Transparent and Fair Machine Learning on Graphs for Humans

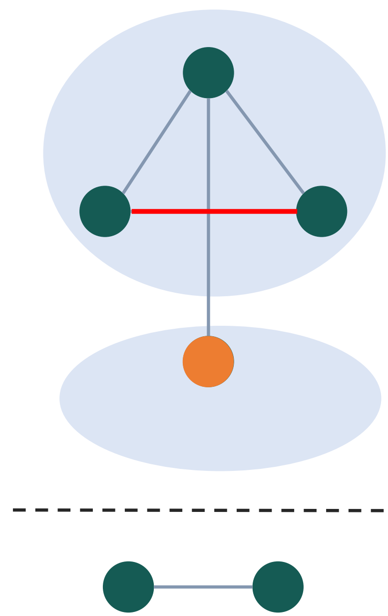
Sihong Xie, Assistant Professor

Computer Science and Enginneering

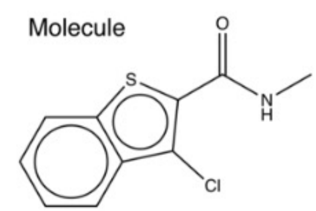
Lehigh University

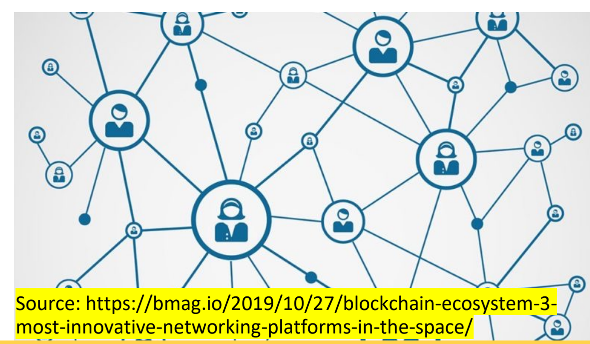
Machine learning on graphs

- 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.



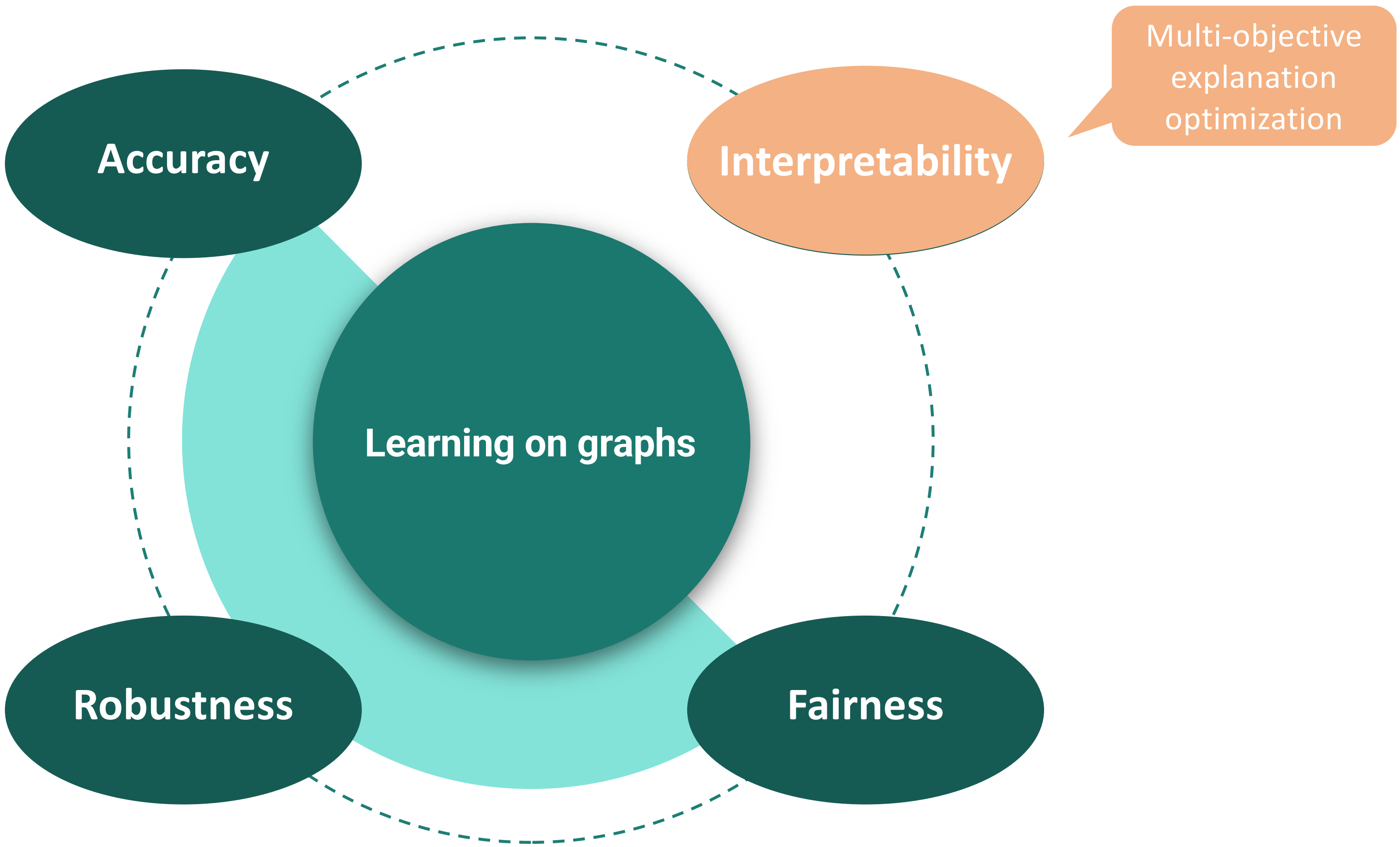
Source: https://exploringyourmind.com/the-human-connectome-project/





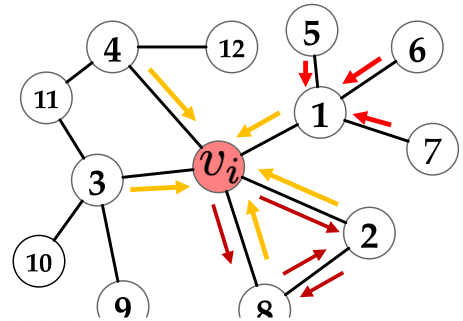
Source: https://bmag.io/2019/10/27/blockchain-ecosystem-3-most-innovative-networking-platforms-in-the-space/

Machine learning on graphs

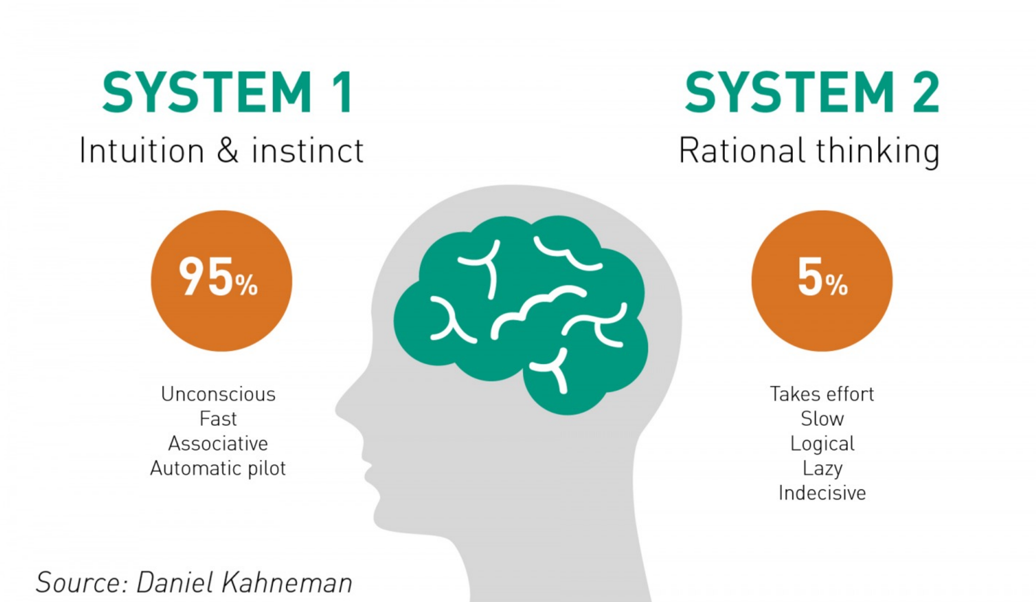


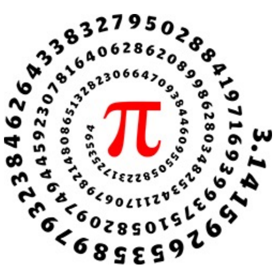
Interpretable ML: just a CS question?

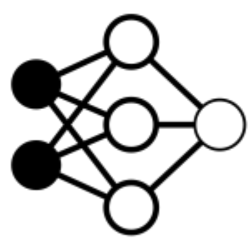
- Graphical models are not easy to be explained
  - Message passing and multiplexing.
  - Multiple steps of transformation.
  - Topology matters: tree vs. cycles.



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



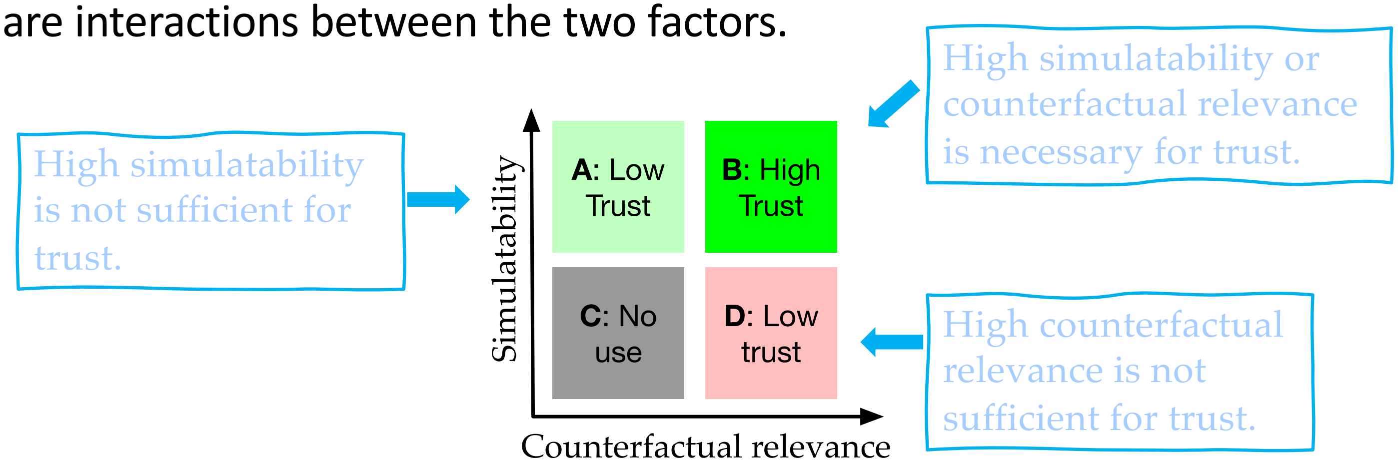




Source: https://news.dartmouth.edu/news/2015/03/pi-day-party-day-mathematical-mavens

Interpretable ML: hypotheses

- 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: `rain ⇒ wet_ground` and `!rain ⇒ !wet_ground`
  - There are interactions between the two factors.



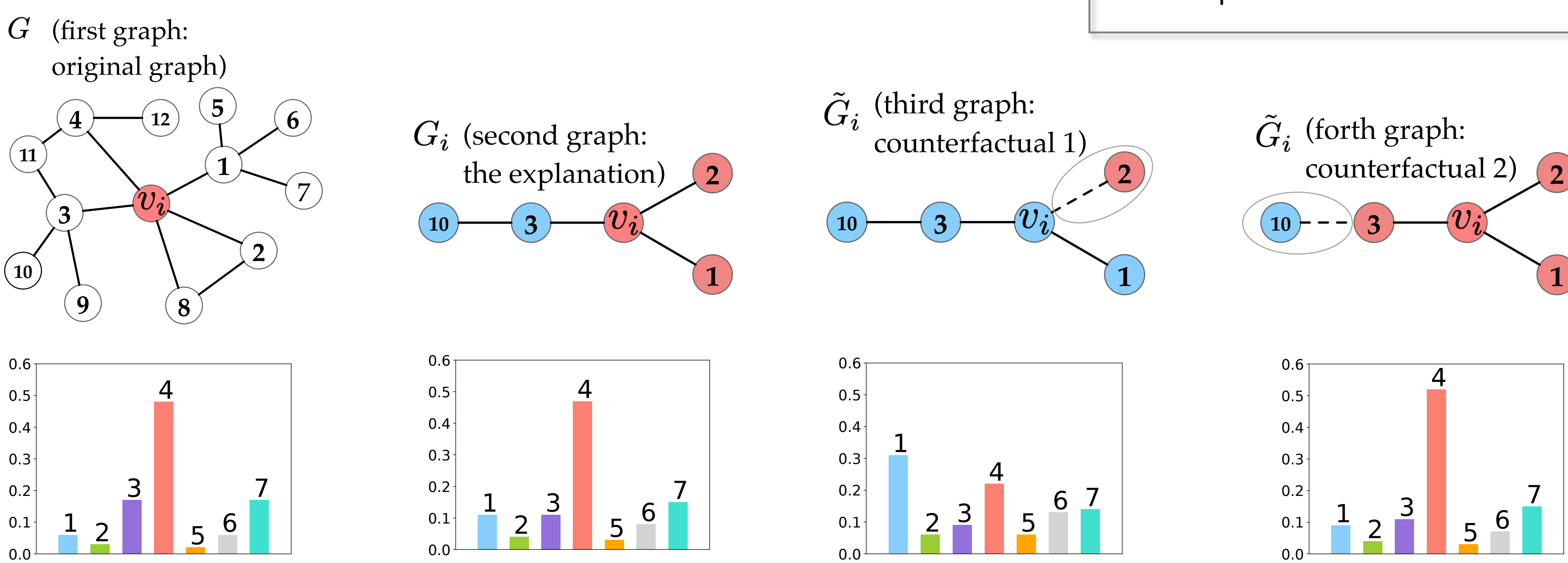
[1] *J.Lee, etc.* Trust in Automation: Designing for Appropriate Reliance. 2004. Human Factors.

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.

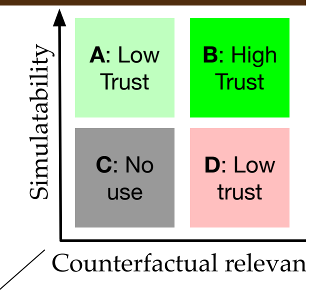
○	○	○	○	○
very little	*little*	*not sure*	*much*	*very much*

- perceived simulatability
- perceived counterfactual relevance
- acceptance



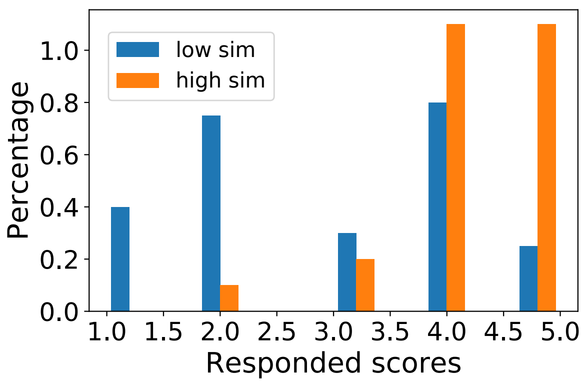
Interpretable ML: a human subject study

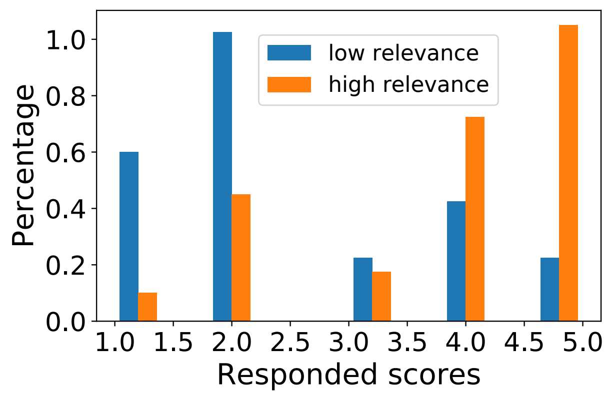
- Measuring simulatability, counterfactual relevance, and their interactions:
  - Collected 400 responses.

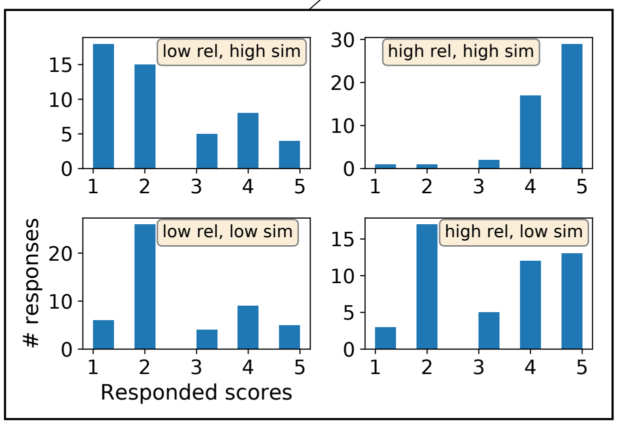


Simulatability helps

Causality helps







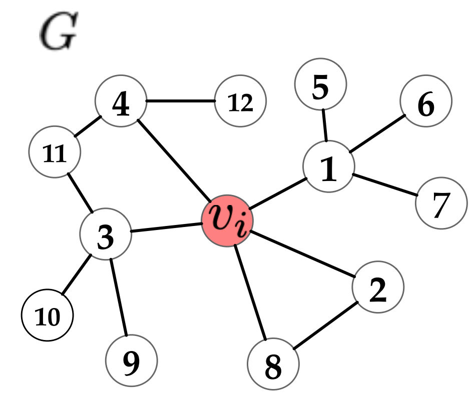
Statistical significance tests conducted to consider the size of samples.

Interpretable ML: a multi-objective approach

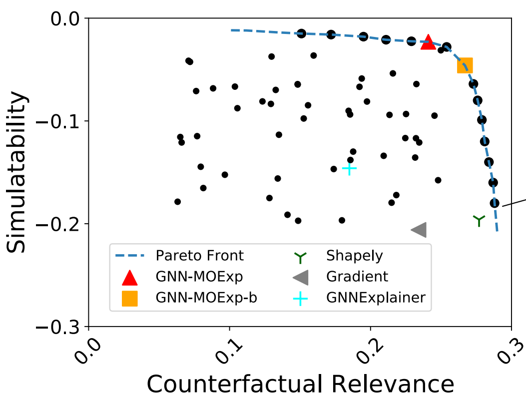
- Multiple objective optimization:

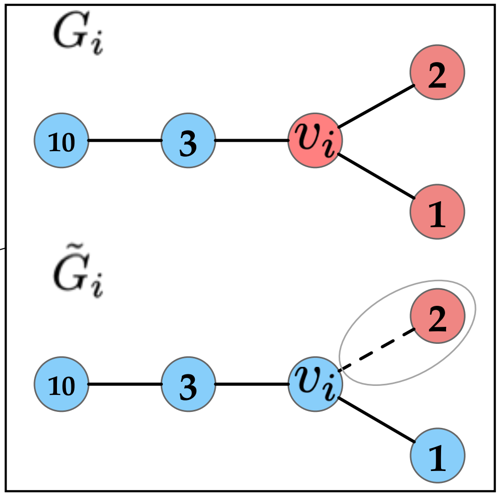
$$\max_{G_i, \tilde{G}_i} \quad F(G_i, \tilde{G}_i) = (\nu(G_i), |\mu(G_i, \tilde{G}_i)|)$$

$$\text{s.t.} \quad v_i \in \tilde{G}_i \subset G_i \subset G, \; |G_i| \leq C, \; G_i \text{ acyclic}$$



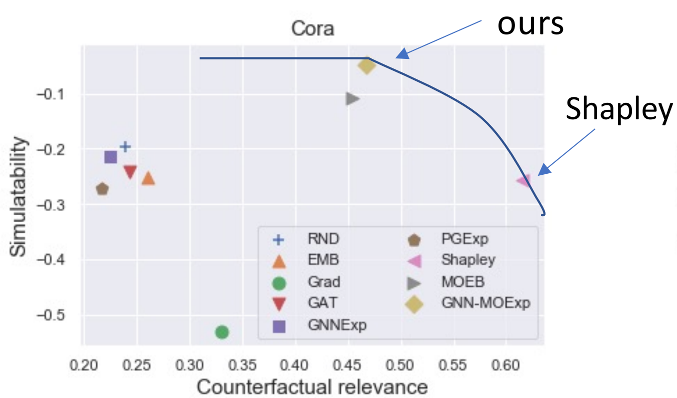
- Large discrete search space and non-differentiable objective functions.
- Need to find the Pareto front for balanced and efficient trade-offs.
- Algorithm:
- 1) BFS search.
- 2) explanation evaluation.
- 3) ranking-based explanations with provable balance and efficiency.

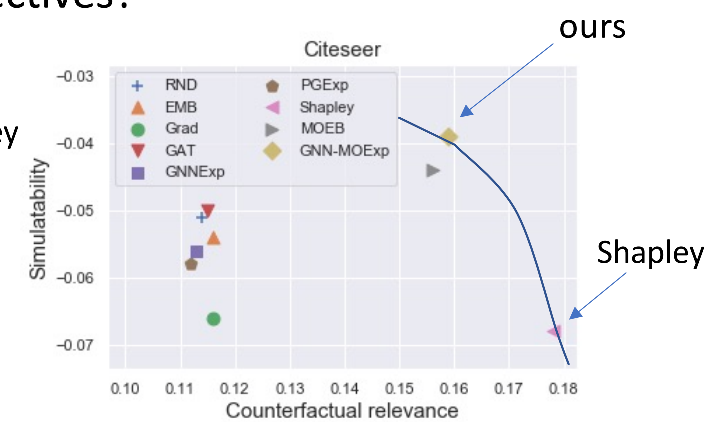




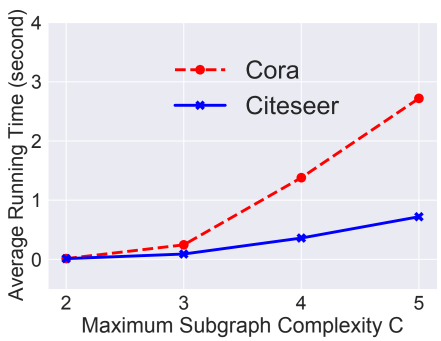
Experimental results

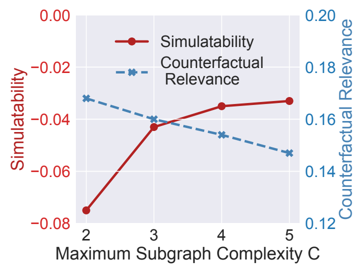
- Average performance: trade-off between the two objectives?



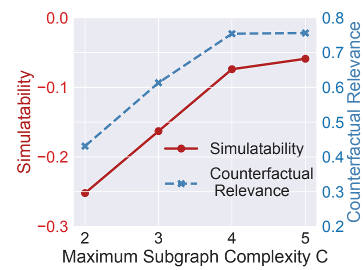


- Running time





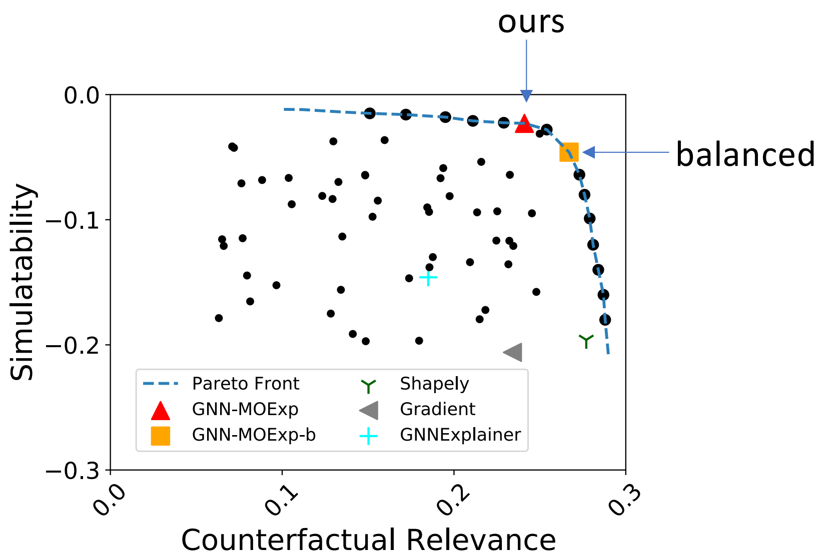
for Cora

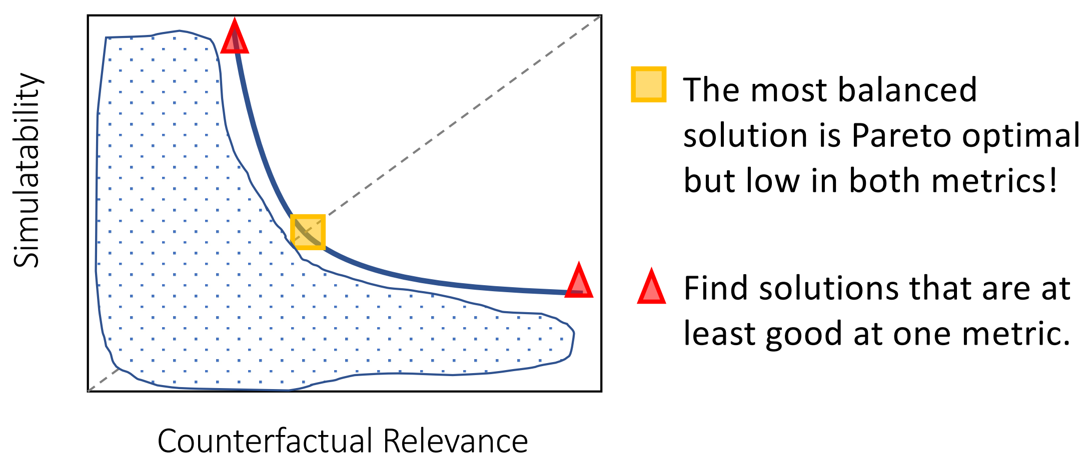


for citeseer

Experimental results

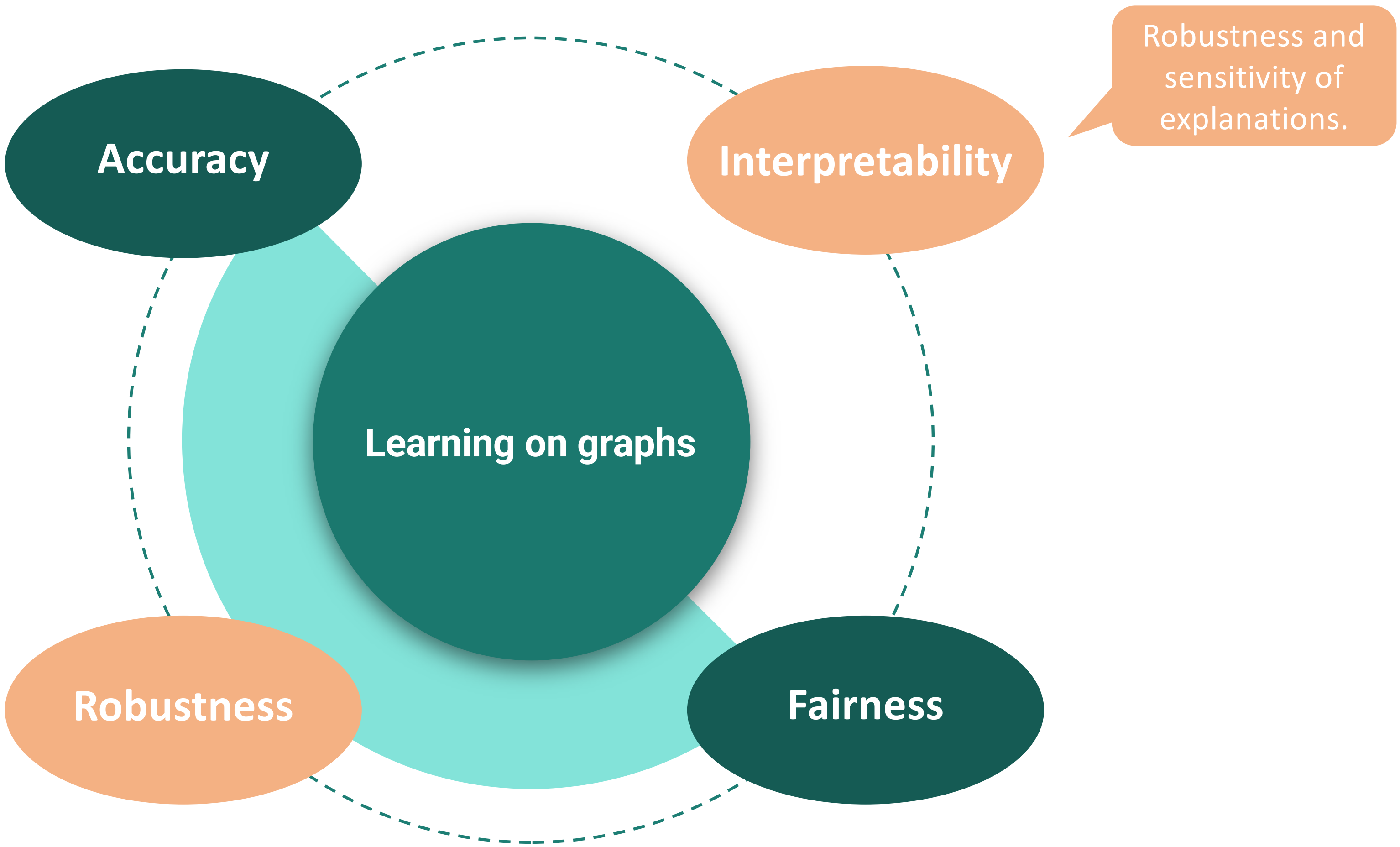
- A pitfall in finding well-balanced Pareto optimal explanations
- the ideal case
- in more cases, the Pareto front is not convex





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

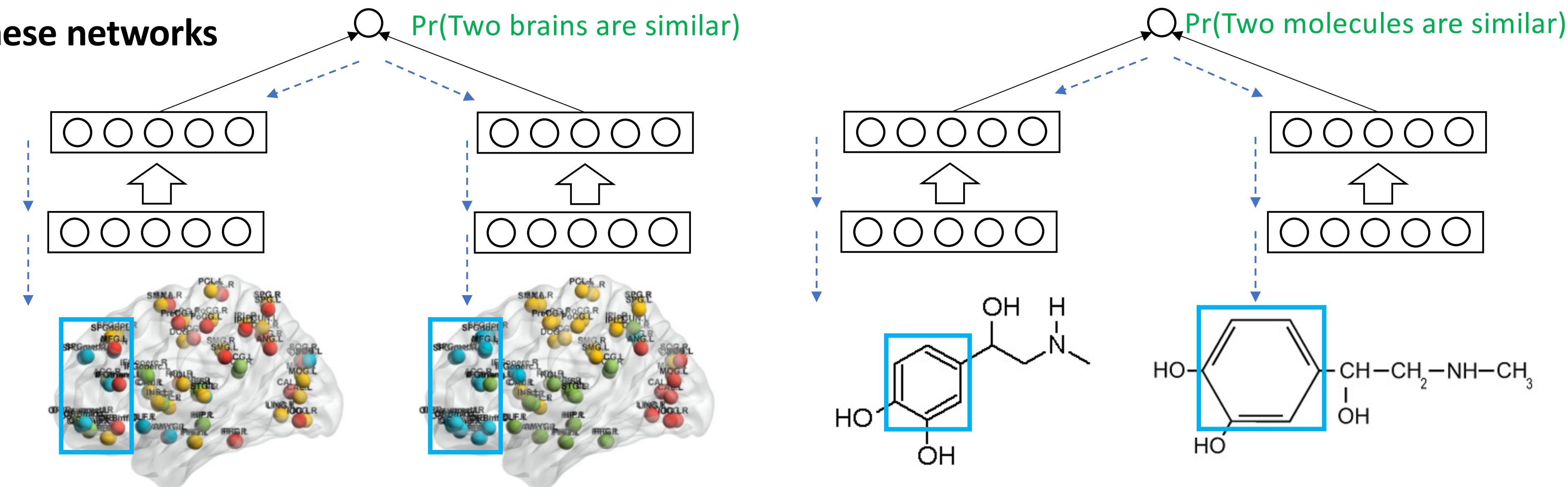
Machine learning on graphs



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



[1] Deep Graph Similarity Learning for Brain Data Analysis. *G. Ma, N.K. Ahmed, T.L. Willke, D. Sengupta.* CIKM, 2019.

Explaining the learned contrastive model

- For the explanations to be trusted, we want

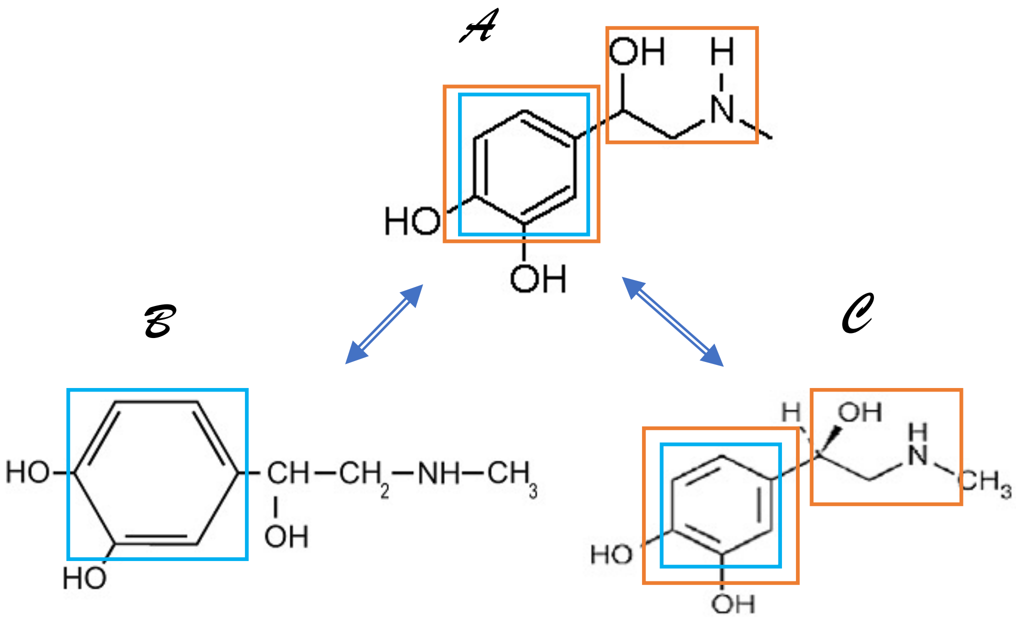
✓ Robustness / stability
Explanations should remain the same with respect to irrelevant changes.

✓ 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.



[1] Smoothed Geometry for Robust Attribution. NeurIPS, 2020.

Explainable contrastive model: self-explanation

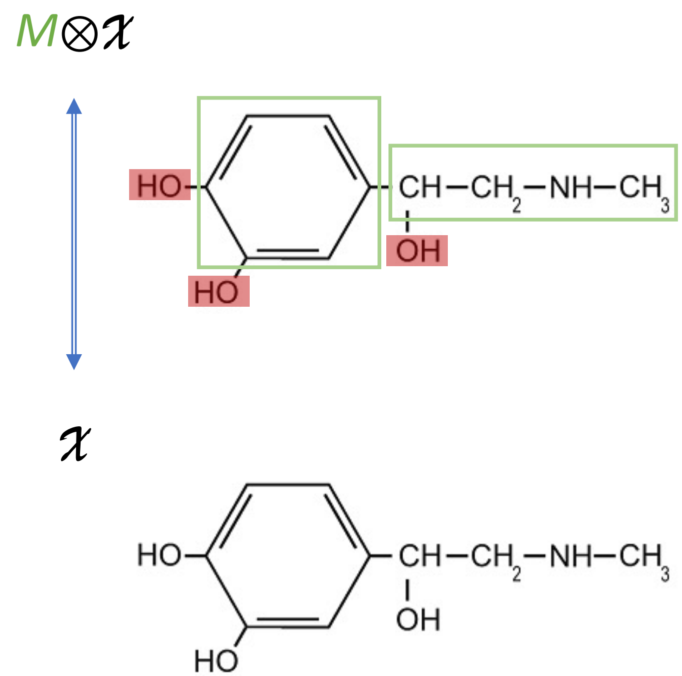
- Learn stable self-explanation for each graph
  - No labeled data is necessary.
- **Stage 1**: learn self-explanations

i) Mask out insignificant parts while preserving self-similarity.

ii) Minimize the retained portions to avoid trivial solution

$$\min_{\mathbf{M}} \quad \ell\left(f(\mathbf{x}, \mathbf{x}), f(\mathbf{x}, \mathbf{M} \otimes \mathbf{x})\right) + \gamma \|a(\mathbf{M})\|,$$
$$\text{s.t.} \quad g_i(\mathbf{M}) \leq 0, i = 1, \ldots, c.$$

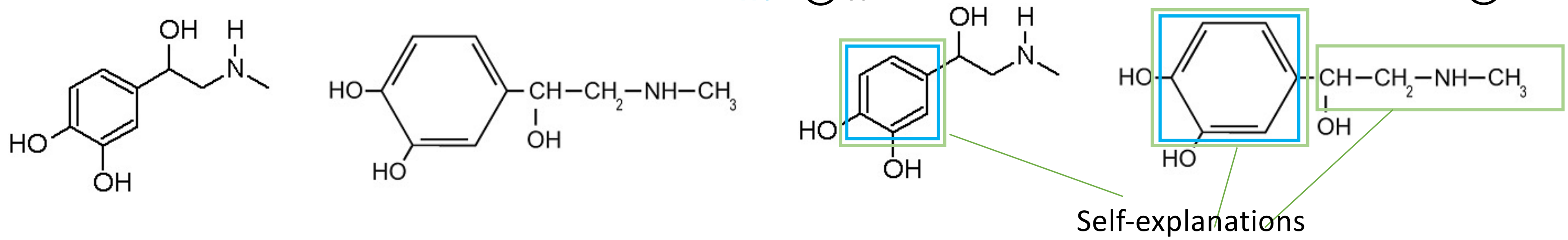
iii) Additional domain constraints



Constrained optimization

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

x^s $\quad$ x^t $\quad$ $m^s \otimes x^s$ $\quad$ $m^t \otimes x^t$



SNX:

$$\min_{\mathbf{m}^s, \mathbf{m}^t} \ell\left(f(\mathbf{x}^s, \mathbf{x}^t), f(\mathbf{m}^s \otimes \mathbf{x}^s, \mathbf{m}^t \otimes \mathbf{x}^t)\right)$$

i) Preserve the comparison results of the input graphs.

$$+\gamma\left(\|a(\mathbf{m}^s)\| + \|a(\mathbf{m}^t)\|\right)$$

ii) Simplicity of the local explanations.

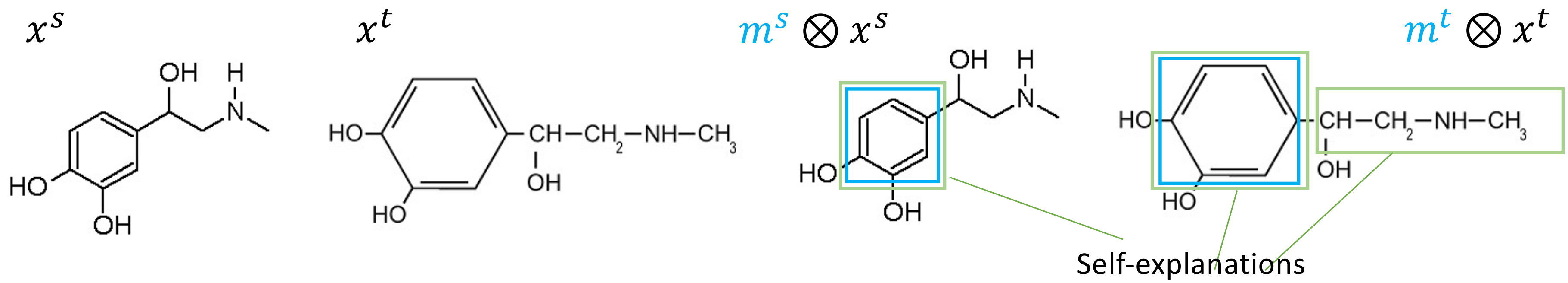
$$\text{s.t.} \quad g_i(\mathbf{m}) = \boxed{a(\mathbf{m}^s)_i} - \boxed{a(\mathbf{M}^s)_i} \leq 0, \; i = 1, \ldots, c_s,$$
$$g_{c_s+i}(\mathbf{m}) = \boxed{a(\mathbf{m}^t)_i} - \boxed{a(\mathbf{M}^t)_i} \leq 0, \; i = 1, \ldots, c_t.$$

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

Solved by gradient descent-ascent: the constraints are enforced softly to allow

Unconstrained optimization

- Adapt a self-explanation when compared with different objects.



SNX-KL:

$$\min_{\mathbf{m}^s,\mathbf{m}^t} \ell\left(f(\mathbf{x}^s,\mathbf{x}^t), f(\mathbf{m}^s\otimes\mathbf{x}^s,\mathbf{m}^t\otimes\mathbf{x}^t)\right)$$

i) Preserve the comparison results of the input graphs.

$$+\gamma(\|a(\mathbf{m}^s)\| + \|a(\mathbf{m}^t)\|)$$

ii) Simplicity of the local explanations.

$$+\beta(\mathrm{KL}(\boxed{\mathbf{m}^s}\|\boxed{\mathbf{M}^s}) + \mathrm{KL}(\boxed{\mathbf{m}^t}\|\boxed{\mathbf{M}^t}))$$

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

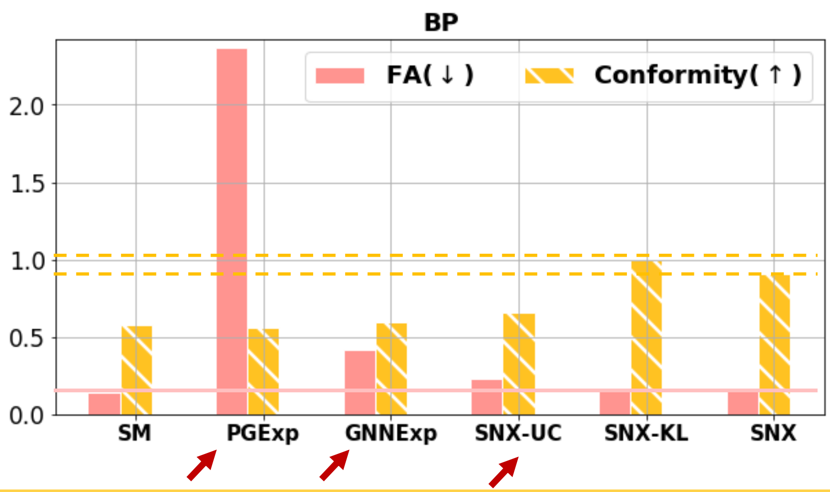
Solved by the regular gradient descent.

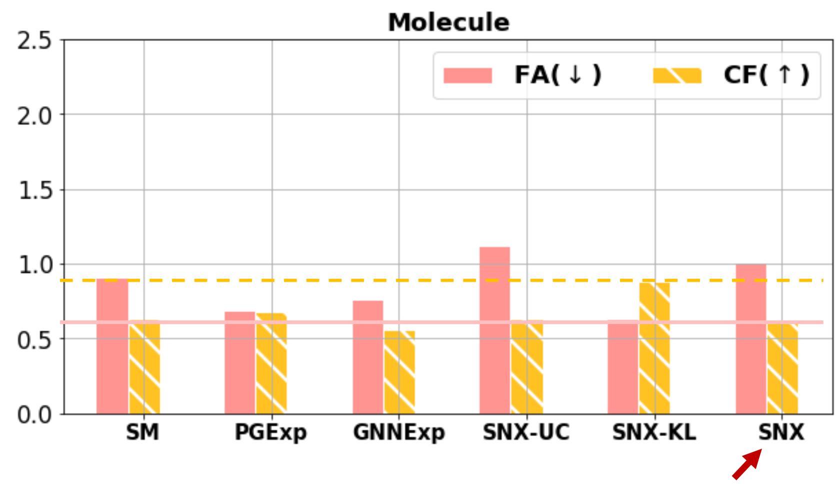
Experimental results

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

Dataset	# graphs	# nodes	# edges	# features	# explain pairs
Molecule	200	10.77	9.77	1068	320
BP	90	82	315.84	82	216

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

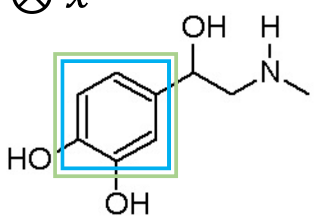




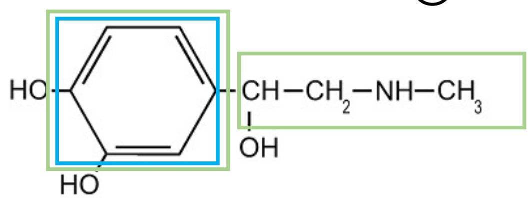
Experimental results

- Convergence of gradient descent ascent.

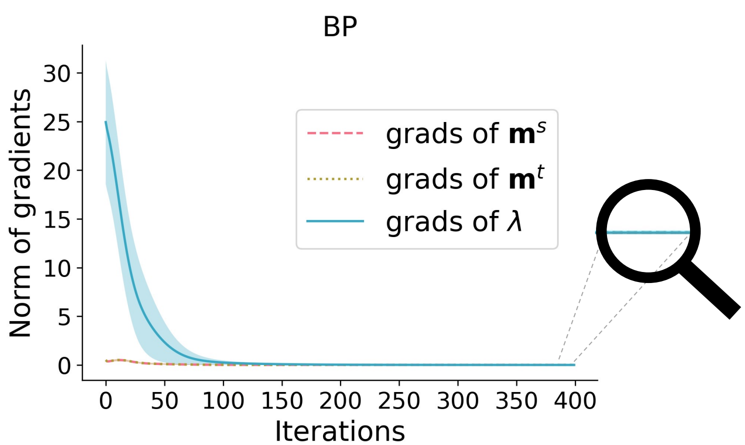
$m^s \otimes x^s$

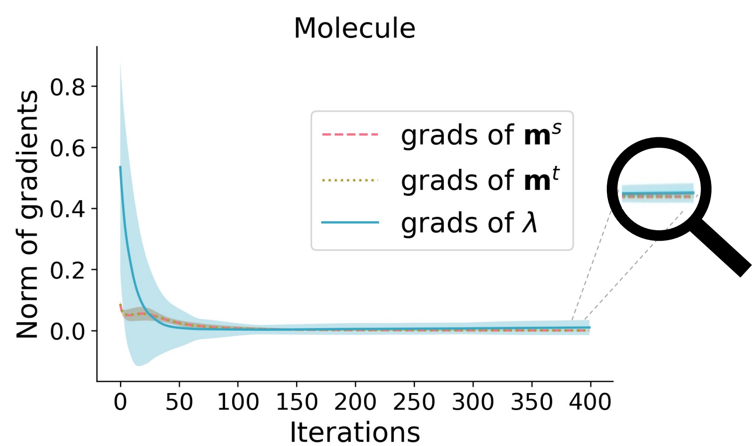


$m^t \otimes x^t$



λ: Lagrangian multipliers for the constraints

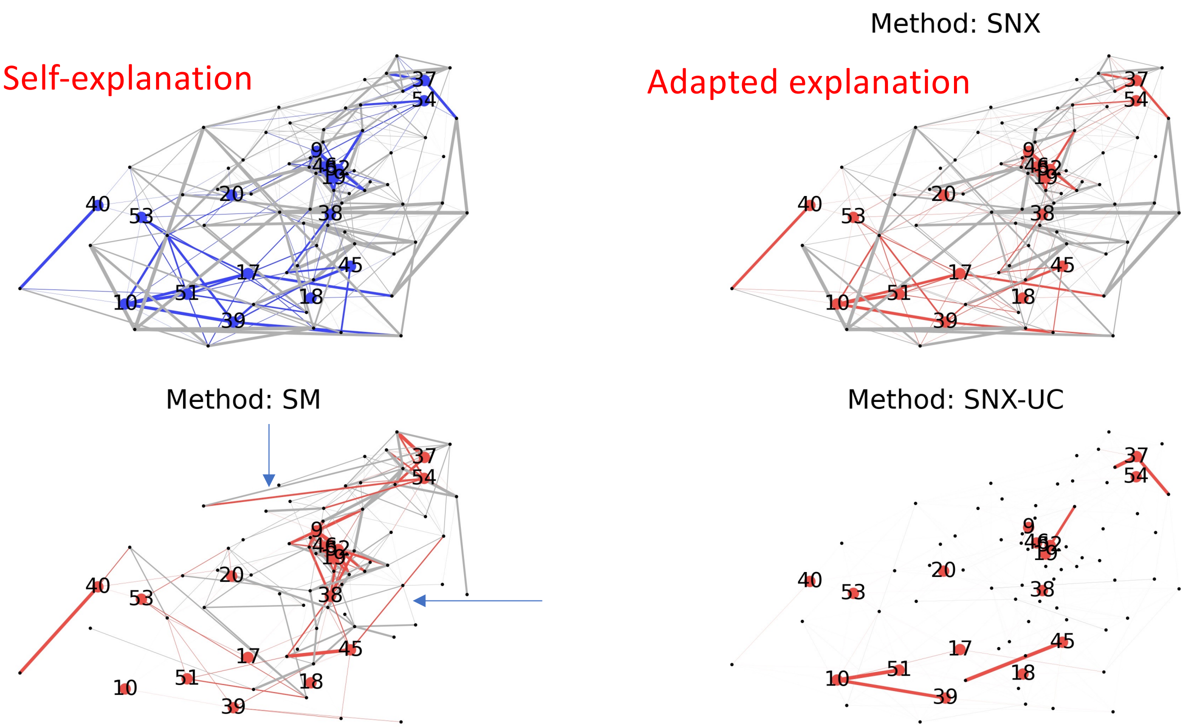




Experimental results

- Case study: bipolar disorder in human brains

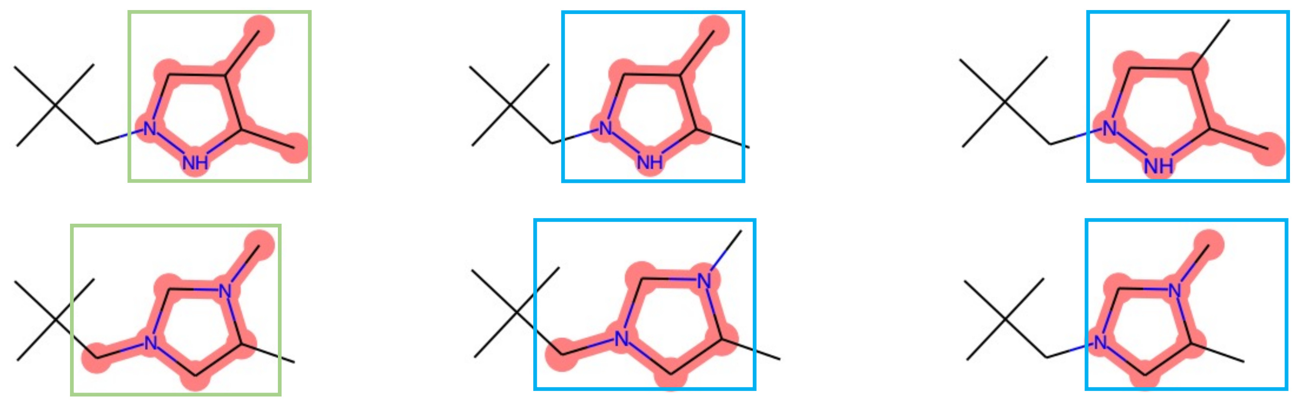
The relevance of the connections between regions of interest is based on neuroscience study [1].



[1] Niccolò Zovetti, etc. Default mode network activity in bipolar disorder. *Epidemiology and Psychiatric Sciences*, 29, 2020.

Experimental results

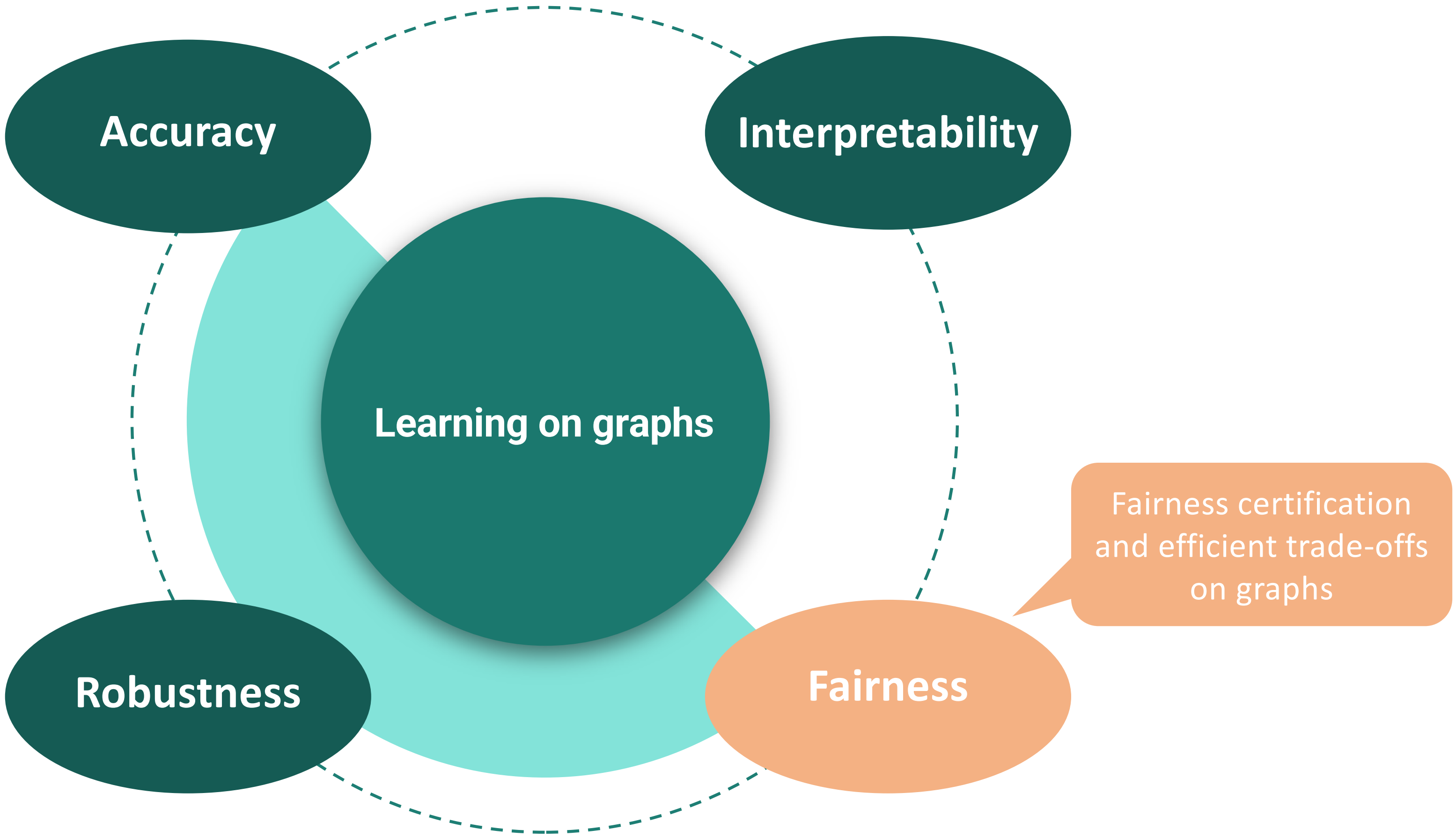
- Case study: molecules



*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.

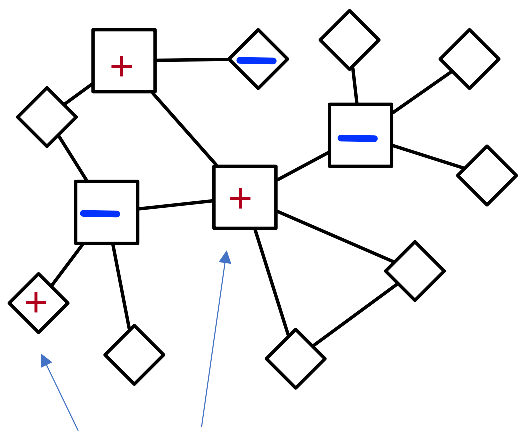
Machine learning on graphs



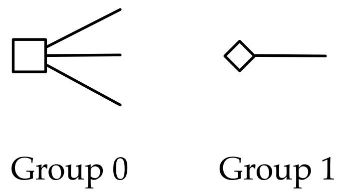
Unfair predictions on graphs

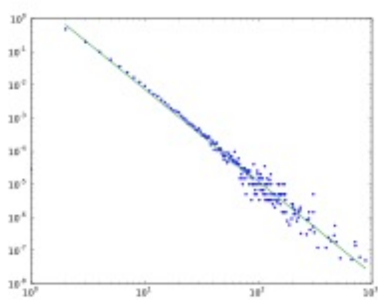
- Privileged group (0) is treated favorably, compared to the protected group (1).

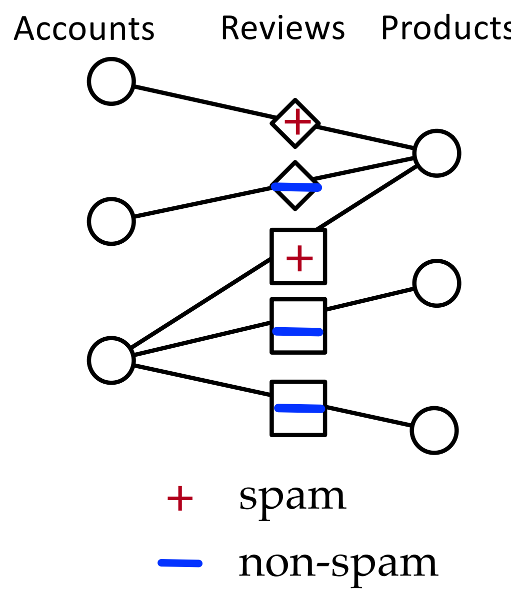
Social network for job recommendation

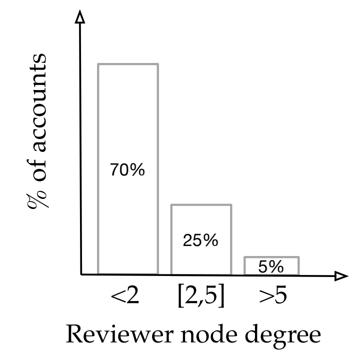


Equally good candidates





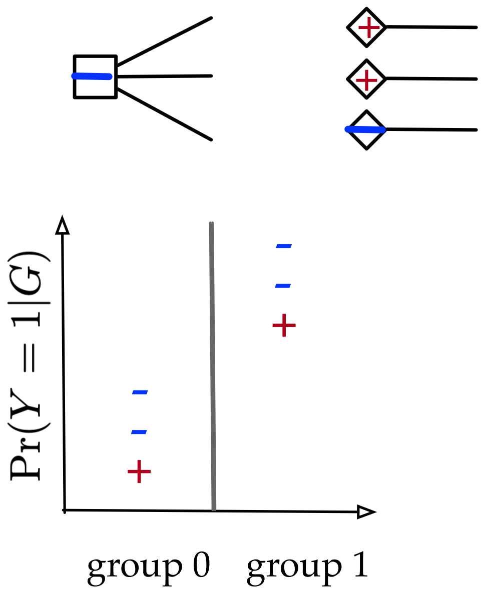




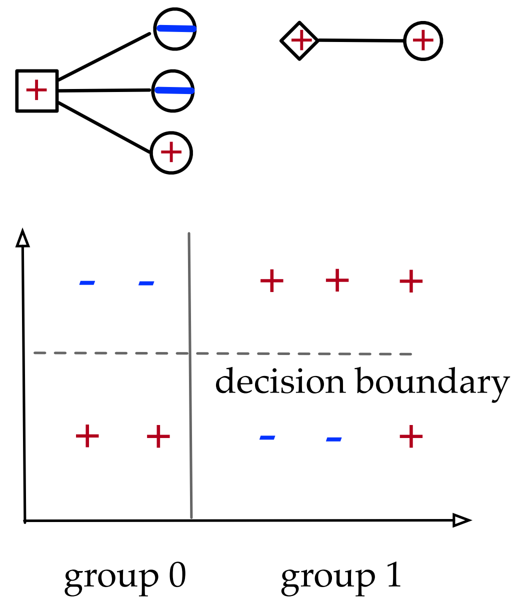
- Fair predictions should treat data from different groups the same.

Measuring fairness

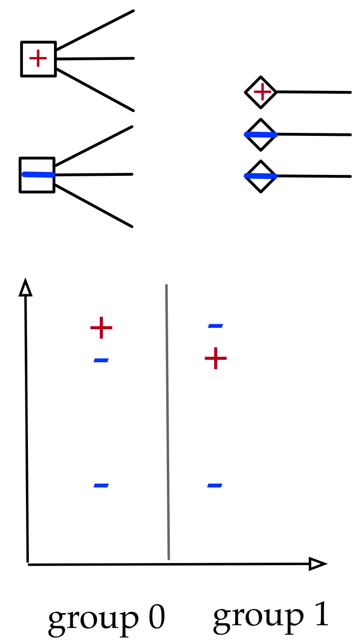
- Different types of unfairness due to different reasons



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



Not equalized True Positive Rates (ETRP)

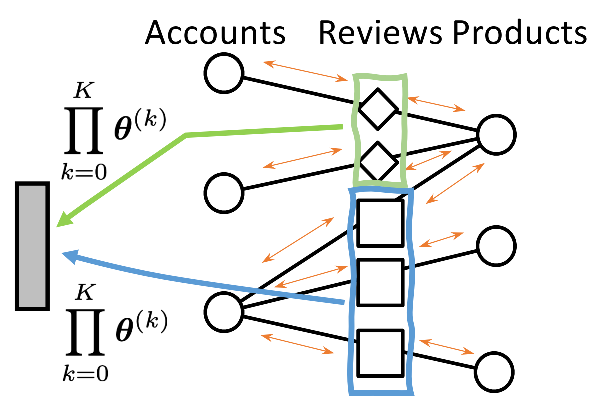


Not equalized ranking performances (e.g., NDCG)

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, $\frac{\sum_{i=1}^{N_0} \mathbf{w}^\top \mathbf{x}_i}{N_0} = \frac{\sum_{j=1}^{N_1} \mathbf{w}^\top \mathbf{x}_j}{N_1}$
  - For node classification, need to take into account of the connections.
  - To simplify the problem, consider the linearized GNN*

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



  - No disparate impact if

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

  - Similar certifications for equalized TRP/TNR/NDCG.

* Wu, Felix, etc. "Simplifying graph convolutional networks." In *International conference on machine learning*, pp. 6861-6871. PMLR, 2019.

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)

$$\min_{\boldsymbol{\theta}} \quad \ell(\boldsymbol{\theta}) = (\ell_1(\boldsymbol{\theta}), \ldots, \ell_m(\boldsymbol{\theta}))^\top,$$

$l_1(\boldsymbol{\theta})$: overall classification loss

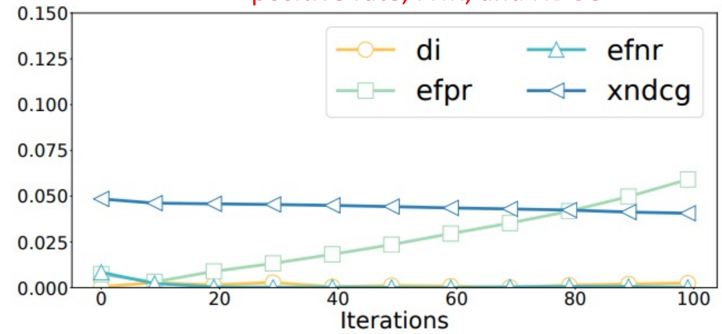
$l_2(\boldsymbol{\theta}) = l^{DI}(\boldsymbol{\theta})$: for removing disparate impact

$l_3(\boldsymbol{\theta}) = l^{FNR}(\boldsymbol{\theta})$: for equalized FNR.

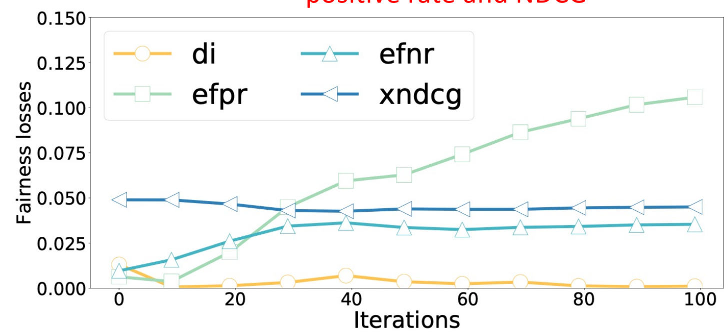
$l_4(\boldsymbol{\theta}) = l^{FPR}(\boldsymbol{\theta})$: for equalized FNR.

$l_5 = l^{XN}(\boldsymbol{\theta})$: for equalized FNR.

Optimize accuracy with equalized positive rate, FNR, and NDCG



Optimize accuracy with equalized positive rate and NDCG



Fair learning with multiple objectives

$$\min_{\boldsymbol{\theta}} \quad \ell(\boldsymbol{\theta}) = (\ell_1(\boldsymbol{\theta}), \ldots, \ell_m(\boldsymbol{\theta}))^\top,$$

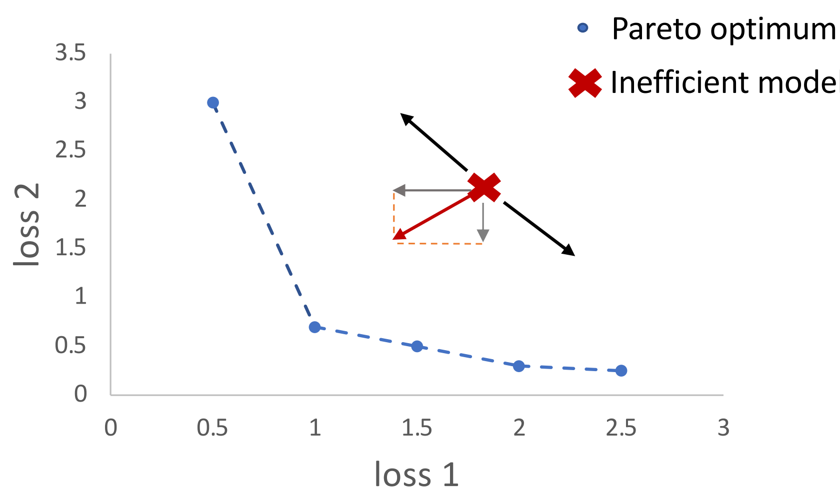
Jacobian $(J(\boldsymbol{\theta}))_{i,j} = \frac{\partial \ell_i}{\partial \theta_j}(\boldsymbol{\theta}).$

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:

$$\max_{\boldsymbol{\lambda}} \quad -\tfrac{1}{2}\|\textstyle\sum_{j=1}^{m} \lambda_j (J(\boldsymbol{\theta}))_j\|^2$$
$$\text{s.t.} \quad \textstyle\sum_{j=1}^{m} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, m.$$

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



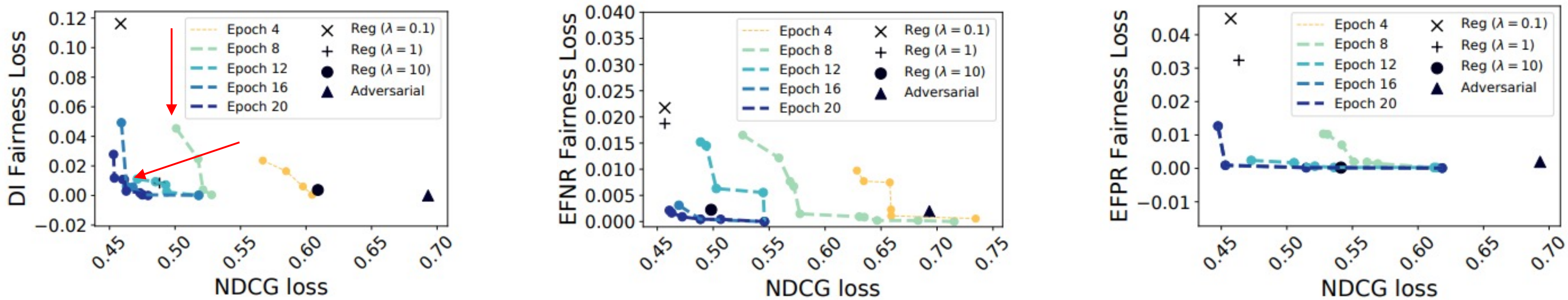
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta_k \sum_{j=1}^{m} \lambda_j^* (J(\boldsymbol{\theta}))_j.$$

η_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

	YelpChi		YelpNYC		YelpZip	
Epochs	# Sol's	#Dom'd	# Sol's	#Dom'd	# Sol's	#Dom'd
2	10	1	9	0	5	0
4	28	0	31	2	21	0
6	117	0	109	1	71	0
8	256	0	289	0	212	1
10	447	0	597	0	345	1

For more details, see
Kai Burkholder, 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.