Global Counterfactual Explainer for Graph Neural Networks

WSDM’23

Mert Kosan[1*], Zexi Huang[1*], Sourav Medya[2], Sayan Ranu[3], Ambuj Singh[1]

[1]Department of Computer Science, University of California, Santa Barbara
[2]Department of Computer Science, University of Illinois, Chicago
[3]Department of Computer Science, Indian Institute of Technology, Delhi
[*]Equal contribution

Research funded by NSF IIS 1817046
Travel funded by SIGIR and NSF

Explainability in Deep Learning

Explainability in Deep Learning

- **Why Explainability Should Be The Core Of Your AI Application**

  Vikas Gupta Forbes Councils Member Jan 23, 2023, 08:30am EST

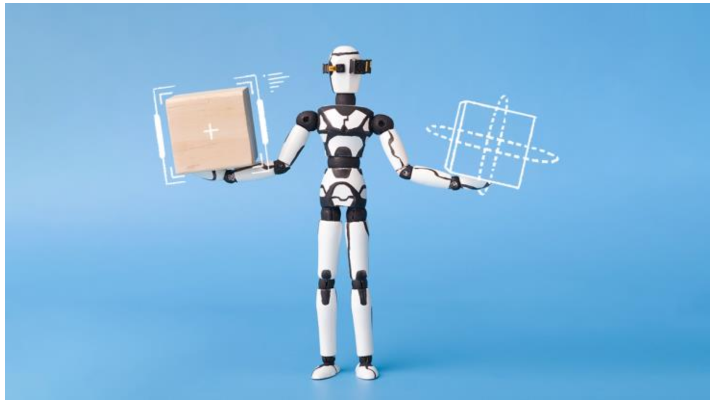
  "When something goes wrong, what do you tell your customer?"

- **Tools such as ChatGPT threaten transparent science** *Nature* **613**, 612 (2023) | 24 January 2023

  "Some scientists were already using chatbots as research assistant…"

- **Chatbots are creating thorny ethical questions about transparency in mental health care**  By Mohana Ravindranath Jan. 23, 2023 STAT

GETTY



HBR Staff/Milkos/Getty Images

AI Decisions and Explanations

AI Decisions and Explanations

- AI Decision System

{Credit Score: Medium, Stable Job Year: 2, Debts: Medium} → Denied

AI Decisions and Explanations

- AI Decision System

  {Credit Score: Medium, Stable Job Year: 2, Debts: Medium} → Denied

- Explanations
  - Feature importance [1, 2]

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?” explaining the predictions of any classifier. In SIGKDD, 2016.
[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.

AI Decisions and Explanations

- AI Decision System

  {Credit Score: Medium, Stable Job Year: 2, Debts: Medium} → Denied

- Explanations
  - Feature importance [1, 2]
  - Local counterfactual [3, 4]

  {Credit Score: Medium, Stable Job Year: 2, Debts: Low} → Approved

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In SIGKDD, 2016.
[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.
[3] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 2018.
[4] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In FAT, 2019.

AI Decisions and Explanations

- AI Decision System

  {Credit Score: Medium, Stable Job Year: 2, Debts: Medium} → Denied

- Explanations
  - Feature importance [1, 2]
  - Local counterfactual [3, 4]

  {Credit Score: Medium, Stable Job Year: 2, Debts: Low} → Approved

  - Global counterfactual [5]

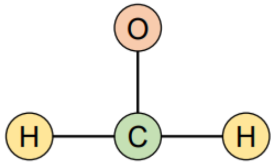
  if Credit Score = Medium and Stable Job Year = 2, then Debts <= Low → Approved

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In SIGKDD, 2016.
[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.
[3] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 2018.
[4] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In FAT, 2019.
[5] Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. NeurIPS, 2020.

Graph Explanations

Graph Explanations

- GNN decisions

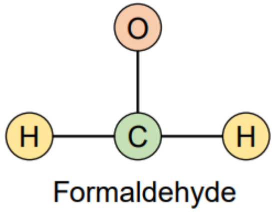


Formaldehyde

Mutagen

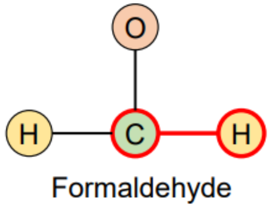
Graph Explanations

- GNN decisions



Mutagen

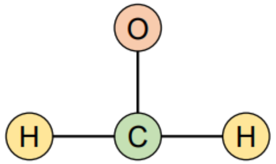
- Explanations
  - Subgraph importance [6, 7]



[6] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. NeurIPS, 2019.
[7] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. NeurIPS, 2020.

Graph Explanations

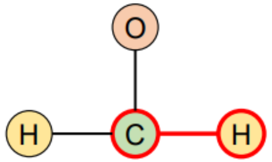
- GNN decisions



Formaldehyde

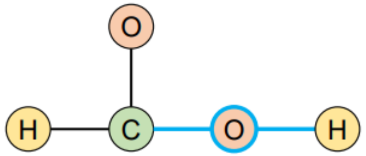
Mutagen

- Explanations
    - Subgraph importance [6, 7]



Formaldehyde

- Local counterfactual [8, 9, 10]



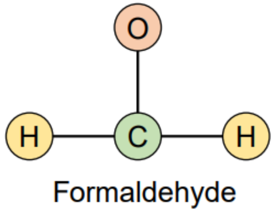
Formic acid

Non-mutagen

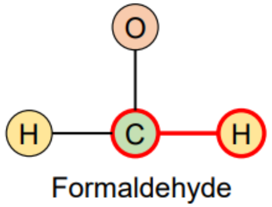
[6] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. NeurIPS, 2019.
[7] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. NeurIPS, 2020.
[8] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. NeurIPS, 2021.
[9] Ana Lucic, Maartje A Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In AISTATS, 2022.
[10] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In WebConf

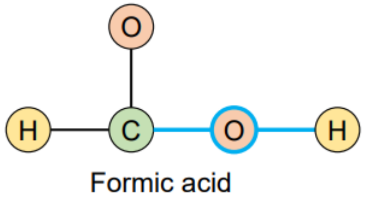
Graph Explanations

- GNN decisions
- Explanations
  - Subgraph importance [6, 7]
  - Local counterfactual [8, 9, 10]
  - Global counterfactual?



Mutagen





Non-mutagen

[6] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. NeurIPS, 2019.
[7] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. NeurIPS, 2020.
[8] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. NeurIPS, 2021.
[9] Ana Lucic, Maartje A Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In AISTATS, 2022.
[10] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In WebConf

Global Counterfactuals

- Global recourse r
  - For any undesired graph $G \in \mathbf{G}$, $r(G)$ should be desired → $\mathbf{GNN}(G) = 0, \mathbf{GNN}\big(r(G)\big) = 1$

Global Counterfactuals

- Global recourse r
  - For any undesired graph $G \in \mathbf{G}$, $r(G)$ should be desired → $\mathbf{GNN}(G) = 0, \mathbf{GNN}\big(r(G)\big) = 1$
  - r should be explainable, i.e., not complex

Global Counterfactuals

- Global recourse r
  - For any undesired graph $G \in \mathbf{G}$, $r(G)$ should be desired → $\mathbf{GNN}(G) = 0, \mathbf{GNN}\big(r(G)\big) = 1$
  - r should be explainable, i.e., not complex
- Counterfactual Summary
  - Let global recourse r be the collection of counterfactual summary graphs $\mathbf{C}$ in the desired class.

Global Counterfactuals

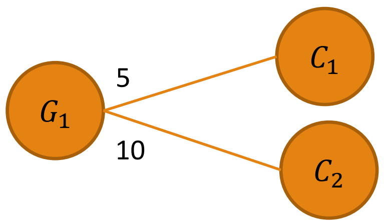
- Global recourse r
  - For any undesired graph $G \in \mathbf{G}$, $r(G)$ should be desired → $\mathbf{GNN}(G) = 0, \mathbf{GNN}\big(r(G)\big) = 1$
  - r should be explainable, i.e., not complex
- Counterfactual Summary
  - Let global recourse r be the collection of counterfactual summary graphs $\mathbf{C}$ in the desired class.
  - The recourse is easy to find for G as the closest counterfactual graph in $\mathbf{C}$.

$$r(G) = \underset{C \in \mathbf{C}}{\operatorname{argmin}}\, d(G, C)$$

Global Counterfactuals

- Global recourse r
  - For any undesired graph $G \in \mathbf{G}$, $r(G)$ should be desired → $\mathbf{GNN}(G) = 0, \mathbf{GNN}\big(r(G)\big) = 1$
  - r should be explainable, i.e., not complex
- Counterfactual Summary
  - Let global recourse r be the collection of counterfactual summary graphs $\mathbf{C}$ in the desired class.
  - The recourse is easy to find for G as the closest counterfactual graph in $\mathbf{C}$.

$$r(G) = \underset{C \in \mathbf{C}}{\operatorname{argmin}}\, d(G, C)$$



Quality of $\mathbf{C}$

Quality of $\mathbf{C}$

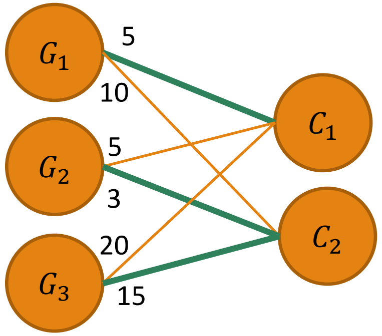
- Cost: minimize recourse cost for all undesired graphs.

$$\mathbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$

Quality of $\mathbf{C}$

- Cost: minimize recourse cost for all undesired graphs.

$$\mathbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$

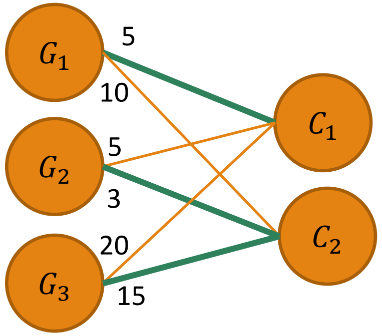


$$\mathbf{cost}(\mathbf{C}) = 5 + 3 + 15$$

Quality of $\mathbf{C}$

- Cost: minimize recourse cost for all undesired graphs.

$$\mathbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$



$$\mathbf{cost}(\mathbf{C}) = 5 + 3 + 15$$

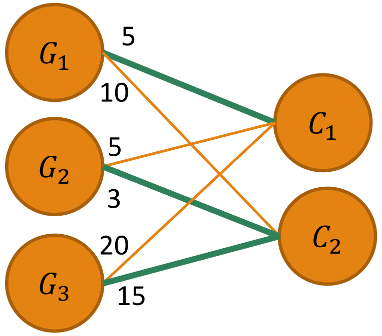
- Coverage: maximize the number of undesired graphs having counterfactual within distance θ.

$$\mathbf{coverage}(\mathbf{C}) = |\{G \in \mathbf{G} \mid \min_{C \in \mathbf{C}} d(G, C) \leq \theta\}| / |\mathbf{G}|$$

Quality of $\mathbf{C}$

- Cost: minimize recourse cost for all undesired graphs.

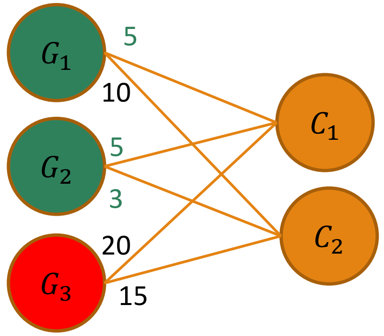
$$\mathbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$



$$\mathbf{cost}(\mathbf{C}) = 5 + 3 + 15$$

- Coverage: maximize the number of undesired graphs having counterfactual within distance θ.

$$\mathbf{coverage}(\mathbf{C}) = |\{G \in \mathbf{G} \mid \min_{C \in \mathbf{C}} d(G, C) \leq \theta\}| / |\mathbf{G}|$$

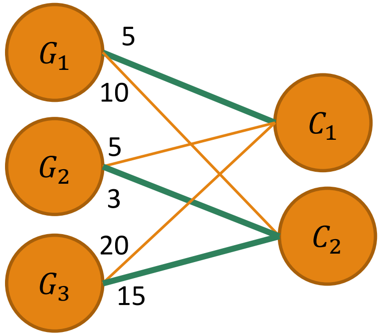


$$\mathbf{coverage}_{\theta=5}(\mathbf{C}) = \frac{|\{G_1, G_2\}|}{3} = \frac{2}{3}$$

Quality of $\mathbf{C}$

- Cost: minimize recourse cost for all undesired graphs.

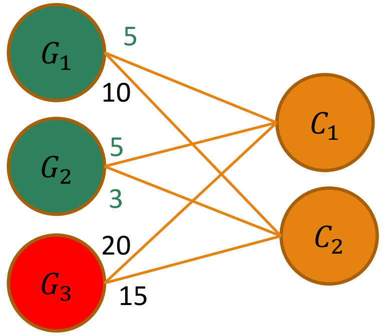
$$\mathbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$



$$\mathbf{cost}(\mathbf{C}) = 5 + 3 + 15$$

- Coverage: maximize the number of undesired graphs having counterfactual within distance θ.

$$\mathbf{coverage}(\mathbf{C}) = |\{G \in \mathbf{G} \mid \min_{C \in \mathbf{C}} d(G, C) \leq \theta\}| / |\mathbf{G}|$$



- Interpretability: minimize the size of the counterfactual summary.

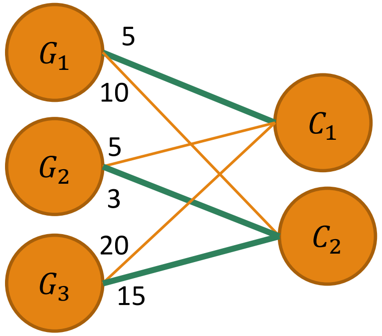
$$\mathbf{size}(\mathbf{C}) = |\mathbf{C}|$$

$$\mathbf{coverage}_{\theta=5}(\mathbf{C}) = \frac{|\{G_1, G_2\}|}{3} = \frac{2}{3}$$

Quality of $\mathbf{C}$

- Cost: minimize recourse cost for all undesired graphs.

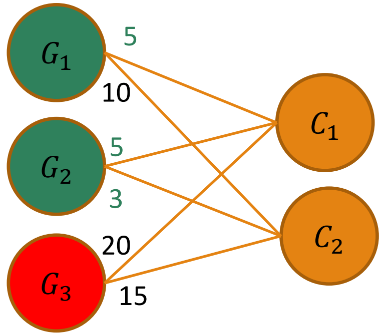
$$\mathbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$



$$\mathbf{cost}(\mathbf{C}) = 5 + 3 + 15$$

- Coverage: maximize the number of undesired graphs having counterfactual within distance θ.

$$\mathbf{coverage}(\mathbf{C}) = |\{G \in \mathbf{G} \mid \min_{C \in \mathbf{C}} d(G, C) \leq \theta\}| / |\mathbf{G}|$$



- Interpretability: minimize the size of the counterfactual summary.

$$\mathbf{size}(\mathbf{C}) = |\mathbf{C}|$$

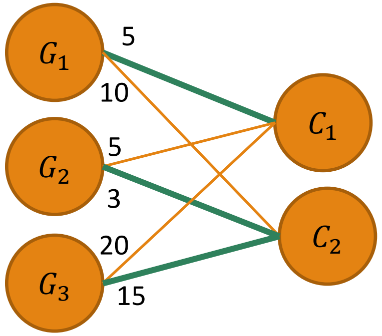
- Problem: $\max_{\mathbf{C}} \mathbf{coverage}(\mathbf{C})$ s.t. $\mathbf{size}(\mathbf{C}) = k$

$$\mathbf{coverage}_{\theta=5}(\mathbf{C}) = \frac{|\{G_1, G_2\}|}{3} = \frac{2}{3}$$

Quality of $\mathbf{C}$

- Cost: minimize recourse cost for all undesired graphs.

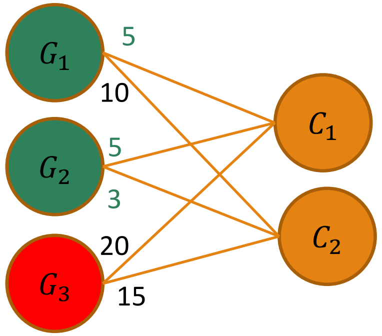
$$\textbf{cost}(\mathbf{C}) = \sum_{G \in \mathbf{G}} \min_{C \in \mathbf{C}} d(G, C)$$



$$\textbf{cost}(\mathbf{C}) = 5 + 3 + 15$$

- Coverage: maximize the number of undesired graphs having counterfactual within distance θ.

$$\textbf{coverage}(\mathbf{C}) = |\{G \in \mathbf{G} \mid \min_{C \in \mathbf{C}} d(G, C) \leq \theta\}| \,/\, |\mathbf{G}|$$



- Interpretability: minimize the size of the counterfactual summary.

$$\textbf{size}(\mathbf{C}) = |\mathbf{C}|$$

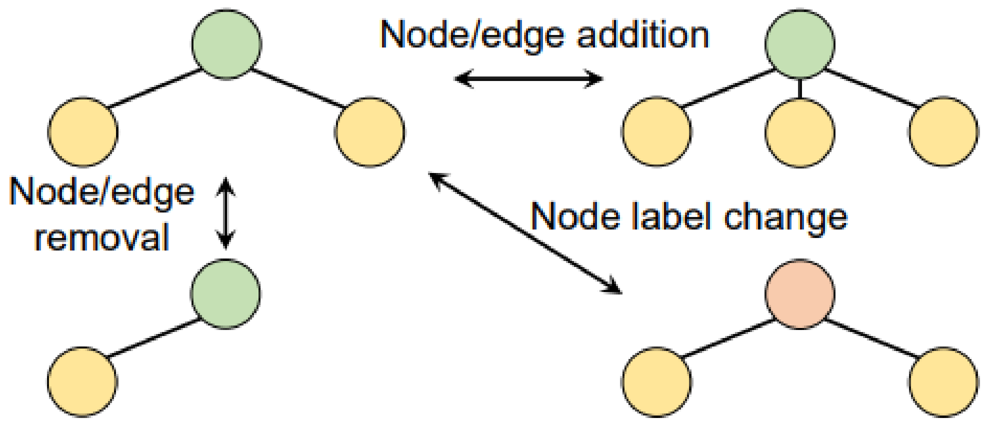
$$\textbf{coverage}_{\theta=5}(\mathbf{C}) = \frac{|\{G_1, G_2\}|}{3} = \frac{2}{3}$$

- Problem: $\max_{\mathbf{C}} \textbf{coverage}(\mathbf{C})$ s.t. $\textbf{size}(\mathbf{C}) = k$
  - NP-hard problem (reduction from Maximum Coverage problem)

Method

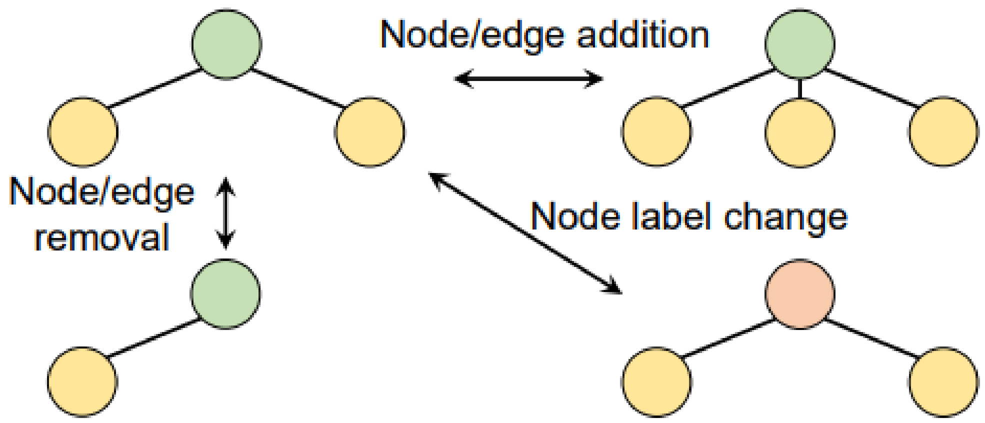
Method

- Generating counterfactual candidates: ***S***
  - Graph edit map: Neighbor graphs with one edit.



Method

- Generating counterfactual candidates: $\boldsymbol{S}$
  - Graph edit map: Neighbor graphs with one edit.

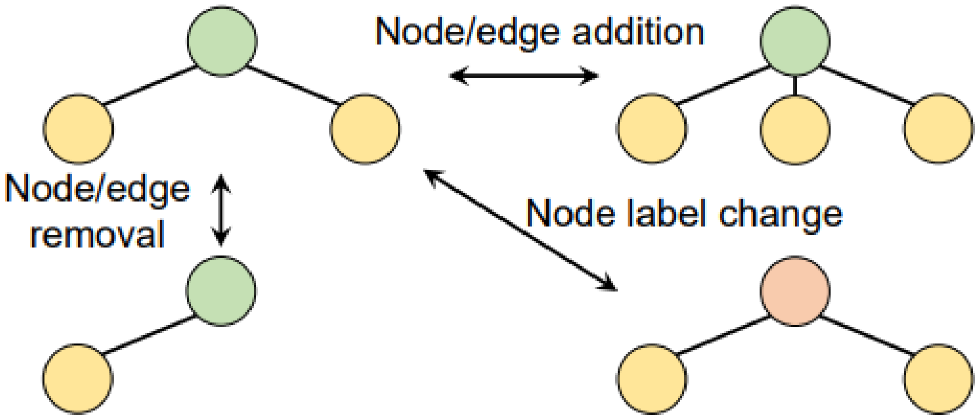


  - Vertex-reinforced random-walk [11]: $Prob(G \rightarrow G') \sim N(G')I(G')$.
    - It converges to representative set [12, 13].

[11] Robin Pemantle. Vertex-reinforced random walk. Probability Theory and Related Fields, 92(1):117–136, 1992.
[12] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In SIGKDD, 2010.
[13] Dheepikaa Natarajan and Sayan Ranu. A scalable and generic framework to mine top-k representative subgraph patterns. In ICDM, 2016.

Method

- Generating counterfactual candidates: $\boldsymbol{S}$
  - Graph edit map: Neighbor graphs with one edit.



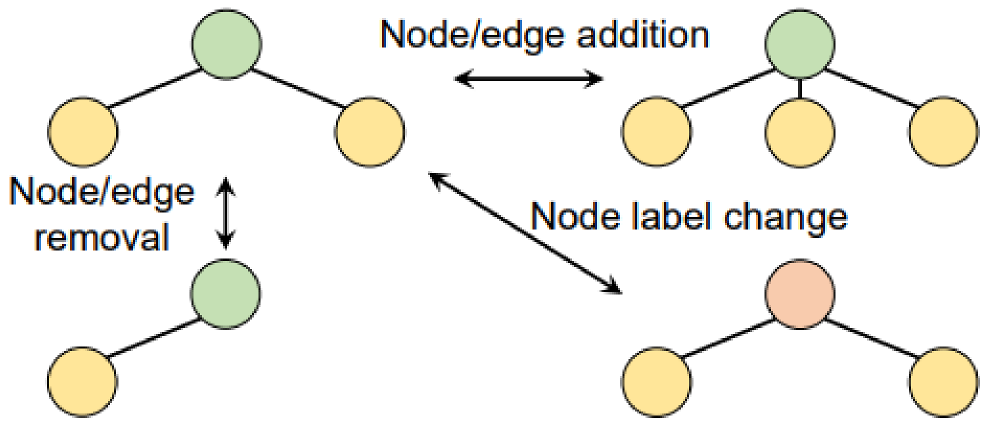
  - Vertex-reinforced random-walk [11]: $Prob(G \rightarrow G') \sim N(G')I(G')$.
    - It converges to representative set [12, 13].

$\mathbf{GNN}(G')$, **coverage**, **gain**

[11] Robin Pemantle. Vertex-reinforced random walk. Probability Theory and Related Fields, 92(1):117–136, 1992.
[12] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In SIGKDD, 2010.
[13] Dheepikaa Natarajan and Sayan Ranu. A scalable and generic framework to mine top-k representative subgraph patterns. In ICDM, 2016.

Method

- Generating counterfactual candidates: $\boldsymbol{S}$
  - Graph edit map: Neighbor graphs with one edit.



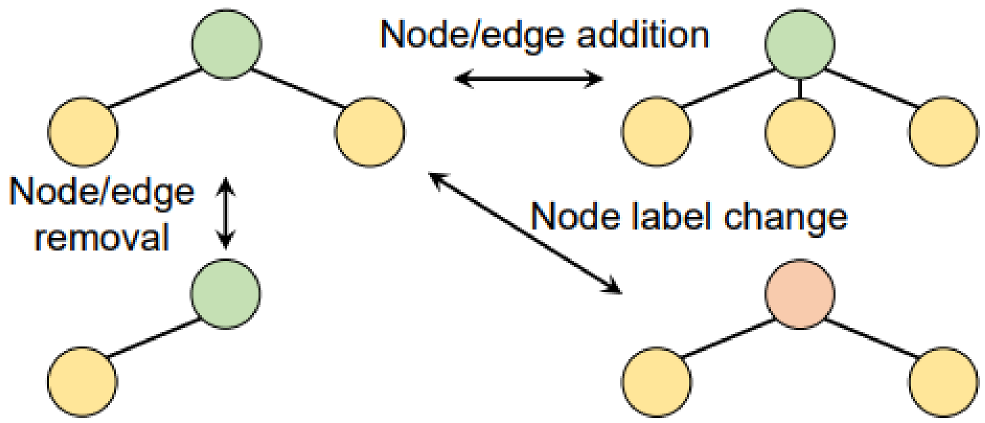
  - Vertex-reinforced random-walk [11]: $Prob(G \rightarrow G') \sim N(G')I(G')$.
    - It converges to representative set [12, 13].
  - Teleportation back to undesired graphs.

$\mathbf{GNN}(G')$, **coverage**, **gain**

[11] Robin Pemantle. Vertex-reinforced random walk. Probability Theory and Related Fields, 92(1):117–136, 1992.
[12] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In SIGKDD, 2010.
[13] Dheepikaa Natarajan and Sayan Ranu. A scalable and generic framework to mine top-k representative subgraph patterns. In ICDM, 2016.

Method

- Generating counterfactual candidates: $\mathbf{S}$
  - Graph edit map: Neighbor graphs with one edit.



  - Vertex-reinforced random-walk [11]: $Prob(G \rightarrow G') \sim N(G')I(G')$.
    - It converges to representative set [12, 13].
  - Teleportation back to undesired graphs.

$\mathbf{GNN}(G')$, $\mathbf{coverage}$, $\mathbf{gain}$

- Greedy summary
  - After finding $\mathbf{S}$, we apply greedy summary based on $\max \mathbf{gain}(C)$ where $C \in \mathbf{S}$ with k iterations.

[11] Robin Pemantle. Vertex-reinforced random walk. Probability Theory and Related Fields, 92(1):117–136, 1992.
[12] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In SIGKDD, 2010.
[13] Dheepikaa Natarajan and Sayan Ranu. A scalable and generic framework to mine top-k representative subgraph patterns. In ICDM, 2016.

Experiments

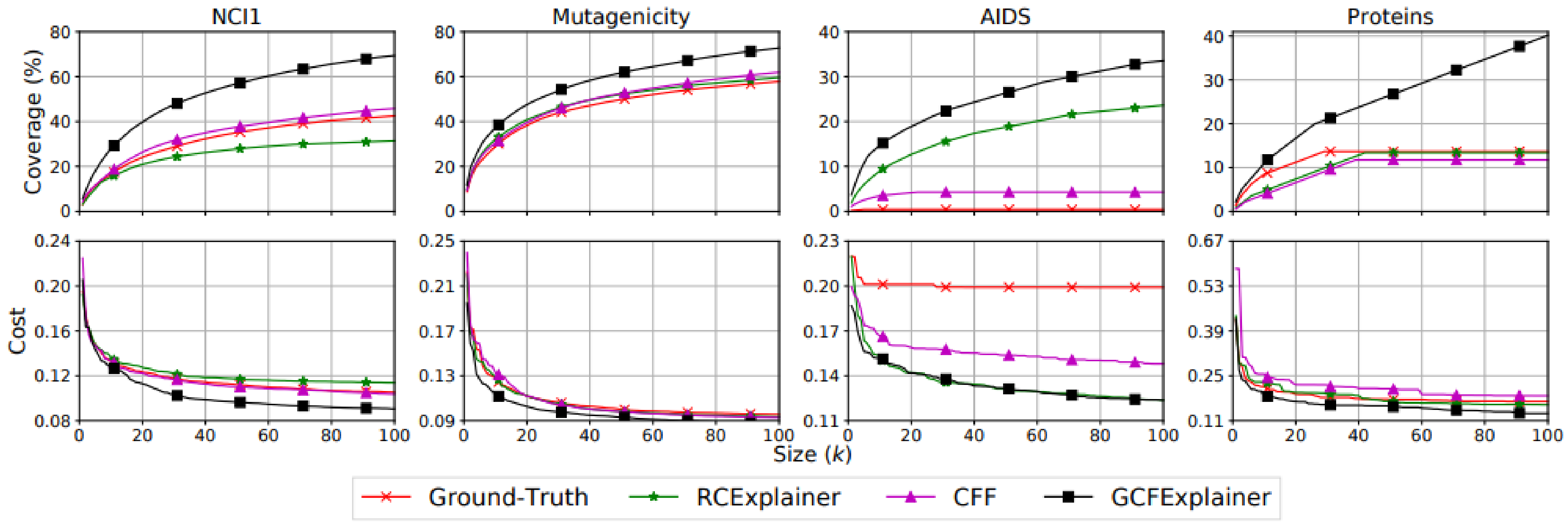
- GCFExplainer consistently outperforms the baselines with 9.5% reduction in recourse cost, 46.9% gain in recourse coverage, for top 10 counterfactuals.

	NCI1		Mutagenicity		AIDS		Proteins	
	Coverage	Cost	Coverage	Cost	Coverage	Cost	Coverage	Cost
Ground-Truth	16.54%	0.1326	28.96%	0.1275	0.41%	0.2012	8.47%	0.2155
RCExplainer	15.22%	0.1370	31.99%	0.1290	8.96%	0.1531	8.74%	0.2283
CFF	17.61%	0.1331	30.43%	0.1327	3.39%	0.1669	3.83%	0.2557
GCFExplainer	**27.85%**	**0.1281**	**37.08%**	**0.1135**	**14.66%**	**0.1516**	**10.93%**	**0.1856**

Experiments

- GCFExplainer consistently outperforms the baselines with 9.5% reduction in recourse cost, 46.9% gain in recourse coverage, for top 10 counterfactuals.

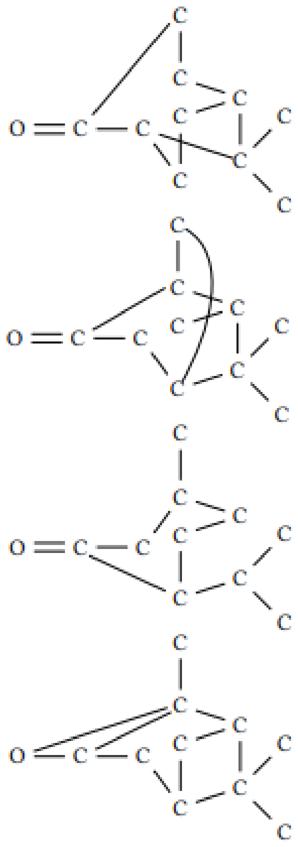
	NCI1		Mutagenicity		AIDS		Proteins	
	Coverage	Cost	Coverage	Cost	Coverage	Cost	Coverage	Cost
Ground-Truth	16.54%	0.1326	28.96%	0.1275	0.41%	0.2012	8.47%	0.2155
RCExplainer	15.22%	0.1370	31.99%	0.1290	8.96%	0.1531	8.74%	0.2283
CFF	17.61%	0.1331	30.43%	0.1327	3.39%	0.1669	3.83%	0.2557
GCFExplainer	**27.85%**	**0.1281**	**37.08%**	**0.1135**	**14.66%**	**0.1516**	**10.93%**	**0.1856**



Case Study

Case Study

- AIDS dataset [14]



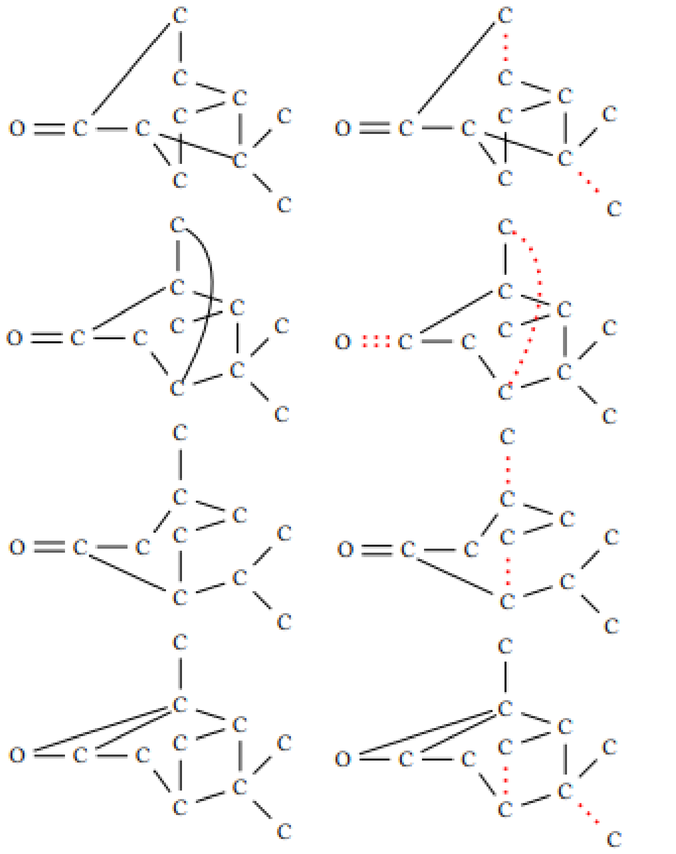
(a) Input graphs

[14] Kaspar Riesen and Horst Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).

Case Study

- AIDS dataset [14]

- Local counterfactuals
  - Hard to generalize

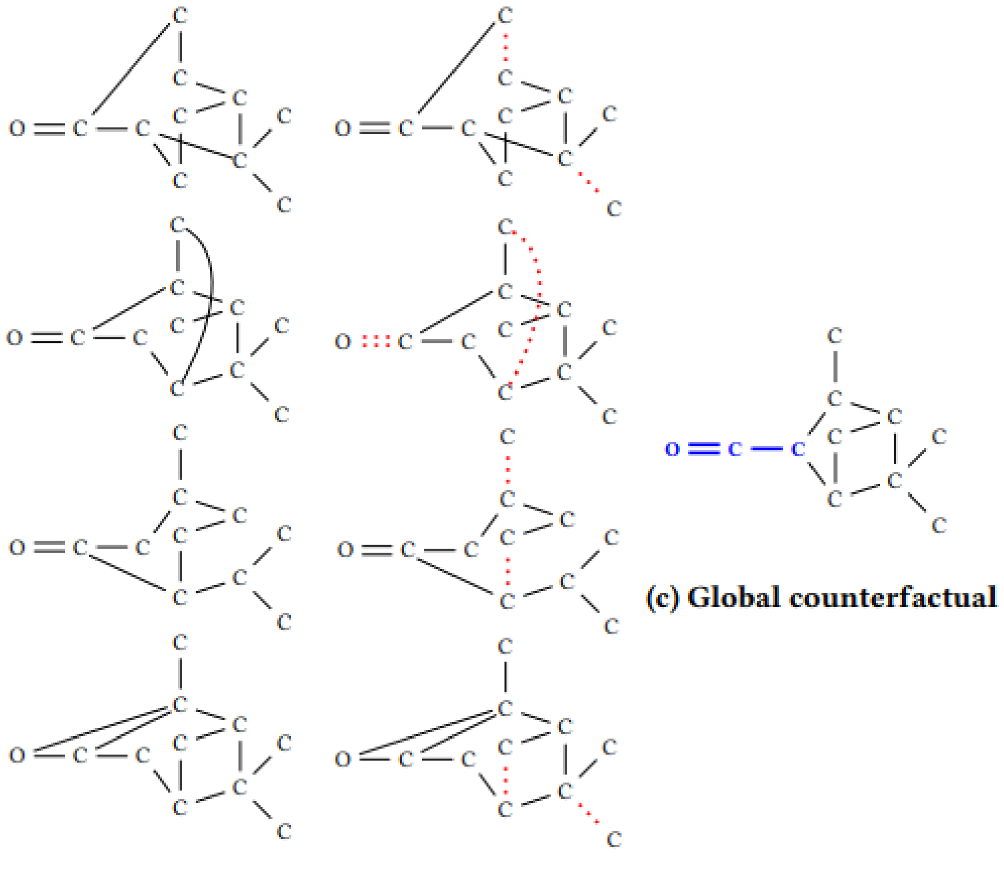


(a) Input graphs (b) Local counterfactuals

[14] Kaspar Riesen and Horst Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).

Case Study

- AIDS dataset [14]
- Local counterfactuals
  - Hard to generalize
- Global counterfactual
  - High-level recourse rule

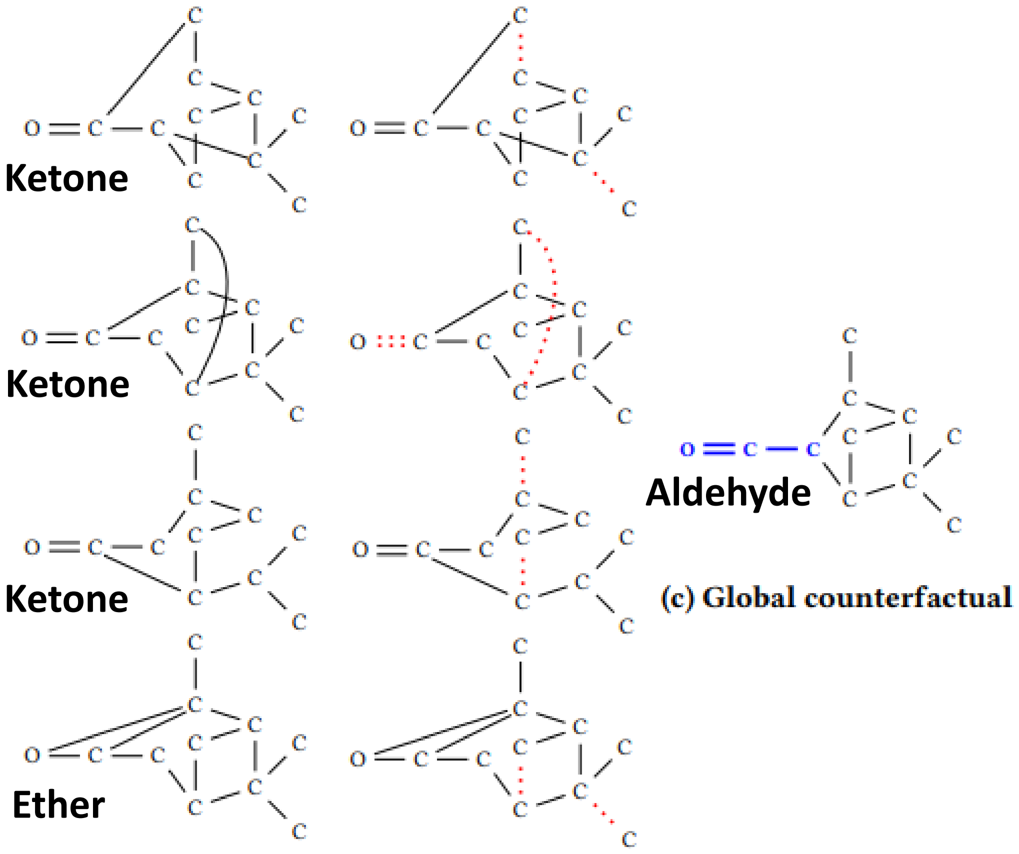


(a) Input graphs (b) Local counterfactuals (c) Global counterfactual

[14] Kaspar Riesen and Horst Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).

Case Study

- AIDS dataset [14]
- Local counterfactuals
  - Hard to generalize
- Global counterfactual
  - High-level recourse rule
  - Ketones and Ethers to Aldehydes

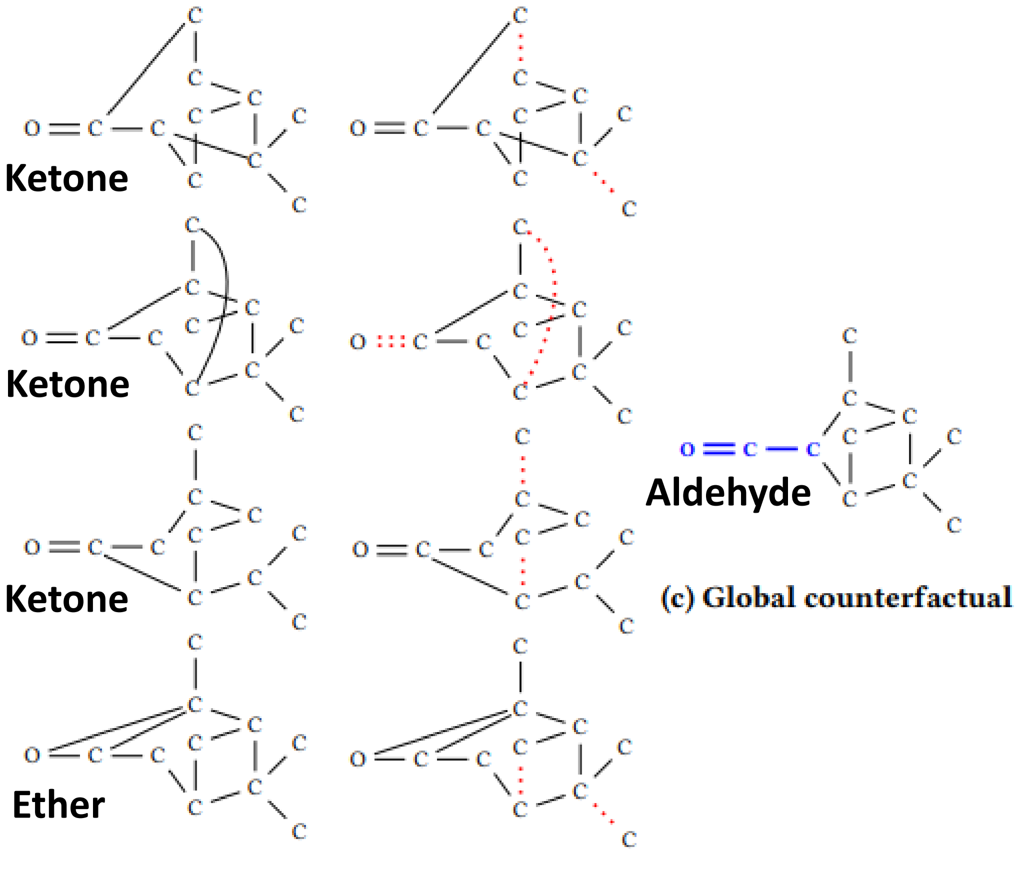


(a) Input graphs (b) Local counterfactuals (c) Global counterfactual

[14] Kaspar Riesen and Horst Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).

Case Study

- AIDS dataset [14]
- Local counterfactuals
  - Hard to generalize
- Global counterfactual
  - High-level recourse rule
  - Ketones and Ethers to Aldehydes
    - Combatting HIV [15]



(a) Input graphs (b) Local counterfactuals (c) Global counterfactual

[14] Kaspar Riesen and Horst Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).
[15] Edoardo Sarubbi, Pier Fausto Seneci, Michael R Angelastro, Norton P Peet, Maurizio Denaro, and Khalid Islam. 1993. Peptide aldehydes as inhibitors of HIV protease. FEBS letters 319, 3 (1993)

Global Counterfactual Explainer

- Contributions

Global Counterfactual Explainer

- Contributions
  - We formulate the novel problem of global counterfactual reasoning/explanation of GNNs for graph classification.

Global Counterfactual Explainer

- Contributions
  - We formulate the novel problem of global counterfactual reasoning/explanation of GNNs for graph classification.
  - While the problem is NP-hard, we present the first global counterfactual explainer generating diverse and representative counterfactual candidates.

Global Counterfactual Explainer

- Contributions
  - We formulate the novel problem of global counterfactual reasoning/explanation of GNNs for graph classification.
  - While the problem is NP-hard, we present the first global counterfactual explainer generating diverse and representative counterfactual candidates.
  - We demonstrate the effectiveness and usefulness of GCFExplainer in experiments.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?” explaining the predictions of any classifier. In SIGKDD, 2016.
[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.
[3] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology, 31(2), 2018.
[4] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In FAT, 2019.
[5] Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. NeurIPS, 2020.
[6] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. NeurIPS, 2019.
[7] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. NeurIPS, 2020.
[8] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. NeurIPS, 2021.
[9] Ana Lucic, Maartje A Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In AISTATS, 2022.
[10] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In WebConf
[11] Robin Pemantle. Vertex-reinforced random walk. Probability Theory and Related Fields, 92(1):117–136, 1992.
[12] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In SIGKDD, 2010.
[13] Dheepikaa Natarajan and Sayan Ranu. A scalable and generic framework to mine top-k representative subgraph patterns. In ICDM, 2016.
[14] Kaspar Riesen and Horst Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).
[15] Edoardo Sarubbi, Pier Fausto Seneci, Michael R Angelastro, Norton P Peet, Maurizio Denaro, and Khalid Islam. 1993. Peptide aldehydes as inhibitors of HIV protease. FEBS letters 319, 3 (1993), 253–256.

Global Counterfactual Explainer for Graph Neural Networks

WSDM'23

Mert Kosan[1*], Zexi Huang[1*], Sourav Medya[2], Sayan Ranu[3], Ambuj Singh[1]

[1]Department of Computer Science, University of California, Santa Barbara
[2]Department of Computer Science, University of Illinois, Chicago
[3]Department of Computer Science, Indian Institute of Technology, Delhi
[*]Equal contribution