Unifying Evolution, Explanation, and Discernment: A Generative Approach for Dynamic Graph Counterfactuals

Oral paper @ KDD'24

Bardh Prenkaj
Chair of Responsible Data Science
Technical University of Munich
bardh.prenkaj@tum.de

October 17, 2024

30th SIGKDD Conference on Knowledge Discovery and Data Mining - Research Track

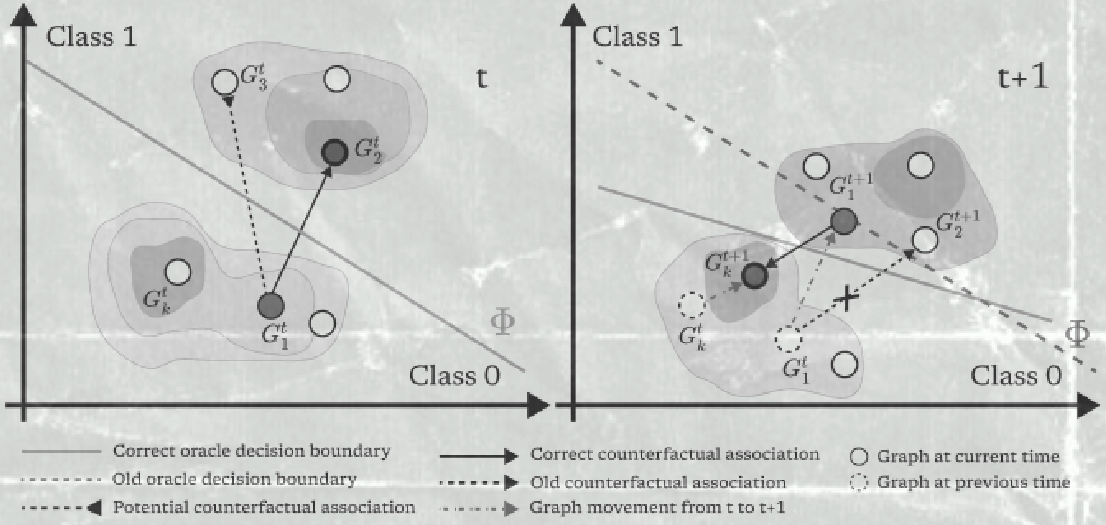
Knowledge Discovery & Data Mining

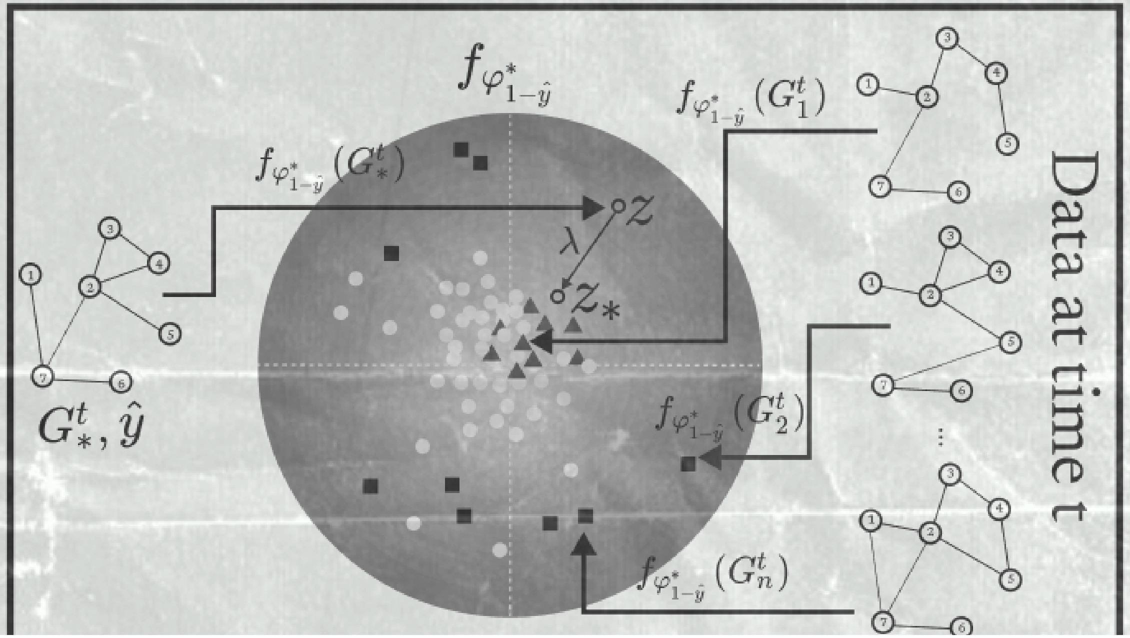
August 25 - 29, 2024
International Barcelona Convention Center
Barcelona, Catalonia

Vol. 30 No. 902 | NEWS FOR TODAY | *FULL PAPER*

UNIFYING EVOLUTION, EXPLANATION, AND DISCERNMENT

A GENERATIVE APPROACH FOR DYNAMIC GRAPH COUNTERFACTUALS





What happens when counterfactuals get

Today's Roadmap

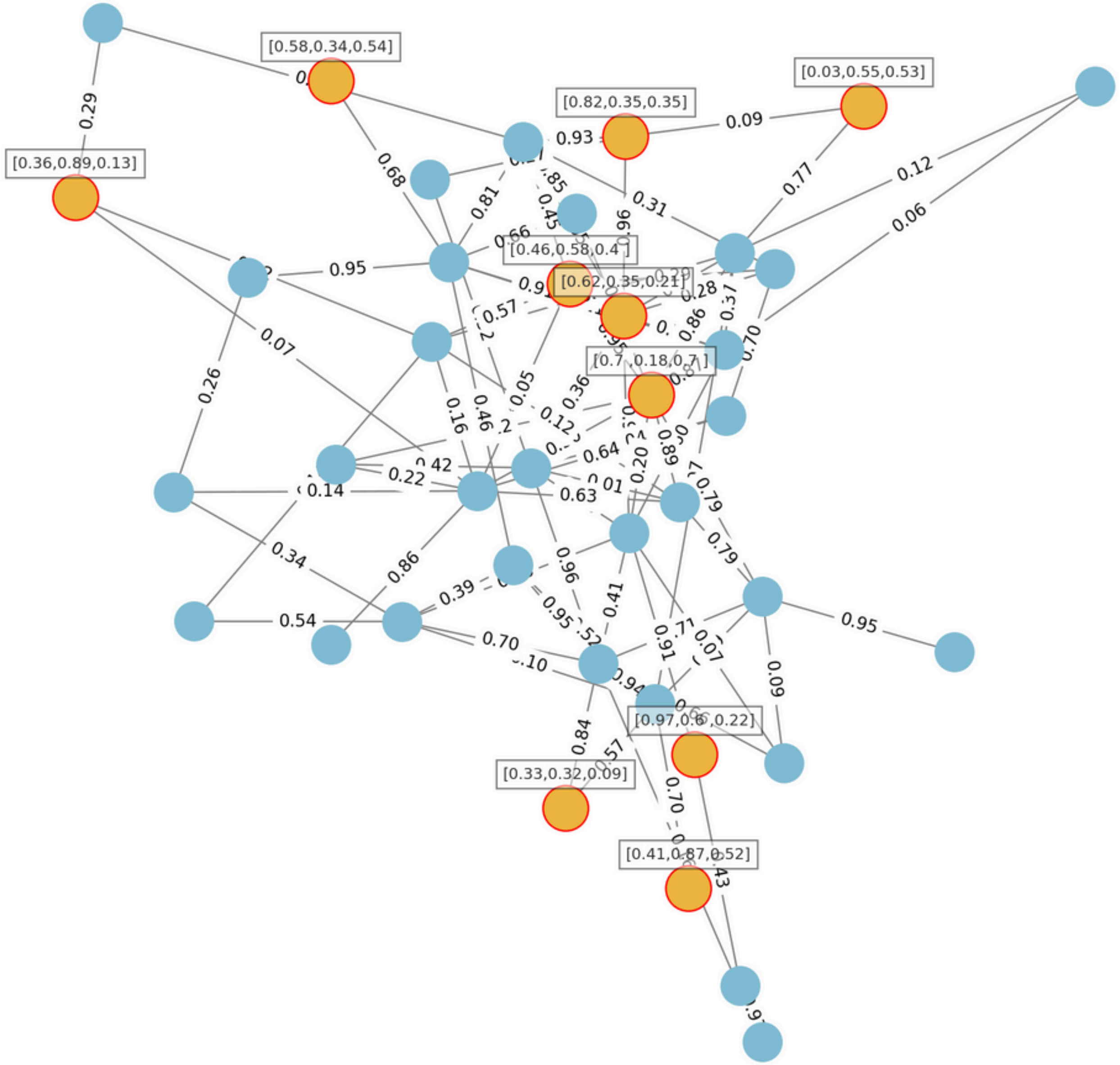
- Introduction
  - Good ol' graphs
  - What are counterfactuals?
  - "*The right to be forgotten*" - Pawelczyk et al.
- Pictorial Problem Statement
- Problem Formulation
- Generative Classification (GC) Perspective
  - Bridging Reconstruction and GC

Today's Roadmap (cont.)

- Fighting out of the **blue corner:** GRACIE!
  - Training
  - Inference and Finding Latent Counterfactuals
  - Dynamic Update
- Experiments
  - Synthetic vs. Real-world Datasets
  - Pulling Factor Trade-Off
  - Qualitative Inspection

Oh yeah... almost forgot about the **Conclusions** 🫠

Good ol' graphs

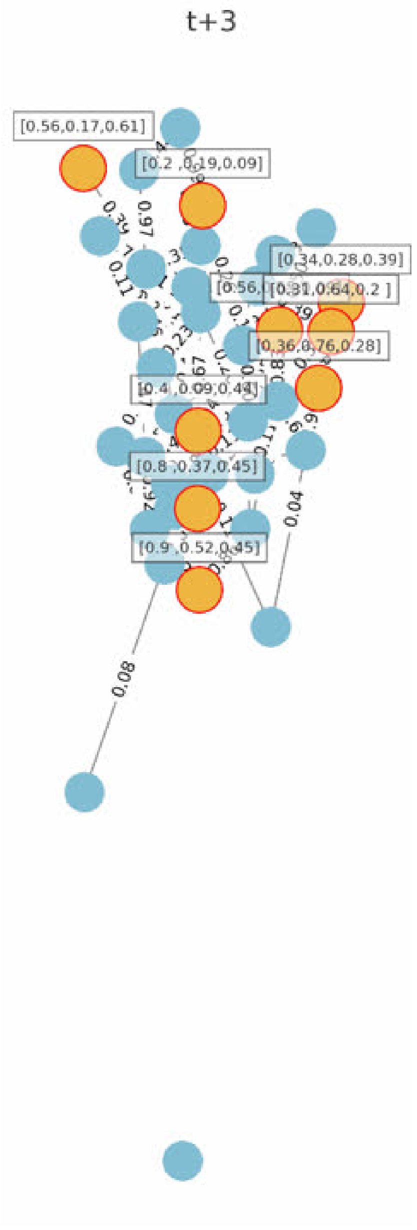


$$G_i = (\mathbf{X}, \mathbf{A}) \in \mathcal{G}$$

$$\mathbf{X} \in \mathbb{R}^{n \times d}$$

$$\mathbf{A} \in \mathbb{R}^{n \times n}$$

Good ol' graphs (contd.)



$$G_i = \{G_i^{t_0}, \ldots, G_i^{t_j}, \ldots, G_i^{t_m}\}$$

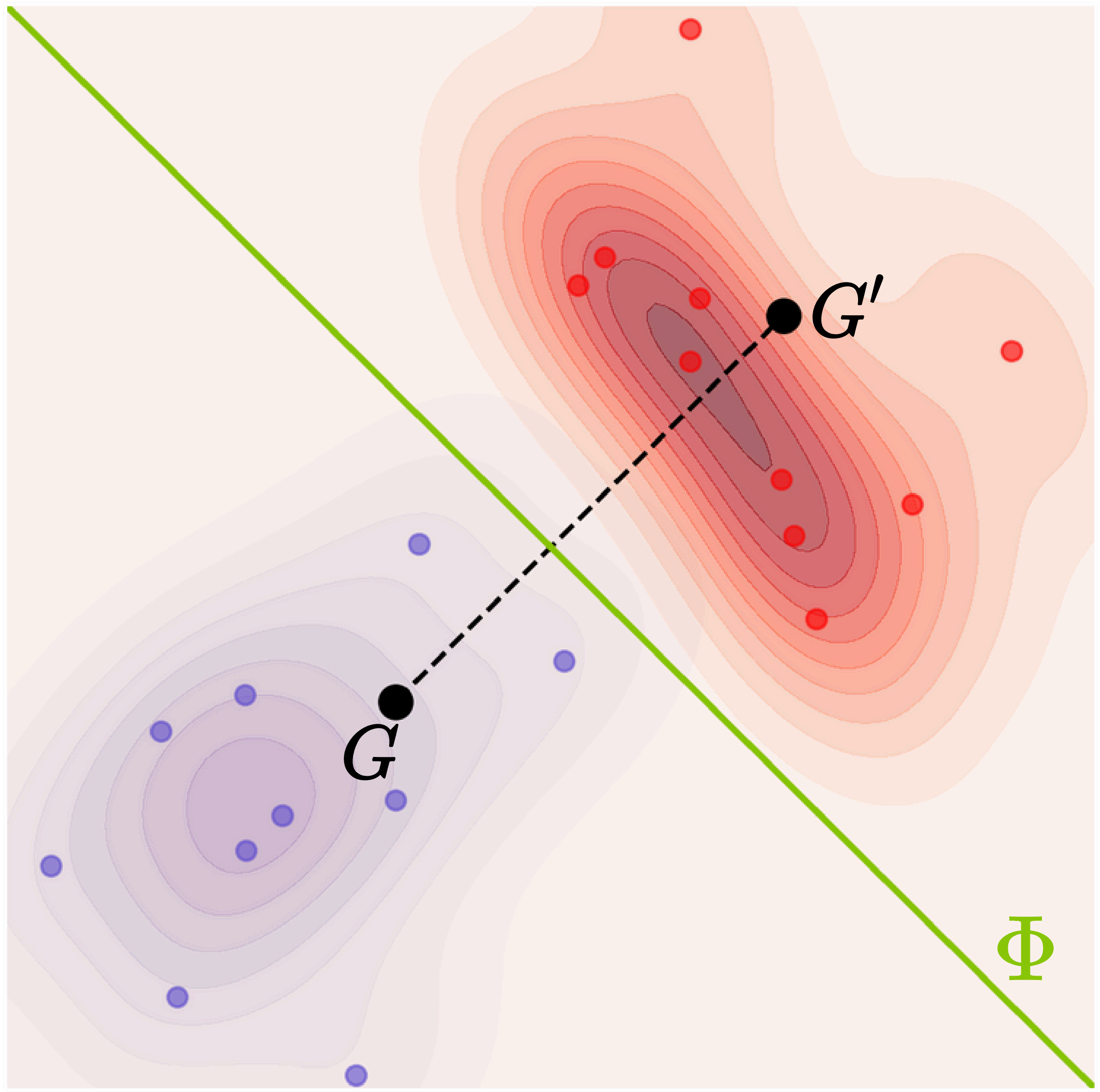
base graph structure

graph mutation in time

Possible modifications in time:

- Node additions/removal
- Edge additions/removal

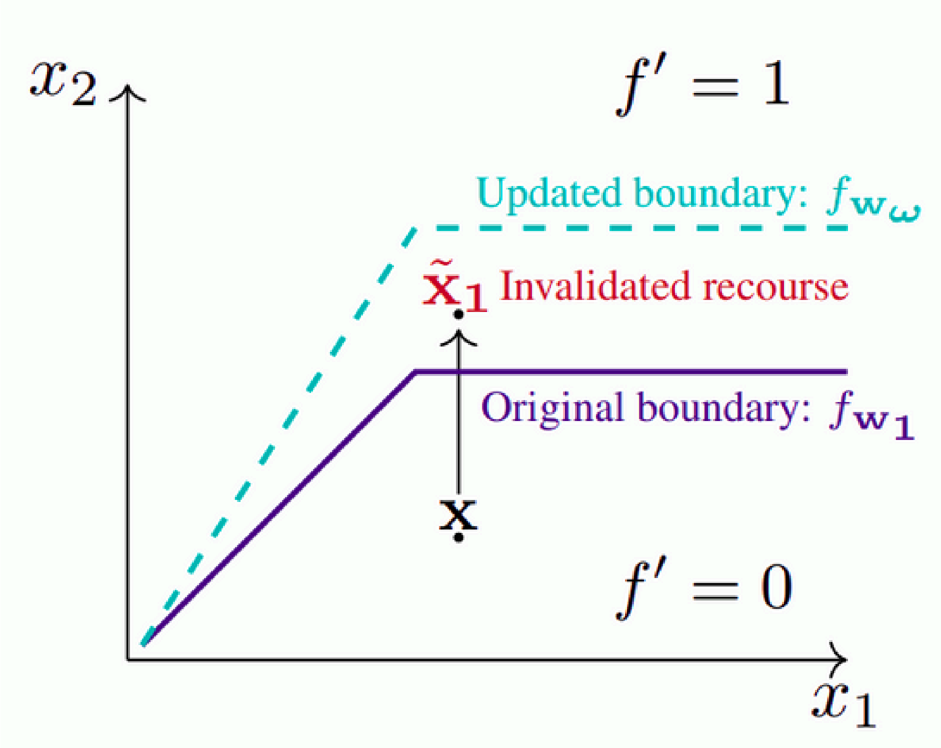
What are Counterfactuals?



$$\Phi(G) \neq \Phi(G')$$

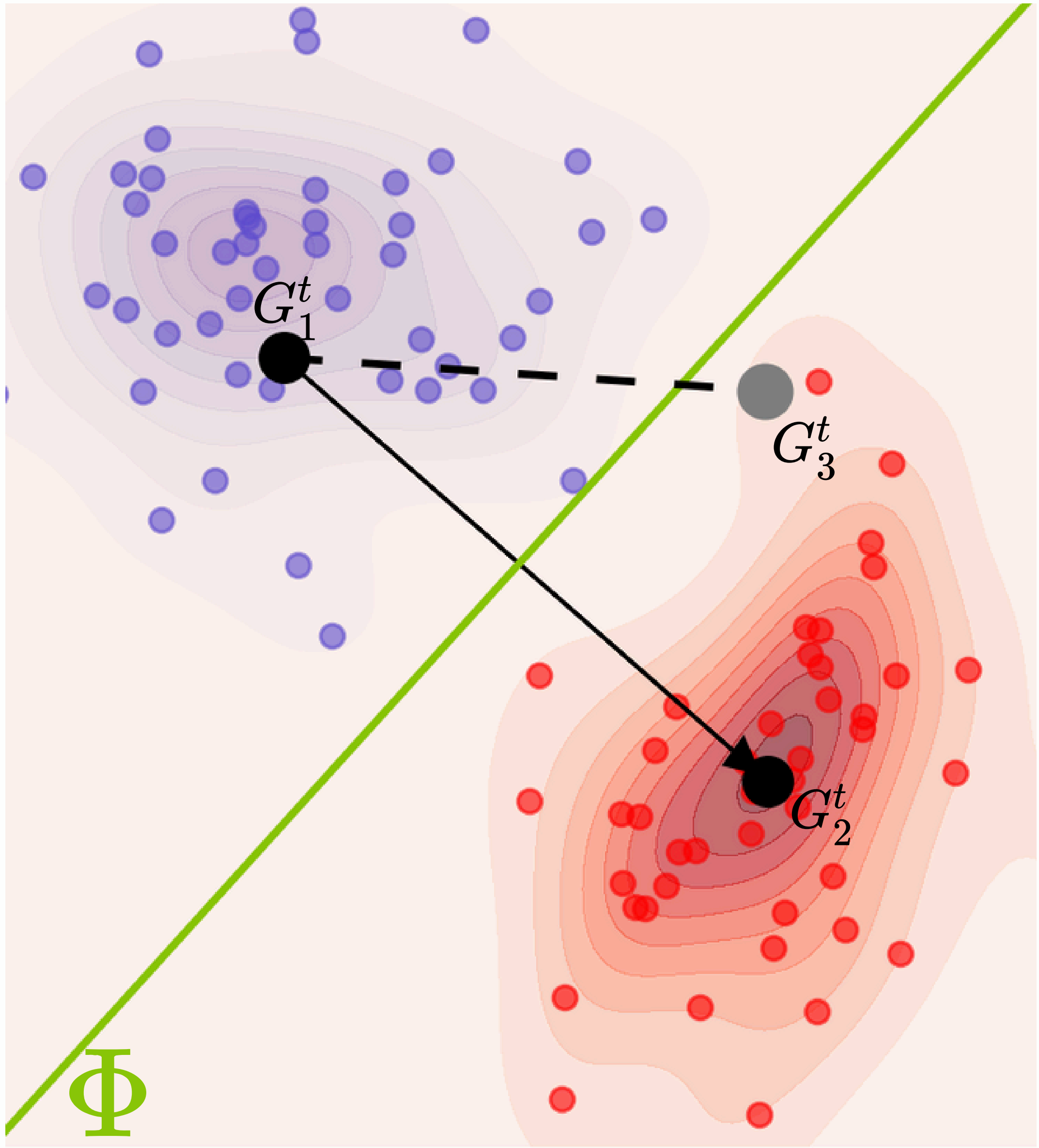
“The right to be forgotten”

- Counterfactuals can become invalidated when data is deleted
- Pawleczyk et al. identify data points that, when deleted at **t + δ**, invalidate the counterfactuals at time **t**

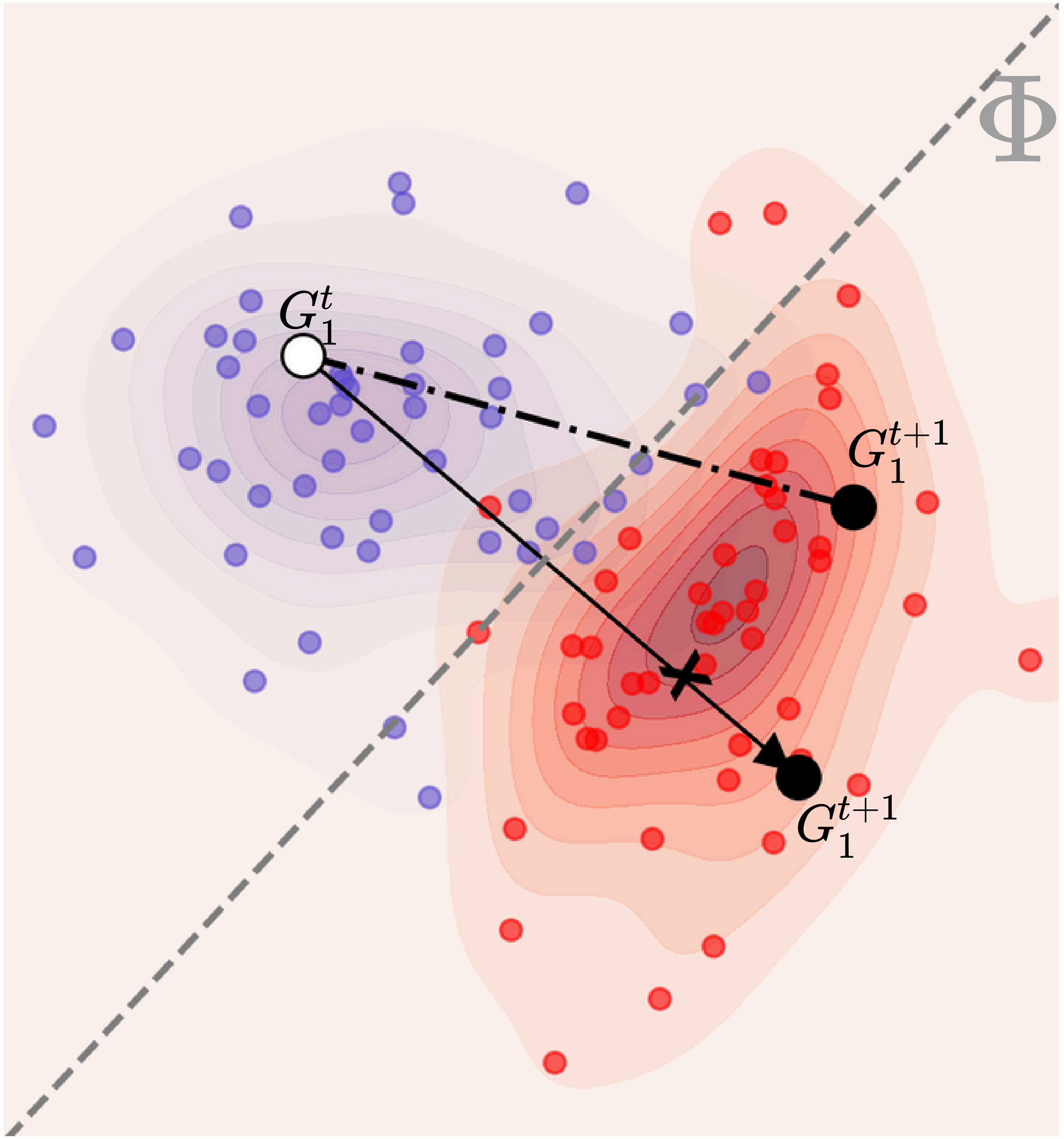


Martin Pawelczyk, Tobias Leemann, Asia Biega, and Gjergji Kasneci. 2023. On the Trade-Off between Actionable Explanations and the Right to be Forgotten. In Proc. of the 11th International Conference on Learning Representations

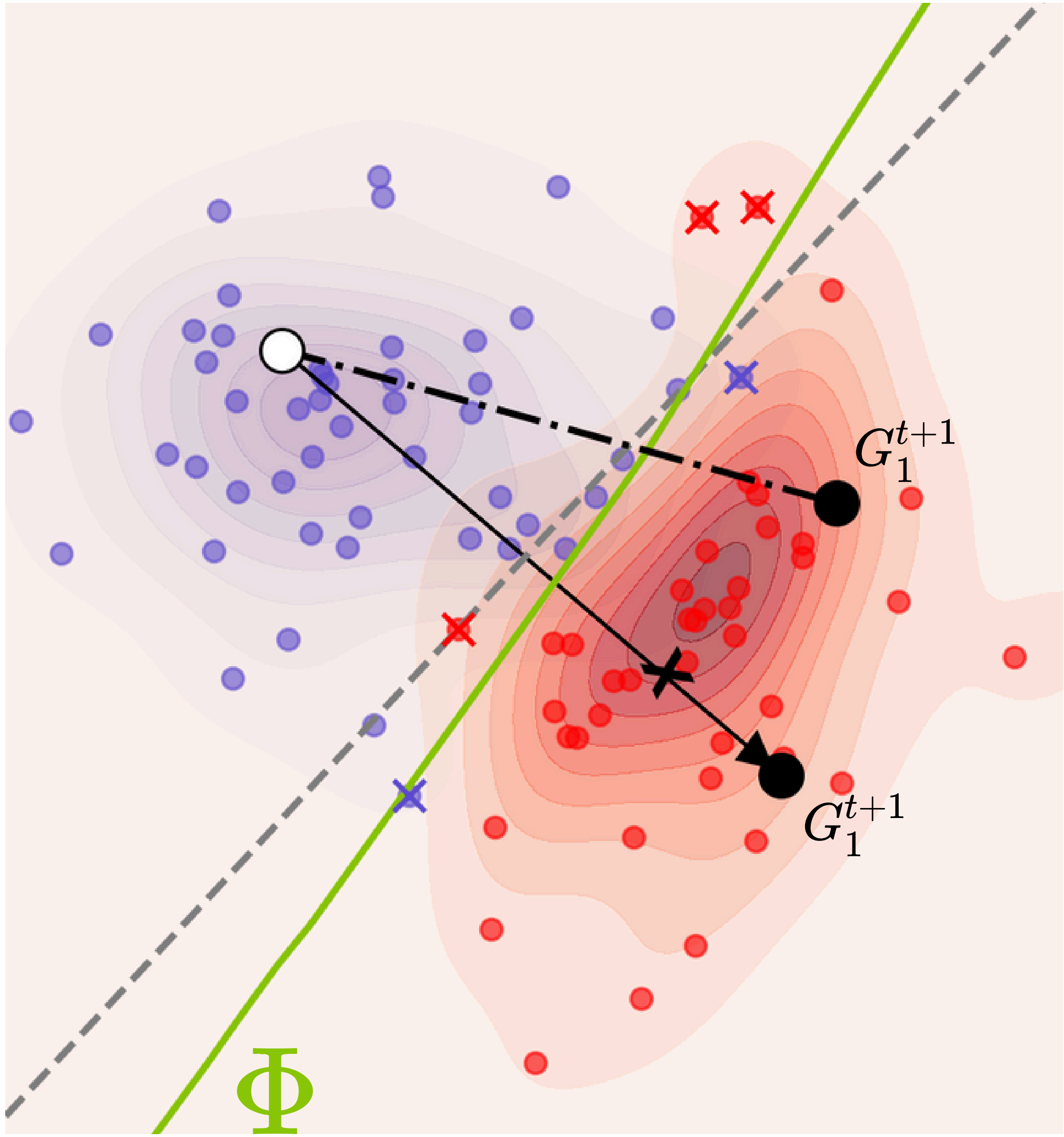
Pictorial Problem Statement



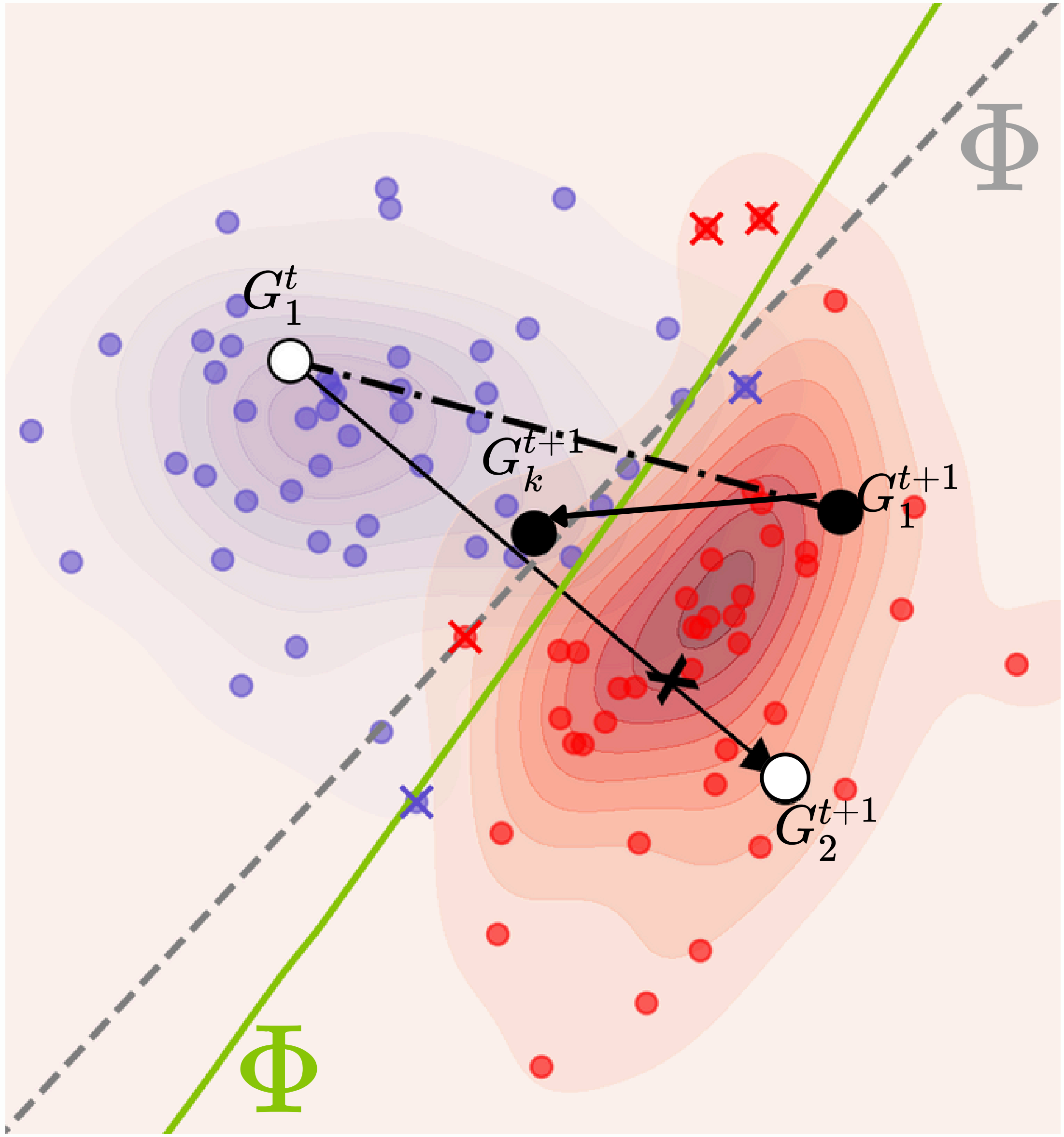
t



$t+1$



$t+1$



$t+1$

Problem Formulation

$$\mathcal{E}_{\Phi}\left(G_i^t\right) = \arg\max_{G_j^t \in \mathcal{G}} P_{cf}^t\left(G_j^t \mid G_i^t, \Phi\left(G_i^t\right), \neg\Phi\left(G_i^t\right)\right)$$

probability of G_j^t being in-distribution counterfactual of G_i^t

Any other class that isn't $\Phi\left(G_i^t\right)$

Differently from previous work, we shift towards a generative classification (GC) perspective

Bardh Prenkaj, Mario Villaizan-Vallelado, Tobias Leemann, and Gjergji Kasneci. 2023. Adapting to Change: Robust Counterfactual Explanations in Dynamic Data Landscapes. arXiv:2308.02353 [cs.LG]

Generative Classification (GC) Perspective

- Generative Classifiers (GCs) perform classification by modeling the full joint distirbution of features x and class labels y

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p(x, y) = \arg\max_{y \in \mathcal{Y}} p(x|y)\, p(y) = \\ = \arg\max_{y \in \mathcal{Y}} \log p(x|y) + \log p(y) .$$

Andrew Ng and Michael Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems 14 (2001)

Generative Classification (GC) Perspective

- Superior generalization capabilities over discriminative classifiers
- Accurately calibrated posteriors
- Increased adversarial robustness

Ilkay Ulusoy and Christopher M Bishop. 2006. Comparison of generative and discriminative techniques for object detection and classification. In Toward Category-Level Object Recognition. Springer, 173–195

Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. 2020. Training normalizing flows with the information bottleneck for competitive generative classification. Advances in Neural Information Processing Systems 33 (2020), 7828–7840

Yingzhen Li, John Bradshaw, and Yash Sharma. 2019. Are generative classifiers more robust to adversarial attacks?. In International Conference on Machine Learning. PMLR, 3804–3814.

Variational Graph Autoencoders (VGAEs)

- We consider the following generative model where the graphs G are generated from factored latent representation $\mathbf{z}$ and the true class label y

$$p\left(G|y\right) = \int_{\mathbf{z}\in\mathcal{Z}} p\left(G|\mathbf{z}, y\right) p\left(\mathbf{z}|y\right) dz$$

VGAEs (Decoder)

- To represent $p(G|y)$, we use a single VGAE for each class , which is dependent on the class where each node has a latent vector and then define

$$p_{\theta_y}(G|\mathbf{z}, y) = p_{\theta_y}(\mathbf{A}, \mathbf{X}|\mathbf{z}, y)$$
$$= p_{\theta_y}(\mathbf{X}|\mathbf{A}, \mathbf{z}, y)\, p_{\theta_y}(\mathbf{A}|\mathbf{z}, y)$$

VGAEs (Encoder)

$$q_{\varphi_y}\left(\mathbf{z}|G,y\right)=\prod_{v_i} q_{\varphi_y}\left(\mathbf{z}_{v_i}|G,y\right)$$

$$q\left(z_{v_i}|G,y\right)=\mathcal{N}\left(z_{v_i}|\mu_{v_i},\gamma^2\mathbf{I}\right),$$

$$\mu=\left[\mu_{v_1},\ldots,\mu_{v_n}\right]=\mathrm{GCN}_{\varphi_y}\left(G\right)$$

> 0 and fixed hyperparameter

Bridging Reconstruction and GC

- We train the VGAEs for each of the classes by optimizing the parameters θ and φ

$$\mathrm{ELBO}_y\left(\theta_y, \varphi_y\right) = \underset{q_{\varphi_y}(z|G,y)}{\mathbb{E}}\left[\log p_{\theta_y}\left(G|z, y\right)\right] - \mathbf{KL}\left[q_{\varphi_y}\left(z|G, y\right) \| \, p\left(z\right)\right]$$

$$\left(\theta_y^*, \varphi_y^*\right) = \underset{\theta_y, \varphi_y}{\arg\max}\, \mathrm{ELBO}_y\left(\theta_y, \varphi_y\right) \;\; \forall y \in \mathcal{Y}$$

Bridging Reconstruction and GC

- Having obtained a generative latent variable model of a specific class, we can now exploit its power to perform generative classification
- If the variational family is expressive enough, the ELBO converges to the logarithm of the true class-conditional probability
- Use the generative models to compare different class probabilities and perform generative classification

Bridging Reconstruction and GC

Proposition 1: Comparing Distance-Augmented Reconstruction Losses performs Implicit GC

If the density model is sufficiently expressive, i.e., it covers the true data generating process, computing

$$\hat{y} = \arg\min_{y \in \mathcal{Y}} \frac{1}{2} \left(\mathbb{E}_{q_{\varphi_y^*}(z|G,y)} \left[\frac{\|g_{\theta_y^*}(z) - G\|_2^2}{\sigma^2} \right] + \|f_{\varphi_y^*}(G)\|_2^2 \right] - \log p(y),$$

is equivalent to performing generative classification for an input graph.

Bridging Reconstruction and GC

Proposition 1: Comparing Distance-Augmented Reconstruction Losses performs Implicit GC

If the density model is sufficiently expressive, i.e., it covers the true data generating process, computing

$$\hat{y} = \arg\min_{y\in\mathcal{Y}} \frac{1}{2}\left(\mathbb{E}_{q_{\varphi_y^*}(z|G,y)}\left[\frac{\|g_{\theta_y^*}(z) - G\|_2^2}{\sigma^2}\right] + \|f_{\varphi_y^*}(G)\|_2^2\right) - \log p(y),$$

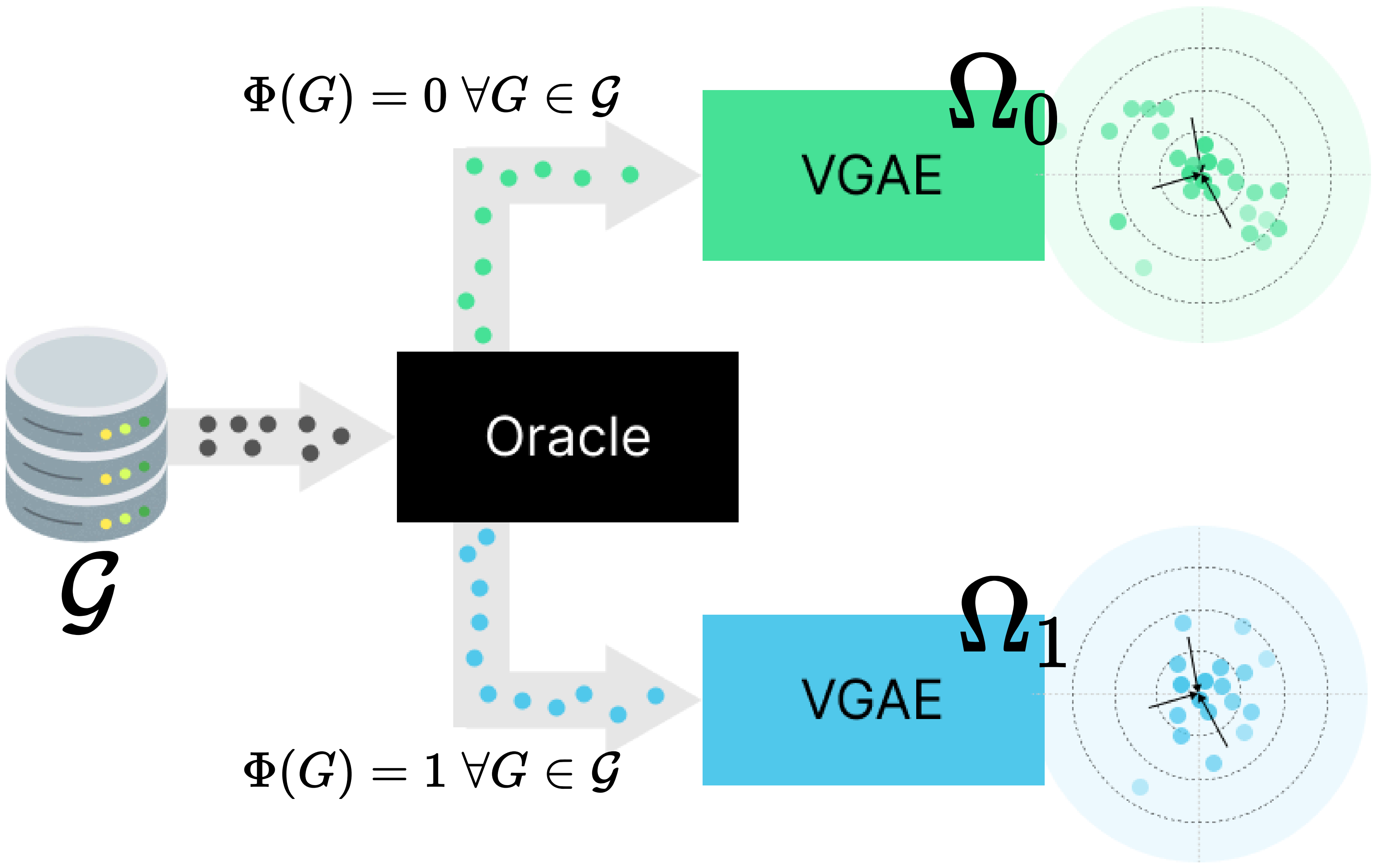
is equivalent to performing generative classification for an input graph.

$g_{\theta_y^*}(z)$ → decoder

$f_{\varphi_y^*}$ → encoder

GRACIE

Class Representation Experts

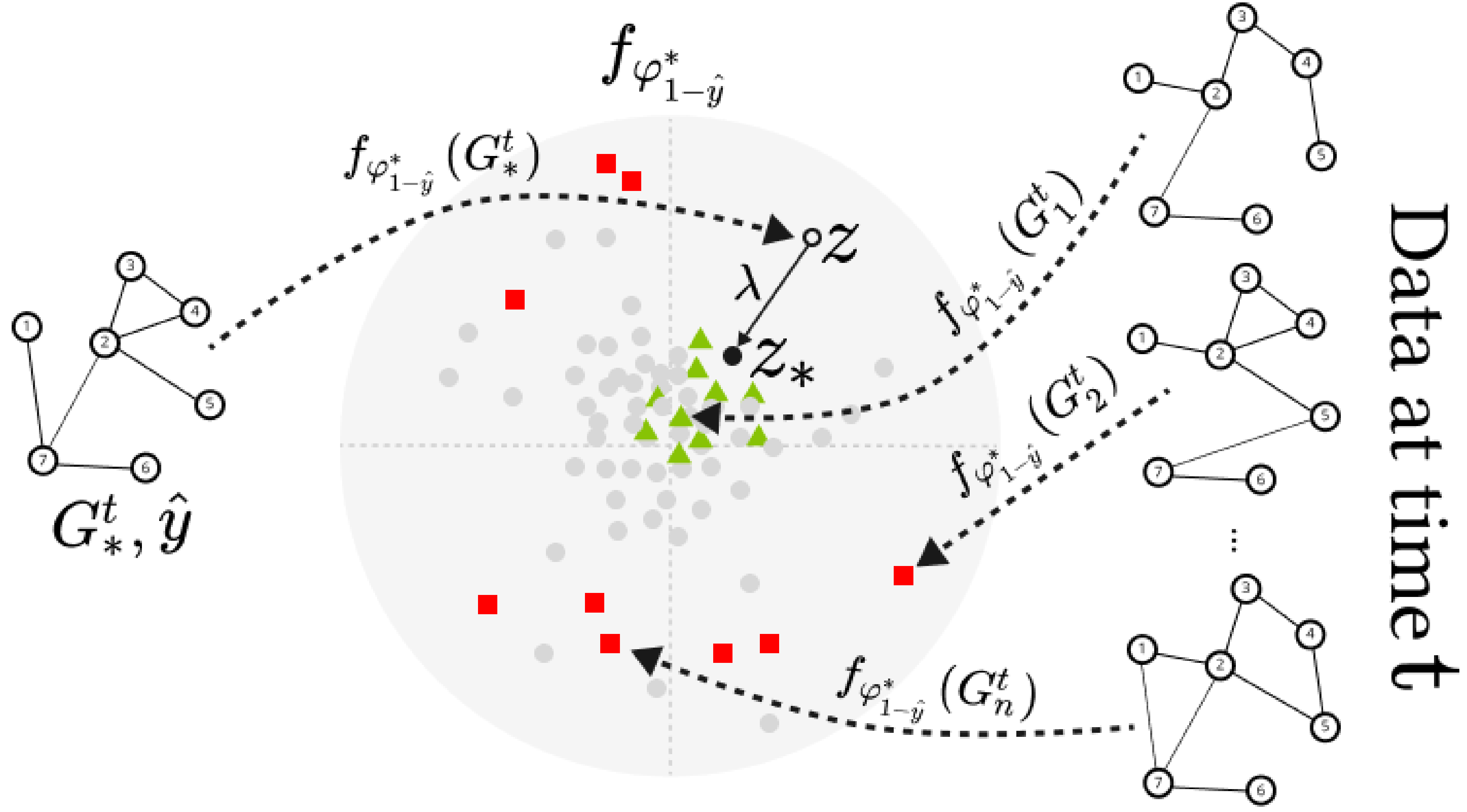


Training

$$-\mathrm{ELBO}_y\left(\theta_y, \varphi_y\right) = \mathcal{L}_{rec} + \mathcal{L}_{dist}$$

$$= \frac{1}{2}\left(\underset{q_{\varphi_y}(\mathbf{z}|G)}{\mathbb{E}}\left[\frac{||g_{\theta_y}(\mathbf{z}) - G||_2^2}{\sigma^2}\right] + ||f_{\varphi_y}(G)||_2^2\right)$$

Inference and Finding Latent Counterfactuals



Dynamic Update

- Use the learned representation of the VGAEs
- For each graph, find k candidate counterfactuals close to the center of the VGAE responsible to learn the counterfactual class
- We can use these counterfactuals to update the counterfactual VGAE and the graph itself to update the factual VGAE
- **GRACIE is semi-supervised in the first snapshot, and completely unsupervised in the next snapshots**

Experiments

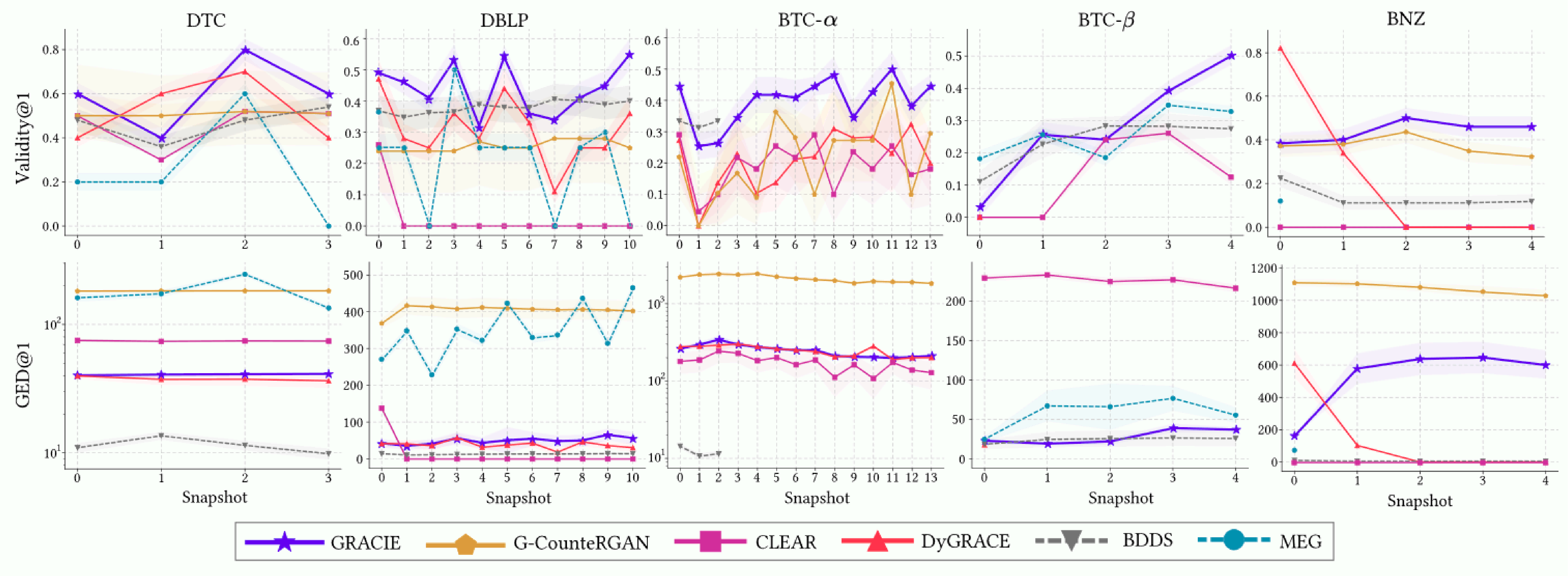
Synthetic vs. Real-world Datasets

	DTC	DBLP	BTC-α	BTC-β	BNZ
BDDS	0.465	<u>0.381</u>	<u>0.360</u>†	0.235	0.136
MEG	0.250	0.209	×	<u>0.260</u>	0.120†
CLEAR	0.458	0.024	0.214	0.125	0.000
G-CounteRGAN	0.507	0.256	0.236	×	<u>0.404</u>
DyGRACE	<u>0.525</u>	0.307	0.232	0.000†	0.232
GRACIE	**0.600**	**0.442**	**0.440**	**0.284**	**0.441**

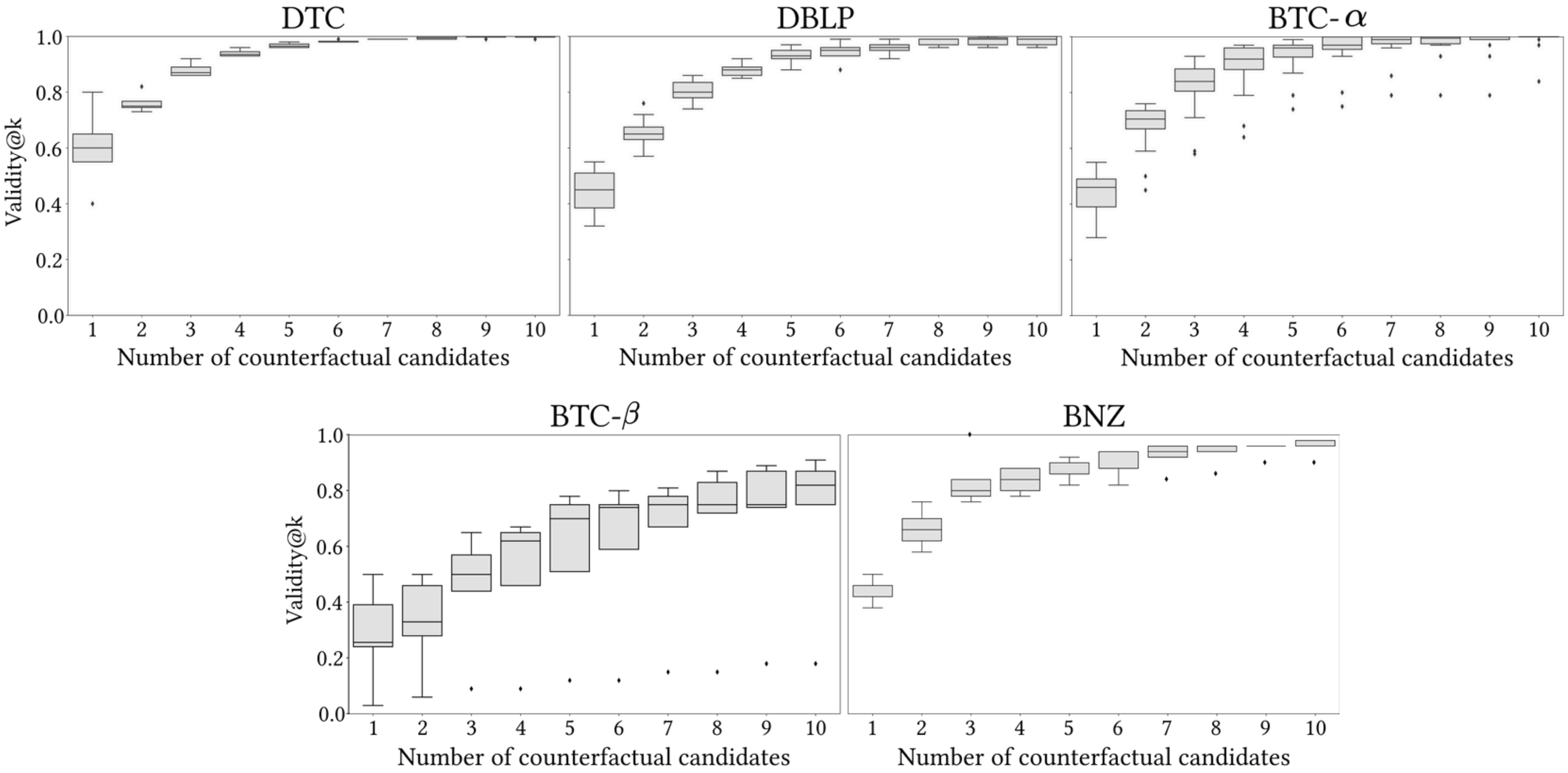
[a]The criterion of non-convergence is to fail to produce at least one counterfactual within 14 days of execution on an HPC SGE Cluster of 6 nodes with 360 cumulative cores, 1.2Tb of RAM, and two GPUs (i.e., one Nvidia A30 and one A100).

	GRACIE	
	w/o Bonferroni (*p-value* .05)	w/ Bonferroni (*p-value* .01)
BDDS	2.472×10^{-6}	3.708×10^{-5}
MEG	1.784×10^{-15}	2.676×10^{-14}
G-CounteRGAN	1.090×10^{-5}	1.635×10^{-4}
CLEAR	9.354×10^{-13}	1.403×10^{-11}
DyGRACE	2.014×10^{-6}	3.021×10^{-5}

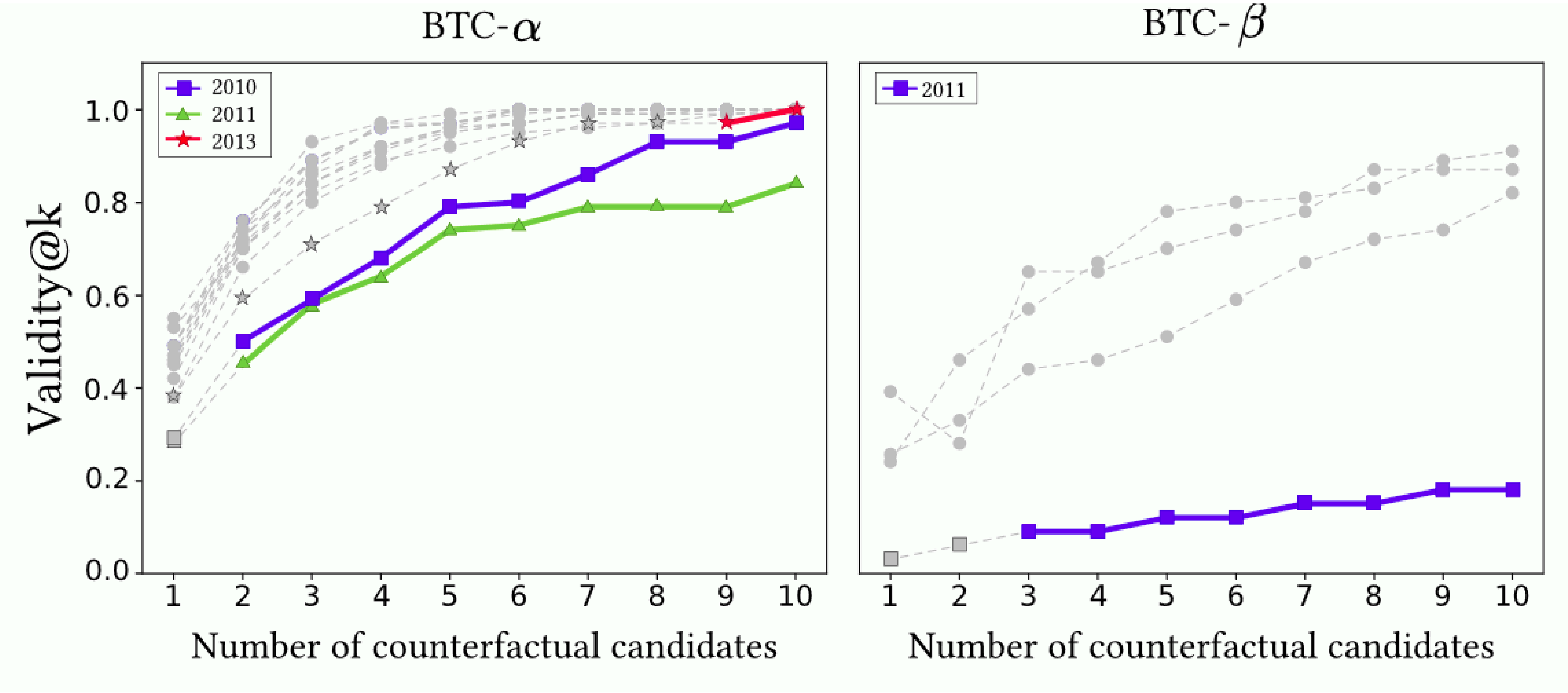
Synthetic vs. Real-world Datasets



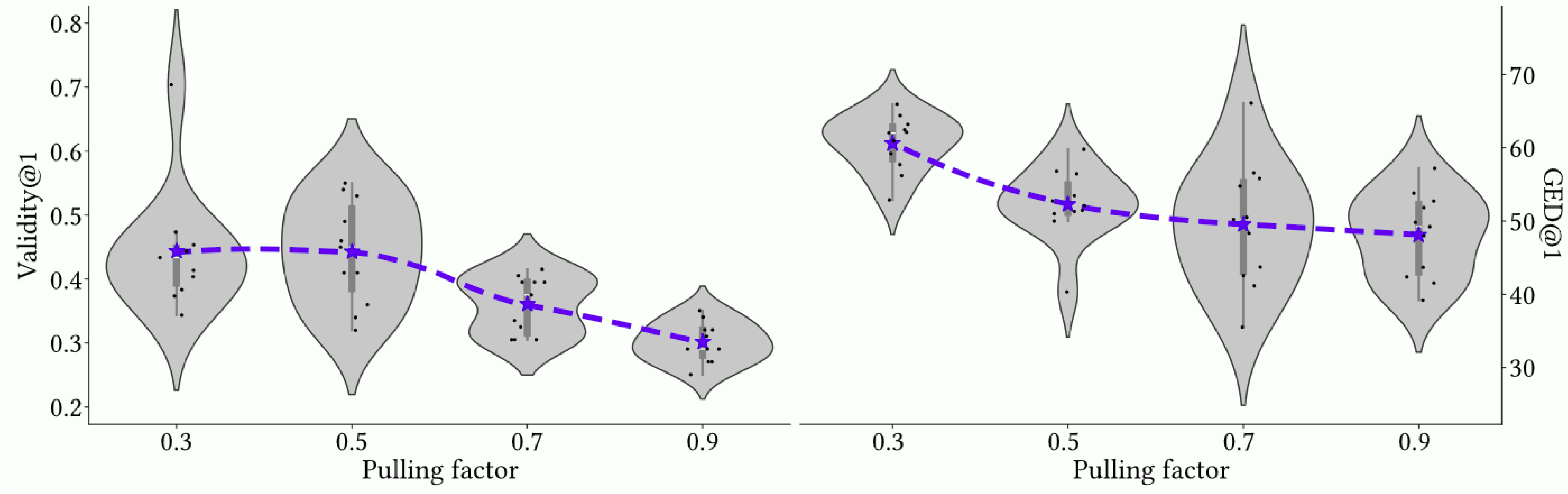
More sampling = more validity



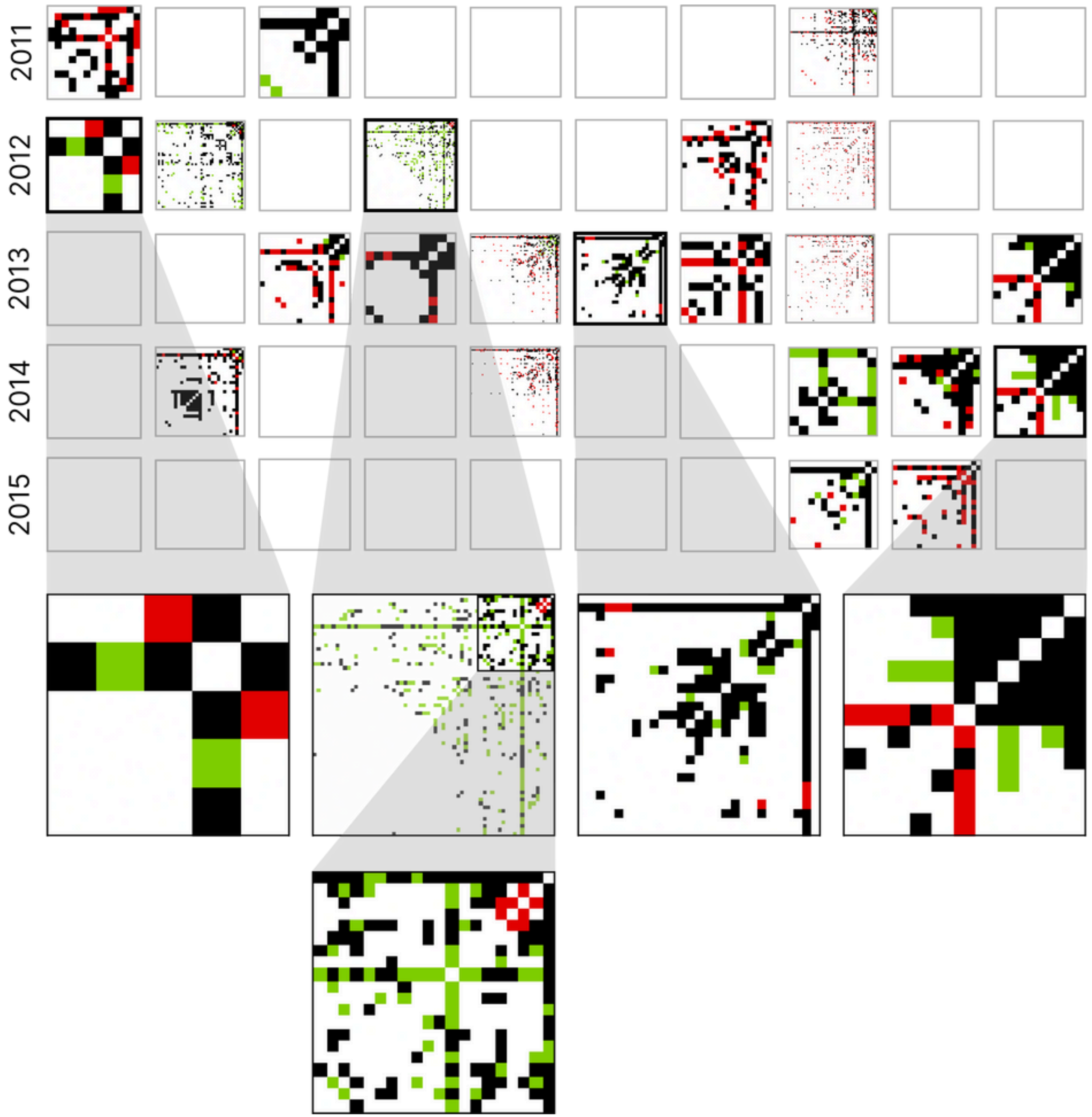
A closer look on sampling

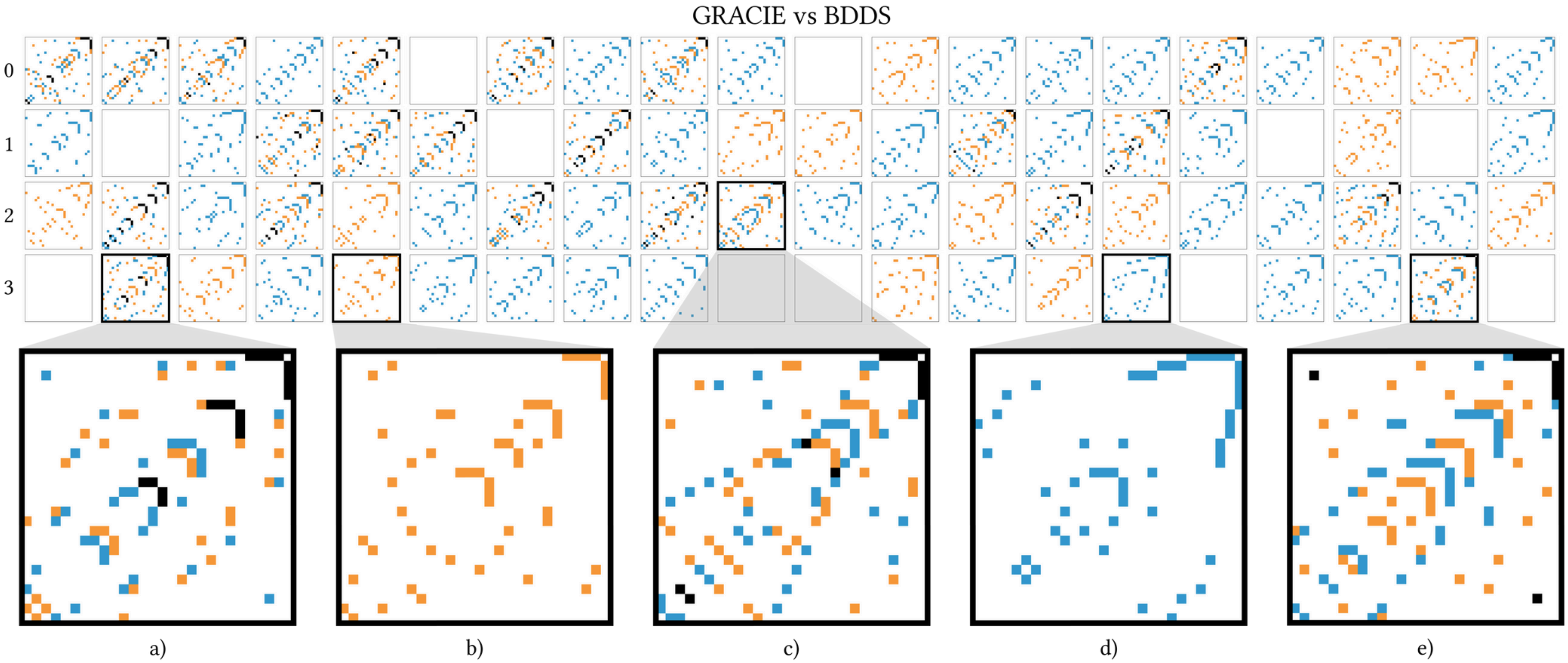


Effect of pulling factor



Qualitative on BTC-β




GRACIE vs BDDS
0
1
2
3
a)
b)
c)
d)
e)

Conclusions

- GRACIE is one of the first generative approaches to address dynamic counterfactual explainability in the context of temporal graphs
- We leverage VGAEs, self-supervisedly, to learn class representations and adapt to data distribution shifts
- Improvement of 13.1% in producing valid counterfactuals than SoTA
- **The center of the latent space of the VGAEs should be used to find valid counterfactual**

Thank you!