CSE 590
Data Mining meets Graph Mining
Course Project Ideas

Leman Akoglu
Stony Brook University

http://www.cs.stonybrook.edu/~cse590
Anomaly detection in review data

- Users write product reviews on many online sites: Yelp, Amazon, TripAdvisor
- Data of the form: `<userID, productID, review-text, timestamp, rating>`

Task:
- How to find fake reviews and reviewers?
- What strange behaviors do fake reviewers have?
- Can you use the network to find anomalies?

Data:
- Amazon: [http://liu.cs.uic.edu/download/data/](http://liu.cs.uic.edu/download/data/)
- Yelp: [http://www.yelp.com/academic_dataset](http://www.yelp.com/academic_dataset)
Event detection in network traffic

- IPs communicating with other IPs
  - `<IP1, IP2, #bytes, protocol, time>`
- Simulated data, over ~10 days

Tasks:
- How to find **events**?
- How to pinpoint **culprits**?
- How can you **explain** the anomalies?
- How to **model** the time series?

Data:
- “Challenge” network (with subtle anomalies)
Consider a large news corpus over time, like all USA Today articles over many years, or opinion platform like Twitter/blogs/forums

Tasks:
- How can we find sentiment (+/-) associated with locations (e.g. Pittsburgh), people (e.g. Obama), and organizations (e.g. IBM)?
  - Exploit a senti-graph (bipartite): nodes-1: words, nodes-2: entities
  - Exploit sentiment associated with words (e.g. bankrupt, success, etc.)
- How does sentiment change over time?

Data:
- Collect your own data. Newspapers sell their database for several hundred dollars.
Mining question-answer sites

- Consider Q&A sites where people ask and/or answer questions. Some sites are focused: e.g. MathOverflow, StackOverflow

Tasks:
- How to automatically identify *experts*?
- How to *detect* whether a user is about to leave?
- How to estimate *quality of answers/questions*?
- How to *estimate* the response times to questions?

Data:
- StackOverflow data available online: [http://blog.stackoverflow.com/category/cc-wiki-dump/](http://blog.stackoverflow.com/category/cc-wiki-dump/)
Time series mining

- Consider the time series of Internet trends, like memes, stock prices, or #online searches.

Tasks:

- How do these time series look like? How to spot change-points?
- Can we characterize the time series (e.g. shape, distribution) so as to differentiate/classify ‘rumor’-based trends from ‘serious’ trends/topics?
  - e.g., searches on celebrities vs home sale prices?

Data:

- Google trends data available to download: http://www.google.com/trends/explore#cmpt=q
Social media mining

- Consider online platforms where people share ‘stuff’, e.g. pictures, links, videos, such as Reddit
- Several questions one can ask:

**Tasks:**
- How do upvoted posts differ from downvoted ones? (control for the same shared link)
- What makes a user more engaged to use such sites? (regular vs sporadic users)
- How to characterize the life-span of a post?

**Data:**
- Collect your own data: http://www.reddit.com/
Mining specialized online forums

- Consider very special-topic online sites for specific group of people, e.g. people interested in gardening, wine, etc., chess players, moms, etc.

**Tasks:**
- What topics are being talked about?
- How often do they ask questions? What are they about? What **type** questions are discussed most? (type: recommendation, opinion, etc.)
- What are most mentioned feelings/reasons/words, for specific concepts like ‘opening’, `divorce’, or ‘wine storage’?

**Data:**
- Collect your own data:  
  http://www.youbemom.com/forum/all
Brain network classification

- Brain networks of 114 human subjects
  - Nodes: brain regions
  - Edges: connection strengths (weighted)
- Small graphs 70x70 nodes (regions)
- Big graphs ~2M nodes

Tasks:
- Classify human: 1- high math vs normal
  2- creative vs normal
  3- male vs female etc.
  - Using/finding discriminative patterns

Data: [http://www.cs.stonybrook.edu/~leman/courses/14CSE590/data/brainnetworks.rar](http://www.cs.stonybrook.edu/~leman/courses/14CSE590/data/brainnetworks.rar)
Link Prediction/Anomalies

- Consider a political forum where users discuss several issues:
  - e.g.: abortion, creation, gay rights, guns, healthcare, re-election of Obama
- Opinions: “in-favor” or “opposed” (signed edges)

Tasks:
- Given a user u and a new issue i
  Predict u’s opinion on i (use network)
- Anomalies? Spam? Conflicts?

Data:
http://www.politicsforum.org/forum/
Wikipedia?
Link Prediction

- Consider a **who-trusts-whom** network
  - Users decide whether to “trust” or “not-trust” each other. (signed edges)

**Task:**

- **Given** a user i and a user j
  - **Predict** whether i trust j and vice versa (use network)

**Data:** Epinions.com at http://snap.stanford.edu/data/soc-sign-epinions.html
Consider **who-follows-whom** Twitter network

**Tasks:**
- **Given** a user i and a user j
  - **Predict** whether i and j follow each other (use network)
- Find community structure
  - Measure quality of communities (conductance, modularity)
  - How dense are they? Are they well separated?
  - What size are they? Communities-within-communities?

NYC Open Data

- Available from: https://nycopendata.socrata.com/browse
- Types of data include
  - Electric consumption by zipcode
  - Emergency (911) or community-concern (311) calls by zipcode
  - Restaurant inspections
  - Noise complaints by zipcode
  - ...
- Tasks:
  - Find anomalies/fraud/events
  - Summarize the data and visualize
Study trends in clothing

- Given photos and tags of what people are wearing (temporal data)
  - Find trends (association rules: what is being worn with what)
  - How do these trends change over time, if at all?
  - What determines the #likes of a photo? (e.g., content, popularity, #friends)

- Data:
  - http://www.cs.stonybrook.edu/~leman/courses/14CSE590/data/chictopia.tar
KDD Cup and Kaggle.com

- KDD is the premier data mining conference
- Every year there is a **competition**
  - KDD-Cup 2010 - Student performance evaluation
  - KDD-Cup 2009 - Customer relationship prediction
  - KDD-Cup 2008 - Breast cancer
  - KDD-Cup 2007 - Consumer recommendations
  - KDD-Cup 2006 - Pulmonary embolisms detection from image data ...

- Similarly check out **Kaggle**:

15

Fall 2014 CSE 590 - Data Mining meets Graph Mining
Hadoop/Map-Reduce projects

- MR: Distributed compute environment
- Hadoop: open-source version of MR
  - Can install on local machine

Tasks:
- How to partition a very large graph? Goal:
  - Many within-partition edges, Few cross-edges
- How to find single-source **shortest paths**?
  - Given node $i$, find all shortest paths to other nodes
  - For **weighted, directed** graphs
  - Modification: with **upper-bound** on shortest path distance
Other resources:

- http://snap.stanford.edu/class/cs224w-2010/datasetsInfo.html
- http://www.stanford.edu/class/cs341/data.html
- http://snap.stanford.edu/data/

Don’t feel limited by these ideas/datasets. You can come up with your own ideas and collect interesting datasets 😊