Action Sequence Augmentation for Early Graph-based Anomaly Detection

Tong Zhao†, Bo Ni†, Wenhao Yu†, Zhichun Guo†, Neil Shah‡, and Meng Jiang†

† Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

‡ Snap Research, Snap Inc., Santa Monica, CA 90405, USA

CIKM 2021 1-5 NOVEMBER

UNIVERSITY OF NOTRE DAME | COLLEGE OF ENGINEERING

Background on Graph Anomaly Detection

- Graph anomaly detection:
  - Identify anomalous nodes in the graph.
- Many anomaly detection applications are better solved with graph anomaly detection approaches. E.g.,
  - Bad buyer detection
  - Bot account detection

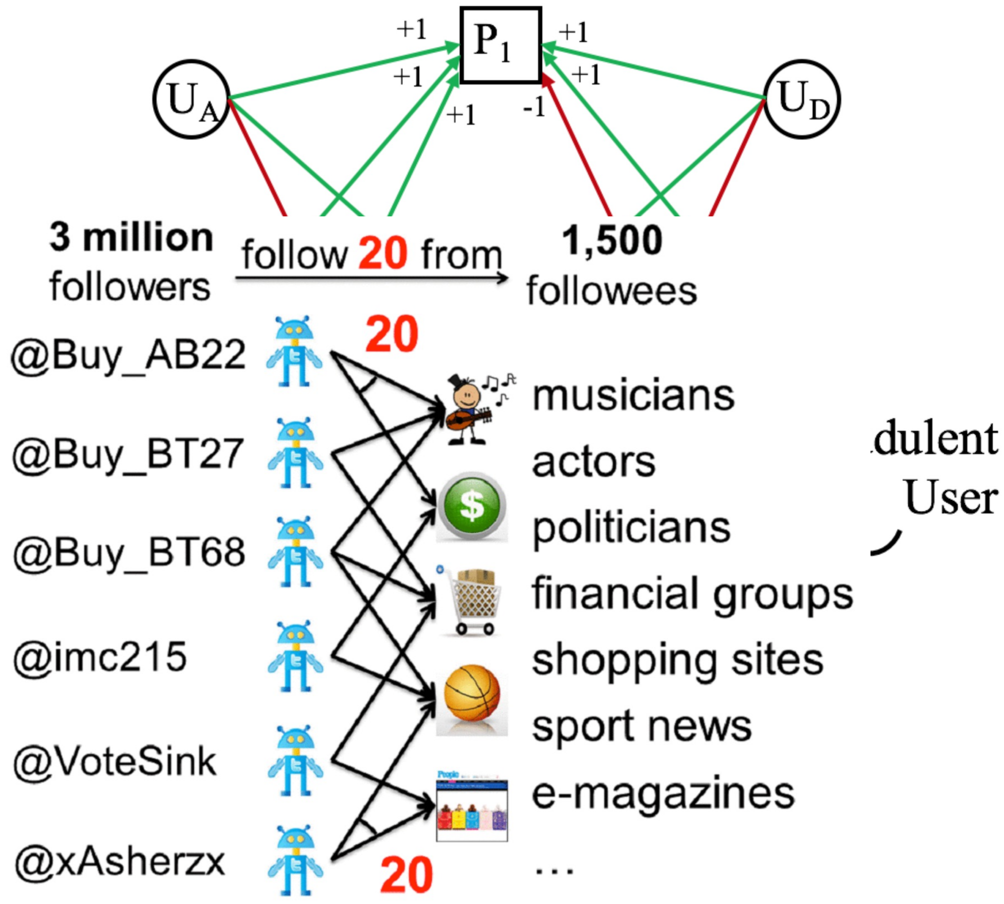
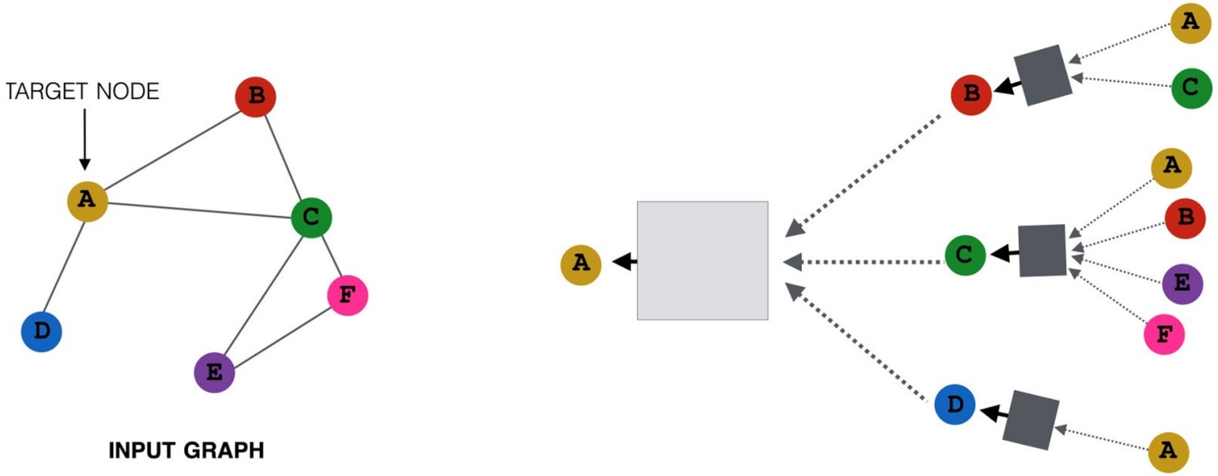


Image sources:
Kumar, et al. "REV2: Fraudulent User Prediction in Rating Platforms." WSDM'18.
Jiang, et al. "CatchSync : Catching Synchronized Behavior in Large Directed Graphs." KDD'14.

Background on Graph Neural Networks

- Given: graph $G = (V, E)$, node features $\mathbf{x}_v \in \mathbb{R}^m, \forall v \in V.$
- Learn: low dimensional node representations $\mathbf{z}_v \in \mathbb{R}^d, \forall v \in V.$

- Neighborhood aggregation: generate node representations based on local neighborhoods.



Layer output embedding

Layer weight matrix

$$h_{v_i}^{l+1} = \sigma\Big(\sum_j \frac{1}{c_{ij}} h_{v_j}^{l} W^{l}\Big)$$

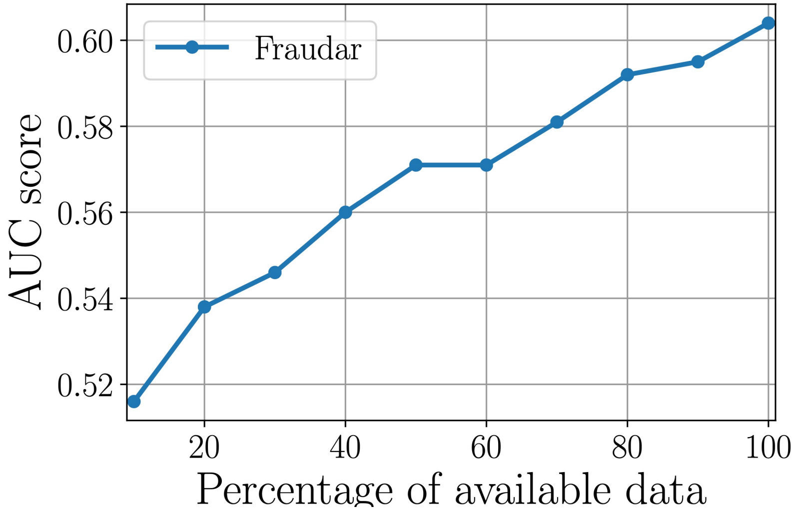
Non-linearity

Aggregated neighbor embeddings normalized

Image sources: Ying, Rex, et al. "Graph convolutional neural networks for web-scale recommender systems." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

Early Graph Anomaly Detection

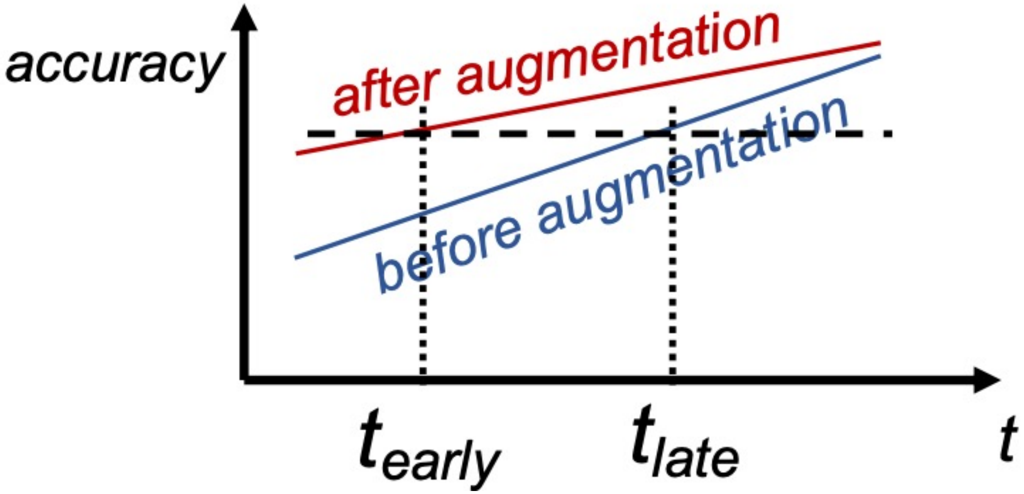
- The performances of existing graph anomaly detection methods might not be satisfactory when observations are limited.
- At the time when existing methods detects the anomalies, they may have already achieved their goals.



Research Problem

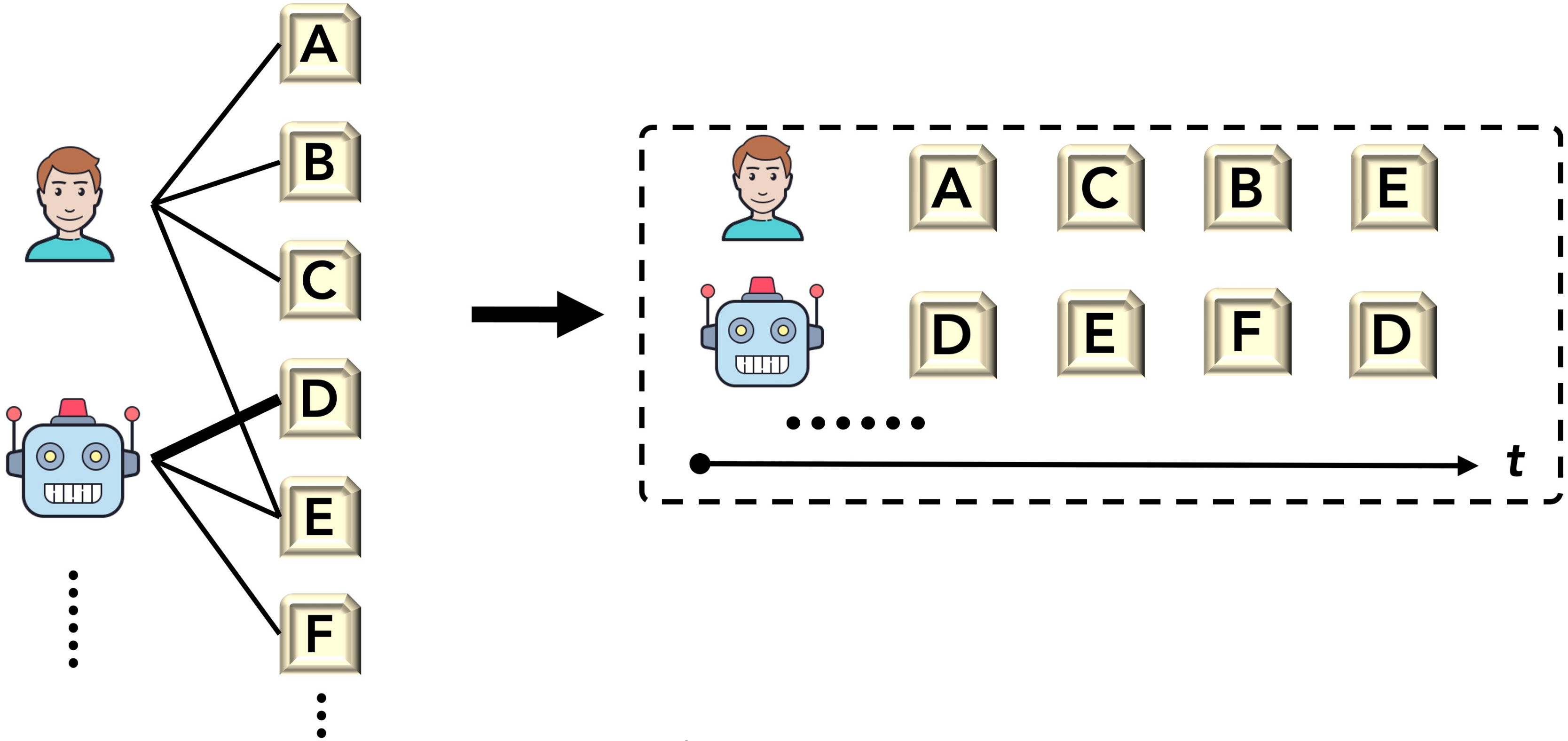
- Can we detect the anomalies before they achieve their goals?

- **Given:** a user-item bipartite graph at an early-time t_{early}.
- **Design:** a framework that can help any anomaly detection methods to achieve a comparable performance at time t_{early}.

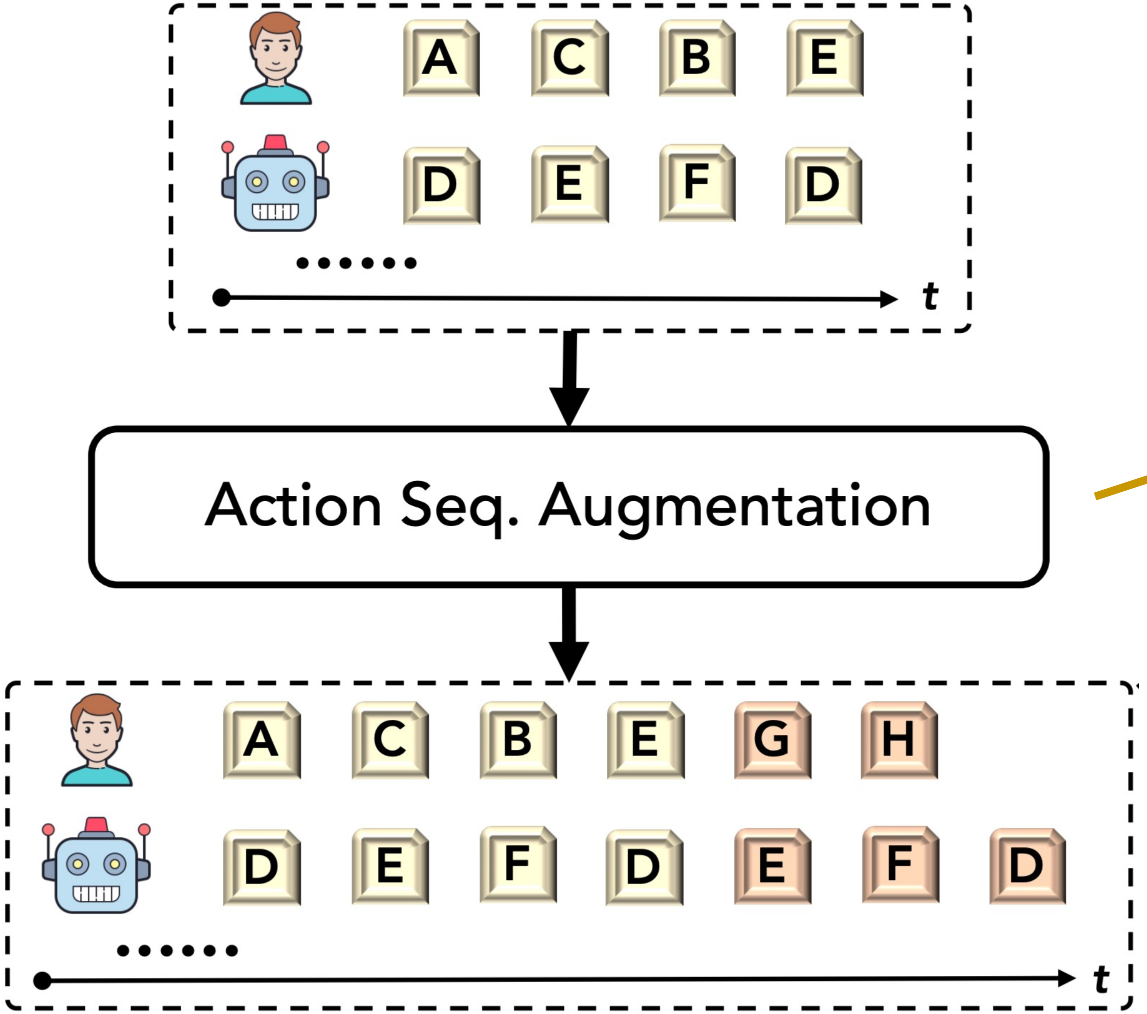


Action Sequence Augmentation

- Idea: Predict the future user actions to augment the data.
  - "Forecast the future"

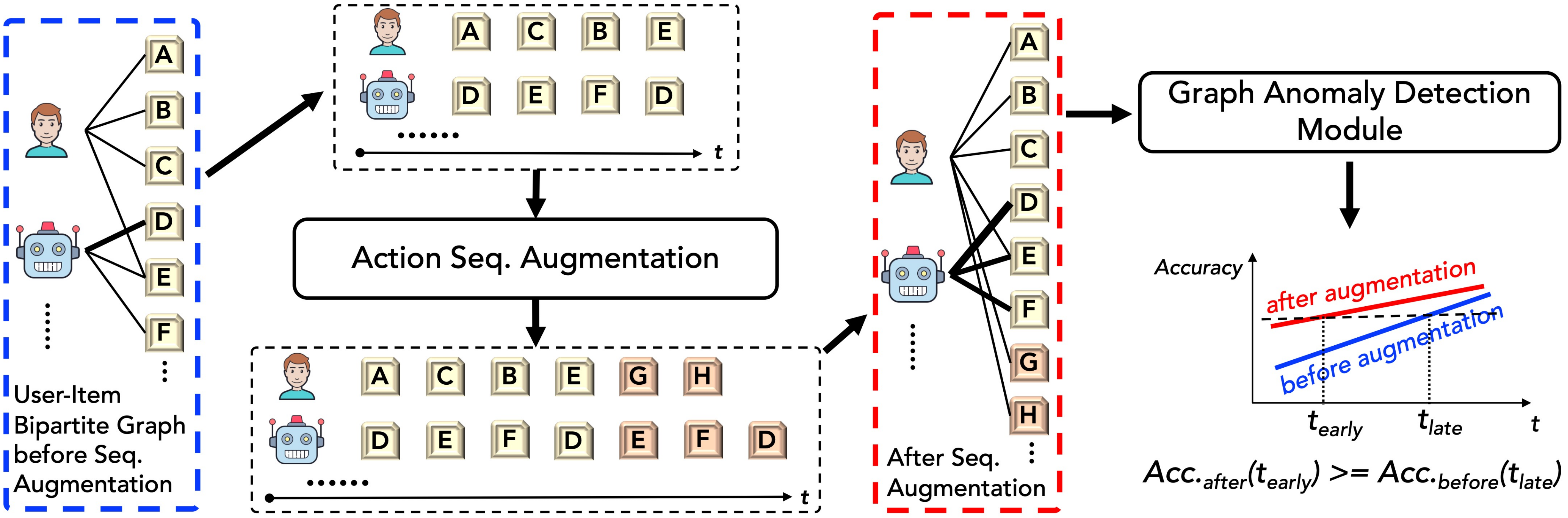


Action Sequence Augmentation



Any Seq2Seq model.
We use GRU for simplicity.

Proposed Framework: Eland



Eland-itr: Bootstrapping iterative training strategy.

- Iteratively train both modules independently.
- Both modules benefit from the output of the other module.
  - Anomaly detection module makes better prediction with the enriched data.
  - Augmentation module use the predicted suspiciousness $\hat{y}$ to determine the number of predicted actions.

Algorithm 1: ELAND-ITR

Input : Adjacency matrix $\mathbf{A}$; node feature matrix $\mathbf{X}$; number of iterations I; anomaly detection module $g \in \{g_{ad}, g_{ad-gnn}\}$; action sequence augmentation module f_{aug}.

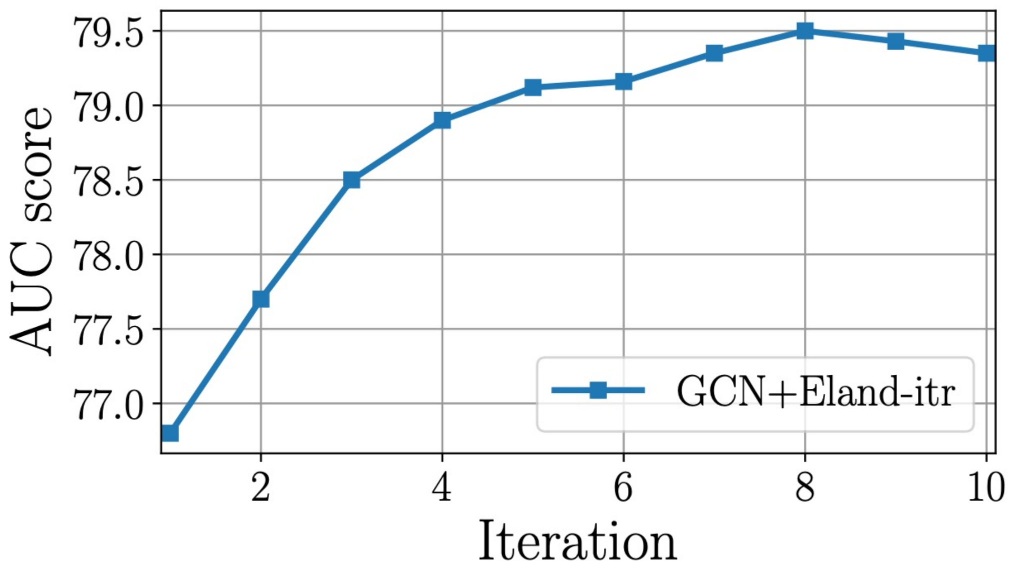
Output : User prediction results $\hat{\mathbf{y}}$.

$\mathbf{X}_{orig} = \mathbf{X}$;

if g *is* g_{ad-gnn} **then**

$\quad \hat{\mathbf{y}}, \mathbf{Z} = g(\mathbf{A}, \mathbf{X})$; `// Defined in Eq.(4)`

$\quad \mathbf{X} = \text{concat}(\mathbf{X}_{orig}, \mathbf{Z})$;



$\qquad \hat{\mathbf{y}} = g(\mathbf{A}', \mathbf{X})$;

$\quad$ **end**

end

return $\hat{\mathbf{y}}$;

Eland-e2e: End-to-End training of the whole framework.

- Enables end-to-end training.
- Avoids potential error propagation issue brought by bootstrapping.
- Anomaly detection module benefits from the graph with augmented actions.
- Augmentation module benefits from anomaly detection module's decisions.
- Train jointly with losses for both modules for stability.

Algorithm 2: ELAND-E2E

Input : Adjacency matrix $\mathbf{A}$; node feature matrix $\mathbf{X}$; action sequence augmentation module $f_{aug-e2e}$; GNN-based anomaly detection module g_{ad-gnn}; number of training epochs n_epochs.

Output : User prediction results $\hat{\mathbf{y}}$.

`/* model training */`

Initialize $\Theta_{aug-e2e}$ in $f_{aug-e2e}$ and Θ_{ad-gnn} in g_{ad-gnn} ;

for *epoch in range(n_epochs)* **do**

  $\mathbf{A}' = f_{aug-e2e}(\mathbf{A}, \mathbf{X})$; `// Defined in Eq.(13)`

  $\hat{\mathbf{y}}, \mathbf{Z} = g_{ad-gnn}(\mathbf{A}', \mathbf{X})$; `// Defined in Eq.(4)`

  Calculate $\mathcal{L}_{e2e}$ with Eq.(16);

  Update $\Theta_{aug-e2e}$ and Θ_{ad-gnn} with $\mathcal{L}_{e2e}$;

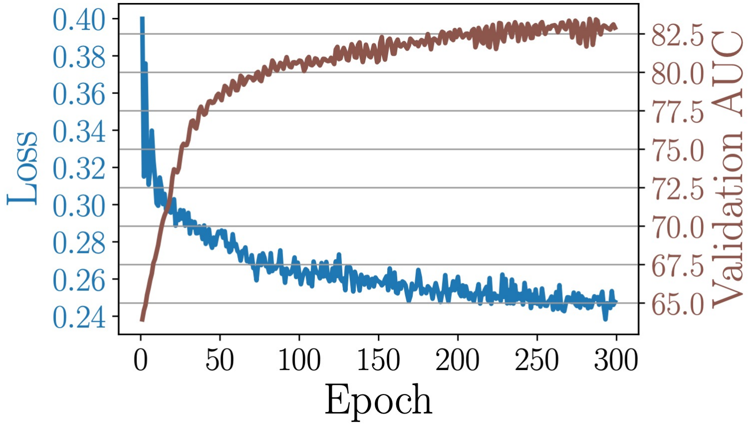
end

`/* model inferencing */`

$\mathbf{A}' = f_{aug-e2e}(\mathbf{A}, \mathbf{X})$;

$\hat{\mathbf{y}}, \mathbf{Z} = g_{ad-gnn}(\mathbf{A}', \mathbf{X})$;

return $\hat{\mathbf{y}}$;



Consistent Gains with Augmentation

Anomaly detection module g_{ad}	Method	Weibo AUC	Weibo AP	Amazon Reviews AUC	Amazon Reviews AP	Reddit AUC	Reddit AP
	RNNFD [3]	54.52±0.12	17.44±0.10	60.22±0.29	28.61±0.11	66.08±0.36	26.45±0.83
	GRAND [15]	82.58±2.11	40.12±2.99	81.71±2.56	57.66±2.98	79.09±0.18	42.37±0.72
GCN [21]	Original	81.78±0.78	41.21±1.36	80.28±0.09	57.73±0.21	78.01±0.71	41.21±0.69
	+JODIE [24]	67.80±1.30	17.12±2.72	–	–	73.12±2.13	31.62±3.98
	+GAUG [53]	82.04±0.40	48.17±0.59	81.91±0.02	60.12±0.15	78.78±0.07	40.74±0.72
	+ELAND-ITR	82.76±0.71	48.51±1.06	80.85±0.67	58.14±0.39	78.94±0.83	43.11±1.22
	+ELAND-E2E	**84.14**±0.50	**54.15**±0.83	**85.54**±0.46	**65.48**±0.14	**79.58**±0.38	**44.60**±0.43
GRAPHSAGE [18]	Original	81.87±0.56	45.26±2.54	78.67±0.09	58.00±0.07	81.06±0.02	47.71±0.01
	+JODIE [24]	69.44±0.95	16.01±2.09	–	–	74.66±0.09	34.70±0.06
	+GAUG [53]	82.10±0.46	47.81±1.29	80.79±0.02	56.38±0.03	81.37±0.01	43.83±0.01
	+ELAND-ITR	82.34±0.50	48.40±0.91	**81.59**±0.23	**59.91**±0.12	**81.62**±0.10	**48.25**±0.11
	+ELAND-E2E	**83.41**±0.37	**50.61**±0.93	79.92±0.19	58.21±0.31	79.83±0.02	44.38±0.02
HETGNN [48]	Original	81.33±0.43	39.66±1.48	86.24±0.13	67.98±0.25	91.51±0.13	67.51±0.17
	+JODIE [24]	68.99±0.44	17.38±1.87	–	–	92.02±0.36	68.16±0.30
	+GAUG [53]	82.09±0.21	47.05±0.51	87.26±0.12	71.76±0.33	91.99±0.02	66.30±0.25
	+ELAND-ITR	81.46±0.57	40.20±1.19	**90.58**±0.86	**75.08**±0.57	**92.44**±0.07	**69.31**±0.29
	+ELAND-E2E	**84.09**±0.55	**54.07**±1.64	87.57±0.26	68.46±0.35	84.24±0.22	55.34±0.88

Anomaly detection module g_{ad}	Method	Weibo AUC	Weibo AP	Reddit AUC	Reddit AP
DOMINANT [7] (Unsupervised)	Original	56.77±1.96	13.73±1.22	61.23±0.35	18.30±0.21
	+JODIE [24]	58.18±0.77	11.09±0.13	61.64±0.09	18.81±0.06
	+GAUG [53]	61.22±1.86	14.15±2.38	62.26±2.70	17.09±1.39
	+ELAND-ITR	**65.44**±1.78	19.42±1.29	**62.96**±0.10	18.90±0.04
	+ELAND-E2E	63.91±0.92	**21.90**±0.87	61.73±0.27	**18.91**±0.14
DEEPAE [55] (Unsupervised)	Original	56.10±2.01	12.65±1.31	61.94±0.39	18.29±0.13
	+JODIE [24]	57.74±0.87	11.16±0.73	61.57±0.32	18.93±0.06
	+GAUG [53]	61.18±2.03	11.58±1.27	61.29±0.82	18.23±0.53
	+ELAND-ITR	**63.34**±0.82	15.88±0.73	**62.87**±0.37	**19.02**±0.11
	+ELAND-E2E	62.80±3.60	**16.99**±3.87	62.47±0.11	18.88±0.04

Up to 15% AUC improvement.

Codes and datasets are available at: https://github.com/DM2-ND/Eland

Achieving Early Graph Anomaly Detection

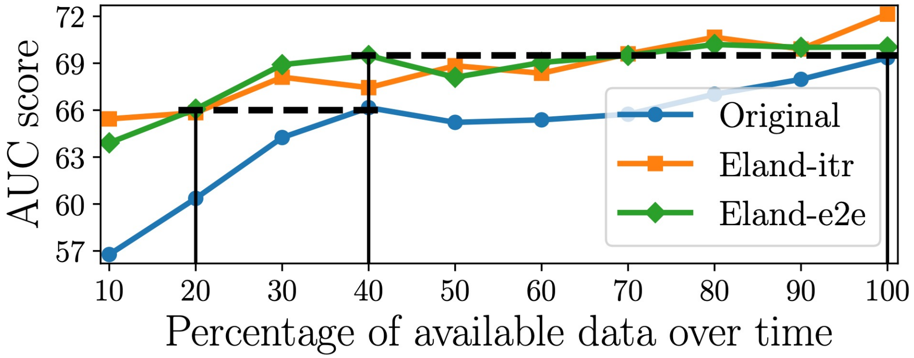
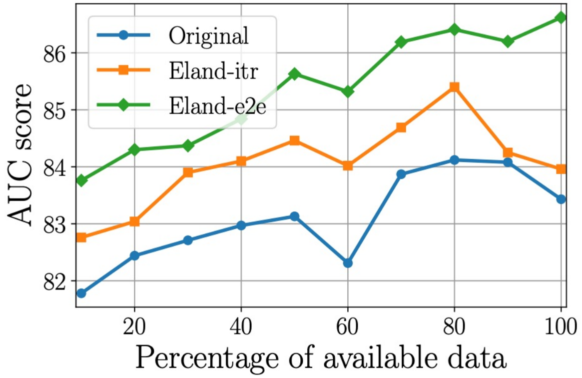
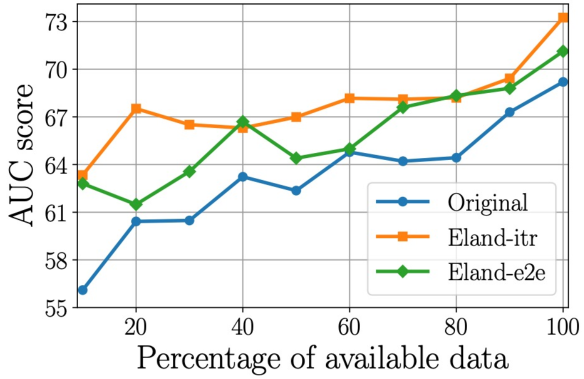


Figure 1: Performance of DOMINANT [7] and ELAND on a social media dataset considering the *earliest* 10%-100% data from each user's action sequence. Both our ELAND-ITR and ELAND-E2E with only 20% (40%) of available data can achieve the same performance as DOMINANT with 40% (100%) of data.



GCN



DeepAE

Thank you for listening!

- Any questions?
- Feel free to email me any further questions at tzhao2@nd.edu

COLLEGE OF ENGINEERING

UNIVERSITY OF NOTRE DAME

Case Study

Early Posts	A	B	C	D
Predicted Future Posts	F	A	A	G
Actual Future Posts	A	C	E	A

Early Posts	H	I		
Predicted Future Posts	J	K	L	
Actual Future Posts	M	N	O	N

A 给力支持啦！@[redacted] 鼎牛竞技收听加转发了，大家来参与！
► Need your support! @··· Follow and Repost.

B 年度大版本《龙脉传说》12月23日公测！新的故事、画面、玩法、伙伴、活动等全新体验等待你来探究！赢奖品，大家一起来！
► Game <Dragon Legend> beta version release on 12/23! ··· !

C 3D动作网游巨制《第九大陆》国服（于12月15日开启不删档测试。首批不删档测试激活码开始发放。每天抽出30名 一起来转发吧！ @[redacted] @[redacted] @[redacted]
► Game <The Ninth Continent> beta version release on 12/15. Repost for activation code. 30 free activation codes every day! @ ···

D 激活动作，立码进化 @[redacted] @[redacted] 活动广场阿狸的台历好想要哦！女朋友最喜欢了
► Upgrade now @··· I want this table calendar. Girls love it.

E 阿狸的台历好想要哦！女朋友最喜欢了 @[redacted]
► I want this table calendar. Girls love it. @···

F 我想得到一部iphone4s，多多支持斗地主 地主爷，送我个过圣诞吧 @[redacted] @[redacted] @[redacted]
► I want an iPhone4s. I support this game. So give me a free one for Christmas gift. @ ··· @ ··· @ ···

G #乐购生活寻找最光的光棍 让伊暖儿来陪伴你 送给你冬日里：11.10-11.11。先睹为快：都是抱枕吗？光棍节福利@[redacted]
► #... Buy a pillow for a warm winter: 11.10-11.11 ···

H 好羡慕这个手机，什么时候咱也弄台，体验店的感觉很不错 @[redacted] @[redacted] @[redacted]
► This phone feels good. We should get one. @ ··· @ ···

I 很酷很喜欢！我支持！ @[redacted] @[redacted]
► Very cool and adorable! I support it! @ ··· @ ···

J 苹果干掉第一 @[redacted]
► Apple products are the best. @ ···

K 支持，转/抱拳/微笑有这样道德的有儿人？【图片】
► I support it! Repost: This guy is noble.

L /傲慢，耶，好酷的造型| @[redacted]:套住/抠鼻狂亲 || @[redacted]: /傲慢 又帽子又大墨镜的。。不过衣服很拽/鼓掌
► Cool! Repost: Big hat, big sunglasses, and cool jacket.

M 价格也不贵，我完全可以接受！ @[redacted] @[redacted] @[redacted]三星梦想计划，草根的平台！
► It is not very expensive. Totally acceptable to me! @ ··· @... @ ··· Repost: Samsung dream plan. Platform for grassroot!

N 三星梦想计划，草根的平台！
► Samsung dream plan. Platform for grassroot!

O 支持三星活动，支持三星手机大屏幕！好喜欢！体验店很真实啊支持三星活动，支持三星手机体验店的感觉很不错很酷很喜欢！我支持！
► Support Samsung activities. Their phones have large screens! I like it! ··· So cool! I like it!