Graph-based ML + VIS

Future directions
Reminder of final project presentation time:
- April 24 (Tuesday) 9:10 - 10:30 a.m
- April 27 (Friday) 8:00 to 10:00 a.m.

Final project report due time:
- April 30 (Monday) 9:10 a.m.

Survey
- What materials would you like to see more/less of?
Graph-Based ML
Plus Visualization
Visualizing Execution of Algorithm
Bubble Sort

Checking if 38 > 44 and swap them if that is true. The current value of swapped = true.

do
    swapped = false
    for i = 1 to indexOfLastUnsortedElement-1
        if leftElement > rightElement
            swap(leftElement, rightElement)
            swapped = true
        while swapped

https://visualgo.net/en/sorting
DFS(0) is completed. Red/grey/blue edge is tree/cross/forward/back edge of the DFS spanning tree, respectively.

DFS (u)
for each neighbor v of u
  if v is unvisited, tree edge, DFS (v)
  else if v is explored, bidirectional / back edge
  else if is visited, forward / cross edge

// ch4_01_dfs.cpp / java, ch4, CP3
Visualizing LP?

Visualization of random walks on graphs?

ML approach to large graph visualization
Machine learning approach to large graph visualization

Cluster graphs by their topological similarities.
Fig. 2. All connected graphlets of 3, 4, or 5 vertices.
Fig. 3. Examples of graphlet frequencies. The x-axis represents connected graphlets of size $k \in \{3, 4, 5\}$ and the y-axis represents the weighted frequency of each graphlet. Four graphs are drawn with sfdp layouts [45]. If two graphs have similar graphlet frequencies, i.e., high topological similarity, they tend to have similar layout results (a and c). If not, the layout results look different (a and b). However, in rare instances, two graphs can have similar graphlet frequencies (b and d), but vary in graph size, which might lead to different looking layouts.
Topological similarities: kernels

**Cosine similarity (COS):** Most existing graphlet kernels use the dot product of two graphlet frequency vectors in Euclidean space, then normalize the kernel matrix. This is equivalent to the cosine similarity of two vectors, which is the $L_2$-normalized dot product of two vectors:

$$\langle x, x' \rangle = \frac{x \cdot x'^T}{\|x\| \|x'\|}$$

**Gaussian radial basis function kernel (RBF):** This kernel is popularly used in various kernelized machine learning techniques:

$$\langle x, x' \rangle = \exp \left( -\frac{\|x - x'\|^2}{2\sigma^2} \right)$$

where $\sigma$ is a free parameter.

**Laplacian kernel (LAPLACIAN):** Laplacian kernel is a variant of RBF kernel:

$$\langle x, x' \rangle = \exp \left( -\frac{\|x - x'\|_1}{\sigma} \right)$$

where $\|x - x'\|_1$ is the $L_1$ distance, or Manhattan distance, of the two vectors.

Treating graphlets frequencies as feature vectors. Define dot product of two graphlet frequency vectors.
Aesthetic criteria for graph VIS

- Crosslessness: Minimizing the number of edge crossings
- Minimum angle metric: maximizing the minimum angle between incident edges on a vertex.
- Edge length variation: Uniform edge lengths.
- Shape-based metric: Mean Jaccard similarity between the input graph and the shape graph.

\[
MJS(G_1, G_2) = \frac{1}{|V|} \sum_{v \in V} \frac{|N_1(v) \cap N_2(v)|}{|N_1(v) \cup N_2(v)|}
\]

https://arxiv.org/abs/1710.04328
What would a graph look like in this layout?

- Compare accuracy of ML predicted similar graph layout vs human chosen layout

https://arxiv.org/abs/1710.04328
Apply ML to Graph Drawing
Existing approaches

- Approaches that learn from human interaction: learn information from users based on their interactions to a graph drawing system.

- Approaches that are not based on human interaction: gather and evolve knowledge about how to draw a graph from the results of other automatic graph drawing algorithms or from the graph structure itself.
Rate by users

(1) distance between nodes  (2) arc direction  (3) line crossing  (4) symmetry  (5) angle between arcs  (6) uniformity

Bad Layout

Better Layout

Figure 1: Constraints used in the layout of directed graphs.

https://www.cse.unr.edu/~sushil/class/gas/papers/masui94evolutionary.pdf
Learning optimal graph drawing for clustered graph

- Construct a handcrafted feature vector of a cluster from a number of graph measures: number of vertices, diameter, and maximum vertex degree.

- Find an optimal layout for each cluster.
Using ML to improve layout quality

- Human in the loop evaluation.
- Using neural network algorithms to optimize a layout for certain aesthetic criteria.
Self-organizing graphs

B. Meyer: Self-organizing graphs - a neural network perspective of graph layout

Not building a general neural network that learns aesthetic criteria or hints about how to draw a graph. Instead, it models the graph structure and the drawing problem as a network coupled with an energy system.
Future directions?

- Using neural network algorithms to optimize a layout for certain criteria
- Really learn how to draw a good graph
The Future of Data Visualization
Current & Next Cool Startups
CUSTOMER STORY
Wells Fargo wrangles data from over 70 million customers to redesign customer banking portal

VISUALIZATION
Test the myth of tech companies' 'rocket-ship' growth

PRODUCT VIDEO
Create rich analyses and share your insights with colleagues in seconds

https://www.tableau.com/
Startup Raised $180M To Take On Tableau Software In $18B Analytics Market

Opinions expressed by Forbes Contributors are their own.

Business intelligence and data analytics are popular buzzwords. But businesses face considerable obstacles to turning the buzz into better decisions. Now Looker, a Santa Cruz, Calif.-based supplier of analytics tools, is taking on publicly-traded Seattle-based Tableau Software in the $18 billion market for data analytics that includes business intelligence (BI) and data visualization, according to Gartner. Is Looker a threat to Tableau or is there room enough for many players in this industry?
Looker Overview

https://www.youtube.com/watch?v=krXaBEi3f1s
https://looker.com/company
Data Wrangler

- Accelerate data manipulation: spend less time fighting with your data and more time learning from it.

- Allow interactive transformation of messy, real-world data into the data tables analysis tools expect.

http://vis.stanford.edu/wrangler/
Our focus is to create radical productivity for people who work with data.

Wrangling data is the most time-consuming and inefficient part of any data project – taking up over 80% of the time and resources. Trifacta enables anyone to more efficiently explore and prepare the diverse data of today by utilizing machine learning to provide a breakthrough user experience, workflow and architecture.

JOIN TOGETHER DISPARATE DATA SOURCES

ONBOARD EXTERNAL OR 3RD-PARTY INFORMATION

CLEAN RAW AND MESSY DATA
"The Future of Data Visualization" - Jeffrey Heer (Strata + Hadoop 2015)

https://www.youtube.com/watch?v=vc1bq0qIKoA
- Intuitive UI
- Real-time visualization
- Line, area, bar, scatter plot to gauges and maps
- Share visualizations…
Build products with game-changing interactive visualization that turns data into insight.

Connected data is all around us. It’s in financial transactions, communications records, IT networks and beyond.

The best way to understand it is to visualize it.

Applications built with KeyLines offer new ways to join the dots in your data and uncover valuable buried insights.

https://cambridge-intelligence.com/keylines/
Opioid Overdose Death Rate (Age-Adjusted)

Top 10 Locations
1. West Virginia 43.4
2. New Hampshire 35.8
3. Ohio 32.9
4. District of Columbia 30.0
5. Maryland 29.7
6. Massachusetts 29.7
7. Rhode Island 26.7
8. Maine 25.2
9. Connecticut 24.5
10. Kentucky 23.6

Bottom 10 Locations
42. Iowa 6.2

https://datausa.io/map/?level=state&key=opioid_overdose_deathrate_ageadjusted

Mining and visualizing US government open data
Drag-and-Drop Chart Creation

Get suggested charts based on your specific data within DataHero or simply create your own customized charts from scratch. Segmenting, filtering, cohort analysis and more are at your fingertips. DataHero makes it easy to create beautiful data visualizations and data dashboards that you’ll want to share with your team and clients.
What are hot topics in advanced data visualization?
What are the next startups in data visualization?

1. Design smart

Smart visualization design
Effective visual encoding
2. Better user interface

Recommend visualization views
New end user exploration users
  No distractions
User-in-the-loop
3. Show data variation, not design variation

By Edward Tufty
4. Scalability

HPC, databases
5. Visualization that support mobile devices
FAST VISUAL ANALYTICS

Fast Visual Analytics & Data Visualization

https://www.zoomdata.com/
6. Effective communication

Not just creative visual
Need substances
Edward Tufte: “Every single pixel should testify directly to content.”

7. Integration with other disciplines
<table>
<thead>
<tr>
<th>The Internet of Things</th>
<th>Network Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tens of billions of devices will be connected to the Internet in the next decade.</td>
<td>Network Theory builds on Graph Theory, which</td>
</tr>
<tr>
<td>From smart appliances and wearables to automobile sensors and environmental</td>
<td>applies algorithms to understand and model pairwise relationships between</td>
</tr>
<tr>
<td>monitors, the Internet of Things will provide unprecedented insight into what's</td>
<td>objects. Network Theory examines relationship symmetry, with the existence</td>
</tr>
<tr>
<td>happening around us. High-throughput, interconnected data streams will help us</td>
<td>of asymmetric relationships providing grounds to predict the likely spread of</td>
</tr>
<tr>
<td>improve safety, drive operational efficiencies and better understand consumer</td>
<td>information (social network analysis), dissect complex disorders (biological</td>
</tr>
<tr>
<td>demand.</td>
<td>network analysis), find the shortest path between two points (network</td>
</tr>
<tr>
<td>In the words of Kevin Ashton, who first coined the term “the Internet of Things” in</td>
<td>optimization) and identify target objects based on their behavior (link</td>
</tr>
<tr>
<td>his seminal 2009 RFID Journal article, “The Internet of Things has the potential</td>
<td>analysis).</td>
</tr>
<tr>
<td>to change the world, just as the Internet did. Maybe even more so.”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complexity Theory</th>
<th>Multidimensional Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity Theory posits that many systems are characterized by complex, non-linear</td>
<td>The adage “a picture is worth a thousand words” gained credence from our</td>
</tr>
<tr>
<td>interactions that evolve dynamically and often unpredictably. Known as the “butterfly</td>
<td>ability to process visuals more easily than text. Visualization has also been</td>
</tr>
<tr>
<td>effect,” small perturbations in one state (“here”) can result in large repercussions</td>
<td>shown to improve learning and recall, and can portray complex concepts and</td>
</tr>
<tr>
<td>in a seemingly unrelated state (“there”). According to Complexity Theory, it’s</td>
<td>relationships more easily than can text. Recent developments in computer</td>
</tr>
<tr>
<td>impossible to predict with certainty a future state, but it is possible to understand</td>
<td>graphics are making possible visualizations that enable the integration,</td>
</tr>
<tr>
<td>the structure and potential states of complex systems.</td>
<td>manipulation and exploration of dynamic multidimensional data sets. Multidimensional</td>
</tr>
<tr>
<td></td>
<td>visualizations allow users to not only examine data from new perspectives but</td>
</tr>
<tr>
<td></td>
<td>also interact with it more effectively.</td>
</tr>
</tbody>
</table>
8. Accurate and contextual vis

How data are connected?

9. Facilitate decision making

10. BI and medical domain
Thanks!

Any questions?

You can find me at: beiwang@sci.utah.edu
CREDITS

Special thanks to all people who made and share these awesome resources for free:

- Presentation template designed by Slidesmash
- Photographs by unsplash.com and pexels.com
- Vector Icons by Matthew Skiles
Presentation Design

This presentation uses the following typographies and colors:

Free Fonts used:
https://www.fontsquirrel.com/fonts/open-sans

Colors used