PART B. GEAR SESSIONS
SESSION 4: LARGE SCALE RECOMMENDATION SYSTEMS AND SOCIAL MEDIA

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

FAQs

• Your disk quota is 20GB (per student)
• If you need more space, please let me know ASAP
Topics of Today's Class

- Part 1: Introduction to Social Network Analysis and Clustering Social Networks
- Part 2: Finding similar nodes: Simrank
- Part 3: Counting Triangles
Social Networks as Graphs

- Social networks are naturally modeled as graphs
  - Social graph
  - Nodes
  - Edge connects two nodes
    - If the nodes are related by the relationship that characterizes the network

Discussions

- "Friends" relationship graph
  - $B$ is a friend with $A$, $C$, and $D$

- Suppose $X$, $Y$, and $Z$ are arbitrary nodes of this graph, with edge $(X, Y)$ and $(X, Z)$
- What would we expect the probability of an edge between $Y$ and $Z$ to be?
Discussions -- continued

- Suppose X, Y, and Z are arbitrary nodes of this graph, with edge (X,Y) and (X, Z)
- \( \binom{3}{2} = 21 \) pairs of nodes that could have had an edge between them
- Currently there are 9 edges (friendships)
- If the graph is very large enough, the probability would be very close to \( \frac{9}{21} = 0.429 \)
- However, the graph is quite small:
  - X, Y, and Z already have 2 edges
  - Therefore among the 19 remaining pairs of nodes
    - 7/19 = 0.368

Locality expected in a social network
Total 9 positive cases and 7 negative cases
Therefore, the fraction of times the third edge exists is \( \frac{9}{16} = 0.563 \)
It is much larger that 0.368 expected values
Varieties of Social Networks

- Telephone Networks
  - Nodes with phone numbers
  - Edge between two nodes if a call has been placed (in some fixed period of time)
- Email Networks
  - Nodes?
  - Edges?
- Facebook Networks
  - Nodes?
  - Edges?
- Collaboration Networks
  - Nodes?
  - Edges?

Graphs with several different node types

- Social phenomena involving entities of different types
- E.g. Collaborative networks
  - Authorship graph
    - Authors
    - Papers
  - One graph? Two graphs?
- How about comments and “likes” for facebook?
  - User
  - Photo
  - Comment
  - Post
- $k$-Partite graph with $k > 1$
A tripartite graph representing users, tags, and photos

- Three sets of nodes
- Users \{U_1, U_2\}
- Tags \{T_1, T_2, T_3, T_4\}
- Web page \{W_1, W_2, W_3\}
- All edges connect nodes from two different sets
- Edge \((U_1, T_2)\) means that user \(U_1\) has placed a tag \(T_2\) on at least one Web page
- This graph cannot tell you the ternary information such as who placed which tags on which photo
  - DB tables can represent it
Clustering of Social Network Graphs

- Social networks contain **entities that are connected by many edges**
  - Group of friends
  - Group of researchers interested in the same topic

Distance Measures for Social-Network Graphs

- How will you define “distance” in a graph?
Distance Measures for Social-Network Graphs

- How will you define “distance” in a graph?

- We can assume that nodes are close if they have an edge between them
  - Distant if not
  - The distance \( d(x, y) \) is 0 if there is an edge \((x, y)\) and 1 if there is no such edge
  - We can use any pair of values
    - Such as 1 and \( \infty \)

- Can this be a valid distance measures?
  - No, they violate the triangle inequality
    - If there are edges \((A, B)\) and \((B, C)\), but no edge \((A, C)\) then the distance from \(A\) to \(A\) exceeds the sum of the distances from \(A\) to \(B\) to \(C\)
Applying Standard Clustering Methods: (1) Hierarchical Clustering

(1) Hierarchical (Agglomerative) and (2) point-assignments clustering

(1) Hierarchical clustering
- Distance based
- intercluster distance the minimum distance between nodes of the two clusters
- Two communities \{A,B,C\} and \{D,E,G,F\}
  - \{D,E,F\} and \{D,F,G\} as two subcommunities of \{D,E,G,F\}
- Problem
  - Chance to combine B and D

\[
\begin{align*}
A & \quad B \\
& \quad C \\
D & \quad E \\
& \quad G \\
& \quad F
\end{align*}
\]

Applying Standard Clustering Methods: (2) Point-assignment approach

(2) k-Means approach
- E.g. \(k=2\)
- If we choose two initial centroids randomly, B and D might be in the same cluster
- If we pick one centroid and then choose another one based on the distance?
  - Still B and D might be in the same cluster

- If we choose two nodes not connected
  - E.g. E and G?

- If we choose B and F?

- Where to place D
  - Can be deferred until we assign some other nodes to the clusters
  - Still chances to make mistakes

\[
\begin{align*}
A & \quad B \\
& \quad C \\
D & \quad E \\
& \quad G \\
& \quad F
\end{align*}
\]
Betweenness

- A method to find **communities** in social networks
- Definition of the betweenness of an edge \((a, b)\)
  - The number of pairs of nodes \(x\) and \(y\) such that the edge \((a, b)\) lies on the shortest path between \(x\) and \(y\)
  - What if there are several possible shortest paths between \(x\) and \(y\)?
    - Edge \((a, b)\) is credited with the fraction of those shortest paths that include the edge \((a, b)\)

- Higher score means
  - Edge \((a, b)\) runs between two different communities
  - \(a\) and \(b\) do not belong to the same community.
Betweenness: Example

- Which edge has the highest betweenness?
  a. (A, B)
  b. (B, D)
  c. (D, E)
  d. (E, F)

![Graph](image.png)

**Betweenness: Example (answer)**

- Which edge has the highest betweenness?
  a. (A, B)
  b. (B, D)
  c. (D, E)
  d. (E, F)

- Edge (B, D) has the highest betweenness
- This edge is on every shortest path between any of A, B, and C to any of D, E, F, and G
- (B, D)'s betweenness is $3 \times 4 = 12$
- Edge (D, F) is on only four shortest paths
  - those from A,B,C, and D to F
Girvan-Newman Algorithm

- **Measuring the betweenness of edges**
  - The number of shortest paths going through each edge

- Girvan-Newman (GN) algorithm
  - Visits each node X once and computes the number of shortest paths from X to each of the other nodes that go through each of the edges

- Starting with BFS

---

- **DAG (directed, acyclic graph)**
  - Edges between levels
  - A DAG edge \((Y, Z)\)
    - \(Y\) is at the level above \(Z\) (i.e., closer to the root)
    - \(Y\) a parent of \(Z\) and \(Z\) a child of \(Y\)
Girvan-Newman Algorithm

- **Step 1**: Perform a breadth-first search (BFS) of the graph

- **Step 2**: Label each node by the number of shortest paths that reach it from the root
  - Starting with 1 for the root
  - From the top down, label each node $Y$ by the sum of the labels of its parents

- **Step 3**: Calculate for each edge $e$ the sum over all nodes $Y$ of the fraction of shortest paths from the root $X$ to $Y$ that go through $e$ (Next slide)

---

### Step 3: Calculate for each edge $e$ the sum over all nodes $Y$ of the fraction of shortest paths from the root $X$ to $Y$ that go through $e$

- **a.** Each leaf DAG node gets a credit of 1
- **b.** Each non-leaf DAG node gets $1 + \text{the sum of the credits of the DAG edges from that node to the level below}$
- **c.** Credit to the parent node(s) is proportionally distributed based on the fraction of shortest paths from the root to the parent node

---

BFS graph starting with E

---

http://www.cs.colostate.edu/~cs535

Spring 2020 Colorado State University, page 13
Girvan-Newman Algorithm

- To complete the betweenness calculation
  1) Repeat this calculation for every node as the root and sum the contributions
  2) Divide by 2 to get the true betweenness
     - Every shortest path will be discovered twice, once for each of its endpoints

Using Betweenness to Find Communities

- The betweenness scores
  - Similar to a distance measure on the nodes of the graph
- It is NOT exactly a distance measure
  - Not defined for pairs of nodes that are unconnected by an edge
  - Might not satisfy the triangle inequality even when defined
- What if we take the edges in order of increasing betweenness and add one at a time
  - Some connected components of the graph form some clusters
  - Higher betweenness will results in the larger cluster
Using Betweenness to Find Communities [2/3]

- (B,D) has the highest betweenness
  - Remove it!
  - \{A,B,C\} vs. \{D,E,G,F\}

- Continue?
  - 5: (A, B) and (B, C)
  - 4.5: (D, E) and (D, G)
  - 4: (D, F)
  - Stop here..

- What do we learn from this?

Using Betweenness to Find Communities [3/3]

- \{A,B,C\} vs. \{D,E,G,F\}
- What do we learn from this?
- B and D were “traitor” to the initial community
  - Have more outside friends compared to others
GEAR Session 4. Large Scale Recommendation Systems and Social Media
Lecture 3. Social Network Analysis
Clustering of Social Network Graphs: Direct Discovery

Direct Discovery of Communities

• Finding Cliques?
  • A subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent
  • Intuitive starting point
• Finding maximal cliques NP-complete
• Even approximating the maximal clique is hard
Complete Bipartite Graphs

- A complete bipartite graph $K_{s,t}$ consists of $s$ nodes on one side and $t$ nodes on the other with all $st$ possible edges between the nodes of one side and the other.

- How do we use CBG to find communities?
  - Divide the nodes into two equal groups at random
  - If a community exists, about half its nodes to fall into each group, and about half its edges would go between groups.

Finding Complete Bipartite Subgraphs

- Suppose we are given a large bipartite graph $G$, and we want to find instances of $K_{s,t}$ within it.

  - We assume,
    - the instance of $K_{s,t}$ we are looking for $t$ nodes on the left side
    - size $t \leq$ size $s$
    - Here, the threshold $s$
      - The number of nodes that the instance of $K_{s,t}$ has on the right side

  - Finding collection(s) of very popular blue nodes
    - Finding frequent itemsets $F$ of size $t$
    - If a set of $t$ nodes on the left side is frequent, then they all occur together in at least $s$ baskets
    - Basket: nodes of the “right” subgraph
Example

- Left side = \{1,2,3,4\}
- Right side = \{a,b,c,d\}
- Basket \textit{a} consists of “items” 1 and 4
  - \(a = \{1,4\}\), \(b = \{2,3\}\), \(c = \{1\}\) and \(d = \{3\}\)

- If \(s = 2\) and \(t = 1\),
- \textbf{WE MUST FIND ITEMSETS of SIZE 1 THAT APPEARS AT LEAST TWO BASKETS}

- 1 and 3

GEAR Session 4. Large Scale Recommendation Systems and Social Media
Lecture 3. Social Network Analysis
Finding Similar Nodes: Simrank
Simrank

- Measure the similarity between nodes of the same type
  - By tracking where random walkers on the graph wind up when starting at a particular node

- Applicable to graphs with different types of nodes

Random Walks in a Social Network Graph

- Imagine that a person randomly “walking” on a social network
- A walker at a node N of an undirected graph will move with equal probability to any of the neighbors of N
- A walker starts out at node T1
  - To U1 or W1
- Starting at T1
  - There is a good chance the walker would visit T2
  - Higher chance than visiting T3 or T4
- Can we infer that tags T1 and T2 are related or similar in some way?
  - Yes
  - E.g. the tag T1 and T2 are used for a common web page W1 and they also have a common user U1 who tag the page using Tag 1 and 2

However, if we allow the walker to continue traversing the graph at random then the probability that the walker will be at any particular node does not depend on where it starts out.
Random Walks with Restart

- Focus on one particular node $N$ of a social network
- Track where the random walker winds up on short walks from that node
- Modify the matrix of transition probabilities to have a probability of transitioning to $N$ from any node

Random Walks with Restart: Example

A simple bipartite social graph

```
Sky
```

```
Tree
```

```
<table>
<thead>
<tr>
<th></th>
<th>Picture 1</th>
<th>Picture 2</th>
<th>Picture 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sky</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tree</td>
<td>1/2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
```

Formally, let $M$ be the transition matrix of the graph $G$. That is, the entry in row $i$ and column $j$ of $M$ is $1/k$ if node $j$ of $G$ has degree $k$, and one of the adjacent nodes is $i$. Otherwise, this entry is 0.
Random Walks with Restart

• Use $\beta$ as the probability that the walker continues at random
• $(1 - \beta)$ is the probability the walker will teleport to the initial node $N$

• Let $e_N$ be the column vector that has 1 in the row for node $N$ and 0’s elsewhere
• If $v$ is the column vector that reflects the probability the walker is at each of the nodes at a particular round
• $v'$ is the probability the walker is at each of the nodes at the next round,
• $v'$ is related to $v$ by:

$$v' = \beta M v + (1 - \beta) e_N$$

Assume that we use same matrix $M$ and $\beta = 0.8$

Also, assume that node N is for Picture 1 -> We want to calculate the similarity of other pictures to “Picture 1”

$$M = \begin{bmatrix}
0 & 0 & 0 & 1/3 & 1/2 \\
0 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 1/3 & 1/2 \\
1/2 & 1 & 1/2 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0
\end{bmatrix}$$

$$v' = \beta M v + (1 - \beta) e_N = \begin{bmatrix}
0 & 0 & 0 & 4/15 & 2/5 \\
0 & 0 & 0 & 4/15 & 0 \\
0 & 0 & 0 & 4/15 & 2/5 \\
2/5 & 4/5 & 2/5 & 0 & 0 \\
2/5 & 0 & 2/5 & 0 & 0
\end{bmatrix} v + \begin{bmatrix}
1/5 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}$$
• If we start with \( v = e_N \), then the sequence of estimates of the distribution of
  the walker that we get is,

\[
\begin{bmatrix}
  1 & 1/5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
  35/75 & 8/75 & 95/375 & 2353/5625 \\
  8/75 & 8/75 & 568/5625 & 0.345 \\
  20/75 & 20/75 & 1228/5625 & 0.066 \\
  6/75 & 142/375 & 786/5625 & 0.145 \\
  6/75 & 110/375 & 690/5625 & 0.249 \\
\end{bmatrix}
\]

• Picture 2’s similarity is 0.066
• Picture 3’s similarity is 0.145