Web as a Network

Prof. Srijan Kumar
Project Resources

- **Compute Resources:**
  - Got everyone access to **PACE COC-ICE cluster**. Powerful machines with several CPUs and GPUs.
  - Queuing mechanism to run code, so expected to be busy before deadlines
  - **Start early, beat the competition**
Project Proposal Expectations

• **We want to make sure your projects have the potential to be successful and complete**

• **Answer the three key questions**

1. **Introduction:** What is the concrete problem definition?
2. **Baselines:** What is the existing technology? What are the shortcomings?
3. **Plan of action:** Which dataset(s) will you use? How do you plan to extend/improve the baselines?

• Make sure your dataset has appropriate ground truth
Project Proposal FAQ

• **Plan of action:** We don’t expect you to know (yet) the exact improvement you will do to the baselines. We want to see potential directions.

• Will we be graded based on our model’s performance? **No**

• Does our model have to improve over the baseline? **No, we will not consider if your model beat the baseline.**
Project Expected Progress

• **Proposal:** plan the problem, dataset, baseline, and potential improvements

• **By midterm:** dataset analysis, baseline(s) implemented, started exploring potential improvements

• **By the final:** completed all baselines and all proposed improvements
Today’s Lecture: Networks

Networks introduction

• Web as a network
• Networks properties
• Random graph model: Erdos-Renyi Model
• Random graph model: Small-world Model

Some slides are inspired by Prof. Jure Leskovec’s CS224W course at Stanford
Networks are Ubiquitous
Two Types of Networks

• Networks (also known as Natural Graphs):
  – Society is a collection of 7+ billion individuals
  – Communication systems link electronic devices
  – Interactions between genes/proteins regulate life

• Information Graphs:
  – Information/knowledge are organized and linked
  – Scene graphs: how objects in a scene relate
  – Similarity networks: take data, connect similar points
Information and Social Networks

- Google
- Cisco
- Facebook
- Amazon
- Pinterest
Networks: Knowledge Discovery

• **Universal language for describing complex data**
  – Networks from science, nature, and technology are more similar than one would expect

• **Shared vocabulary between fields**
  – Computer Science, Social Science, Physics, Economics, Statistics, Biology

• **Data availability & computational challenges**
  – Web/mobile, bio, health, and medical

• **Impact!**
  – Social networking, Drug design, AI reasoning
Why Study Networks

Learn how to process large scale networks to discover knowledge
Ways to Analyze Networks

- **Predict the type/color of a given node**
  - Node classification

- **Predict whether two nodes are linked**
  - Link prediction

- **Identify densely linked clusters of nodes**
  - Community detection

- **Measure similarity of two nodes/networks**
  - Network similarity
Application: Modeling Epidemics

- Infrastructure networks are crucial for modeling epidemics

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040961
Application: Blog Network Polarization

Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]
Application: Drug Repurposing

- **Question:** Can we predict therapeutic uses of a drug?
- **Insight:** Proteins are worker molecules in a cell. Protein interaction networks capture how the cell works.
Networks Really Matter

- If you want to understand the spread of diseases, **you need to figure out who will be in contact with whom**
- If you want to understand the structure of the Web, **you have to analyze the ‘links’**
- If you want to understand dissemination of news or evolution of science, **you have to follow the flow**
Today’s Lecture: Networks

- Networks introduction
- Web as a network
- Networks properties
  - Random graph model: Erdos-Renyi Model
  - Random graph model: Small-world Model

Some slides are inspired by Prof. Jure Leskovec’s slides
Structure of the Web

• Observations and models for the Web graph:
  – 1) We will take a real system: the Web
  – 2) We will represent it as a directed graph
  – 3) We will use the language of graph theory
    • Strongly Connected Components
  – 4) We will design a computational experiment:
    • Find In- and Out-components of a given node \( v \)
  – 5) Answer: what is the structure of the Web?
The Web as a Graph

- What does the Web “look like” at a global level?
- Web as a graph:
  - Nodes = web pages
  - Edges = hyperlinks
  - Side issue: What is a node?
    - Dynamic pages and edges created on the fly
    - “dark matter” – inaccessible database generated pages
Structure of the Web

- **Broder et al.: Altavista web crawl (Oct ’99)**
  - Web crawl is based on a large set of starting points accumulated over time from various sources, including voluntary submissions.
  - 203 million URLs and 1.5 billion links
    - **Computer:** Server with 12GB of memory
What Does the Web Look Like?

• How is the Web linked?
• What is the “map” of the Web?
• Web as a directed graph [Broder et al. 2000]:
  – Given node $v$, what can $v$ reach?
  – What other nodes can reach $v$?

\[
\begin{align*}
In(v) &= \{w \mid w \text{ can reach } v\} \\
Out(v) &= \{w \mid v \text{ can reach } w\}
\end{align*}
\]

For example:
\[
\begin{align*}
In(A) &= \{A, B, C, E, G\} \\
Out(A) &= \{A, B, C, D, F\}
\end{align*}
\]
Reasoning about Directed Graphs

- **Two types of directed graphs:**
  - **Strongly connected:** Any node can reach any node via a directed path:
    \[ \text{In}(A) = \text{Out}(A) = \{A, B, C, D, E\} \]
  - **Directed Acyclic Graph (DAG):** Has no cycles: if \( u \) can reach \( v \), then \( v \) cannot reach \( u \)

- **Any directed graph (the Web) can be expressed in terms of these two types!**
  - Is the Web a big strongly connected graph or a DAG?
A Strongly Connected Component (SCC) is a set of nodes $S$ so that:

- Every pair of nodes in $S$ can reach each other
- There is no larger set containing $S$ with this property

Strongly connected components of the graph:

- $\{A, B, C, G\}$
- $\{D\}$
- $\{E\}$
- $\{F\}$
**Strongly Connected Component**

- **Fact: Every directed graph is a DAG on its SCCs**
  - (1) SCCs partitions the nodes of $G$
    - That is, each node is in exactly one SCC
  - (2) If we build a graph $G'$ whose nodes are SCCs, and with an edge between nodes of $G'$ if there is an edge between corresponding SCCs in $G$, then $G'$ is a DAG

![Graph G with Strongly Connected Components](image)

- Strongly connected components of graph $G$: \{A, B, C, G\}, \{D\}, \{E\}, \{F\}
- $G'$ is a DAG:

![DAG G'](image)
• **Question:** How is the Web linked?
• **Method:** Take a large snapshot of the Web and try to understand how its SCCs “fit together” as a DAG
Graph Structure of the Web

- **Computational issue:**
  - Want to find a SCC containing node $v$?

- **Observation:**
  - $\text{Out}(v)$ … nodes that can be reached from $v$
  - **SCC containing $v$ is:** $\text{Out}(v) \cap \text{In}(v)$
    $= \text{Out}(v, G) \cap \text{Out}(v, G')$, where $G'$ is $G$ with all edge directions flipped
Out(A) \cap In(A) = SCC

- Example:
  - Out(A) = \{A, B, D, E, F, G, H\}
  - In(A) = \{A, B, C, D, E\}
  - So, SCC(A) = Out(A) \cap In(A) = \{A, B, D, E\}
Graph Structure of the Web

• **There is a single giant SCC**
  – That is, there won’t be two SCCs

• **Why only 1 big SCC? Heuristic argument:**
  – Assume two equally big SCCs.
  – It just takes 1 page from one SCC to link to the other SCC.
  – If the two SCCs have millions of pages the likelihood of this not happening is very very small.
Structure of the Web

- **Directed version of the Web graph:**
  - Altavista crawl from October 1999
    - 203 million URLs, 1.5 billion links
- **Computation:**
  - Compute IN(v) and OUT(v) by starting at random nodes.
  - **Observation:** The BFS either visits many nodes or gets quickly stuck.

![Diagram](reachability.png)
Result: Based on IN and OUT of a random node $v$:

- $\text{Out}(v) \approx 100$ million (50% nodes)
- $\text{In}(v) \approx 100$ million (50% nodes)
- Largest SCC: 56 million (28% nodes)

• What does this tell us about the conceptual picture of the Web graph?

The cumulative distributions of the nodes covered in these BFS runs are summarized in Fig. 7. They reveal that the true structure of the Web graph must be somewhat subtler than a ‘small world’ phenomenon in which a browser can pass from any Web page to any other with a few clicks. We explicate this structure in Section 3.
Bowtie Structure of the Web

203 million pages, 1.5 billion links [Broder et al. 2000]
What did We Learn/Not Learn?

- **What did we learn:**
  - Conceptual organization of the Web (i.e., the bowtie)
- **Unanswered questions and challenges:**
  - Model treats all pages as equal
    - Google’s homepage == my homepage
  - What are the most important pages?
    - How many pages have $k$ in-links as a function of $k$?
      - The degree distribution: $\sim k^{-2}$
  - Internal structure inside giant SCC
    - Clusters, implicit communities?
  - How far apart are nodes in the giant SCC:
    - Distance = number of edges in shortest path
    - Avg. = 16 [Broder et al.]
Today’s Lecture: Networks

✓ Networks introduction
✓ Web as a network
✓ Networks properties
   • Random graph model: Erdos-Renyi Random Graph Model
   • Random graph model: Small-world Model

Some slides are inspired by Prof. Jure Leskovec’s slides
Plan: Key Network Properties

- **Degree distribution:** $P(k)$
- **Path length:** $h$
- **Clustering coefficient:** $C$
• **Degree distribution** $P(k)$: Probability that a randomly chosen node has degree $k$
  – $N_k = \# \text{ nodes with degree } k$
• Normalized histogram: $P(k) = \frac{N_k}{N}$
• **Example: 11 nodes**
  – Degree 1: 6 nodes
  – Degree 2: 1 node
  – Degree 3: 2 nodes
  – Degree 4: 1 node
Degree Distribution in Real Networks

- **Power-law distribution**: $k$ vs $N_k$ is linear in log-log scale
- **Long-tail distribution**: Most nodes have a small degree, very few nodes have a high degree

$N_k$: number of nodes with degree $k$

$k$: degree
Path Length and Network Diameter

- **Diameter:** The maximum (shortest path) distance between any pair of nodes in a graph

- **Average path length** for a connected graph (component) or a strongly connected (component of a) directed graph

  \[ \bar{h} = \frac{1}{2E_{\text{max}}} \sum_{i,j \neq i} h_{ij} \]

  where \( h_{ij} \) is the distance from node \( i \) to node \( j \)

  - We compute the average only over the connected pairs of nodes
Finding Shortest Paths

• **Breadth First Search:**
  – Start with node $u$, mark it to be at distance $h_u(u) = 0$, add $u$ to the queue
  – While the queue not empty:
    • Take node $v$ off the queue, put its unmarked neighbors $w$ into the queue and mark $h_u(w) = h_u(v) + 1$
Clustering Coefficient

- **Clustering coefficient:**
  - What portion of $i$’s neighbors are connected?
  - Node $i$ with degree $k_i$
  - $C_i \in [0, 1]$
  - $C_i = \frac{2e_i}{k_i(k_i - 1)}$ where $e_i$ is the number of edges between the neighbors of node $i$

- **Average clustering coefficient:**
  \[ C = \frac{1}{N} \sum_{i=1}^{N} C_i \]
Clustering Coefficient: Example

- **Node b:** \( k_B = 2, \ e_B = 1, \ C_B = \frac{2}{2} = 1 \)

- **Node d:** \( k_D = 4, \ e_D = 2, \ C_D = \frac{4}{12} = \frac{1}{3} \)
Real-World Network Example

• Let’s measure $P(k)$, $h$ and $C$ on a real-world network!

• **MSN Messenger activity in June 2006:**
  – 245 million users logged in
  – 180 million users engaged in conversations
  – More than 30 billion conversations
  – More than 255 billion exchanged messages
Communication Network

• Network: 180M people, 1.3B edges
Messaging as a Multi-Graph

- **Messaging as an undirected graph**
- **Edge** \((u,v)\) if users \(u\) and \(v\) exchanged at least 1 message
- \(N=180\) million people
- \(E=1.3\) billion edges
MSN Network: Connectivity

![Graph showing the distribution of weakly connected component sizes in the MSN network](image)

- **Count** vs. **Weakly connected component size**
- The largest component (99.9% of the nodes) is indicated in the graph.

Srijan Kumar, Georgia Tech, CSE6240 Spring 2020: Web Search and Text Mining
MSN Network: Degree Distribution
Note: We plotted the same data as on the previous slide, just the axes are now logarithmic.
**MSN Network: Clustering**

\[ C_k: \text{average } C_i \text{ of nodes } i \text{ of degree } k \]

\[ C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i \]

**Average clustering of the MSN network** \( C = 0.1140 \)
MSN Network: Diameter

Number of links between pairs of nodes

Average path length 6.6
90% of the nodes can be reached in < 8 hops

<table>
<thead>
<tr>
<th>Steps</th>
<th>#Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>3,96</td>
</tr>
<tr>
<td>4</td>
<td>8,648</td>
</tr>
<tr>
<td>5</td>
<td>3,299,252</td>
</tr>
<tr>
<td>6</td>
<td>28,395,849</td>
</tr>
<tr>
<td>7</td>
<td>79,059,497</td>
</tr>
<tr>
<td>8</td>
<td>52,995,778</td>
</tr>
<tr>
<td>9</td>
<td>10,321,008</td>
</tr>
<tr>
<td>10</td>
<td>1,955,007</td>
</tr>
<tr>
<td>11</td>
<td>518,410</td>
</tr>
<tr>
<td>12</td>
<td>149,945</td>
</tr>
<tr>
<td>13</td>
<td>44,616</td>
</tr>
<tr>
<td>14</td>
<td>13,740</td>
</tr>
<tr>
<td>15</td>
<td>4,476</td>
</tr>
<tr>
<td>16</td>
<td>1,542</td>
</tr>
<tr>
<td>17</td>
<td>536</td>
</tr>
<tr>
<td>18</td>
<td>167</td>
</tr>
<tr>
<td>19</td>
<td>71</td>
</tr>
<tr>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
</tbody>
</table>
Plan: Key Network Properties

- **Degree distribution**: $P(k) = 14.4$
- Heavily skewed average degree
- **Path length**: $h = 6.6$
- **Clustering coefficient**: $C = 0.11$

Are these values “expected”? Are they “surprising”? To answer this we need a null-model!
Next Lecture: Random Graphs

- Networks introduction
- Web as a network
- Networks properties
- Random graph model: Erdos-Renyi Random Graph Model
- Random graph model: Small-world Model

Some slides are inspired by Prof. Jure Leskovec’s slides