Finding Patterns in Complex Social Networks

Lennon Ganz

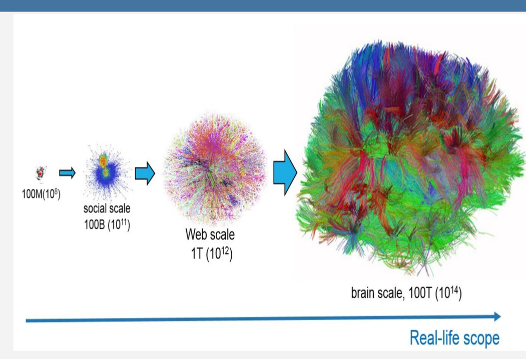
Mentor: Dr. Yinghui Wu

Faculty Advisor: Dr. Xifeng Yan

UCSB UNIVERSITY OF CALIFORNIA SANTA BARBARA

Overview

Motivation



- Networks are all around us
- Large networks are difficult to analyze and understand

Challenge

- Networks quickly become large and complex
- Questions (queries) are difficult to write and often require learning a special query language
- Algorithms become slow as size of data increases

Goals

- Develop a simple, intuitive platform to analyze complex social networks
- Improve the efficiency of existing search algorithms and develop an efficient pattern matching algorithm

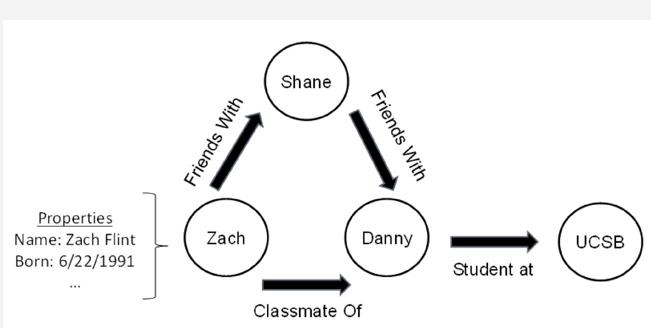
Key Terms

Query - A question or request for information

Algorithm - A set of instructions to answer a question or complete a process

"Social" Network - A network of social interactions

Graph - Data structure that represents entities (nodes) and the relationships (edges) between them



Node - An entity in a graph (e.g. person, place event)

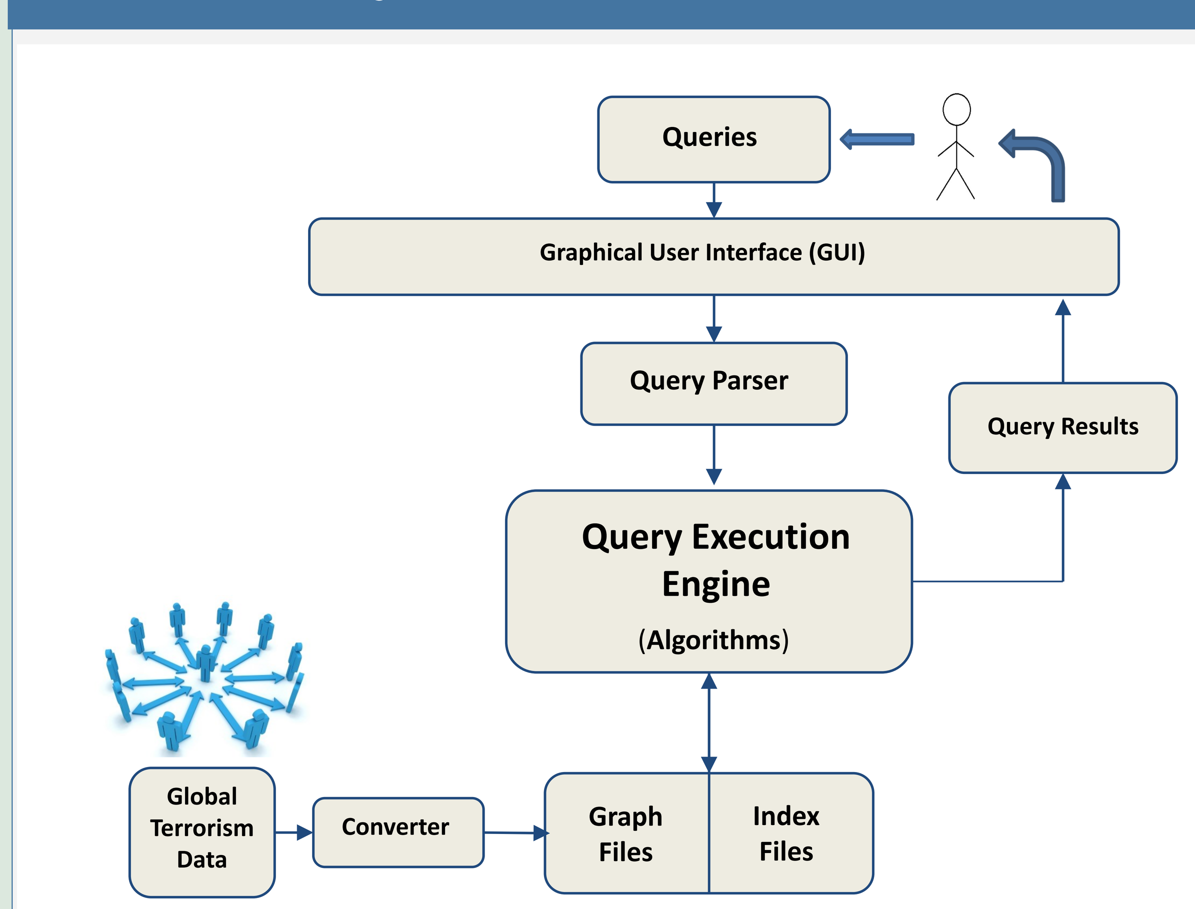
Edge - A connection between two nodes in a graph, representing a relationship between them

Density - # of edges divided by # of nodes

Types of Queries

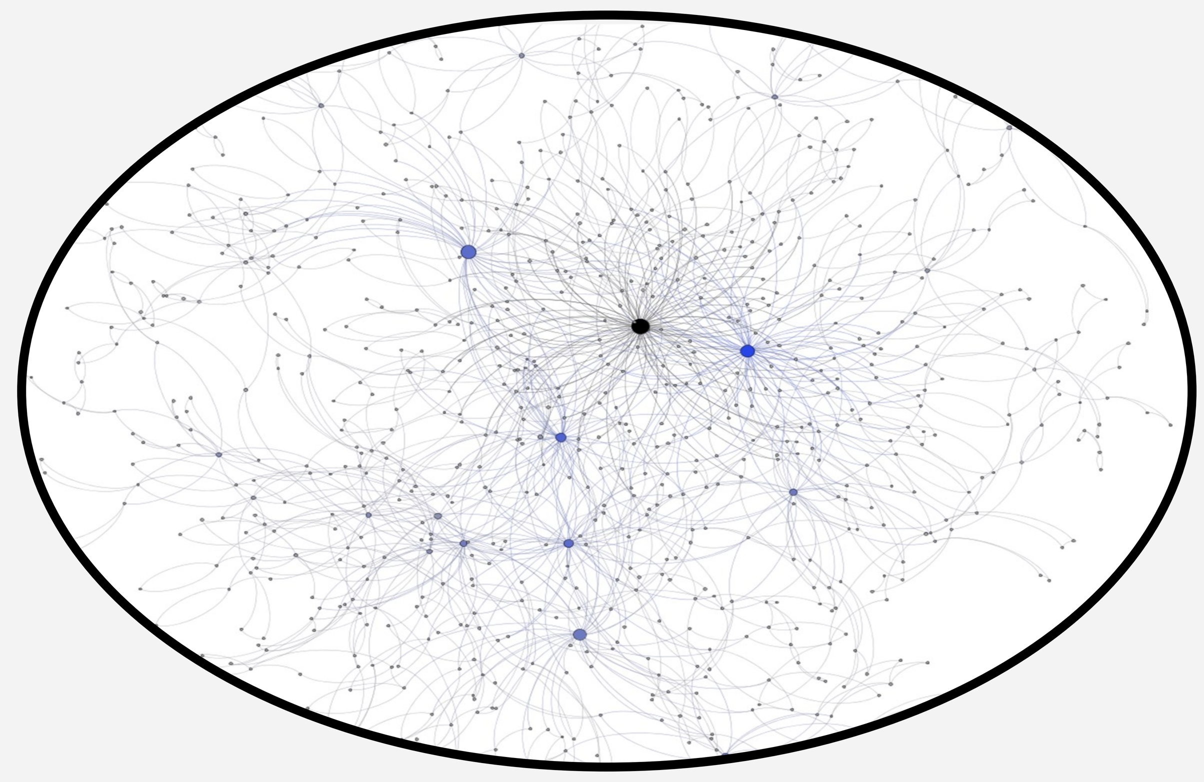
Type	Example (Dataset = Global Terrorism Database)
Aggregation	"How many incidents were there in 2005?" "What was the most common type of weapon used?"
Reachability	"Have these two terrorist groups ever collaborated on an attack?"
Pattern Matching	"Find instances where a terrorist group targeted a government official who had also been attacked by a different group earlier that year"

System Overview



Visualization

GLOBAL TERRORISM DATABASE



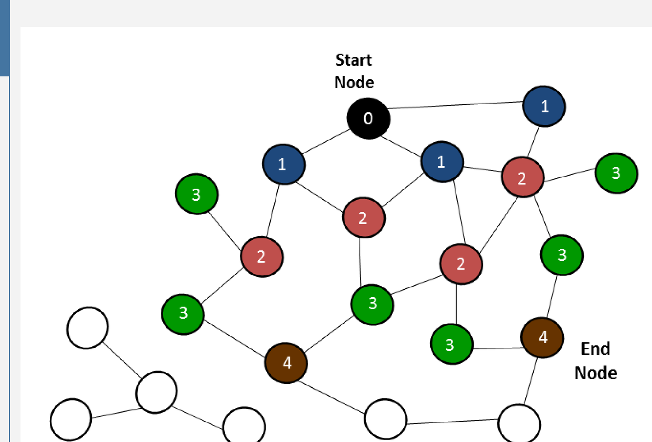
~500 Incidents represented out of 113,000 in GTD

Nodes: Perpetrators, Incidents, and Targets

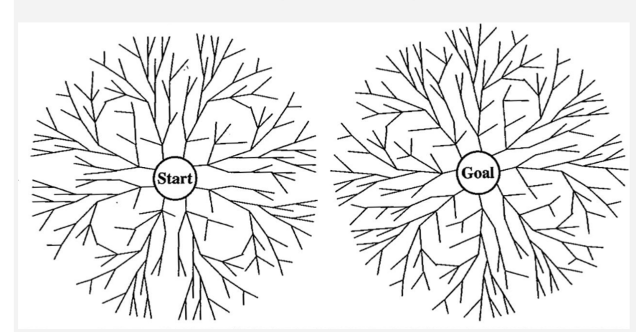
Edges: Perpetrated/Targeted

Reachability Algorithms

Breadth-First Search (BFS)



Bidirectional Breadth-First Search (BIBFS)

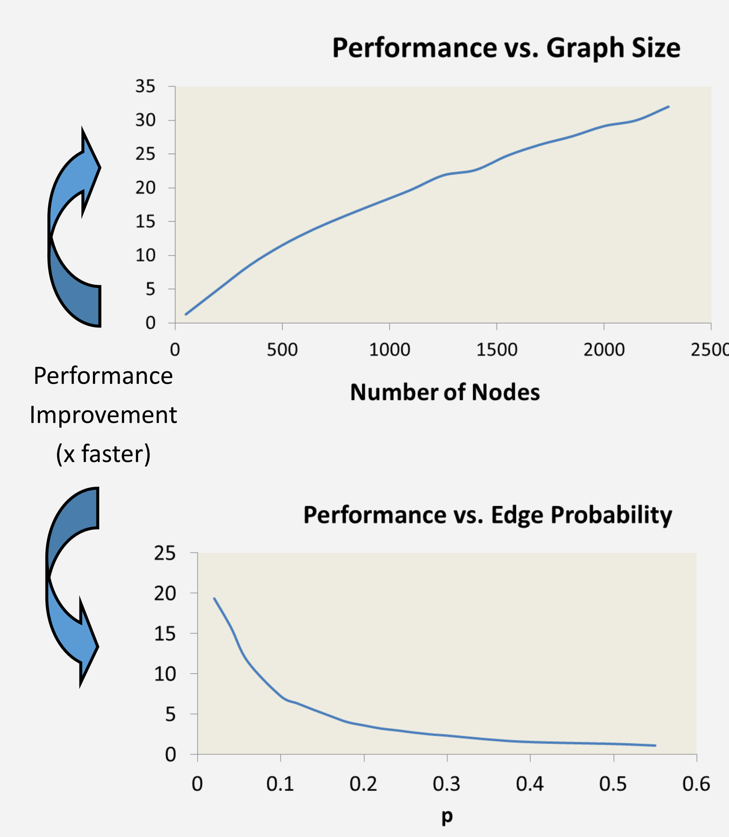


How much faster is BIBFS?

$$Improvement = \frac{BFS\ Running\ Time}{BIBFS\ Running\ Time}$$

Improvement

- As graph size increases with constant density, BIBFS convincingly outperforms the standard BFS approach by over 30x
- As edge density increases with constant number of nodes, BIBFS remains faster but approaches the speed of BFS

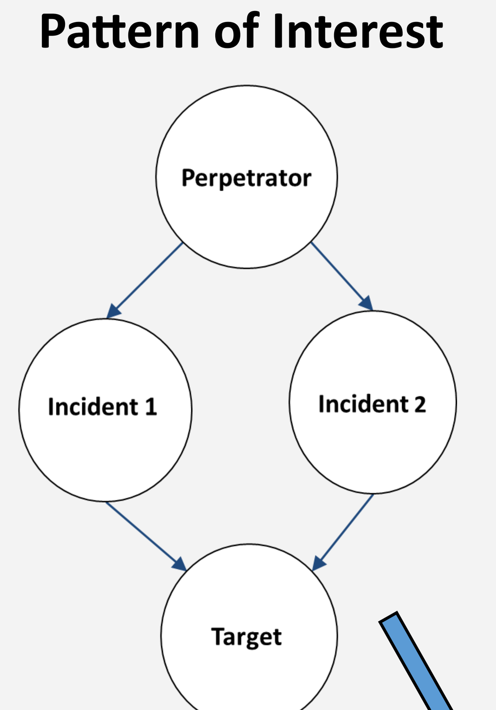


Drawbacks

- Larger overhead (resource-heavy)
- Must know starting and ending nodes (unable to freely explore graph)

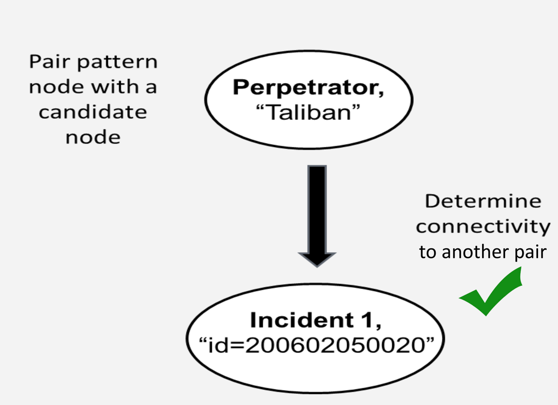
Pattern Matching Algorithm

Pattern of Interest



First, candidate nodes (nodes that might fit in the pattern) are quickly found using inverted indexes

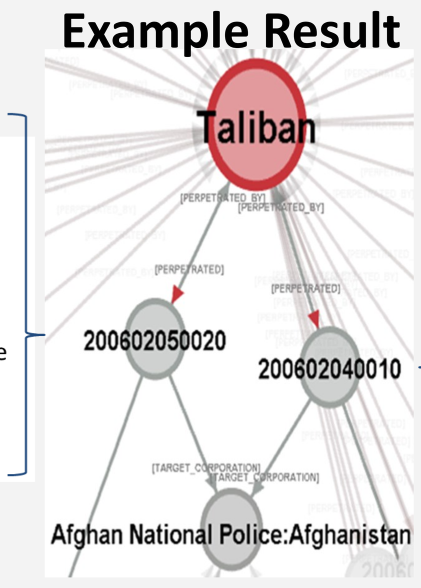
Second, pattern and candidate nodes are paired and their connectivity is inspected to find valid matches.



This is repeated for every pattern/candidate node pair and returns a match if the entire pattern is found

Example Result

200602050020
Target: Afghan National Police
Date: 2/5/06
Country: Afghanistan
City: Khakrez
Weapon Type: Land Mine
Casualties: 6
Wounded: 2
Summary: ....



200602040010
Target: Afghan National Police
Date: 2/4/06
Country: Afghanistan
City: Kandahar
Weapon Type: IED
Casualties: 2
Wounded: 2
Summary: ....

Indexing

Inverted Index

A mapping of properties to the nodes that contain that property.

Similar to the index at the back of a book.

Country	Incidents
United States	ID=23422, ID=66430, ...
Iraq	ID=72230, ID=94102, ...
Nepal	ID=40660, ID=19430, ...

Prevents repeated, expensive searches of entire dataset

Summary

Results

- Bidirectional searching is almost always faster than one-way searching, upwards of 30x or more
- Inverted indexing greatly improves performance for finding specific nodes at the cost of memory/storage. Access time is reduced from seconds to milliseconds
- The pattern matching algorithm successfully matches patterns of nodes and edges each with specified properties

Accomplishment

This system makes the process of analyzing large networks and graphs faster and more intuitive. This allows for a better understanding of each specific dataset and ultimately a better understanding of the complex connections and relationships present in the world around us

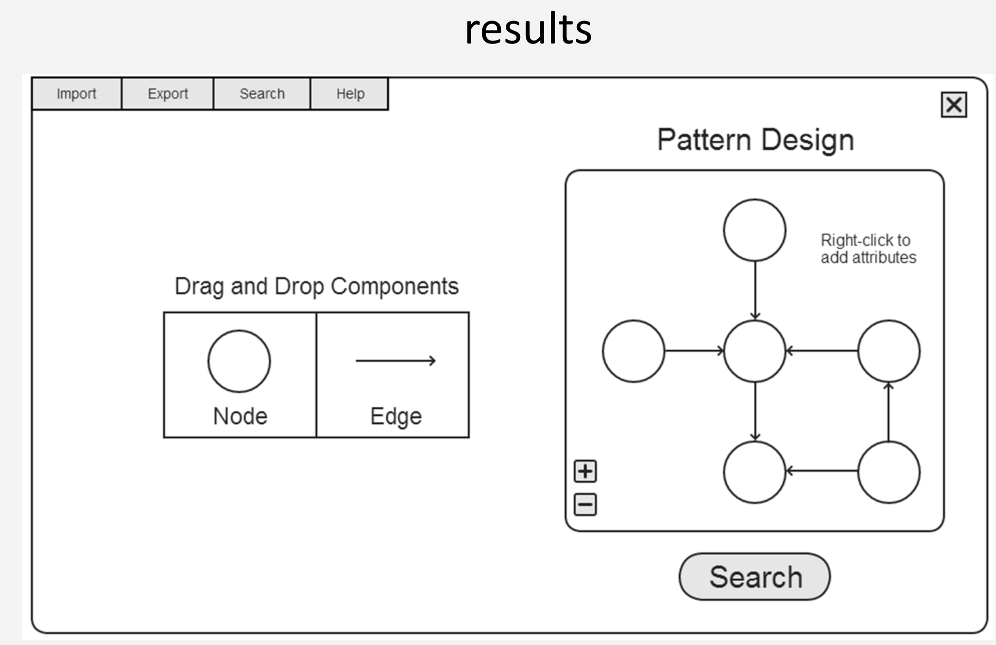
Future Development and Goals

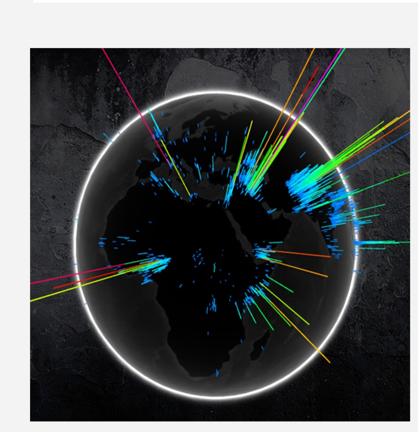
Vision

For this system to be used as a network analysis tool for many different types of data; social networks, cyber security networks, neural networks and more.

Graphical User Interface (GUI)

Allow users to easily and visually write queries, create search patterns, and view results





Geographic Visualization

Data and search results can be presented visually on a globe

Application Programming Interface (API)

Allow other developers/users to import and analyze other sets of data

Application Programming Interface