Learning to rank and compare graph layouts

Toby Dylan Hocking
toby@sg.cs.titech.ac.jp
joint work with Supaporn Spanurattana and Masashi Sugiyama

6 Aug 2013
Introduction: what makes a graph layout good or bad?

Learning to rank and compare graph layouts
Biology is full of networks (graphs)

Source: Kyoto encyclopedia of genes and genomes (KEGG).
Biology is full of networks (graphs)

Goal: find a good layout for a particular graph

Two categories of methods for graph layout

▶ Heuristic layout algorithms:
  ▶ Force-directed
  ▶ Hierarchical clustering (trees/dendrograms)
  ▶ Hive plots
  ▶ ...

▶ Manual layout using programs such as:
  ▶ Cytoscape/cytoscape.js
  ▶ Gephi
  ▶ Image processing: gimp/inkscape
  ▶ ...

Force-directed layout has many tuning parameters

Source: Data-Driven Documents (D3) JavaScript visualization library (Bostock 2011).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Min</th>
<th>Default</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>?</td>
<td>1 x 1</td>
<td>?</td>
</tr>
<tr>
<td>link distance</td>
<td>0</td>
<td>20</td>
<td>∞</td>
</tr>
<tr>
<td>link strength</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>friction</td>
<td>0</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>charge</td>
<td>−∞</td>
<td>-30</td>
<td>∞</td>
</tr>
<tr>
<td>theta</td>
<td>0</td>
<td>0.8</td>
<td>∞</td>
</tr>
<tr>
<td>gravity</td>
<td>0</td>
<td>0.1</td>
<td>∞</td>
</tr>
</tbody>
</table>

Question: how to tune these parameters for a specific graph?
Manual layout using a GUI is time-consuming

- Try default parameters of several different algorithms.
- Play with tuning parameters, select a combination that looks good.
- Finally, refine the algorithm’s layout by dragging nodes to positions that look better.

Goal: learn from a database of manually labeled graphs.
Manual layout using a GUI is time-consuming

- Try default parameters of several different algorithms.
- Play with tuning parameters, select a combination that looks good.
- Finally, refine the algorithm’s layout by dragging nodes to positions that look better.

Goal: learn from a database of manually labeled graphs.
Pairwise comparison in the graph layout literature

Figure 7: US airlines graph (235 nodes, 2101 edges) (a) not bundled and bundled using (b) FDEB with inverse-linear model, (c) GBE, and (d) FDEB with inverse-quadratic model.

Pairwise comparison in the graph layout literature

Fig. 8. Space filling curve layouts versus existing layouts. Existing methods (a-c) range from very slow to very fast, and produce layouts of various qualities, with the faster ones generally producing less aesthetic or less useful layouts, and devoting smaller regions of the screen to the majority of nodes, which obscures details such as the number of subclusters and their contents. The treemap layout (g) is both fast and effective at showing structures such as the three distinct clusters in the graph, but introduces problems such as poor aspect ratios. The space filling curve-based layouts (h) solve this by guaranteeing good aspect ratios, while also clearly showing the distinct clusters and maintaining speed.

Pairwise comparison in the graph layout literature

Introduction: what makes a graph layout good or bad?

Learning to rank and compare graph layouts
Learning a comparison function

We are given $n$ training pairs $(G_i, x_i, x'_i, y_i)$ where we have

- a graph $G_i$,
- two layouts $x_i, x'_i \in \mathbb{R}^p$ of that graph (feature vectors),
- a comparison $y_i =\begin{cases} -1 & \text{if } x_i \text{ is better} \\ 0 & \text{if } x_i \text{ is as good as } x'_i \\ 1 & \text{if } x'_i \text{ is better.} \end{cases}$

Goal: find a comparison function $g : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \{-1, 0, 1\}$

- Symmetry: $g(x, x') = -g(x', x)$.
- Good prediction with respect to the zero-one loss $E$:

$$\min_g \sum_{i \in \text{test}} E \left[ y_i, g(x_i, x'_i) \right]$$
Learning to rank and compare

We will learn a

- Ranking function $f : \mathbb{R}^p \to \mathbb{R}$. Bigger means a better layout.
- Threshold $t \in \mathbb{R}^+$. A small difference $|f(x') - f(x)| \leq t$ is not significant.
- Comparison function $g_t(x, x') = \begin{cases} 
-1 & \text{if } f(x') - f(x) < -t \\
0 & \text{if } |f(x') - f(x)| \leq t \\
1 & \text{if } f(x') - f(x) > t. 
\end{cases}$

The problem becomes

$$\minimize_{f,t} \sum_{i=1}^{n} E[y_i, g_t(x_i, x'_i)]$$
Some labeled layouts of a 2-node graph

<table>
<thead>
<tr>
<th></th>
<th>good 1</th>
<th></th>
<th>good 2</th>
<th></th>
<th>good 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image" alt="Diagram of good 1 layout" /></td>
<td></td>
<td><img src="image" alt="Diagram of good 2 layout" /></td>
<td></td>
<td><img src="image" alt="Diagram of good 3 layout" /></td>
<td></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Diagram of bad 1 layout" /></td>
<td></td>
<td><img src="image" alt="Diagram of bad 2 layout" /></td>
<td></td>
<td><img src="image" alt="Diagram of bad 3 layout" /></td>
<td></td>
</tr>
</tbody>
</table>

The diagrams show different layouts of a 2-node graph, with 'good' layouts minimizing edge crossings and 'bad' layouts having more crossings.
Map 20 layouts $x_i \in \mathbb{R}^2$ to a feature space
Generate 10 pairwise constraints $x'_{i} - x_{i} \in \mathbb{R}^{2}$
10 labeled difference vectors $x'_i - x_i \in \mathbb{R}^2$
All 190 labeled difference vectors $x'_i - x_i \in \mathbb{R}^2$
Max margin comparison function

\[ f(x') - f(x) = -1 \]

\[ f(x') - f(x) = 1 \]
Invariance of \( \hat{g} \) when switching train direction \( x_i, x_i' \)

\[
f(x') - f(x) = 1
\]

\[
f(x') - f(x) = -1
\]

\begin{align*}
\text{comparison} & \quad y_i
\end{align*}

\begin{itemize}
  \item [\text{line}] \\
    \begin{itemize}
      \item [-] margin
      \item [-] decision
    \end{itemize}
  \\
  \begin{itemize}
    \item [-1] blue
    \item [0] red
  \end{itemize}

\begin{itemize}
  \item [\text{constraint}] \\
    \begin{itemize}
      \item [\text{active}] black
      \item [\text{inactive}] white
    \end{itemize}
\end{itemize}
Defining the margin

Recall: for all pairs $i \in \{1, \ldots, n\}$ we have

- features $x_i, x'_i \in \mathbb{R}^p$ and
- comparisons $y_i \in \{-1, 0, 1\}$.

We define

- Ranking function $f(x) = w^\top x \in \mathbb{R}$.
- Threshold $t = 1$.
- Comparison function $g_1(x, x') \in \{-1, 0, 1\}$.

![Graph showing predicted rank difference $f(x'_i) - f(x_i)$ and margin $\mu$.]
Max margin comparison is a linear program (LP)

For $y \in \{-1, 0, 1\}$, let $I_y = \{i \mid y_i = y\}$ be the corresponding training indices.

$$\max_{\mu \in \mathbb{R}, w \in \mathbb{R}^p} \mu$$

subject to

$\mu \leq 1 - |w^T(x'_i - x_i)|$, $\forall i \in I_0$

$\mu \leq -1 + w^T(x'_i - x_i)y_i$, $\forall i \in I_1 \cup I_{-1}$.

Note: if the optimal $\mu > 0$ then the data are separable.
Related work: reject, rank, and rate

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Inputs</th>
<th>single items $x$</th>
<th>pairs of items $x, x'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y \in {-1, 1} )</td>
<td>SVM</td>
<td>SVMrank</td>
<td></td>
</tr>
<tr>
<td>( y \in {-1, 0, 1} )</td>
<td>Reject option</td>
<td>this work</td>
<td></td>
</tr>
</tbody>
</table>

- T Joachims. Optimizing search engines using clickthrough data. KDD 2002. (SVMrank)
- K Zhou et al. Learning to rank with ties. SIGIR 2008. (boosting, ties are more effective with more output values)
- R Herbrich et al. TrueSkill: a Bayesian skill rating system. NIPS 2006. (generalization of Elo for chess)
SVMrank is a quadratic program (QP)

\[
\begin{align*}
\text{minimize} & \quad w^T w \\
\text{subject to} & \quad w^T (x'_i - x_i) y_i \geq 1, \quad \forall i \in I_1 \cup I_{-1}.
\end{align*}
\]
Conclusions and future work

Learned a function $f(x)$ for ranking a graph layout $x$.

- **Features** for good performance on real graphs?
- **Tune** layout algorithm parameters to maximize $f$.
- **SVMrank** is sufficient under what assumption?
Thank you!

Supplementary slides appear after this one.
Layout evaluation metrics (features $x_i, x'_i$)

- Number of crossing edges (smaller is better)
- Aspect ratio (closer to 1:1 is better?)
- Symmetry (more is better when the graph has symmetries)
- Edge length (small and less variable is better?)
- Angle between edge pairs (big is better?)
- Area of smallest bounding box (smaller is better to let small features be more legible)

Source: http://en.wikipedia.org/wiki/Graph_drawing#Quality_measures