An Exploration of Spatial Interaction on Twitter

**Objective / Purpose**

Twitter
- Short messages of 140 characters or less ("tweets")
- Public unless specified otherwise by user
- Users "follow" each other
- Messages of all followees are displayed on the follower’s feed
- Profile has user-contributed information about themselves
- 7 million accounts
- Over 3 billion tweets: 600,000 to 850,000 per day

Importance
- Extremely large human sensor network
- Can be mined for many interesting and practical applications
  - e.g. trends in sickness, topical maps, disaster situations

**Spatial Interaction**
- A cornerstone of geographic theory
- Models linkages and flow between locations
  - e.g. flow of commuters, raw materials, ideas

Our Goals
- Crawl and extract location-based information for 1 million users
- Construct a communication network
- Analyze location distribution
- Compare interaction at various granularities

**Methodology**

**The Communication Network**
- Following or being followed does not indicate any type of real relationship
- Relationships are determined as the interaction (@ mentions) among users
- An @ mention is directional, but for our purpose interaction will be undirected

**Processing the Data**
- List communication from one user to any other users (done by another student)
- Extract location of each user as given in their profile
- Determine a path (e.g. "Bryan, Brazos, Texas, United States") for each location
- Compile a list of residents (users) for each path A (resA)
- Determine a list of contacts for each path A (conA)
- Interaction between path A and path B = \(|\text{ConA} \cap \text{ResB}| + |\text{ConB} \cap \text{ResA}|\)

**Analyzing Available Information**
- Discover various ways users embed geographic information
- Communication among users: @ mentions
  - placing an @ before a username directs the message to them

**Mining Locations**
- Not all locations are machine-readable
- Any geocoding service is susceptible to syntax
  - e.g. misspelling, nicknames, reverse order, name changes

**Processing the Data**
- List communication from one user to any other users (done by another student)
- Extract location of each user as given in their profile
- Determine a path (e.g. "Bryan, Brazos, Texas, United States") for each location
- Compile a list of residents (users) for each path A (resA)
- Determine a list of contacts for each path A (conA)
- Interaction between path A and path B = \(|\text{ConA} \cap \text{ResB}| + |\text{ConB} \cap \text{ResA}|\)

**Communication Network**

**Invalid Locations**
- Distribution
  - Largely English-speaking
  - Surprisingly diverse
  - Iran is placed 4th, misrepresenting actuality

**Effects of Distribution**
- Interaction not normalized
- Underrepresented areas do not get displayed
- Hard to get the real picture

**Summary**
- Since location is provided by the user, it may be incorrect
- Geographic distance has some (but little) effect on interaction
- Interaction is strongly affected by language barriers

**Possibilities for the Future**
- Compare spatial interaction with other distances
  - e.g. economy, values, religious beliefs
- Model spatial interaction with a gravitational model
- Analyze usage patterns and their correlation to other variables

**Acknowledgments**

I would like to thank both of my mentors for guiding the project, providing lots of ideas, and helping troubleshoot my code.

Paths were determined using MetaCarta’s online geocoding service. Maps were obtained on Wikimedia Commons under the CC License.

**References**