpha < 1$, where $\mathbf{I}$ is the identity matrix.

- The matrix $\mathbf{I} - \alpha\mathbf{W}$ is invertible, since it is strictly diagonally dominate.
- $\mathbf{r}^* = \frac{1-\alpha}{n}\mathbf{e}(\mathbf{I} - \alpha\mathbf{W})^{-1} = \frac{1-\alpha}{n}\mathbf{e}\sum_{\ell=0}^{\infty} \alpha^\ell \mathbf{W}^\ell$.
- By induction, k steps of scaled PageRank update rule obtain

$$\mathbf{r}^{(k)} = \mathbf{r}^{(0)}\alpha^k\mathbf{W}^k + \frac{1-\alpha}{n}\mathbf{e}\sum_{\ell=0}^{k-1} \alpha^\ell \mathbf{W}^\ell \geq \frac{1-\alpha}{n}\mathbf{e}\sum_{\ell=0}^{k-1} \alpha^\ell \mathbf{W}^\ell$$

  - If the ℓ_1-norm of $\mathbf{r}^{(0)}$, $\|\mathbf{r}^{(0)}\|_1$, is 1, then $\|\mathbf{r}^{(k)}\|_1 = 1$.
  - If $\|\mathbf{r}^{(0)}\|_1 = 0$, then $\|\mathbf{r}^* - \mathbf{r}^{(k)}\|_1 = \alpha^k$.

The Limit of the Scaled PageRank Update Rule

- Repeatedly applying the scaled PageRank update rule converges to a set of limiting PageRank values $\frac{1-\alpha}{n}\mathbf{e}(\mathbf{I} - \alpha\mathbf{W})^{-1}$ as the number of updates, k, goes to infinity
  - These limiting values form the unique equilibrium for the scaled PageRank update rule
- But, these limiting values depend on the value of α
  - In effect, we can consider the update rule for each value of α as a different rule
  - In practice, PageRank uses the rule with a scaling factor α between 0.8 and 0.9

The Random Walk Perspective

Recall that the scaled PageRank update rule is $\mathbf{r}^{(t+1)} = \alpha \mathbf{r}^{(t)} \mathbf{W} + (1-\alpha)\frac{1}{n}\mathbf{e}$

- At each time step, the random surfer has two options
  - With probability α, follow a link at random
  - With probability $1-\alpha$, jump to a random page

Outline

PageRank

Scaled PageRank

Personalized PageRank

Generalized PageRank (Optional)

Personalized PageRank

- (Scaled) PageRank is used as a network centrality measure
  - Yields the importance of each node in light of the entire graph structure
  - At each time step, the random surfer has two options
    - With probability α, follow a link at random
    - With probability $1 - \alpha$, jump to a random page among all pages

- Personalized PageRank is used to illuminate a region of a large graph around a target set S of interest.
  - At each time step, the random surfer has two options
    - With probability α, follow a link at random
    - With probability $1 - \alpha$, jump to a random page among a set S of pre-selected pages

Revisit the Spectral Analysis of PageRank

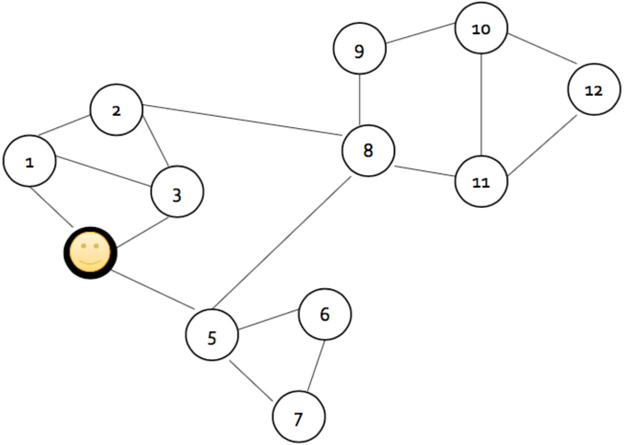
- In (Scaled) PageRank $\mathbf{r}^{(t+1)} = \alpha \mathbf{r}^{(t)} \mathbf{W} + (1-\alpha)\mathbf{s}$, where $\mathbf{s} = \frac{1}{n}\mathbf{e} = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$.

- Personalized PageRank is customizing $\mathbf{s}$, which represents the preference of a user
  - E.g., to compute the relevance/importance of all nodes to nodes 1 and 2, we can use $\mathbf{s} = (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$
  - Equilibrium: $\mathbf{r}^*_{\mathbf{s},\alpha} = (1-\alpha)\mathbf{s}(\mathbf{I} - \alpha\mathbf{W})^{-1} = (1-\alpha)\mathbf{s}\sum_{\ell=0}^{\infty} \alpha^{\ell}\mathbf{W}^{\ell}$.
  - $\mathbf{r}^{(0)}$ is typically also set as $\mathbf{s}$, and then

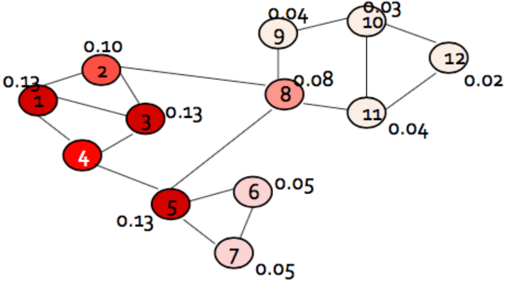
$$\mathbf{r}^{(k)} = \mathbf{s}\alpha^k \mathbf{W}^k + (1-\alpha)\mathbf{s}\sum_{\ell=0}^{k-1} \alpha^{\ell}\mathbf{W}^{\ell}$$

- $(1-\alpha)(\mathbf{I} - \alpha\mathbf{W})^{-1}$ is called the PPR matrix.

Example



Example



	Node 4
Node 1	0.13
Node 2	0.10
Node 3	0.13
Node 4	/
Node 5	0.13
Node 6	0.05
Node 7	0.05
Node 8	0.08
Node 9	0.04
Node 10	0.03
Node 11	0.04
Node 12	0.02

Nearby nodes, higher scores

More red, more relevant

Ranking vector

A Different Random Walk Interpretation (Optional)

- Let $(X_0, X_1, \ldots, X_L)$ be a random walk that starts from $X_0 = s$ and is of length $L \sim \text{Geometric}(\alpha)$, i.e., $\text{prob}(L = \ell) = (1-\alpha)\alpha^{\ell}$.
  - The random walk starts at node s and does the following at each step
    - with probability α, move to a random neighbor
    - with probability $1-\alpha$, terminate at the current node
  - Then, $r_s(t) = \text{prob}(X_L = t)$.
- We can do ω such random walks, and estimate $r_s(t)$ as the fraction of random walks that end at t
  - By Chernoff bound, for $\omega \geq \frac{(2\epsilon/3+2)\ln(2/\delta)}{\epsilon^2\theta}$ and a fixed t with $r_s(t) \geq \theta$, with probability at least $1-\delta$ the relative error of the estimation is at most ϵ.

Forward Push (Optional) [2]

Algorithm: ForwardPush$(\boldsymbol{s}, \alpha, q_{\max})$

$\boldsymbol{p} \leftarrow 0$; $\boldsymbol{q} \leftarrow \boldsymbol{s}$;
while $\exists u, q(u) \geq q_{\max} d(u)$ **do**
    Pick any vertex u where $q(u) \geq q_{\max} d(u)$;
    Push(u);
return $(\boldsymbol{p}, \boldsymbol{q})$;

Procedure Push(u)
$p(u) \leftarrow p(u) + (1-\alpha) q(u)$;
for each $v \in N(u)$ **do** $q(v) \leftarrow q(v) + \alpha \frac{q(u)}{d(u)}$;
$q(u) \leftarrow 0$;

- $N(u)$ is the set of (out)-neighbors of u
- $r_s(t) = p(t) + \sum_{v \in V} q(v) r_v(t), \forall t \in V$
- $\|\mathbf{r}_s - \mathbf{p}\|_1 = \sum_{v \in V} q(v) < q_{\max} \cdot m$
- The time complexity of ForwardPush$(\mathbf{s}, \alpha, q_{\max})$ is $O(\frac{\|\mathbf{s}\|_1}{(1-\alpha) \cdot q_{\max}})$.

[2] Efficient Algorithms for Personalized PageRank Computation: A Survey. TKDE, 2024

Outline

PageRank

Scaled PageRank

Personalized PageRank

Generalized PageRank (Optional)

Generalized PageRanks

Recall (personalized) PageRank is $\mathbf{r}^*_{\mathbf{s},\alpha} = \sum_{\ell=0}^{\infty}(1-\alpha)\alpha^\ell \mathbf{s}\mathbf{W}^\ell$

- (Personalized) PageRank uses a geometric distribution of random walk lengths, assigning weight $(1-\alpha)\alpha^\ell$ to walks of length ℓ for some parameter α.
  - Note that $\sum_{\ell=0}^{\infty}(1-\alpha)\alpha^\ell = 1$.
- Heat Kernel uses Poisson distribution to assign weight $e^{-\alpha}\frac{\alpha^\ell}{\ell!}$ to walks of length ℓ.
  - $\mathbf{r}^*_{\mathbf{s},\alpha} = \sum_{\ell=0}^{\infty} e^{-\alpha}\frac{\alpha^\ell}{\ell!}\mathbf{s}\mathbf{W}^\ell$
- We can also use (or learn) more general weights $\{\gamma_\ell\}_{\ell\geq 0}$ such that $\mathbf{r}^*_{\mathbf{s}} = \sum_{\ell=0}^{\infty}\gamma_\ell \mathbf{s}\mathbf{W}^\ell$. More information can be found at [3] and [4]

---

[3] Pan Li, I Chien, and Olgica Milenkovic. Optimizing generalized pagerank methods for seed-expansion community detection. In *Advances in Neural Information Processing Systems*, pp. 11705–11716, 2019.

[4] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive Universal Generalized PageRank Graph Neural Network. In *International Conference on Learning Representations (ICLR)*, 2021.

PageRank beyond the Web

- Personalized PageRank as a relevance/proximity measure is more robust than others, e.g., shortest distance, #disjoint paths
- (Personalized) PageRank applications [5]
  - Chemistry
  - Biology: GeneRank, ProteinRank, IsoRank
  - Neuroscience
  - Engineered systems: MonitorRank
  - Mathematical systems
  - Sports
  - Literature: BookRank
  - Bibliometrics: CiteRank, AuthorRank
  - Database & knowledge systems: PopRank, FactRank, ObjectRank, FolkRank
  - Recommender systems: ItemRank
  - Social networks: BuddyRank

[5] D. F. Gleich. Pagerank beyond the web. SIAM Review, vol. 57, no. 3, 2015.

Conclusion

- Link analysis of networks is important
  - Relies on
    - Hubs and authorities
    - Directly passing of endorsement
    - Repeated improvement technique
  - Allows to rank pages, journals, cases ... (any information that is networked)
- Search engine is probably the mostly used application
- PageRank is considered to be one of the top-10 algorithms for Data Mining

Reading

- Reading for this week
  - Chapter 14 of the textbook
  - Chapter 5 (link analysis) of Mining of Massive Datasets (3rd edition)
    `http://www.mmds.org/`
- Next week
  - Machine learning on graphs