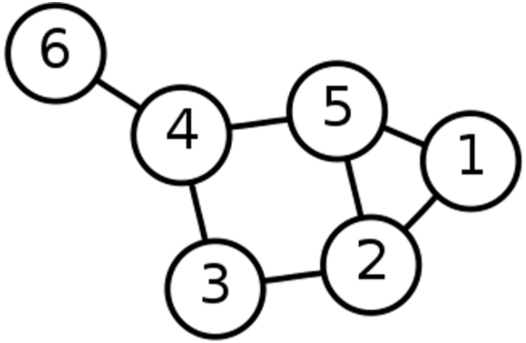
CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Graph Analysis and Page Rank

What is a Graph?

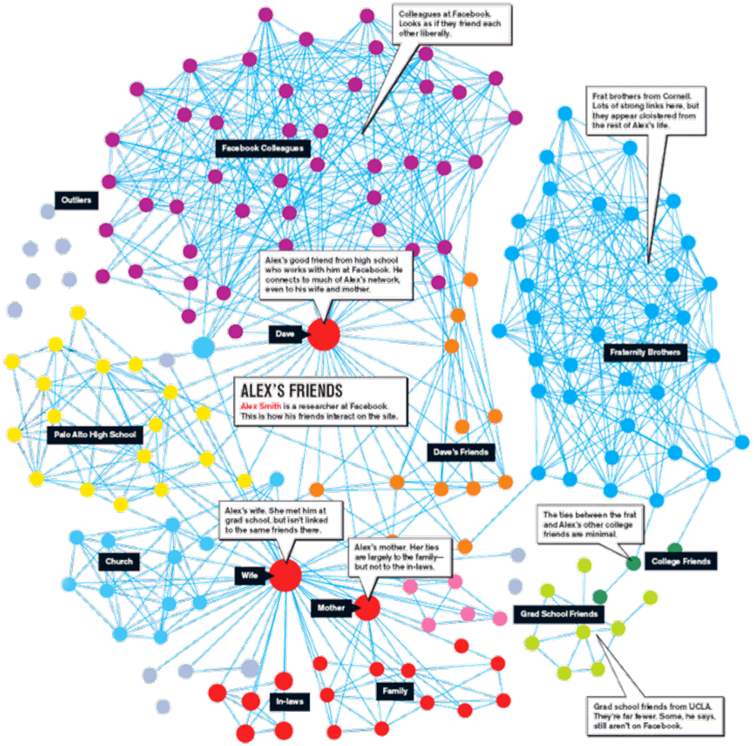


- A ***graph*** is a structure made up of a set of objects, where some pairs of the objects are "related"
- Mathematically, objects are represented with ***vertices*** (or nodes or points) and the relations between two vertices are represented with ***edges*** (or links or lines)
- Typically, a graph is depicted in diagrammatic form as a set of dots or circles for the vertices, joined by lines or curves for the edges
- Edges can be directed or undirected (a relationship can go both ways)
- Edges can be weighted to show the “strength”, distance, etc

Graph Structure is Everywhere

- Any social media application that you are a part of can be modeled as a graph
  - Vertices are users, posts or images
  - Edges are any social relationship between them.

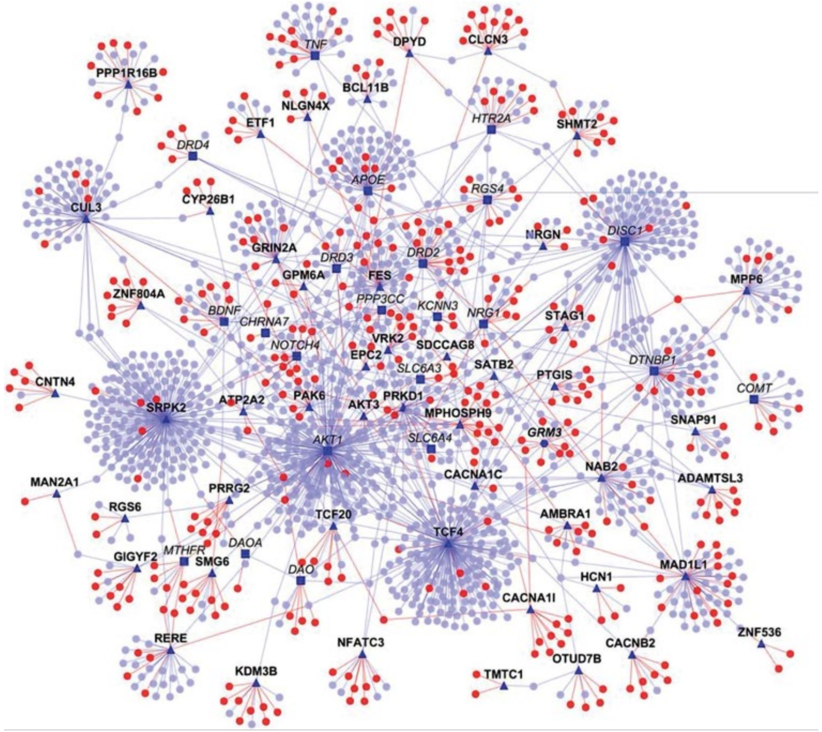
ie: a person hitting 'like' on a particular post/image can be considered an edge



Ref :https://www.pinterest.com/pin/490470215639647556/

Graph Structure is Everywhere

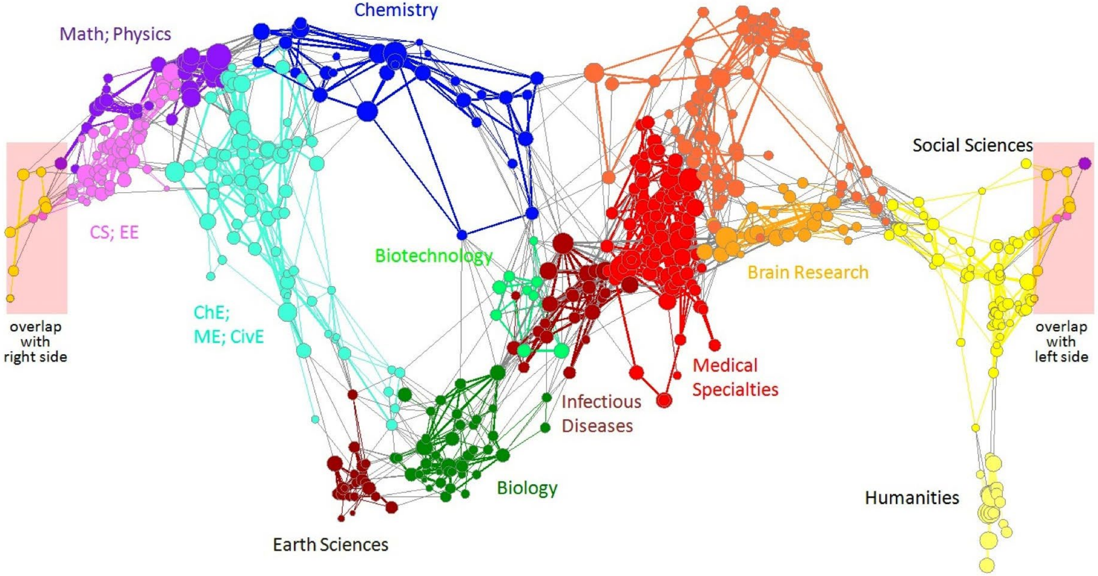
- A Biological network can be modelled as a graph
- For example, a Protein-Protein interaction graph
  - Vertices are proteins
  - Edges are interaction between them



Ref :https://www.genengnews.com/insights/protein-protein-interactions-get-a-new-groove-on/

Graph Structure is Everywhere

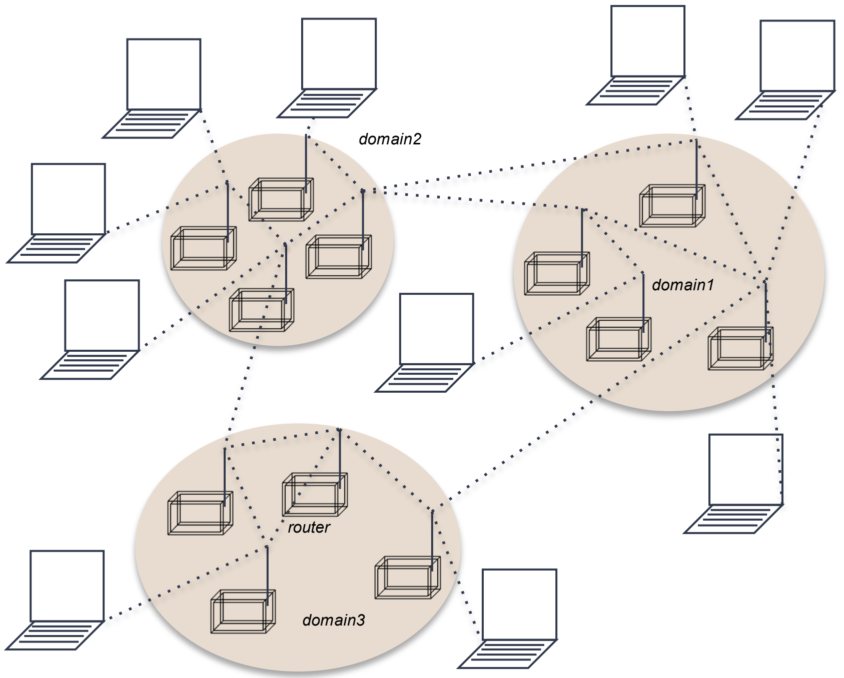
- Graph Data: Information Nets



Citation networks and Maps of science
[Börner et al., 2012]

Graph Structure is Everywhere

- Graph Data: Communications Nets

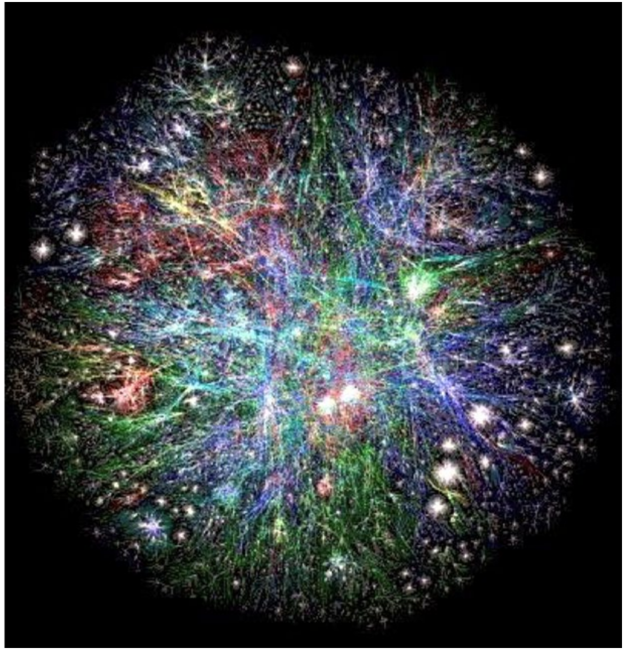


Internet

Graph Structure is Everywhere

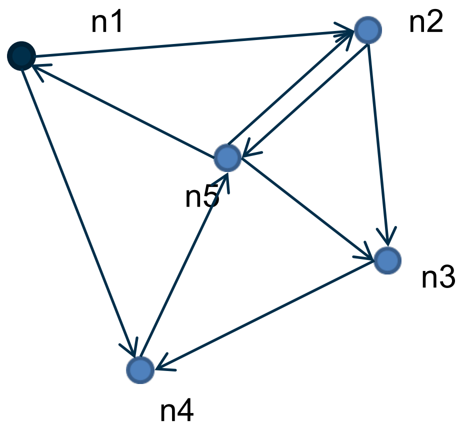
- The internet can be modeled as a graph
  - Vertices are webpages
  - Edges are the links between them
- This modeling can be used to compute the importance of each webpage in the network

This will be the topic of the next few lectures



Ref : http://www.vlib.us/web/worldwideweb3d.html

Graph Representations



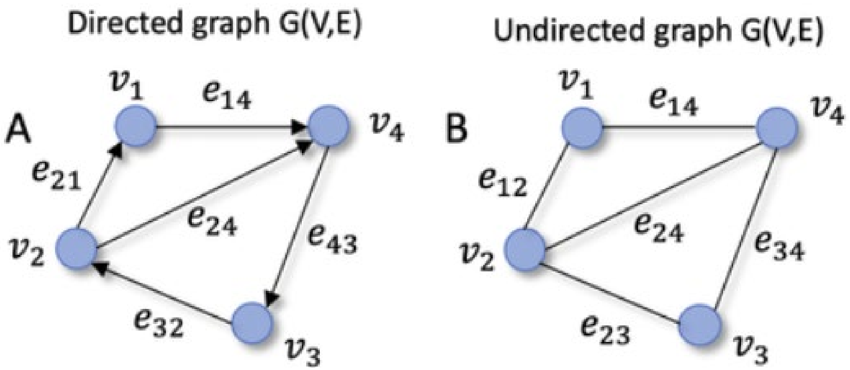
How do you represent this visual diagram as data?

Graph Representations

There are two standard ways to represent a graph G(V,E) *[V is the set of vertices, E is the set of edges]*

1. adjacency list representation
2. adjacency matrix

An adjacency matrix is 2-Dimensional Array of size V x V, where V is the number of vertices in the graph.



	v_1	v_2	v_3	v_4
v_1	0	0	0	1
v_2	1	0	0	1
v_3	0	1	0	0
v_4	0	0	1	0

	v_1	v_2	v_3	v_4
v_1	0	1	0	1
v_2	1	0	1	1
v_3	0	1	0	1
v_4	1	1	1	0

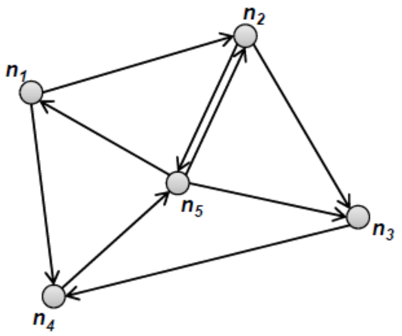
Two types of Graph:

1. Directed Graph
2. Undirected Graph

Graph Representations

There are two standard ways to represent a graph G(V,E)
[V is the set of vertices, E is the set of edges]

1. adjacency list representation
2. adjacency matrix



An adjacency matrix is 2-Dimensional Array of size $V \times V$, where V is the number of vertices in the graph.

An adjacency list is an array of linked lists, where the array size is same as number of vertices in the graph. Every vertex has a linked list. Each node in this linked list represents the reference to another vertex that shares an edge with the current vertex.

	n_1	n_2	n_3	n_4	n_5
n_1	0	1	0	1	0
n_2	0	0	1	0	1
n_3	0	0	0	1	0
n_4	0	0	0	0	1
n_5	1	1	1	0	0

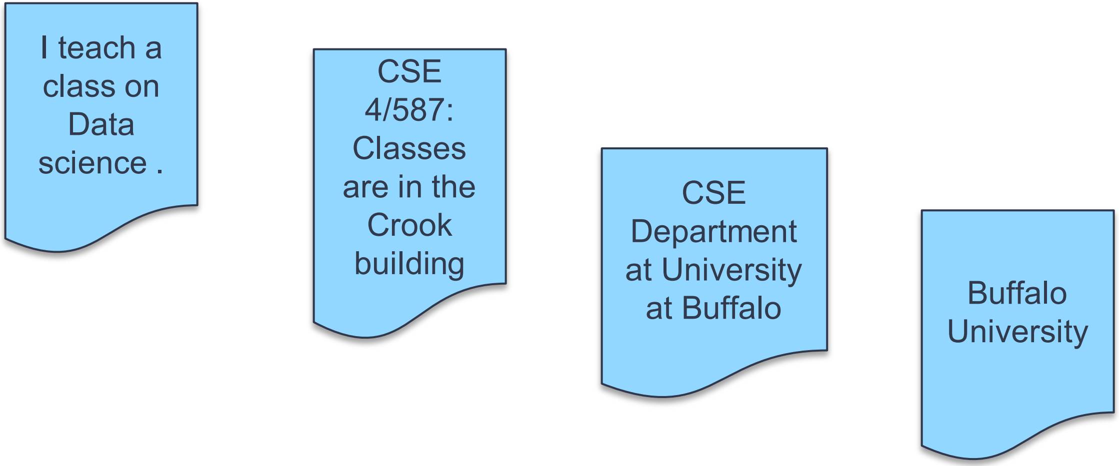
adjacency matrix

n_1	$[n_2, n_4]$
n_2	$[n_3, n_5]$
n_3	$[n_4]$
n_4	$[n_5]$
n_5	$[n_1, n_2, n_3]$

adjacency lists

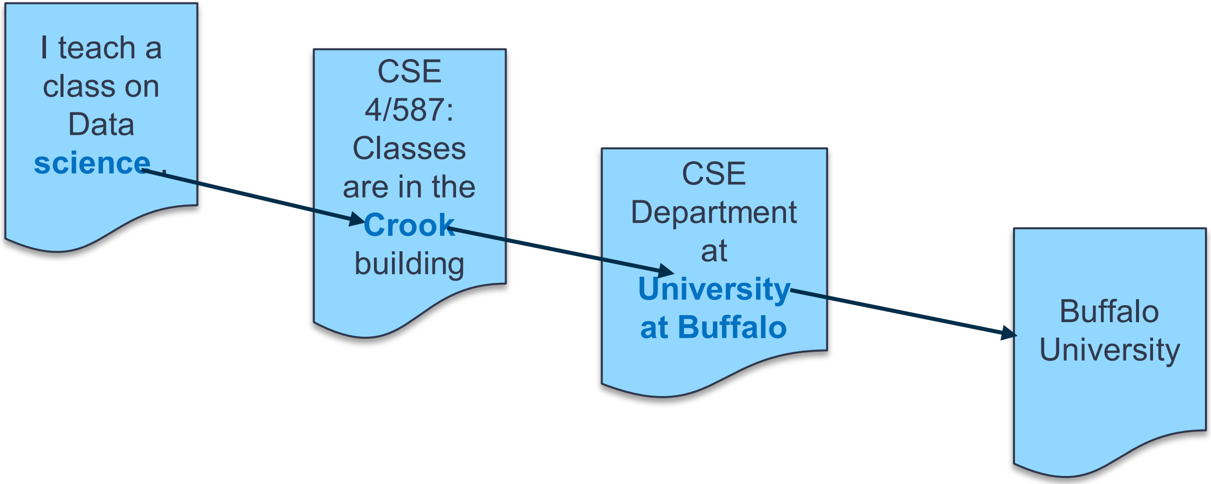
Web As a Graph

- Web as a directed graph:
  - Nodes: Webpages
  - Edges: Hyperlinks

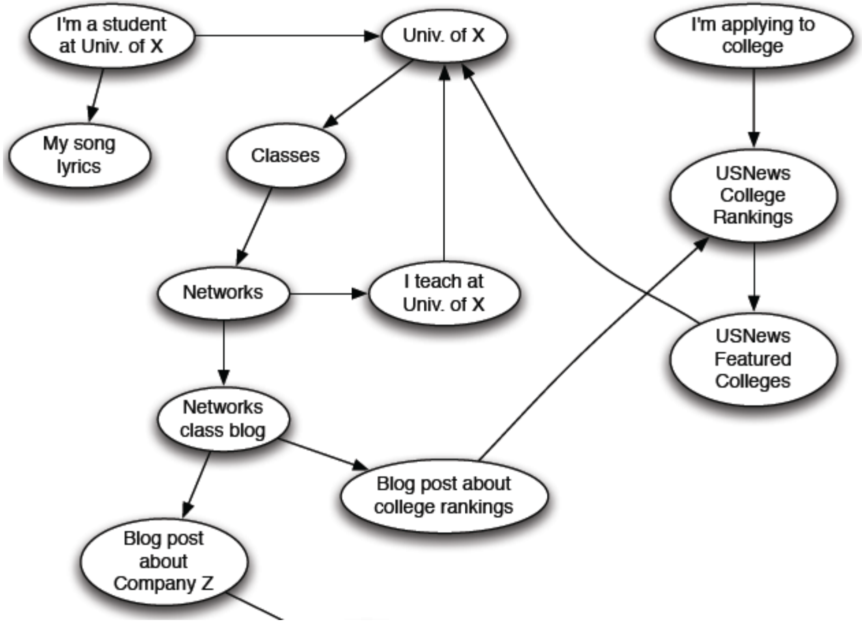


Web As a Graph

- Web as a directed graph:
  - Nodes: Webpages
  - Edges: Hyperlinks



Web As a Directed Graph



Broad Question

How can we organize the internet?

Broad Question

How can we organize the internet?

First try: Human Curated

- Web directories
- Yahoo, DMOZ, LookSmart (1996)

Broad Question

How can we organize the internet?

First try: Human Curated

- Web directories
- Yahoo, DMOZ, LookSmart

Second try: Web Search

- Information retrieval investigates to find relevant docs in a small and trusted set of newspaper articles, patents, etc.

Broad Question

How can we organize the internet?

First try: Human Curated

- Web directories
- Yahoo, DMOZ, LookSmart

Second try: Web Search

- Information retrieval investigates to find relevant docs in a small and trusted set of newspaper articles, patents, etc.

But: Web is huge, full of untrusted documents, random things, web spam, etc

Web Search Challenges

Two challenges of web search:

- (1) Web contains many sources of information
  Which can we "trust"?
  - Trick: If we know one trustworthy page, it may point to another.

- (2) What is the "best" answer to query a"newspaper"?
  - No single right answer
  - Trick: Pages that actually know about newspapers might all be pointing to many newspapers

Ranking Nodes in a Graph

All web pages are not equally "important"

Some websites may provide more trustworthy information

Consider the following websites :

<my homepage> *vs* www.buffalo.edu *or* www.stanford.edu

Which one is important?

Ranking Nodes in a Graph

All web pages are not equally "important"

Some websites may provide more trustworthy information

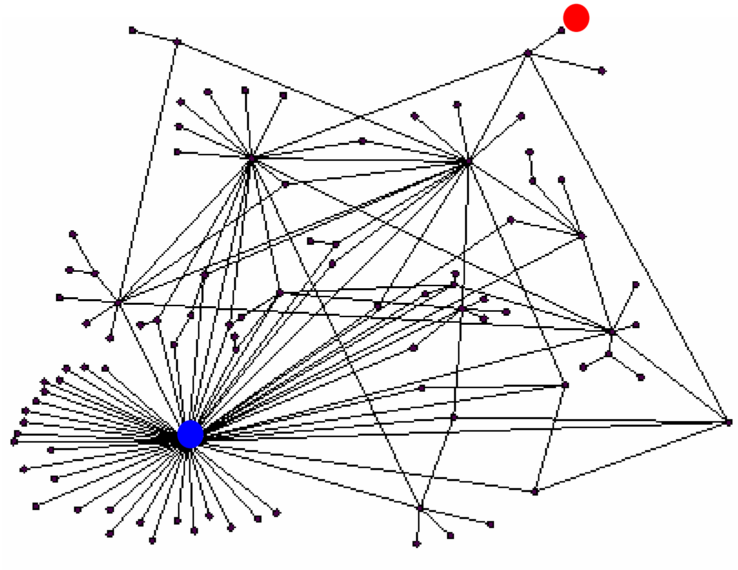
Consider the following websites :

<my homepage> *vs* www.buffalo.edu *or* www.stanford.edu

The university websites are more important than the other website

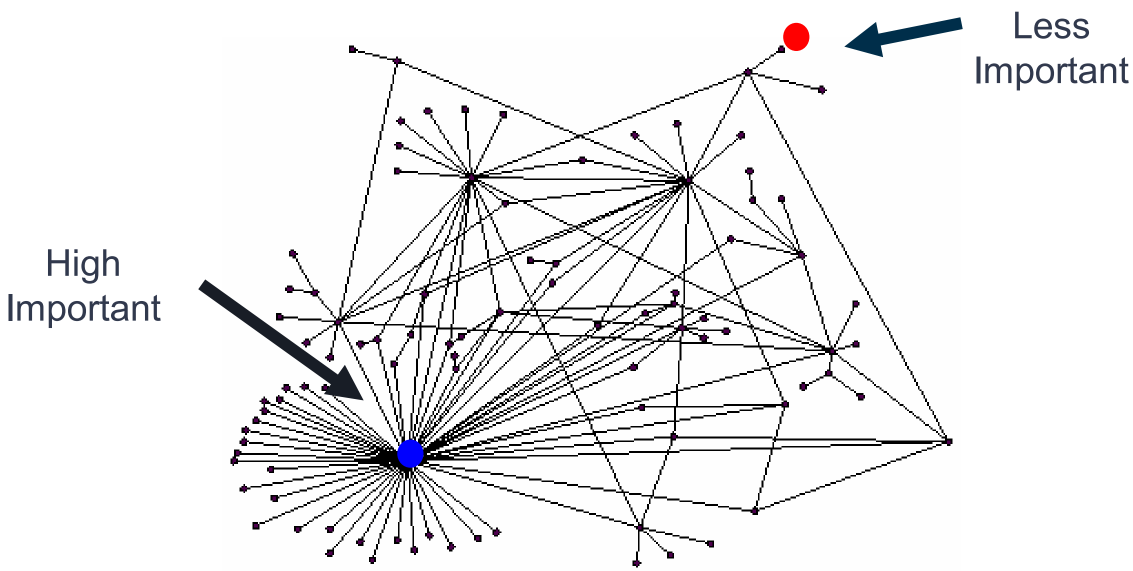
Ranking Nodes in a Graph

Large Diversity of web graph in terms of node connectivity



Ranking Nodes in a Graph

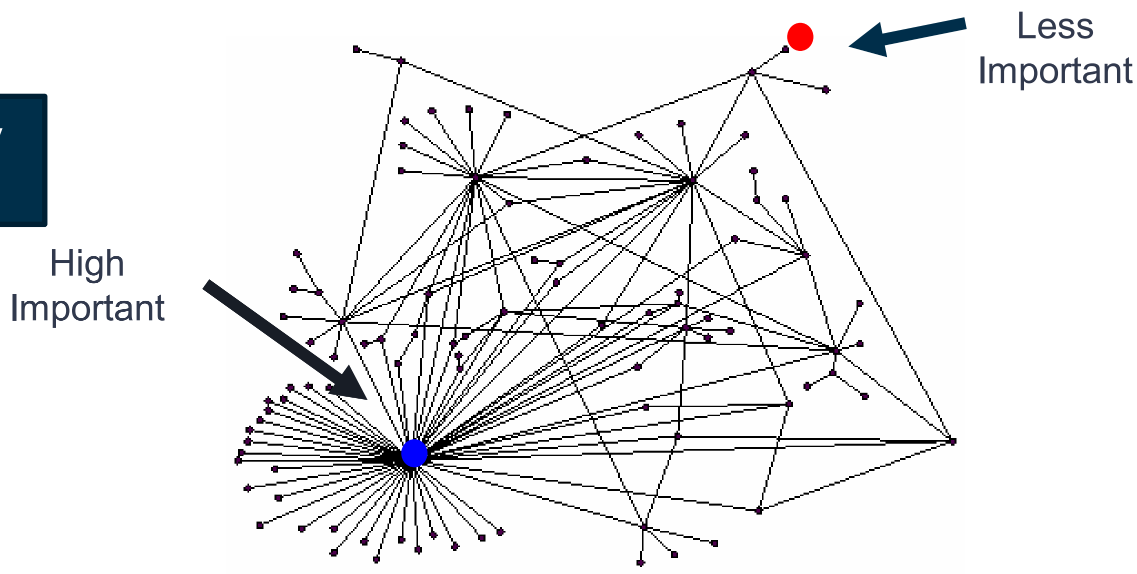
Large Diversity of web graph in terms of node connectivity



Ranking Nodes in a Graph

Large Diversity of web graph in terms of node connectivity

Lets rank the pages by
Link Structure



Link Analysis Algorithm

Key idea is to use links between pages as ***votes***

A page is more important if it has more links associated with it

What kind of links are more important? Incoming or outgoing?

Link Analysis Algorithm

Key idea is to use links between pages as ***votes***

A page is more important if it has more links associated with it

What kind of links are more important?Incoming or outgoing?

The incoming links are more important!

Link Analysis Algorithm

Key idea is to use links between pages as ***votes***

A page is more important if it has more links associated with it

What kind of links are more important? *Incoming or outgoing?*

The incoming links are more important!

www.buffalo.edu is referred to in lot of other pages. So it must be a pretty influential page.

So do all incoming links have equal weightage?

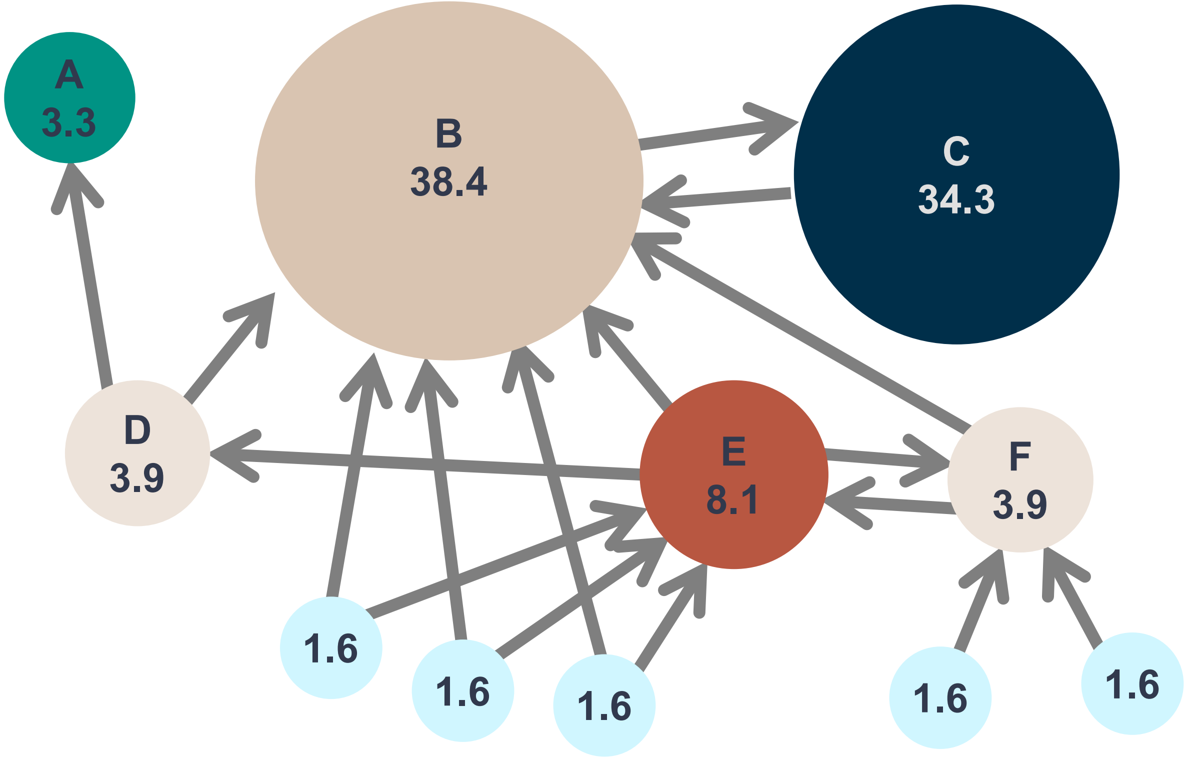
Link Analysis Algorithm

- Think of in -links as votes:
  - www.buffalo.edu has 23,400 in-links
  - www.myhomepage.com has 3 in-link

www.buffalo.edu is referred to in lot of other pages. So it must be a pretty influential page.

- So Do all in-links are equal?
  - Links from important pages count more

Example: PageRank Scores

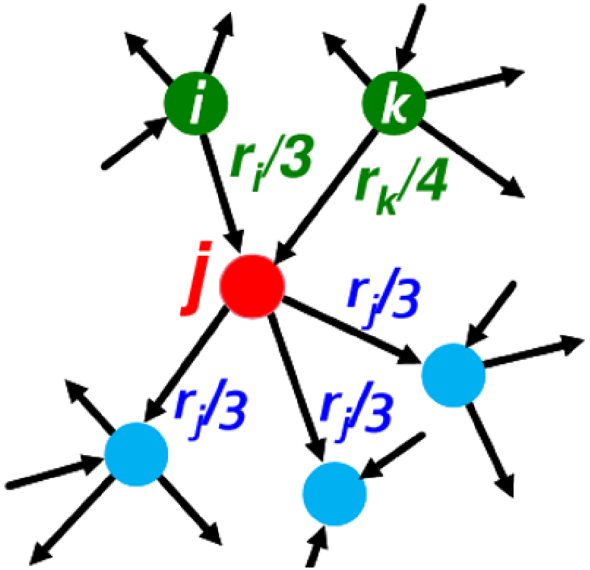


Recursive Formulation

Each link's vote is proportional to the importance of its source page

If page j with importance r_j has n out-links, each link gets r_j / n votes

Page j's own importance is the sum of the votes on its in-links



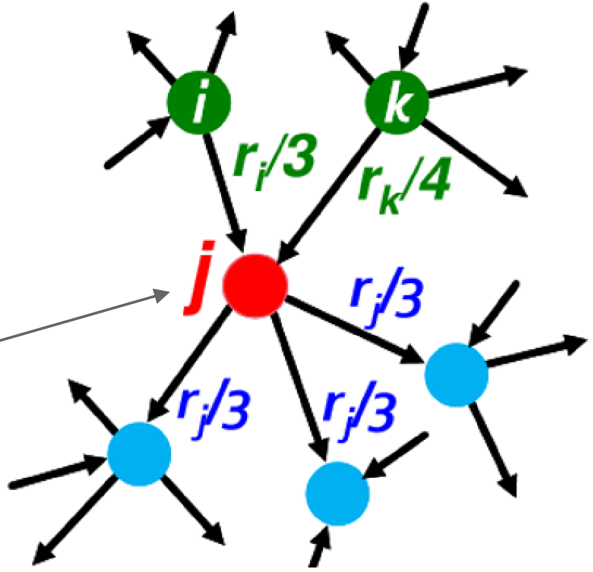
Recursive Formulation

Each link's vote is proportional to the importance of its source page

If page j with importance r_j has n out-links, each link gets r_j / n votes

Page j's own importance is the sum of the votes on its in-links

$$r_j = (r_i / 3) + (r_k / 4)$$



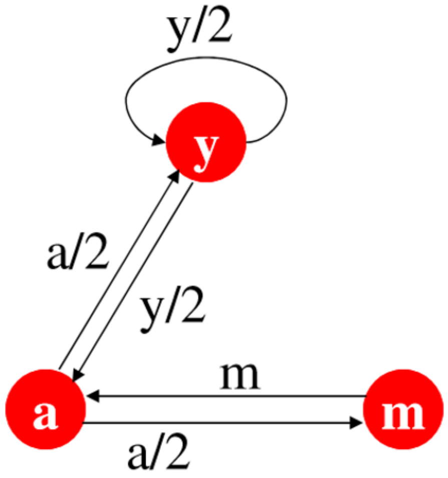
Page Rank: The Flow Model

A link from an *important page* (higher ranking page) is worth more

A page is *important* if it is pointed to by other important pages

Define a "rank" r_j for page j as:

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$



"Flow" equations:

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

Solving the Flow Equation

3 equations, 3 unknowns, no constants

No unique solution: All solutions equivalent modulo the scale factor

Adding an additional constraint forces uniqueness:

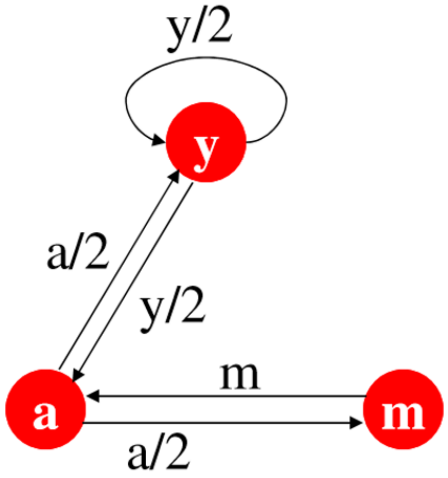
$$r_y + r_a + r_m = 1$$

Solution: $r_y = \frac{2}{5}, \; r_a = \frac{2}{5}, \; r_m = \frac{1}{5}$

Gaussian Elimination can be used to find the solution.

This method will work for small graphs, but won't scale for larger graphs

We need a new formulation!



"Flow" equations:

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

Page Rank: Matrix Formulation

Stochastic Adjacency matrix ***M,***

Let page I has di out links

$$\text{If } i \rightarrow j, \text{then } M_{ji} = \frac{1}{d_i} \quad \text{else } M_{ji} = 0$$

$\boldsymbol{M}$ is a **column stochastic matrix**

Columns sum to 1

Page Rank: Matrix Formulation

Stochastic Adjacency matrix M

$M_{ji} = 1/(d_i)$ if there is a link from i to j, else value is 0

If r is vector with the initial importance of a page and

If Rank vector , r vector with the initial importance of a page then we can write

$$\sum_i r_i = 1$$

Page Rank: Matrix Formulation

Stochastic Adjacency matrix M

$M_{ji} = 1/(d_i)$ if there is a link from i to j, else value is 0

If r is vector with the initial importance of a page and

$$\sum_i r_i = 1$$

Then the flow equation can be written as

$$r = M \cdot r$$

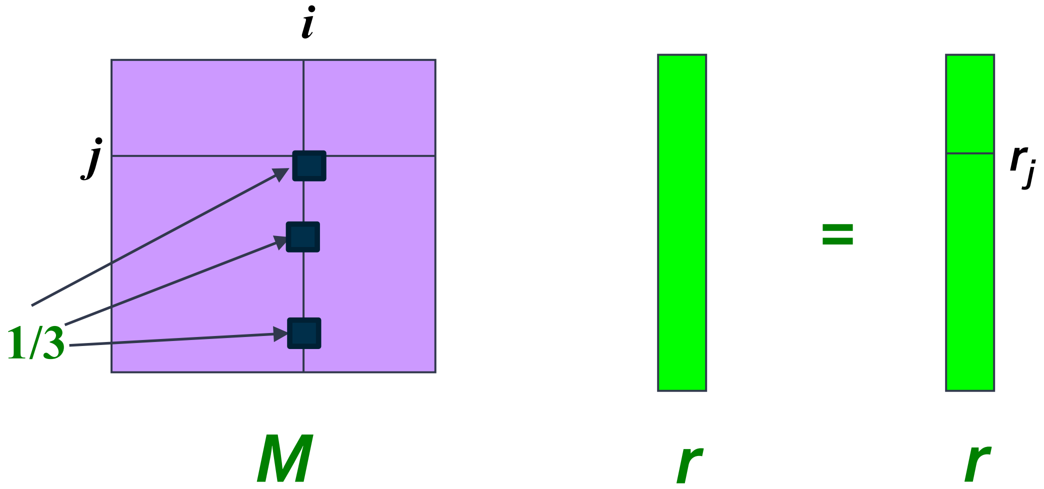
$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_i}$$

Example

- Remember the flow equation: $r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d_i}}$
- Flow equation in the matrix form

$$\boldsymbol{M} \cdot \boldsymbol{r} = \boldsymbol{r}$$

  - Suppose page *i* links to 3 pages, including *j*



Eigenvector Formulation

- The flow equations can be written

$$r = M \cdot r$$

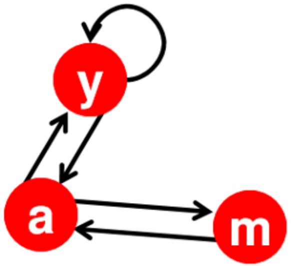
- So the **rank vector** *r* is an **eigenvector** of the stochastic web matrix *M*. In fact, its first or principal eigenvector, with corresponding eigenvalue *1*
  - Largest eigenvalue of *M* is 1 since *M* is column stochastic (with non-negative entries)
    - *We know r is unit length and each column of M sums to one, so* $Mr \leq 1$

- We can now efficiently solve for *r*! The method is called Power iteration

NOTE: ***x*** is an eigenvector with the corresponding eigenvalue **λ** if:

$$Ax = \lambda x$$

Solving with Power Iteration



$$\begin{array}{c|c|c|c|}
 & y & a & m \\
\hline
y & \frac{1}{2} & \frac{1}{2} & 0 \\
\hline
a & \frac{1}{2} & 0 & 1 \\
\hline
m & 0 & \frac{1}{2} & 0 \\
\hline
\end{array}$$

$r = M \cdot r$

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

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

Solving with Power Iteration

Given a web graph with n nodes, where the vertices are pages and edges are hyperlinks

Power iteration: a simple iterative scheme

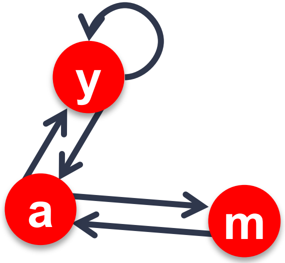
Suppose there are N web pages

1. **Initialize:** $r(0) = [1/N, \ldots, 1/N]^T$
2. **Iterate:** $r(t+1) = M \cdot r(t)$
3. **Stop when:** $\|r(t+1) - r(t)\|_1 < \varepsilon$

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

PageRank: How to solve?

- **Power Iteration:**
  - Set $r_j = 1/N$
  - 1: $r'_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - 2: $r = r'$
  - Goto 1
- **Example:**



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

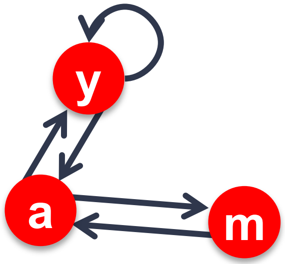
$$\mathbf{r_y} = \mathbf{r_y}/2 + \mathbf{r_a}/2$$
$$\mathbf{r_a} = \mathbf{r_y}/2 + \mathbf{r_m}$$
$$\mathbf{r_m} = \mathbf{r_a}/2$$

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{matrix} 1/3 & 1/3 & 5/12 & 9/24 & & 6/15 \\ 1/3 & 3/6 & 1/3 & 11/24 & \ldots & 6/15 \\ 1/3 & 1/6 & 3/12 & 1/6 & & 3/15 \end{matrix}$$

Iteration 0, 1, 2, …

PageRank: How to solve?

- **Power Iteration:**
  - Set $r_j = 1/N$
  - 1: $r'_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - 2: $r = r'$
  - Goto 1
- **Example:**



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

$$\mathbf{r_y} = \mathbf{r_y}/2 + \mathbf{r_a}/2$$
$$\mathbf{r_a} = \mathbf{r_y}/2 + \mathbf{r_m}$$
$$\mathbf{r_m} = \mathbf{r_a}/2$$

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \quad \begin{matrix} 1/3 \\ 1/3 \\ 1/3 \end{matrix} \quad \begin{matrix} 1/3 \\ 3/6 \\ 1/6 \end{matrix} \quad \begin{matrix} 5/12 \\ 1/3 \\ 3/12 \end{matrix} \quad \begin{matrix} 9/24 \\ 11/24 \\ 1/6 \end{matrix} \quad \ldots \quad \boxed{\begin{matrix} 6/15 \\ 6/15 \\ 3/15 \end{matrix}} \rightarrow \boxed{\begin{matrix} 2/5 \\ 2/5 \\ 1/5 \end{matrix}}$$

Iteration 0, 1, 2, …

References

[1] http://www.mmds.org

[2] Chapter 5 Lin and Dyer