Graphs and the VAST Challenges
New title: Visualization of Graphs – Key Issues
April 17, 2013

Georges Grinstein
Professor - University of Massachusetts Lowell
Director - Institute for Visualization and Perception Research
Director - Center for Biomolecular and Medical Informatics

Past funders on visualization, analysis, software
- NVAC
- NSF
- CDC
- USAF
- Open Indicators Consortium

and BBN, Biogen, Charles River Analytics, MITRE, ...

Our Visual Analytics Research
Integrated Visualization Systems

Enhance data perception (1980s)
Visualize millions of variables (1990s)
Evaluate insight (2000s)
Measure cognition (2010 on)

Web-based Analysis and Visualization

The Visualization Pipeline (1980-2013)

The Visualization Pipeline
User interaction/collaboration and more control (2000s)
Analysis, Computational and Synchronization Tools
Visual Analytics Pipeline

Tightly coupled analysis and visualization (2010s)
When should we do visualization?

- Almost always
- To have a human in the loop to augment, not replace, human cognition for problems that cannot be (completely) automated
- To see different views with multiple linked visualizations as each is often an aggregation or sampling as each presents a different views
- To remember, to overcome bias, to harness other memories, to reason, to …

Anscombe’s Quartet

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of X</td>
<td>9</td>
</tr>
<tr>
<td>Variance of X</td>
<td>11</td>
</tr>
<tr>
<td>Mean of Y</td>
<td>7.5</td>
</tr>
<tr>
<td>Variance of Y</td>
<td>4.1</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.816</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>( y = 3.0 + 0.5 \times x )</td>
</tr>
</tbody>
</table>

And in fact

Given any \( m \) statistics for real numbers
There exist \( \infty \)-ly many data sets each having these statistics

Which state has the highest income? Most college degrees?

<table>
<thead>
<tr>
<th>State</th>
<th>Per Capita Income</th>
<th>College Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>57,480</td>
<td>62%</td>
</tr>
<tr>
<td>New York</td>
<td>60,550</td>
<td>50%</td>
</tr>
<tr>
<td>Texas</td>
<td>47,350</td>
<td>29%</td>
</tr>
<tr>
<td>Florida</td>
<td>48,100</td>
<td>35%</td>
</tr>
<tr>
<td>New Jersey</td>
<td>55,000</td>
<td>43%</td>
</tr>
<tr>
<td>Georgia</td>
<td>51,200</td>
<td>40%</td>
</tr>
</tbody>
</table>

Courtesy - Chris North lecture
However there are problems:

- data
- algorithms
- users

Large x and y scale

Large x scale

Large y scale

Data determined scale
Identical patient data presented 4 different ways
Will a clinical trail be stopped if significance

The decision to stop varied significantly depending on the presentation of the data

Correct decisions were
- 56% with bar charts
- 56% pie charts
- 68% with tables
- 82% with icon displays

In actual clinical practice, up to 25% of the patients treated based on the bar or pie charts would have received inappropriate treatment

From Elting et al., BMJ 1999, 318:1527-31

Four Types of Visualizations

- **Exploratory**
  - Have no hypotheses about the data
  - Explore data interactively as undirected searches

- **Confirmatory**
  - Have specific hypotheses about the data
  - Goal-oriented examination of the hypotheses

- **Presentation**
  - Facts to be presented are fixed a priori
  - Select appropriate presentation techniques

- **Interactive Presentation**
  - Interactions with a pre-defined animation

Graphs, Networks, … examples
15 seconds … flythrough

Harnessing your associative pre-attentive memory (in other words you will remember them)
Network Clustering

How to do Visualization
Key Examples of Good Approaches

Visual analytics is not a static map, not information retrieval, not data mining

Visual analytics IS intelligent interactive integrated analysis and coordinated visualization

Overlays: Graphs + Heatmaps
Possible integrated visualization environment of the future

A visual canonical adjacency matrix
- An algorithm producing a canonical visual matrix
  only depends on the graph's topological information
  no node or label information required
- Thus two structurally identical graphs have exactly the same visual matrix representation

Hongli Li – thesis and Pacific Vis (2009)

Zoom in and Subgraph View

Scalability: Millions of Variables
RadViz

Chemical Compound Reactions
Associations Between 16,000 Variables

Class Discrimination Layout Algorithms
Newt Gene Expression Data

Salamander Expression Levels

Institute for Visualization and Perception Research

Alternative Visualizations
IEEE VAST Challenges (8 years: 2006-2013)

- 2006 – Mad Cow Disease
- 2007 – Exotic Animal Smuggling
- 2008 – Boat People, Bombs, Phones
- 2009 – Insider Threat, Video
- 2010 – Mutating virus pandemic
- 2011 – Possible Pandemic
- 2012 – Cyberthreat Monitoring
- 2013 – Box Office RT and Cyberthreat

VAST Challenge Overview

- Part of the VisWeek Conference
- A series of domain-specific mini-challenge scenarios
- Integrated also into a grand challenge
- A mixture of synthetic and real-life datasets with ground truth
- Evaluate solutions using quantitative and qualitative techniques
- 2006-2013 data sets/questions/solutions/videos/docs
  Available on VAST web site
  For training, class projects, building better test sets, …
- NSF: Scientific Evaluation Methods for Visual Analytics Science and Technology (SEMVAST)
Examples of Previous Challenges

- **VAST 2009**: cyber-security (building access and file upload), social network analysis (macro blog hashtag analysis).
- **VAST 2008 Challenge**: spatial-temporal analysis (migration patterns), social network analysis (telephone network), path analysis (evacuation).
- **VAST 2006 and 2007**: analysis of news reports (~ criminal investigation).

Synthetic Datasets

- **Characteristics**
  - Realistic enough for detailed analysis
  - Cleaner than real data
- **Ground truth**
  - Embedded or injected
- **Data Generation**
  - Statistical, rule-based, or re-sampling algorithms
  - Or transformed real data + synthetic
  - Iterative validation and refinement

Data Set

- **Data set consisted of about 1200 news stories from the Alderwood Daily News**
- **Some images, maps, voter registry, phone call log**
- **Background information**

FOR MORE INFO:

http://www.cs.umd.edu/hcil/VASTcontest06/

Sample text

Alderwood to probe voting machines - Story by Ellie Olmsen, Data Published to Web 11/16/2004

Republicans in Alderwood joined Democrats yesterday in criticizing the performance of the city's costly new high-tech voting system, saying that it may have disenfranchised voters in the Nov. 4 election.

The Republican commission scolded the city board of elections for minimizing problems with the touch-screen machines that the city purchased this year for $1.5 million and asked Mayor Rex Luthor to investigate what went wrong before the machines are pressed into service again.

Alderwood's touch-screen voting machines, which resemble laptop computers without keyboards, were supposed to simplify voting and tabulating results. But in a debut that mirrored many of the problems experienced last year in areas across the country, some voters found the machines confusing, and the reporting of vote tallies was delayed almost a day.

Luthor responded that he would try to address the board's concerns. He said he has called for a public meeting of the three-member board of elections to go over the requests at 5 p.m. today.

"I pledge that I will answer every question as soon as I possibly can in the proper fashion," he said.

Sample image
### City Hall Phone Logs

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Phone Number</th>
<th>Caller ID</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/22/13</td>
<td>12:00</td>
<td>555-1234</td>
<td>555-5678</td>
<td>Town Hall</td>
</tr>
<tr>
<td>04/22/13</td>
<td>13:00</td>
<td>555-9876</td>
<td>555-4321</td>
<td>Switzerland</td>
</tr>
<tr>
<td>04/22/13</td>
<td>14:00</td>
<td>555-5432</td>
<td>555-3210</td>
<td>New Lab</td>
</tr>
<tr>
<td>04/22/13</td>
<td>15:00</td>
<td>555-6789</td>
<td>555-1234</td>
<td>Politician</td>
</tr>
<tr>
<td>04/22/13</td>
<td>16:00</td>
<td>555-3210</td>
<td>555-4321</td>
<td>Voting Machines</td>
</tr>
</tbody>
</table>

### Voter Registry

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Phone Number</th>
<th>Caller ID</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/22/13</td>
<td>12:00</td>
<td>555-1234</td>
<td>555-5678</td>
<td>Town Hall</td>
</tr>
<tr>
<td>04/22/13</td>
<td>13:00</td>
<td>555-9876</td>
<td>555-4321</td>
<td>Switzerland</td>
</tr>
<tr>
<td>04/22/13</td>
<td>14:00</td>
<td>555-5432</td>
<td>555-3210</td>
<td>New Lab</td>
</tr>
<tr>
<td>04/22/13</td>
<td>15:00</td>
<td>555-6789</td>
<td>555-1234</td>
<td>Politician</td>
</tr>
<tr>
<td>04/22/13</td>
<td>16:00</td>
<td>555-3210</td>
<td>555-4321</td>
<td>Voting Machines</td>
</tr>
</tbody>
</table>

### Tasks

- Determine the inappropriate activities taking place
- Report hypotheses and conclusions including people, places, events
- Identify the associated relevant documents
- Answer form (answers and process)
- Video
- 2 page summary (see proceedings or archive)

### A taste of the activities

- Suspicious weekend calls from town hall to Switzerland
- Mad cow disease outbreak
- New lab created, headed by strange character
- Politician hanky-panky
- Voting machines irregularities
- And many more…

### Three winning entries
VAST 2008 Challenge
Challenge themes included geospatial, activity and behavior, text processing, and social network analyses
1. wiki edits
2. cell phone call logs
3. Coast Guard interdiction records
4. RFID tags in an evacuation of a DOH site after an explosion

Almost all teams visualized the social network with a node/link diagram

Other Representations
Nyenrode Business Universiteit
University College Dublin

Almost all teams visualized the social network with a node/link diagram

Visualization/Analysis of the geospatial components of the data
CORE
Prajna Project
Palantir

Visualization/Analysis of the geospatial components of the data
ICAVE
AttributeLayoutViz

Coordinated Displays
VisWhiz
SocialDynamicsVis
ComVis
NEVAC
Coordinated Displays

University of Illinois
National Center for Supercomputing

Representing Time

• 3D
• Hybrid Projection

Sorting it all out
Staining by Southern Illinois

Who Escaped?

Looking at the Exits

UCD Protovis Timeline

Fraunhofer Institute
VAST 2009 Challenges
Mock scenario & synthetic datasets involving an embassy
1. badge and network traffic data
2. social network including geospatial information
3. security video

VAST 2010 Challenges
1. Text records and an investigation into arms dealing (PNNL)
2. Hospital records & characterization of a viral pandemic spread (PNNL)
3. Genetic sequences and tracing the mutations of a disease (UML)
Grand Challenge - Combination of all data sets and tasks

Intuitive traffic Visualization and Video Description of the Analysis Process
Palantir Technologies

Intuitive Visual Representation of Alibis
Best one screenshot of the solution
- HRL Laboratories

Representation of Uncertainty in Rules & Visualization
- LaBRI, INRIA Bordeaux (student team)

Dataset Generation
- Virus
  - Rapidly mutating RNA virus
  - Single viral gene (surface protein)
  - Simple genetic distance used
  - Random point mutations (no insertions, deletions)
Results – Question 1 and 2

Visual Analytics Benchmark Repository

Results – Question 3 and 4

Visual Analytics Benchmark Repository

Datasets

- **Microblogs**
  Messages collected from various devices with GPS capabilities

- **Map Information**
  Maps for the entire metropolitan area and a satellite image with labeled highways, hospitals, important landmarks, and water bodies

- **Supplemental tables**
  Population statistics, observed weather data and additional information in a README file

VAST Challenges 2011

1. Epidemic Spread: microblog messages and map information
2. Computer Network Operations: several network security datasets
3. Criminal Activity: news reports

Grand Challenge: Analyze all 3 - Causes and Effects

MC 1 Award: Outstanding Integration of Computational and Visual Methods

- MTA-SZTAKI (the Computer and Automation Research Institute of the Hungarian Academy of Sciences)

MC 1 Award: Unique Integration of Tag Clouds in Geospatial Visualizations

- University of Stuttgart

Institute for Visualization and Perception Research

information visualized
knowledge discovered
decisions made

MC 1 Award: Outstanding Integration of Computational and Visual Methods

© 2013 VAST

Harshavardhan Achrekar - thesis and papers
MC 1 Award: Outstanding Analysis Using Custom Tools
- Bangor University

MC 1 Award: Novel Extension of Visual Analytics to Mobile Devices
- Purdue University

Datasets
- The computer network architecture including a list of priority computers which are essential to All Freight’s ability to conduct business
- Security policy rules
- A firewall log
- An IDS log
- An aggregated file of syslogs for all the hosts on the network
- A Nessus Network Vulnerability Scan Report

MC 2 Award: Outstanding Integrated Overview Display
- University of Buenos Aires

MC 2 Award: Innovative Tool Adaptation
- Penn State

MC 2 Award: High Potential for Scalability
- UNC-Charlotte
Questions

- **MC 3.1 Potential Threats**
- Identify any imminent terrorist threats in the Vastopolis metropolitan area. Provide detailed information on the threat or threats (e.g. who, what, where, when, and how) so that officials can conduct counterintelligence activities with the evidential documents supporting your answer.

Datasets

- A text corpus containing news reports with each news report a plaintext file containing a headline, the date of publication, and the content of the article.
Mini-Challenge 1

- The imaginary BankWorld's largest financial institution, the (fictitious) Bank of Money needs your best situation awareness visualizations to understand the health of its global corporate network.

- But how do you visualize data out of a network containing nearly a million computers in a way that you can perceive its state and identify problems?

Mini-Challenge 2

- Unusual events are occurring in one of the Bank of Money's regional offices. Some of them may very well wreak havoc across the institution if they turn out to be malicious.

- What are these unusual events?

- And if you were in charge of computer security, what actions should be taken to safeguard the network and quite possibly save the Bank of Money from disaster?

VAST Challenge 2013

1. For a selected set of movies:
   - Predict their opening weekend box office take
   - Predict the critics' rating

   Early start: Kicked off before the Thanksgiving/Christmas releases

   Submit throughout the year – predictions due before opening weekends

2. Cyber very very large

   Prospective Analyses:
   - Cinema
   - Sound Interesting?

   Email BoxOfficeVAST@yahoo.com and we will make sure to add you to our contact list!