Outline

• Graphs problems and representations
• Structure of social networks
• Applications of structural analysis
What is a graph?

- $G = (V,E)$
  - $V$ represents the set of vertices (nodes)
  - $E$ represents the set of edges (links)
  - Both vertices and edges may contain additional information

- Different types of graphs:
  - Directed vs. undirected edges
  - Presence or absence of cycles

- Graphs are everywhere:
  - Hyperlink structure of the Web
  - Highway system
  - Social networks
Some graph problems

• Finding shortest paths
  • Routing Internet traffic and UPS trucks
• Finding minimum spanning trees
  • Telco laying down fiber
• Finding Max Flow
  • Airline scheduling
• Identify “special” nodes and communities
  • Breaking up terrorist cells, spread of avian flu
• Bipartite matching
  • Tinder
• PageRank
Graphs are hard!

- Poor locality of memory access
- Very little work per vertex
- Changing degree of parallelism
- Running over many machines makes the problem worse

Graph storage:
- Flat Files: no query support
- RDBMS: can store the graph with limited support for graph query

State of the art today:
- Write your own infrastructure
- MapReduce – tends to be inefficient
Distributed Graph Processing

- Google’s Pregel
  - Large-scale graph processing
  - Vertex centered computation
- Apache Giraph
  - Open source
  - Iterative graph processing
  - Used at Facebook
- Twitter’s Cassovary
  - In-memory computation
  - Used for: “Who to Follow” and “Similar to”
  - Very simple to use (no need for persistence, databases or partitions)
- Neo4j Graph Database
  - Flexible schema
  - Powerful query language, ACID
Representing graphs

Two common representations:

• Adjacency matrix

• Adjacency list
Adjacency matrices

Represent a graph as an $n \times n$ square matrix $M$

- $n = |V|$
- $M_{ij} = 1$ means a link from node $i$ to $j$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Adjacency matrices: critique

Advantages:
  • Easy mathematical manipulation
  • Iteration over rows and columns corresponds to computations on outlinks and inlinks

Disadvantages:
  • Lots of zeros for sparse matrices
  • Lots of wasted space
Adjacency lists

Take adjacency matrices... and throw away all the zeros

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

1: 2, 4
2: 1, 3, 4
3: 1
4: 1, 3
Adjacency lists: critique

Advantages:
  • Much more compact representation
  • Easy to compute over outlinks

Disadvantages:
  • Much more difficult to compute over inlinks
Social graphs
Social graphs

- Asymmetric follow relationship: very skewed graphs

- Very valuable “interest graphs”

- Huge graphs:
What can networks tell us?

• The strength of weak ties [Granovetter ’73]

• Motivating question: How do people find new jobs?
  • Through acquaintances rather than close friends
  • Surprising fact: discovery is enabled by weak ties

• Understanding structure affords deep insights

• Interplay between sociology and graph theory
Question: What are the mechanisms by which node arrive and depart and by which edges form and vanish?

If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.
Over time…

(a) *Before new edges form.*  
(b) *After new edges form.*

… new edges are forming.

But not all due to triadic closure (e.g. DG)
Clustering Coefficient

• The probability that two randomly selected friends of A are friends with each other.

• The fraction of pairs of A’s friends that are connected to each other by edges.

• For node A:
  • at a) 1/6
  • at b) 1/2

• The more strongly triadic closure is operating in the neighborhood of the node, the higher the clustering coefficient will tend to be.
Reasons for Triadic Closure

- Opportunity
- Trust
- Incentive
Strength of weak ties

- Definition: a **bridge** in a graph is an edge whose removal disconnects the endpoints.

Bridges are presumably extremely rare in real social networks!
Strength of weak ties

• Definition: a **local bridge** in a graph is an edge whose endpoints have no common neighbor.
Types of edges

- **Structural approach:**
  - Local bridges or not

- **Interpersonal approach:**
  - Weak or strong

**Challenge:** how to link them?
Strong Triadic Closure

- **Strong Triadic Closure Property:** if the node has strong ties to two neighbors, then these neighbors must have at least a weak tie between them.
Claim: If a node A in a network satisfies the Strong Triadic Closure Property and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.

Consequence: all local bridges are weak ties!
Strength of weak ties

Proof: by contradiction

Strong Triadic Closure says the B-C edge must exist, but the definition of a local bridge says it cannot.
• **Discovery** is enabled by weak ties
  - Surprising strength of weak ties!

• Simple *structural model* explains this cleanly

• Applies to Twitter/Facebook
Tie strength on Facebook

All Friends

Maintained Relationships

One-way Communication

Mutual Communication

Active Network Sizes

- Maintained Relationships
- One-way communication
- Reciprocal communication

# of People vs. Network Size
Tie strength on Twitter

• **Stronger…**
  • Directed tweets: @someone

• … and **weaker ties**
  • Followers

• **The number of strong ties remains relatively modest**
  • Below 50 even for users with over 1000 followers.