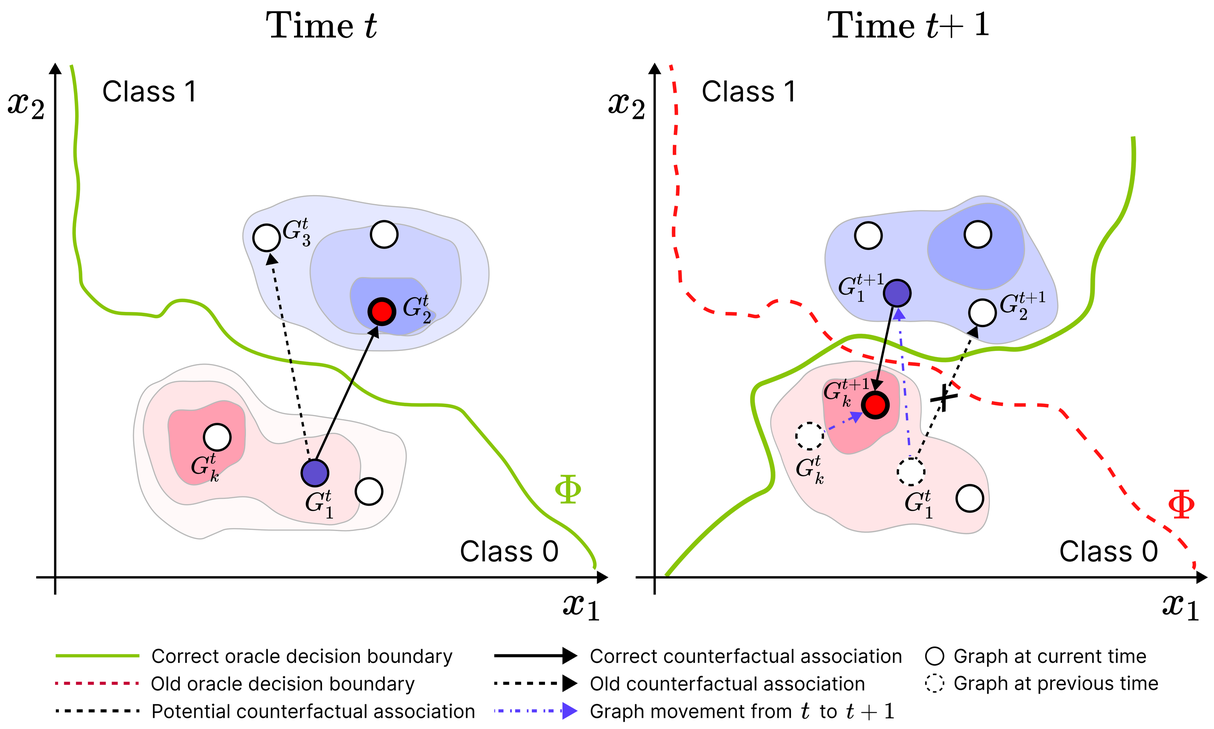
Adapting to Change: Robust Counterfactual Explanations in Dynamic Data Landscapes

Bardh Prenkaj, Mario Villaizán-Vallelado, Tobias Leemann, Gjergji Kasneci



1 What's the problem?

- Timed-Graph Counterfactual Explainability (TGCE)
- Data undergoes constant changes and distribution shifts
- Impact on robustness, relevance, and the validity of counterfactuals [1]
- No approaches to address counterfactual generation in dynamic data landscapes: at $t+\delta$ most counterfactuals generated at time t are obsolete

2 How the literature approached TGCE?

- Aligning counterfactual explanations in the presence of distributional drifts has become a relevant subject of study [2]
- This is the first approach in TGCE where data can undergo any change: node and edge insertions/deletions, and graph removals/additions
- Other time-unrelated counterfactual explainers are search- [3,4], heuristic - [5,6], and learning-based [7-9]

[1] Pawelczyk, M.; Leemann, T.; Biega, A.; and Kasneci, G.2023. On the Trade-Off between Actionable Explanations and the Right to be Forgotten. In Proceedings of the Eleventh International Conference on Learning Representations.

[2] Haug, J.; and Kasneci, G. 2021. Learning parameter distributions to detect concept drift in data streams. In 2020 25th International Conference on Pattern Recognition (ICPR), 9452–9459. IEEE

[3] Abrate, C.; and Bonchi, F. 2021. Counterfactual graphs for explainable classification of brain networks. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2495–2504

[4] Faber, L.; Moghaddam, A. K.; and Wattenhofer, R. 2020. Contrastive Graph Neural Network Explanation. In Proc.of the 37th Graph Repr. Learning and Beyond Workshop at ICML 2020, 28. Int. Conf. on Machine Learning

[5] Bajaj, M.; Chu, L.; Xue, Z. Y.; Pei, J.; Wang, L.; Lam, P. C.-H.; and Zhang, Y. 2021. Robust counterfactual explanations on graph neural networks. Advances in Neural Information Processing Systems, 34: 5644–5655

[6] Wellawatte, G. P.; Seshadri, A.; and White, A. D. 2022. Model agnostic generation of counterfactual explanations for molecules. Chemical science, 13(13): 3697–3705

[7] Ma, J.; Guo, R.; Mishra, S.; Zhang, A.; and Li, J.2022. CLEAR: Generative Counterfactual Explanations on Graphs. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho,K., eds., Advances in Neural Information Processing Systems

[8] Prado-Romero, M. A.; Prenkaj, B.; and Stilo, G. 2023. Revisiting CounteRGAN for Counterfactual Explainability of Graphs. In Maughan, K.; Liu, R.; and Burns, T. F., eds., The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR2023, Kigali, Rwanda, May 5, 2023. OpenReview.net

[9] Tan, J.; Geng, S.; Fu, Z.; Ge, Y.; Xu, S.; Li, Y.; and Zhang, Y.2022. Learning and Evaluating Graph Neural Network Explanations Based on Counterfactual and Factual Reasoning.In Proceedings of the ACM Web Conference 2022, WWW'22, 1018–1027.

[10] Kipf, T.N., Welling, M.: Variational graph auto-encoders. In: NeurIPS Workshop on Bayesian Deep Learning (2016)

3 TGCE definition

- We have a predictor (oracle) $\Phi : \mathcal{G}^t \to \{0,1\}$ s.t. $\mathcal{G}^t = \{G_1^t, \dots, G_k^t\}$
- At any $t+1$, Φ might wrongly predict G_i^t, i.e., $\Phi(G_i^t) \neq \Phi(G_i^{t+1})$
- We expect that the counterfactual for G_i^{t+1} changes w.r.t. that of G_i^t

$$\mathcal{E}_\Phi(G_i^t) = \underset{G_j^t \in \mathcal{G}^t}{\arg\max}\ P_{cf}^t\big(G_j^t \mid G_i^t, \Phi(G_i^t), \neg\Phi(G_i^t)\big)$$

4 Proposed approach: DyGRACE

- Semi-supervised approach that uses Φ at t_0 and avoids its outdated decision function for $t_i > t_0$
- We rely on two GAEs [10] $f_y, f_{\neg y} : \mathcal{G} \to \mathcal{G}$ responsible to learn each class
- At t_0, $G_i^0 \in \mathcal{G}^0$ gets directed through the GAEs via $y = \Phi(G_i^0)$
- The objective of the GAEs is to minimize the rec error $h : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$

Property: $\forall (G_i, G_j),\ y = \Phi(G_i),\ y \neq \Phi(G_j) \to h(G_j, f_y(G_j)) \geq h(G_j, f_{\neg y}(G_j))$

- Train a logistic regressor to maximize the following parametric density function

$$\underset{G_j \in \mathcal{G}}{\arg\max}\ P_{cf}\big(G_j \mid G_i, \Phi(G_i), \neg\Phi(G_i)\big)$$

$$= \underset{G_j \in \mathcal{G}}{\arg\max}\Big(\alpha \underbrace{h(G_j, f_y(G_j))}_{\text{the factual GAE should have a big rec error}} - \beta \underbrace{h(G_j, f_{\neg y}(G_j))}_{\text{the counterfactual GAE should have a smaller rec error}} + \gamma \underbrace{g(G_i, G_j)}_{\text{graph similarity function}}\Big)$$

- Continual adaptation: find k counterfactuals via the logistic regressor and use them to update the knowledge of the GAEs

5 Experiments

		Runtime (s) ↓	Correctness ↑		Sparsity ↓		GED ↓		Oracle Calls ↓
			@1	@k	@1	@k	@1	@k	
DTC	t_0	$24.73^{\pm 70.419}$	$0.40^{\pm 0.490}$	$1.00^{\pm 0.000}$	$0.65^{\pm 0.109}$	$0.72^{\pm 0.114}$	$36.90^{\pm 5.839}$	$40.04^{\pm 6.272}$	$9000.00^{\pm 0.00}$
	t_1	$73.35^{\pm 3.622}$	$0.60^{\pm 0.490}$	$1.00^{\pm 0.000}$	$0.69^{\pm 0.087}$	$0.73^{\pm 0.085}$	$38.60^{\pm 4.652}$	$40.46^{\pm 4.725}$	$0.00^{\pm 0.00}$
	t_2	$74.62^{\pm 6.225}$	$0.70^{\pm 0.458}$	$1.00^{\pm 0.000}$	$0.73^{\pm 0.038}$	$0.75^{\pm 0.079}$	$40.50^{\pm 2.291}$	$41.68^{\pm 4.366}$	$0.00^{\pm 0.00}$
	t_3	$65.99^{\pm 8.409}$	$0.40^{\pm 0.490}$	$1.00^{\pm 0.000}$	$0.76^{\pm 0.065}$	$0.74^{\pm 0.090}$	$42.50^{\pm 3.775}$	$41.16^{\pm 4.983}$	$0.00^{\pm 0.00}$
DBLP	t_0	$12.83^{\pm 32.599}$	$0.38^{\pm 0.484}$	$1.00^{\pm 0.000}$	$0.67^{\pm 0.186}$	$0.71^{\pm 0.194}$	$28.50^{\pm 8.617}$	$30.04^{\pm 9.499}$	$9900.00^{\pm 0.00}$
	t_1	$140.21^{\pm 53.821}$	$0.25^{\pm 0.433}$	$0.88^{\pm 0.331}$	$0.54^{\pm 0.315}$	$0.62^{\pm 0.238}$	$35.25^{\pm 47.872}$	$40.24^{\pm 48.264}$	$0.000^{\pm 0.00}$
	t_2	$121.29^{\pm 57.151}$	$0.12^{\pm 0.331}$	$0.88^{\pm 0.331}$	$1.52^{\pm 2.709}$	$3.59^{\pm 6.178}$	$18.88^{\pm 15.070}$	$34.82^{\pm 33.389}$	$0.000^{\pm 0.00}$
	t_3	$150.73^{\pm 20.020}$	$0.38^{\pm 0.484}$	$0.88^{\pm 0.331}$	$0.85^{\pm 0.489}$	$0.81^{\pm 0.384}$	$63.12^{\pm 51.033}$	$57.86^{\pm 44.450}$	$0.000^{\pm 0.00}$
	t_4	$161.89^{\pm 10.609}$	$0.25^{\pm 0.433}$	$0.62^{\pm 0.484}$	$0.63^{\pm 0.261}$	$0.58^{\pm 0.263}$	$21.62^{\pm 15.337}$	$20.51^{\pm 15.186}$	$0.000^{\pm 0.00}$
	t_5	$162.77^{\pm 13.782}$	$0.50^{\pm 0.500}$	$1.00^{\pm 0.000}$	$1.01^{\pm 1.388}$	$0.75^{\pm 0.835}$	$21.12^{\pm 10.361}$	$21.71^{\pm 12.553}$	$0.000^{\pm 0.00}$
	t_6	$165.39^{\pm 15.160}$	$0.50^{\pm 0.500}$	$1.00^{\pm 0.000}$	$2.26^{\pm 3.125}$	$2.21^{\pm 4.238}$	$55.00^{\pm 39.528}$	$53.36^{\pm 42.852}$	$0.000^{\pm 0.00}$
	t_7	$81.22^{\pm 82.797}$	$0.12^{\pm 0.331}$	$0.50^{\pm 0.500}$	$1.44^{\pm 3.244}$	$3.13^{\pm 5.679}$	$11.62^{\pm 19.118}$	$23.80^{\pm 32.100}$	$0.000^{\pm 0.00}$
	t_8	$76.85^{\pm 75.487}$	$0.12^{\pm 0.331}$	$0.50^{\pm 0.500}$	$0.45^{\pm 0.535}$	$0.84^{\pm 0.671}$	$17.88^{\pm 23.861}$	$33.32^{\pm 24.627}$	$0.000^{\pm 0.00}$
	t_9	$66.97^{\pm 67.959}$	$0.25^{\pm 0.433}$	$0.50^{\pm 0.500}$	$0.44^{\pm 0.450}$	$0.67^{\pm 0.257}$	$33.75^{\pm 34.387}$	$51.85^{\pm 20.576}$	$0.000^{\pm 0.00}$
	t_{10}	$63.07^{\pm 65.097}$	$0.25^{\pm 0.433}$	$0.38^{\pm 0.484}$	$0.90^{\pm 1.310}$	$0.97^{\pm 0.685}$	$57.62^{\pm 69.611}$	$77.09^{\pm 62.334}$	$0.000^{\pm 0.00}$

Average of DyGRACE's performance on DynTree-Cycles (DTC) and DBLP-Coauthors (DBLP) on 10-fold cross-validation. k = 10