PART 1. BATCH COMPUTING MODEL FOR BIG DATA ANALYTICS
2. WEB SCALE LINK ANALYSIS

Sangmi Lee Pallickara
Computer Science, Colorado State University
http://www.cs.colostate.edu/~cs535

FAQs
- Programming Assignment 1
- We discuss link analysis in this week
- Installation/configuration guidelines for Hadoop and Spark has been uploaded
- Port assignment has been posted

Today's topics
- Link analysis
  - HITS
  - Network Centrality

Web Scale Link Analysis
1. PageRank
2. Hyperlink-Induced Topic Search (HITS)
3. Network Centrality

This material is built based on

2. Hyperlink-Induced Topic Search (HITS)
Types of Web queries

- Yes/No queries
  - Does Chrome support .ogv video format?

- Broad topic queries
  - Find information about "eclipse"

- Similar-page query
  - Find pages similar to 'fandango.com'

Ranking algorithm to find the most "authoritative" pages

- To find the small set of the most authoritative pages that are relevant to the query

  - Authority
    - "python"
      - https://www.python.org
    - "Colorado State University"
      - https://www.colostate.edu
    - "iPhone"
      - https://www.apple.com/iphone

Challenge of content-based ranking

- Most useful pages do not have the keyword
  - "hardware" in the APPLE page? Or the IBM page?

- Pages are not sufficiently descriptive
  - "computer hardware manufacturers" in IBM, or APPLE

HITS (Hypertext-Induced Topic Search)

- PageRank captures simplistic view of a network

  - Authority
    - A Web page with good, authoritative content on a specific topic
    - A Web page that is linked by many hubs

  - Hub
    - A Web page pointing to many authoritative Web pages
    - e.g. portal pages (Yahoo)

Understanding Authorities and Hubs

- A.K.A. Hubs and Authorities
  - Jon Kleinberg 1997
  - Topic search
  - Automatically determine hubs/authorities

  - In practice
    - Performed only on the result set (PageRank is applied on the complete set of documents)
    - Developed for the IBM Clever project
    - Used by Teoma (later Ask.com)
Understanding Authorities and Hubs [2/2]
- A good hub page points to many good authoritative pages
- A good authoritative page is pointed to by many good hub pages
- Authorities and hubs have a mutual reinforcement relationship

Calculating Authority/Hub scores [1/3]
Let there be $n$ Web pages
Define the $n \times n$ adjacency matrix $A$ such that,
\[ A_{uv} = 1 \] if there is a link from $u$ to $v$.
Otherwise $A_{uv} = 0$

Graph with pages

Calculating Authority/Hub scores [2/3]
Each Web page has an authority score $a_i$ and a hub score $h_i$.
We define the authority score by summing up the hub scores that point to it,
\[ a_i = \sum_{j \neq i} h_j A_{ji} \]
This can be written concisely as,
\[ a = A^T h \]

Graph with pages

Calculating Authority/Hub scores [3/3]
Similarly, we define the hub score by summing up the authority scores,
\[ h_i = \sum_{j \neq i} a_j A_{ij} \]
This can be written concisely as,
\[ h = A a \]

Graph with pages

Implementing Topic Search using HITS
- Step 1.
  - Constructing a focused subgraph based on a query
- Step 2.
  - Iteratively calculate the authority value and hub value of the page in the subgraph
Step 1. Constructing a focused subgraph (root set)
- Generate a root set from a text-based search engine
  - e.g., pages containing query words

Step 2. Initial values

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Hub</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Ranks
Hub: P1=P2=P3=P4
Authority: P1=P2=P3=P4

Step 2. After the first iteration (without normalization)

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Hub</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>3</td>
<td>0.125</td>
</tr>
<tr>
<td>P2</td>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>P3</td>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>P4</td>
<td>1</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Ranks
Hub: P1>P2=P3>P4
Authority: P1=P2<P3<P4

Step 2. After the first iteration (after normalization)

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Hub/Normalized</th>
<th>Authority/Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.875</td>
<td>0.25</td>
</tr>
<tr>
<td>P2</td>
<td>0.75</td>
<td>0.375</td>
</tr>
<tr>
<td>P3</td>
<td>0.5</td>
<td>0.625</td>
</tr>
<tr>
<td>P4</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Ranks
Hub: P1>P2>P3>P4
Authority: P1=P2=P3=P4

Normalization
- Original paper: using squares sum (to 1)
- You can use sum (to 1)
- value = value/sum of all values

Step 2. After the second iteration (without normalization)

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Hub</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.875</td>
<td>0.25</td>
</tr>
<tr>
<td>P2</td>
<td>0.75</td>
<td>0.375</td>
</tr>
<tr>
<td>P3</td>
<td>0.5</td>
<td>0.625</td>
</tr>
<tr>
<td>P4</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Ranks
Hub: P1>P2>P3>P4
Authority: P1>P2>P3>P4
Step 2. After the second iteration (after normalization)

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Hubs/Normalized</th>
<th>Authority/Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.875/0.333</td>
<td>0.25/0.111</td>
</tr>
<tr>
<td>P2</td>
<td>0.75/0.273</td>
<td>0.375/0.187</td>
</tr>
<tr>
<td>P3</td>
<td>0.625/0.227</td>
<td>0.625/0.278</td>
</tr>
<tr>
<td>P4</td>
<td>0.5/0.182</td>
<td>1/0.444</td>
</tr>
</tbody>
</table>

Ranks
- Hub: P2 > P1 > P3 > P4
- Authority: P1 < P2 < P3 < P4

Step 2. Convergence of scores
- Repeat the calculation (step 2) until the scores converge
  - You should specify your threshold

Web Scale Link Analysis
1. PageRank
2. Hyperlink-Induced Topic Search (HITS)
3. Network Centrality

This material is built based on
- Robert A. Hanneman and Mark Riddle, Introduction to social network methods

Understanding Social Media
- Advertising on Facebook
  - From $0.16 to $1.00+ per click

https://redespresso.com/academy/blog/facebook-ads-cost/
Network Centrality

- A common goal in SNA is to identify the “central” nodes of a network
- What does “central” mean?
  - active?
  - important?
  - non-redundant?
- Koschutzki et al. (2005) attempted a classification of centrality measures:
  - Reach: ability of ego to reach other vertices
  - Flow: quantity/weight of walks passing through ego
  - Vitality: effect of removing ego from the network
  - Feedback: a recursive function of alter centralities

What is the Network Centrality?

- “There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement.” — Linton Freeman

Network Centrality Measures

- Local measure
  - Out-degree/in-degree
  - Degree Centrality
- Relative to the rest of network
  - Closeness (based on the average distance)
  - Betweenness (based on geodesics)
  - Eigenvector ranking (PageRank, and Katz centrality)

Who is important based on the Network Position?

- In each of the following networks, X has higher centrality than Y according to a particular measure

Degree Centrality (undirected)

- He or she who has many friends is most important

How equal are the nodes? — measure of Degree Centrality

- How much variation exist in the centrality scores among the nodes?
- Freeman’s general formula for centralization (can use other metrics, e.g. gini coefficient or standard deviation):
  \[ C(D) = \frac{\sum_{i,j} (d_i - d_j)^2}{(n^2-n)} \]
  where \( d_{iv} \) is the maximum value in the graph
  \( n \) is the number of vertices
Examples

\[ C_D = \frac{g_{jk}(i)}{g_{jk}} \]

High centralization
- A few investors are trading with many others

Low centralization
- Trades are more evenly distributed

Is “Degree” enough to describe centrality?
- Graph with multiple sub-graphs
  - Who can pass information across the different sub-graphs?

Betweenness Centrality
- How many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?
- Who has the higher betweenness, Ann (A) or Bob (B)?

\[ C_B(i) = \sum_{j,k} g_{jk}(i)/g_{jk} \]

Where \( g_{jk} \) = the number of geodesics connecting \( j \) and \( k \), and \( g_{jk}(i) \) = the number of geodesics that actor \( i \) is on.

- Usually these values are normalized by:
  \[ C_B(i) = C_B(i)/(n - 1)(n - 2) \]

Real-world example of Betweenness Centrality
- Facebook subgraph
  - Are there any individuals with a low degree but connecting large Groups?
Measuring Betweenness Centrality

A and E are not located between any two other vertices
B and D are located between 3 possible pairs of vertices
C is located between 4 possible pairs of vertices

Worksheet

What is the Betweenness Centrality measures of A, B, C, D, and E?

Measuring Betweenness Centrality

A, B, G, and F are not located between any pair of other vertices
C and E are located between 8 possible pairs of vertices
D is located between 9 possible pairs of vertices

Measuring Betweenness Centrality

A, B, C, D, and E are not located between any pair of other vertices
F is located between 10 possible pairs of vertices

Why do E and C have a betweenness of 1?
They are both on the shortest paths for pairs (B-D, A-D).
They will share credits: 0.5+0.5 = 1
B’s betweenness = 3.5
D’s betweenness = .5
Finding all-pairs shortest paths (APSP)

- Requires large memory for computing APSP over a massive graph
  - You can use GraphX, or Pregel for Single source shortest path (SSSP)

- Floyd-Warshall algorithm
  - Iterative algorithm

Floyd-Warshall algorithm

Matrix representation

\[
\begin{align*}
D_0 &= \begin{bmatrix}
8 & 13 & -1 & 0 & -6 & 6 & 12 \\
-9 & 0 & -6 & 12 & -6 & -1 & 0 \\
7 & 0 & 0 & -1 & -9 & 0 & -6 \\
12 & -1 & 0 & 0 & 13 & 0 & -6 \\
7 & 0 & 0 & -1 & 0 & -9 & -9 \\
8 & -9 & -1 & 0 & -1 & 0 & 0 \\
11 & 0 & 0 & 0 & 0 & 0 & -6
\end{bmatrix}
\end{align*}
\]

1. For \( k = 1 \to n \) // intermediate vertices considered
2. for \( i = 1 \to n \) // the “from” vertex
   for \( j = 1 \to n \) // the “to” vertex
   \( d_{ij}^{(k)} = \text{min}(d_{ij}^{(k-1)}, d_{ik}^{(k-1)} + d_{kj}^{(k-1)}) \)

Pseudocode for Floyd-Warshall

- Input: \( D^0 = (d_{ij}^{(0)}) \) (the initial edge-cost matrix)
- Output: \( D^n = (d_{ij}^{(n)}) \) (the final path-cost matrix)

for \( k = 1 \) to \( n \) // intermediate vertices considered
for \( i = 1 \) to \( n \) // the “from” vertex
  for \( j = 1 \) to \( n \) // the “to” vertex
    \( d_{ij}^{(k)} = \text{min}(d_{ij}^{(k-1)}, d_{ik}^{(k-1)} + d_{kj}^{(k-1)}) \)

Closeness Centrality

- Based on the length of the average shortest path between a vertex and all vertices in the graph
- You can normalize the Closeness Centrality with:
  \[ C_C(i) = \frac{1}{\sum_{j \neq i} d(i,j)} \]
Closeness Centrality

\[ C_c(A) = \frac{1}{N-1} \sum_{j=1}^{N} d(A,j)^{-1} = \frac{1 + 2 + 3 + 4}{4} = 0.4 \]