Integrating Statistics and Visualization: Case Studies of Gaining Clarity During Exploratory Data Analysis

Adam Perer, Ben Shneiderman

Bo Liu
Paul Rosenfeld
CMSC734
Overview

- Vizster: Visualizing Online Social Networks by Jeffrey Heer and Danah Boyd
  - Brief overview
  - Demo!
- SocialAction
  - Summary of design goals
  - Overview of features
  - Case studies
    1. Senatorial Voting Patterns
    2. Knowledge Discovery for Medical Research
    3. Engaging Hospital Trustee Networks
    4. Group Dynamics in Terrorist Networks
  - Discussion/Conclusion
Vizster: Overview

- Tool built to visualize the Friendster social network
- Collects data about users and their friends' friends' friends' friends and builds a force-directed node-link diagram
- And why should we care? (According to the authors)
  - Social aspect
    - "...facilitate better discovery of people, connections, and communities..."
  - Technological aspect:
    - "...to promote increased awareness of community structure and information exposure..."
  - Design aspect:
    - "...use of ethnographic techniques and visualization design to craft a domain-specific visualization tool..."
Vizster: Demo
Vizster: Conclusion/Criticism

- Definitely neat to play with (looks good, cross platform, fast animations), but seems rather limited in actual usefulness
  - One of the usability experiments was to put the software on a projector and leave it running during a party and let people play with it
  - The comment "What!? She's not single!" is referred to by the authors as "ad-hoc analysis"
    - Procrastinating college students perform this ad-hoc analysis on facebook without visualization tools
  - Friendster is so 2003
  - No real statistical capabilities
SocialAction: Introduction

- Social Network Analysis (SNA) Challenges:
  - complex network
  - node attributes and relationships
  - overlapping nodes and tangled edges (hairballs)
- Previous Work/Tools:
  - either statistical analysis - difficult to find outliers and patterns
  - or visualization - hard to interpret
- SocialAction Strategy: tightly integrate both statistics and visualization to derive the benefits of both
SocialAction: Evaluation

- Traditional: lab-based controlled experiments
  - 20-60 participants
  - 10-30 min training
  - 2-20 tasks during a 1-3 hr session
  - t-test or ANOVA

- Drawbacks:
  - real work processes are impossible to reconstruct
  - exploratory tasks are poorly defined
  - inter-user variations undermine summary statistics

- Solution: 4 long-term case studies
  - interview (1 hour)
  - training (2 hours)
  - early use (2-4 weeks)
  - mature use (2-4 weeks)
  - outcome (1 hour)
Even though we see the typical "hairball", highlights statistically significant nodes in the overall view.
Use a statistical method to determine "gatekeepers" to find interesting portions of the network. Rank by feature.
Clicking on a node presents details on demand (left pane)
Case 1: Senatorial Voting Patterns

180 vote threshold. Republican senators who are most likely to vote with Democrats are evident.
Case 1: Senatorial Voting Patterns

290 vote threshold. Democrats' relationships are much more intact.
Case 2: Knowledge Discovery for Medical Research

Relevant documents cluster together. A number of isolated islands of documents.
Case 3: Engaging Hospital Trustee Networks

Scatterplot quickly finds the outlier - a key gatekeeper.
Case 4: Terrorist Networks

Analyze different relationship types without reloading data sets.

Friendship

Religious ties
Conclusion

- Four case studies have provided evidence that exploratory data analysis improves with integrated statistics and visualization.
  - load all data at once, can handle large datasets
  - optimized statistical methods for real-time interaction
  - filter network by node and edge attributes
  - filter network by statistical principles of social network analysis, e.g., betweenness and degree
  - many data analysis methods, e.g., scatterplot, community algorithm
Future Improvements

- More statistical measures
- Comprehensive map-editing for nodes (e.g., size, label)
- Save working environment for future editing
- Export results as vector graphics
Thanks!