Transparent and Fair Machine Learning on Graphs for Humans

Sihong Xie, Assistant Professor
Computer Science and Engineering
Lehigh University
Machine learning on graphs

- Graph
  - Nodes: variables
  - Edges: relationship between variables

- Applications
  - Human brain networks
  - Chemical compounds: drug discovery
  - Social networks
  - Fraudster networks

- Graphical models: ML on graphs
  - Node clustering
  - Nodes and edges property prediction
  - Graph classification or clustering.
Interpretable ML: just a CS question?

• Graphical models are not easy to be explained
  o Message passing and multiplexing.
  o Multiple steps of transformation.
  o Topology matters: tree vs. cycles.

• The human factors
  • Limited memory capacity
  • Background knowledge
  • Fast and slow thinking.

Source: https://news.dartmouth.edu/news/2015/03/pi-day-party-day-mathematical-mavens

Source: Daniel Kahneman
Establishing human trust in intelligent agents is non-trivial [1]. Explanations can help.

But what kind of explanations are more likely to help establish human trust?

Hypotheses

- Simulatability helps: \(1+1=2\) but not \(1.1+101.9=103\)
- Counterfactual helps: \(\text{rain} \Rightarrow \text{wet\_ground} \) and \(!\text{rain} \Rightarrow !\text{wet\_ground}\)
- There are interactions between the two factors.

Interpretable ML: a human subject study

• Settings of the study
  • GNN on a citation network (CORA) to predict a paper’s area.
  • Extract explaining subgraphs, with different simulatabilities.
  • Extract two subgraphs with different counterfactual relevance.

\[ G \] (first graph: original graph)

\[ G_i \] (second graph: the explanation)

\[ \tilde{G}_i \] (third graph: counterfactual 1)

\[ \tilde{G}_i \] (forth graph: counterfactual 2)

• perceived simulatability
• perceived counterfactual relevance
• acceptance
Interpretable ML: a human subject study

- Measuring simulatability, counterfactual relevance, and their interactions:
  - Collected 400 responses.

Statistical significance tests conducted to consider the size of samples.
Multiple objective optimization:

\[
\max_{G_i, \tilde{G}_i} F(G_i, \tilde{G}_i) = (\nu(G_i), |\mu(G_i, \tilde{G}_i)|)
\]

s.t. \( v_i \in \tilde{G}_i \subset G_i \subset G, \ |G_i| \leq C, \ G_i \text{ acyclic} \)

Large discrete search space and non-differentiable objective functions.

Need to find the Pareto front for balanced and efficient trade-offs.

Algorithm:

1) BFS search.

2) explanation evaluation.

3) ranking-based explanations with provable balance and efficiency.
Experimental results

• Average performance: trade-off between the two objectives?

• Running time
Experimental results

- A pitfall in finding well-balanced Pareto optimal explanations
- the ideal case
- in more cases, the Pareto front is not convex

For more details, see Yifei Liu, Chao Chen, Yazheng Liu, Xi Zhang, and Sihong Xie. *Multi-objective Explanations of GNN Predictions.* ICDM 2021.
Machine learning on graphs

- Accuracy
- Interpretability
- Robustness
- Fairness

Robustness and sensitivity of explanations.
Interpretable contrastive ML

- Contrasting two graphs using a Siamese network:
  - Graph comparisons: human brains (healthy vs. ADHD) [1]
    chemical molecules (soluble vs. non-soluble).
  - Contrastive learning: representation learning with scarce labeled data.

Siamese networks

спор с вероятностью [Two brains are similar]

спор с вероятностью [Two molecules are similar]

Explaining the learned contrastive model

- For the explanations to be trusted, we want
  - Robustness / stability
    Explanations should remain the same with respect to irrelevant changes.
  - Sensitivity
    Explanations should be different when the compared object differs.

- Challenges:
  - The gradient-based explanations are not robust [1]
  - the boundary between robustness/stability and sensitivity is hard to know beforehand.

Explainable contrastive model: self-explanation

• Learn stable self-explanation for each graph
  o No labeled data is necessary.

• **Stage 1**: learn self-explanations
  
  i) Mask out insignificant parts while preserving self-similarity.

  \[
  \min_M \ell(f(x, x), f(x, M \otimes x)) + \gamma \|a(M)\|,
  \]
  
  s.t. \( g_i(M) \leq 0, i = 1, \ldots, c. \)

  ii) Minimize the retained portions to avoid trivial solution

  iii) Additional domain constraints
Constrained optimization

- **Stage 2**: adapt a self-explanation when compared with different objects.

\[
\begin{align*}
    x^s & \quad x^t \\
    \text{SNX:} & \quad \min_{m^s, m^t} \ell \left( f(x^s, x^t), f(m^s \otimes x^s, m^t \otimes x^t) \right) \\
    & \quad + \gamma \left( \|a(m^s)\| + \|a(m^t)\| \right) \\
    \text{s.t.} & \quad g_i(m) = a(m^s)_i - a(M^s)_i \leq 0, \quad i = 1, \ldots, c_s, \\
    & \quad g_{c_s+i}(m) = a(m^t)_i - a(M^t)_i \leq 0, \quad i = 1, \ldots, c_t.
\end{align*}
\]

\( m^s \otimes x^s \) \quad \( m^t \otimes x^t \)

i) Preserve the comparison results of the input graphs.

ii) Simplicity of the local explanations.

iii) Restrict local explanations to subset of the self-explanations for robustness.

Solved by gradient descent-ascent: the constraints are enforced softly to allow...
• Adapt a self-explanation when compared with different objects.

\[
x^s \quad x^t
\]

\[
m^s \otimes x^s \quad m^t \otimes x^t
\]

**SNX-KL:**

\[
\min_{m^s, m^t} \ell \left( f(x^s, x^t), f(m^s \otimes x^s, m^t \otimes x^t) \right) + \gamma (\|a(m^s)\| + \|a(m^t)\|) + \beta (\text{KL}(m^s || M^s) + \text{KL}(m^t || M^t))
\]

Solved by the regular gradient descent.

- i) Preserve the comparison results of the input graphs.
- ii) Simplicity of the local explanations.
- iii) Restrict local explanations to subset of the self-explanations for robustness.
Experimental results

• Datasets
  • Bipolar disorder (BP) classification of human brains.
  • Chemical molecule in material discovery.

• Overall explanation performance
  • faithfulness loss: simulate the target prediction (↓)
  • conformity: agreement with the self-explanation (↑).

<table>
<thead>
<tr>
<th>Dataset</th>
<th># graphs</th>
<th># nodes</th>
<th># edges</th>
<th># features</th>
<th># explain pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecule</td>
<td>200</td>
<td>10.77</td>
<td>9.77</td>
<td>1068</td>
<td>320</td>
</tr>
<tr>
<td>BP</td>
<td>90</td>
<td>82</td>
<td>315.84</td>
<td>82</td>
<td>216</td>
</tr>
</tbody>
</table>

[Graphs showing experimental results for BP and molecule datasets]
Experimental results

- Convergence of gradient descent ascent.

\[ m^s \odot x^s \]

\[ m^t \odot x^t \]

\( \lambda \): Lagrangian multipliers for the constraints
Experimental results

- Case study: bipolar disorder in human brains

The relevance of the connections between regions of interest is based on neuroscience study [1].

Experimental results

- Case study: molecules

Self-explanations

\[ x^s \]

\[ x^t \]

Adapted explanations

\[ m^s \otimes x^s \]

\[ m^t \otimes x^t \]

Other adapted explanations

\[ m^s \otimes x^s \]

\[ m^t \otimes x^t \]

*The relevance of the identified sub-structure of the molecules is confirmed by a bio-chemist.

For more details, see
Chao Chen, Yifan Shen, Guixiang Ma, Xiangnan Kong, Srinivas Rangarajan, Xi Zhang, and Sihong Xie. Self-learn to Explain Siamese Networks Robustly. ICDM 2021.
Machine learning on graphs

- Accuracy
- Interpretability
- Robustness
- Fairness

Fairness certification and efficient trade-offs on graphs
Unfair predictions on graphs

• Privileged group (0) is treated favorably, compared to the protected group (1).

Social network for job recommendation

Equally good candidates

Accounts Reviews Products

+ spam
+ non-spam

• Fair predictions should treat data from different groups the same.
Different types of unfairness due to different reasons

- **Disparate impact (DI):** different probabilities of being positive.

- **Not equalized True Positive Rates (ETRP):**

- **Not equalized ranking performances (e.g., NDCG):**
Certificating fairness on graphs

• With multiple fairness metrics, can we certify that they are satisfied?
  • For linear model on IID data, it is a simple equation.
  • for example, to certify statistical parity,
    \[ \frac{\sum_{i=1}^{N_0} w^T x_i}{N_0} = \frac{\sum_{j=1}^{N_1} w^T x_j}{N_1} \]
  • For node classification, need to take into account of the connections.
  • To simplify the problem, consider the linearized GNN*

\[
\Pr(\hat{Y}_j = 1|G; \theta) = \sigma \left( (\tilde{W})^K H^{(0)} \prod_{k=0}^{K} \theta^{(k)} \right)
\]

• No disparate impact if

\[
\left[ \frac{1}{N_0} \mathbb{1}[G_0]^\top (\tilde{W})^K H^{(0)} - \frac{1}{N_1} \mathbb{1}[G_1]^\top (\tilde{W})^K H^{(0)} \right] \prod_{k=0}^{K} \theta^{(k)} = 0
\]

• Similar certifications for equalized TRP/TNR/NDCG.

Fair learning with multiple objectives

- Optimizing one metric can harm the others.
- Find all efficient trade-offs and let the end-users select the suitable trade-off, possibly using additional domain knowledge.
- Multi-Objective Optimization (MOO)

\[
\min_{\theta} \ell(\theta) = (\ell_1(\theta), \ldots, \ell_m(\theta))^T,
\]

\(\ell_1(\theta)\): overall classification loss

\(\ell_2(\theta) = l^{DI}(\theta)\): for removing disparate impact

\(\ell_3(\theta) = l^{FNR}(\theta)\): for equalized FNR.

\(\ell_4(\theta) = l^{FPR}(\theta)\): for equalized FNR.

\(\ell_5 = l^{XN}(\theta)\): for equalized FNR.
Fair learning with multiple objectives

\[ \min_{\theta} \ell(\theta) = (\ell_1(\theta), \ldots, \ell_m(\theta))^T, \]

Jacobian \( (J(\theta))_{i,j} = \frac{\partial \ell_i}{\partial \theta_j}(\theta). \)

Descent in one objective can lead to ascend in another. How to combine the multiple gradients to ensure descent in all objectives?

Solve the dual problem:

\[
\begin{align*}
\max_{\lambda} & \quad -\frac{1}{2} \| \sum_{j=1}^{m} \lambda_j (J(\theta))_j \|^2 \\
\text{s.t.} & \quad \sum_{j=1}^{m} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, m.
\end{align*}
\]

\( \lambda = [\lambda_1, \ldots, \lambda_m] \): relative learning rates of the \( m \) objective functions.

\( \theta \leftarrow \theta - \eta_k \sum_{j=1}^{m} \lambda^*_j (J(\theta))_j. \)

\( \eta_k \): overall learning rate.

Remarks: 1) it converge to a single Pareto optimum; 2) multiple starting points can lead to multiple optimal solutions.
Experimental results

When optimizing one fairness metric with prediction accuracy

Only adversarial fair learning can efficiently optimize many metrics.
• MOO dominates adversarial fair training

<table>
<thead>
<tr>
<th></th>
<th>YelpChi</th>
<th></th>
<th></th>
<th>YelpNYC</th>
<th></th>
<th></th>
<th>YelpZip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td># Sol’s</td>
<td>#Dom’d</td>
<td></td>
<td># Sol’s</td>
<td>#Dom’d</td>
<td># Sol’s</td>
</tr>
<tr>
<td>Epochs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1</td>
<td></td>
<td>9</td>
<td>0</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>0</td>
<td></td>
<td>31</td>
<td>2</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>117</td>
<td>0</td>
<td></td>
<td>109</td>
<td>1</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
<td>0</td>
<td></td>
<td>289</td>
<td>0</td>
<td></td>
<td>212</td>
</tr>
<tr>
<td>10</td>
<td>447</td>
<td>0</td>
<td></td>
<td>597</td>
<td>0</td>
<td></td>
<td>345</td>
</tr>
</tbody>
</table>

For more details, see
Kai Burkhother, Kenny Kwock, Sheldon Xu, Jiaxin Liu, Chao Chen, and Sihong Xie.
Certification and Trade-off of Multiple Fairness Criteria in Graph-based Spam Detection.
CIKM 2021.
Conclusions

More connections between humans and ML

• Individual and collective perception of fairness and how that influence fairness evaluation.

• Human provide constraints for the learning of fair and transparent ML.

Systematic study

• All aspects of ML are not isolated.

• Dynamics are abundant.