Traversed Internship

Frank Sanchez

The Company

- Founded in 2014
- Big data analytics
  - Product: Proximity
    - a high-performance platform for analyzing social media and unstructured text in real-time
    - “Finding the what, when, and where in social media”

TRAVERSED

CONNECTING YOU TO YOUR DATA

What I Worked On

- GUI
  - Worked on existing web application
  - Carrot2 – clustering plugin
  - Cluster tweets based on phrases
  - Created a table to display clustered tweets
- Data Science: Investigatory Exercise
  - Find data sources for a certain event
  - Reddit API: retrieving Json data
  - Use data to attempt accurate prediction

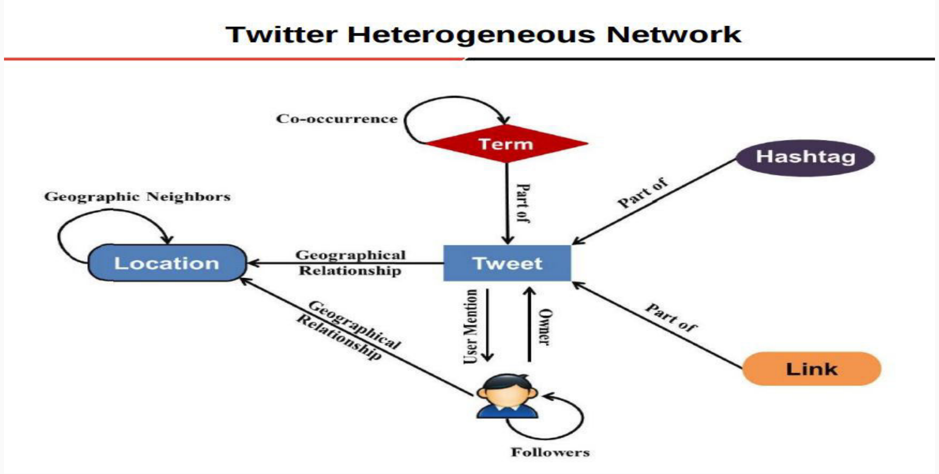
Main Project

- Social media event detection and forecasting program
  - Implementation of a research paper
- Goal
  - To identify highly anomalous subgraphs within a twitter heterogeneous graph
    - Graph loader
    - Empirical calibration
    - Scan

neo4j

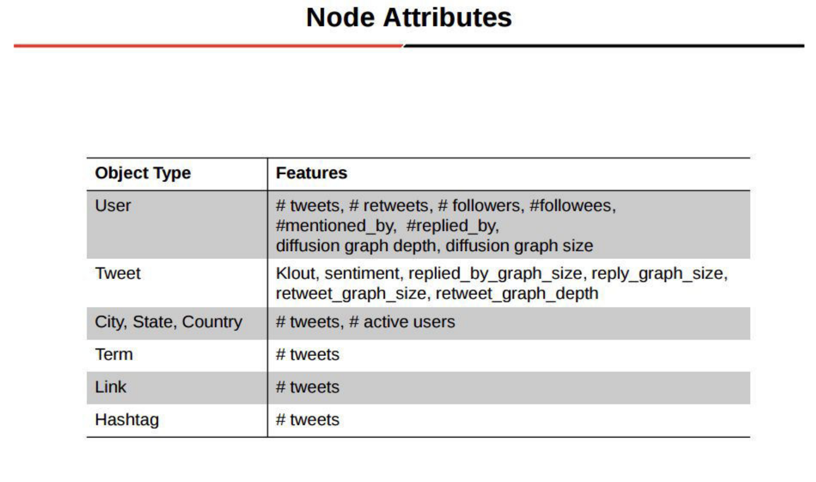
Graph Loader

- Heterogeneous graph
  - Composed of nodes , attributes, and relationship of different types
- Graph Loader
  - Twitter4j status objects
    - Uses Twitter 1% stream
    - Multiple days
  - Neo4j-OGM



Empirical Calibration Process

- Historical datasets
  - Day to day time span
- Calibrate each node with a pvalue
  - Score of anomalousness
  - Compare attributes of nodes
- Cypher query language

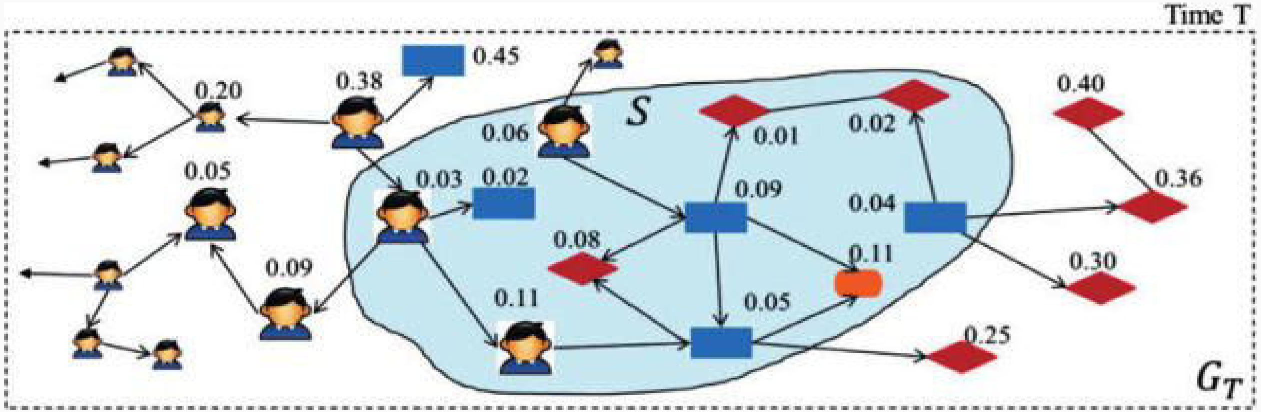


Node Attributes

Object Type	Features
User	# tweets, # retweets, # followers, #followees, #mentioned_by, #replied_by, diffusion graph depth, diffusion graph size
Tweet	Klout, sentiment, replied_by_graph_size, reply_graph_size, retweet_graph_size, retweet_graph_depth
City, State, Country	# tweets, # active users
Term	# tweets
Link	# tweets
Hashtag	# tweets

Graph Scan

- Scan the graph for connected subgraph
  - Subgraph consists of nodes with pvalue less than a given max(α)
  - The resulting subgraph may contain valuable information pertaining to an occurring event
  - Manually evaluate the returned subgraph



Challenges

- Learning Github and working on other people's code
- Dealing with new libraries and learning their APIs
- Translating a technical paper into code
  - Understanding equations/algorithms
- Working independently with little direction

Skills Used From School

- Basic Java programming knowledge
- Logic and problem solving skills from programming classes
- Starting a large program from scratch
- Discrete Math
    - Graphing terminology

What I've Learned

- Java concepts and software development practices
  - OO Design/Unit Testing
- Maven
  - Project structure
- Github
- Minor JavaServer Faces concepts
- Libraries: Carrot2, Reddit, Twitter, Twitter4j, Neo4j, Neo4j-OGM
- Graph Databases
  - Query language