Sociology and CS

Philip Chan

Sociology Problems

- Problem 1
  - How close are people connected?
  - “Small World”
- Problem 2
  - Who is the most connected?
  - “Connector”

Small World

Problem 1

How close are people connected? (Problem Understanding)

- Are people
  - closely connected,
  - not closely connected,
  - isolated into groups,
  - …

Degree of Separation

- The number of connections to reach another person

Milgram’s Experiment

- Stanley Milgram, psychologist
- Experiment in the late 1960’s
  - Chain letter to gather data
  - Stockbroker in Boston
  - 160 people in Omaha, Nebraska
    - Given a packet
    - Add name and forward it to another person who might be closer to the stockbroker
- Partial “social network”

Small World

- Six degrees of separation
- Everyone is connected to everyone by a few people—about 6 on the average.
  - Obama might be 6 connections away from you
- "Small world" phenomenon

Bacon Number

- Number of connections to reach actor Kevin Bacon
- http://oracleofbacon.org/
- Is a connection in this network different from the one in Milgram's experiment?

Problem Formulation

- Given (input)
- Find (output)
- Simplification
  - don't care about …

Problem Formulation

- Given (input)
  - People
  - Connections/links/friendships
- Find (output)
  - the average number of connections between two people
- Simplification
  - don't care about how long/strong/… the friendships are

Problem Formulation

- Formulate it into a graph problem (abstraction)
- Given (input)
  - People
  - Connections
- Find (output)
  - the average number of connections between two people

Problem Formulation

- Formulate it into a graph problem (abstraction)
- Given (input)
  - People -> vertices
  - Connections -> edges
- Find (output)
  - the average number of connections between two people -> ?

Problem Formulation

- Formulate it into a graph problem (abstraction)
- Given (input)
  - People -> vertices
  - Connections -> edges
- Find (output)
  - the average number of connections between two people -> average shortest path length

Algorithm

- Ideas?

Algorithm

- Shortest Path
  - Dijkstra's algorithm
- Limitations?

Algorithm

- Shortest Path
  - Dijkstra's algorithm
- Limitations? Single-source
- All-pair Shortest Path
  - Floyd's algorithm
- This could be an overkill, why?

Algorithm

- Unweighted edges
  - Each edge has the same weight of 1
- Simpler algorithm?

Algorithm

- Breadth-first search (BFS)

Algorithm

- Breadth-first search (BFS)
  - Data structure to remember visited vertices

Algorithm

- Breadth-first search (BFS)
  - Data structure to remember visited vertices
  - Single source; repeat for each vertex to start

Algorithm

- Breadth-first search (BFS)
  - Data structure to remember visited vertices
  - Single source; repeat for each vertex to start
  - ShortestPath(x,y) = shortestPath(y,x)

Implementation

- Which data structure to represent a graph (vertices and edges)?

Implementation

- Which data structure to represent a graph (vertices and edges)?
  - Adjacency matrix
  - Adjacency list
  - Tradeoffs?

Implementation

- Which data structure to represent a graph (vertices and edges)?
  - Adjacency matrix
  - Adjacency list
  - Tradeoffs?
    - Time
    - Space

Adjacency Matrix vs List

- Time
  - Speed of what?

Adjacency Matrix vs List

- Time
  - Speed of key operations in the algorithm
    - Algorithm:
    - Key operation:

Adjacency Matrix vs List

- Time
  - Speed of key operations in the algorithm
    - Algorithm: BFS
    - Key operation: identifying children

Adjacency Matrix vs List

- Time
  - Speed of key operations in the algorithm
    - Algorithm: BFS
    - Key operation: identifying children
- Space
  - Amount of data in the problem

Adjacency Matrix vs List

- Time
  - Speed of key operations in the algorithm
    - Algorithm: BFS
    - Key operation: identifying children
- Space
  - Amount of data in the problem
    - Number of people/vertices
    - Number of friends/edges each person has

Connector

Problem 2

Revolutionary War

- Spreading the word that the British is going to attack
- Paul Revere vs William Dawes
  - Revere was more successful than Dawes
  - History books remember Revere more

Who is the most connected? (Problem understanding)

- What does that mean?

Who is the most connected?

- What does that mean?
  - The person with the most friends?

Who is the most connected?

- What does that mean?
  - The person with the most friends?
    - Phone book experiment
      - 250 random surnames
      - Number of friends with those surnames

Who is the most connected?

- What does that mean?
  - The person with the most friends?
    - Phone book experiment
      - 250 random surnames
      - Number of friends with those surnames
    - Number of friends have a wide range
      - Random sample: 9 -118
      - Conference in Princeton: 16 - 108

Who is the most connected?

- What does that mean?
  - The person with the most friends?
    - How to formulate it into a graph problem?

Who is the most connected?

- What does that mean?
  - The person with the most friends?
    - How to formulate it into a graph problem?
      - Output: the vertex with the highest degree

Who is the most connected?

- What does that mean?
  - The person with the most friends?
  - Are all friends equal?

Who is the most connected?

- What does that mean?
  - The person with the most friends?
  - Are all friends equal?
    - You have 100 friends
    - Michelle Obama has only one friend:
      - Barack Obama, who has a lot of friends
    - Not just how many, but who you know

Milgram's Experiment

- 24 letters get to the stockbroker at home
  - 16 from Mr. Jacobs
- The rest get to the stockbroker at work
  - Majority from Mr. Brown and Mr. Jones
- Overall, half of the letters came through the three people
- But Milgram started from a random set of people
- What does this suggest?

Milgram's Experiment

- Average degree of separation is six, but:
  - A small number of special people connect to many people in a few steps
    - Small degree of separation
  - The rest of us are connected to those special people
- Called "Connectors" by Gladwell

Getting a Job experiment

- Mark Granovetter, sociologist
- Experiment in 1974
  - 19%: formal means—advertisements, headhunters
  - 20%: apply directly
  - 56%: personal connection

Getting a Job experiment

- Personal connection
  - 17%: see often
  - 56%: see occasionally
  - 28%: see rarely
- What does this suggest?

Getting a Job experiment

- Personal connection
  - 17%: see often → friends
  - 56%: see occasionally → acquaintances
  - 28%: see rarely → almost strangers?
- What does this suggest?
  - Getting jobs via acquaintances
    - Why?

Getting a Job experiment

- Personal connection
  - 17%: see often → friends
  - 56%: see occasionally → acquaintances
  - 28%: see rarely → almost strangers?
- What does this suggest?
  - Getting jobs via acquaintances
    - connect you to a different world
    - might have a lot connections
  - "The Strength of Weak Ties"

Who is the most connected?

- "Connector"
  - How many friends does one have?
  - What kind of friends does one have?
- How do you find Connectors?
- How do you formulate it into a graph problem?

Problem Formulation

- Given (input)
  - People -> vertices
  - Connections -> edges
- Find (output)
  - Person with the "best" Connector score
    - Part of the algorithm is to define the Connector score
- Simplification
  - Don't care about how strong/long/… the friendships/connections are

Algorithm 1: Connector Score

- Motivation:
  - "Friends" who are closer have higher scores
- Friends of
  - distance 1, score = ?
  - distance 2, score = ?
  - distance 3, score = ?
  - distance *d*, score = ?

Algorithm 1: Connector Score

- Motivation:
  - "Friends" who are closer have higher scores
- Friends of
  - distance 1, score = ?
  - distance 2, score = ?
  - distance 3, score = ?
  - distance *d*, score = 1/*d*, $1/d^2$, …

Algorithm 1: Adding the scores

- How to enumerate the people so that we can add the scores?

Algorithm 1: Adding the scores

- How to enumerate the people so that we can add the scores?
  - BFS

Algorithm 1: Adding the scores

- How to enumerate the people so that we can add the scores?
  - BFS
  - Is score(x, y) the same as score(y, x)?

Algorithm 2: Connector Score

- Motivation:
  - Degree of separation (number of connections) to other people is small
- Connector score:
  - Ideas?

Algorithm 2: Connector Score

- Motivation:
  - Degree of separation (number of connections) to other people is small
- Connector score:
  - Average degree of separation from a person to every other person

Algorithm 2: Adding the scores

- How to enumerate the people so that we can add the scores?

Algorithm 1 vs 2

- How do you compare the two algorithms?

Algorithm 1 vs 2

- How do you compare the two algorithms?
  - Changing Algorithm 1 slightly will yield Algorithm 2, how?

Algorithm 1 vs 2

- How do you compare the two algorithms?
  - Changing Algorithm 1 slightly will yield Algorithm 2, how?
  - Algorithm 1 is more flexible, why? But?

Algorithm 3: Connector Score

- Motivation:
  - "bridge"
  - If the person is not there, it takes a longer path for two people to connect
- Connector Score:
  - Ideas?

Algorithm 3: Connector Score

- Motivation:
  - "bridge"
  - If the person is not there, it takes a longer path for two people to connect
- Connector Score ("betweenness"):
  - Number of times the person appears on the shortest path between all pairs
  - For one pair, what if multiple shortest paths of the same length (ties)?

Algorithm 3: Connector Score

- Motivation:
  - "bridge"
  - If the person is not there, it takes a longer path for two people to connect
- Connector Score ("betweenness"):
  - Number of times the person appears on the shortest path between all pairs

Algorithm 3: Connector Score

- Motivation:
  - "bridge"
  - If the person is not there, it takes a longer path for two people to connect
- Connector Score ("betweenness"):
  - Number of times the person appears on the shortest path between all pairs
  - For one pair, what if multiple shortest paths of the same length exist (ties)?
    - Fractional score for each person/vertex

Summary

- Problem 1: Degree of Separation—how close are we from each other?
- Problem 2: Connector—who is the most connected?
  - Algorithm 1: score = 1/d
  - Algorithm 2: score = degree of separation
    - Length (not vertices) of shortest path is needed
  - Algorithm 3: score = betweenness
    - Vertices (not length) on the shortest path are needed

Can Sociology Help CS?

Search Engines

- How do they rank pages?

Search Engines

- How do they rank pages?
  - Content: words
    - Most search engines in 1990's

Search Engines

- How do they rank pages?
  - Content: words
    - Most search engines in 1990's
  - Link structure: incoming and outgoing links
    - PageRank algorithm (1998) → Google

Search Engines

- How do they rank pages?
  - Content: words
    - Most search engines in 1990's
  - Link structure: incoming and outgoing links
    - PageRank algorithm (1998) → Google
  - User data: click data

Key Ideas of PageRank

- How to use link structure to score web pages?

Key Ideas of PageRank

- How to use link structure to score web pages?
  - If a web page is important
    - What can we say about the number of incoming links?

Key Ideas of PageRank

- How to use link structure to score web pages?
  - If a web page is important
    - What can we say about the number of incoming links?
    - Are all incoming links equal?

Key Ideas of PageRank

- How to use link structure to score web pages?
  - If a web page is important
    - What can we say about the number of incoming links?
    - Are all incoming links equal?
      - How important are they? (recursive)

Key Ideas of PageRank

- How to use link structure to score web pages?
  - If a web page is important
    - What can we say about the number of incoming links?
    - Are all incoming links equal?
      - How important are they? (recursive)
      - How many outgoing links do they have?

Key Ideas of PageRank

- How to use link structure to score web pages?
  - If a web page is important
    - What can we say about the number of incoming links?
    - Are all incoming links equal?
      - How important are they? (recursive)
      - How many outgoing links do they have?
- A link is similar to a vote/recommendation

Key Ideas of PageRank

- How to use link structure to score web pages?
  - If a web page is important
    - What can we say about the number of incoming links?
    - Are all incoming links equal?
      - How important are they? (recursive)
      - How many outgoing links do they have?
- A link is similar to a vote/recommendation
- Is this similar to finding the “Connector?”

PageRank

- PageRank(p) ~=
  - $\text{Sum}_{\,i=incoming(p)}$ PageRank(i) / #outgoing(i)

Reading Assignment

- How and why does the Dijkstra’s shortest path algorithm work?

Reading Assignment

- Handout on “Representation of Spatial Objects”
  - P. Rigaux, M. Scholl & A. Voisard
  - *Spatial Databases with Application to GIS*
  - Morgan Kaufmann, 2002
- How and why does the Dijkstra’s shortest path algorithm work?