THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

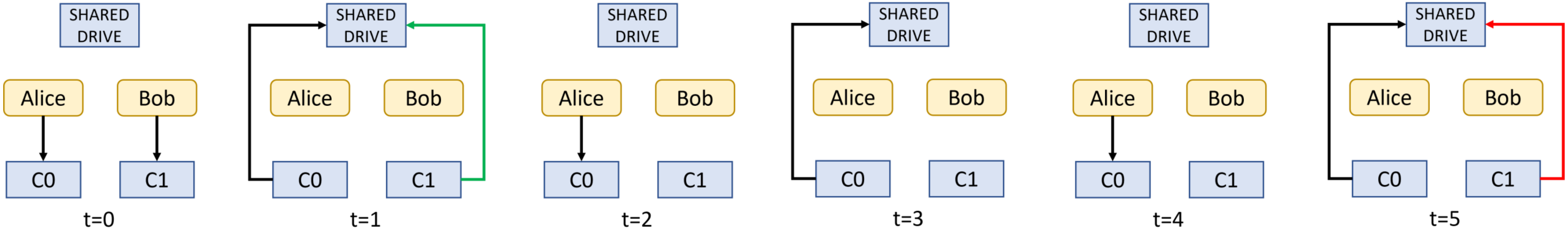
Evaluating EULER: Experimental Results of Network Anomaly Detection Models

Isaiah J. King & H. Howie Huang

Learning from Authoritative Security Experiment Results (LASER), 2022

San Diego, CA

Networks as a Temporal Graphs

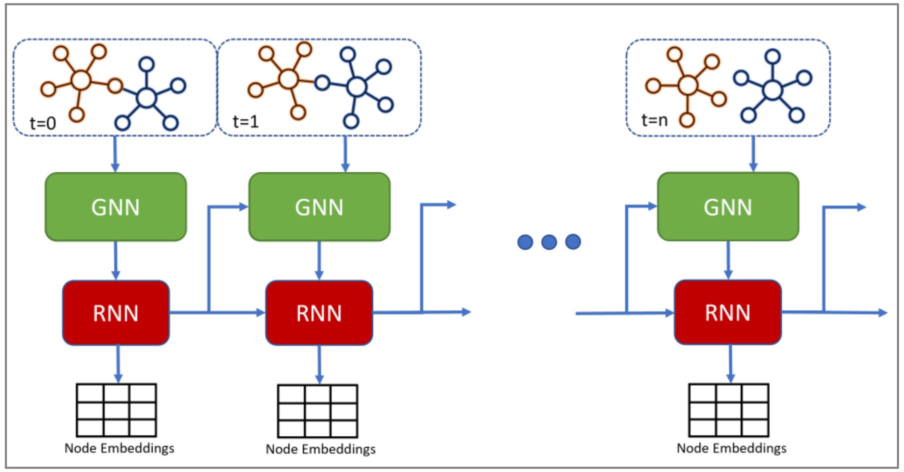


- Interactions on a network are relational, and temporal
- Given a series of graphs $\mathrm{G} = \{G_0, \ldots, G_T\}$ where $G_t = \{V_t, E_t\}$ anomalous edges correlate to lateral movement
- Can we detect anomalous edges using a **temporal link predictor**?

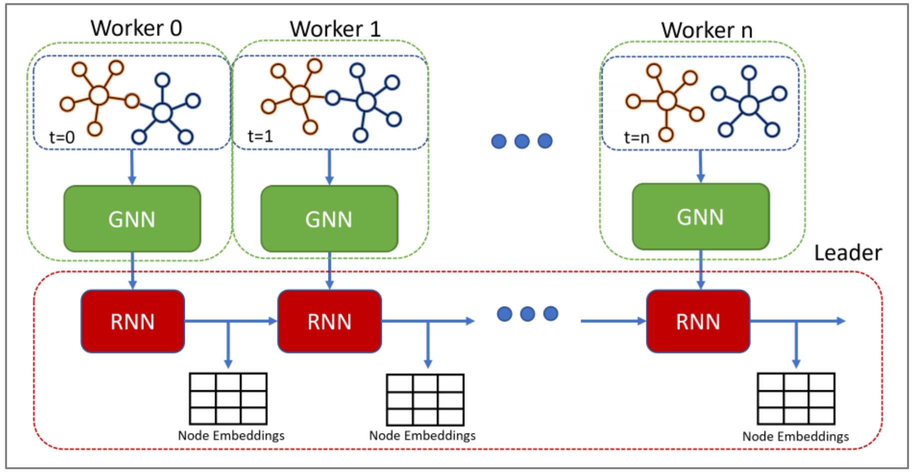
Temporal Link Prediction

- In the past, TLP has been accomplished by running GNN output through a sequence encoder
- Highly engineered models prone to overfitting
- Forces process to be sequential
- Cannot scale to large graphs (i.e. network logs)

- We propose uncoupling the RNN and GNN
- GNN is most complex portion of the approach
- Amdahl’s law—distribute the hard parts

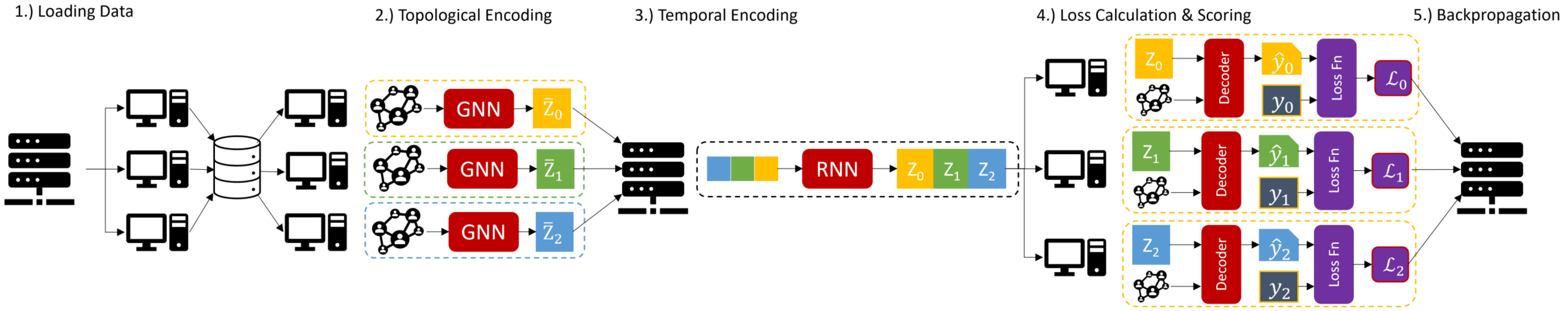


SoTA



Our Approach

The Distributed Framework



The Encoder-Decoder

- The EULER framework is a generic extension of the traditional GAE model
- It stacks a model-agnostic GNN upon a model-agnostic RNN
- Aims to find a low-dimensional encoding function $f(\cdot)$ of G
- And a decoding function $g(\cdot)$ of those encodings
- As a result of IP decoding,
  $\Pr[(u,v) \in E_{t+n}] \propto \mathbf{Z}_t[u]\mathbf{Z}[v]^{\mathrm{T}}$

$$f(G) = \mathbf{Z} = \mathrm{RNN}(\,[\mathrm{GNN}(\mathbf{X}_0, \mathbf{A}_0), \ldots, \mathrm{GNN}(\mathbf{X}_t, \mathbf{A}_t)]\,)$$

$$g(\mathbf{Z}_t) = \Pr[\mathbf{A}_{\mathrm{t+n}} = 1 \mid \mathbf{Z}_t] = \sigma\left(\mathbf{Z}_t \mathbf{Z}_t^{\mathrm{T}}\right)$$

Classifier

- Though most evaluation metrics used are for quality of scoring (AUC & AP) it’s useful to automate finding a cutoff
- An additional 5% of snapshots are held out of training for this
- Given TPR and FPR at threshold τ, optimal threshold is

$$\underset{\tau}{\operatorname{argmin}} \quad \|(1-\lambda)\mathrm{TPR}(\tau) - \lambda \mathrm{FPR}(\tau)\|$$

- $\lambda \in (0,1)$ is a user-defined hyperparameter, biasing against high FPR

Experiments & Challanges

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Replicating Prior Work

- (SI-)VGRNN
  - GCN on GRNN
  - GRNN output used as GCN input next snapshot
  - Currently #1 ranked Temporal LP model on PapersWithCode.com
- EGCN
  - RNN aims to find *parameters* of GCN
  - Very unique method, excellent at low info LP (guessing 10+ snapshots in the future)
- DynGraph2Vec (DynAE, DynRNN, DynAERNN)
  - MLP on RNN (no message passing or spectral convs)
  - Uses adj matrix as input & output vectors (not scalable)

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Sets

TABLE I: Data set metadata

Data Set	Nodes	Edges	Avg. Density	Timestamps
FB	663	23,394	0.00591	9
COLAB	315	5,104	0.01284	10
Enron10	184	4,784	0.00514	11

All data sets provided by VGRNN authors

- Facebook (FB)
    - Graph of users commenting on others' walls
    - Each snapshot is 1 day
- COLAB
    - Citation network in order of publication date
    - Each snapshot is 1 year
- Enron10
    - Emails between Enron employees between 1999-2000
    - Snapshots are 1 week

Tests

- Dynamic Link *Detection*
  - Inductive
  - Find $\mathrm{Pr}[\mathbf{A}_{\mathrm{t}} = 1 \mid \mathbf{Z}_{\mathrm{t}}]$ given $\mathbf{Z} = f\left(\left\{\hat{G}_0, \ldots, \hat{G}_t\right\}\right)$
- Dynamic Link *Prediction*
  - Transductive
  - Find $\mathrm{Pr}[\mathbf{A}_{\mathrm{t+1}} = 1 \mid \mathbf{Z}_{\mathrm{t}}]$ given $\mathbf{Z} = f(\{G_0, \ldots, G_t\})$
- Dynamic *New* Link Prediction
  - Same as above, but set of positive samples is only

    $\{(u, v) \mid (u, v) \in \mathcal{E}_{t+1} \wedge (u, v) \notin \mathcal{E}_t\}$

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Results

TABLE II: Comparison of EULER to related work on dynamic link detection

Metrics	Methods	Enron	COLAB	Facebook
AUC	VGAE	88.26 ± 1.33	70.49 ± 6.46	80.37 ± 0.12
	DynAE	84.06 ± 3.30	66.83 ± 2.62	60.71 ± 1.05
	DynRNN	77.74 ± 5.31	68.01 ± 5.50	69.77 ± 2.01
	DynAERNN	91.71 ± 0.94	77.38 ± 3.84	81.71 ± 1.51
	EGCN-O	93.07 ± 0.77	**90.77 ± 0.39**	86.91 ± 0.51
	EGCN-H	92.29 ± 0.66	87.47 ± 0.91	85.95 ± 0.95
	VGRNN	94.41 ± 0.73	88.67 ± 1.57	88.00 ± 0.57
	SI-VGRNN	95.03 ± 1.07	89.15 ± 1.31	88.12 ± 0.83
	EULER	**97.34 ± 0.41**	**91.89 ± 0.76**	**92.20 ± 0.56**
AP	VGAE	89.95 ± 1.45	73.08 ± 5.70	79.80 ± 0.22
	DynAE	86.30 ± 2.43	67.92 ± 2.43	60.83 ± 0.94
	DynRNN	81.85 ± 4.44	73.12 ± 3.15	70.63 ± 1.75
	DynAERNN	93.16 ± 0.88	83.02 ± 2.59	83.36 ± 1.83
	EGCN-O	92.56 ± 0.99	**91.41 ± 0.33**	84.88 ± 0.52
	EGCN-H	92.56 ± 0.72	88.00 ± 0.85	82.56 ± 0.91
	VGRNN	95.17 ± 0.41	89.74 ± 1.31	87.32 ± 0.60
	SI-VGRNN	**96.31 ± 0.72**	89.90 ± 1.06	87.69 ± 0.92
	EULER	**97.06 ± 0.48**	**92.85 ± 0.88**	**91.74 ± 0.71**

TABLE III: Comparison of EULER to related work on dynamic link prediction

Metrics	Methods	Enron	COLAB	Facebook
AUC	DynAE	74.22 ± 0.74	63.14 ± 1.30	56.06 ± 0.29
	DynRNN	86.41 ± 1.36	75.7 ± 1.09	73.18 ± 0.60
	DynAERNN	87.43 ± 1.19	76.06 ± 1.08	76.02 ± 0.88
	EGCN-O	84.28 ± 0.87	78.63 ± 2.14	77.31 ± 0.58
	EGCN-H	88.29 ± 0.87	80.80 ± 0.95	75.88 ± 0.32
	VGRNN	93.10 ± 0.57	**85.95 ± 0.49**	89.47 ± 0.37
	SI-VGRNN	**93.93 ± 1.03**	85.45 ± 0.91	**90.94 ± 0.37**
	EULER	**93.15 ± 0.42**	**86.54 ± 0.20**	**90.88 ± 0.12**
AP	DynAE	76.00 ± 0.77	64.02 ± 1.08	56.04 ± 0.37
	DynRNN	85.61 ± 1.46	78.95 ± 1.55	75.88 ± 0.42
	DynAERNN	89.37 ± 1.17	81.84 ± 0.89	78.55 ± 0.73
	EGCN-O	86.55 ± 1.57	81.43 ± 1.69	76.13 ± 0.52
	EGCN-H	89.33 ± 1.25	83.87 ± 0.83	74.34 ± 0.53
	VGRNN	93.29 ± 0.69	87.77 ± 0.79	89.04 ± 0.33
	SI-VGRNN	**94.44 ± 0.85**	**88.36 ± 0.73**	**90.19 ± 0.27**
	EULER	**94.10 ± 0.32**	**89.03 ± 0.08**	**89.98 ± 0.19**

TABLE IV: Comparison of EULER to related work on dynamic new link prediction

Metrics	Methods	Enron	COLAB	Facebook
AUC	DynAE	66.10 ± 0.71	58.14 ± 1.16	54.62 ± 0.22
	DynRNN	83.20 ± 1.01	71.71± 0.73	73.32 ± 0.60
	DynAERNN	83.77 ± 1.65	71.99 ± 1.04	76.35 ± 0.50
	EGCN-O	84.42 ± 0.82	79.06 ± 1.60	75.95 ± 1.15
	EGCN-H	87.00 ± 0.85	78.47 ± 1.27	74.85 ± 0.98
	VGRNN	**88.43 ± 0.75**	77.09 ± 0.23	87.20 ± 0.43
	SI-VGRNN	**88.60 ± 0.95**	**77.95 ± 0.41**	87.74 ± 0.53
	EULER	87.92 ± 0.64	**78.39 ± 0.68**	**89.02 ± 0.09**
AP	DynAE	66.50 ± 1.12	58.82 ± 1.06	54.57 ± 0.20
	DynRNN	80.96 ± 1.37	75.34 ± 0.67	75.52 ± 0.50
	DynAERNN	85.16 ± 1.04	77.68 ± 0.66	78.70 ± 0.44
	EGCN-O	86.92 ± 0.39	81.36 ± 0.85	73.66 ± 1.25
	EGCN-H	86.46 ± 1.42	79.11 ± 2.26	73.43 ± 1.38
	VGRNN	87.57 ± 0.57	79.63 ± 0.94	86.30 ± 0.29
	SI-VGRNN	**87.88 ± 0.84**	**81.26 ± 0.38**	**86.72 ± 0.54**
	EULER	**88.49 ± 0.55**	**81.34 ± 0.62**	**87.54 ± 0.11**

- EULER out-performs prior work on all detection tests
  - Though only with *statistical significance* on FB and Enron AUC
- Prior works are not statistically significantly better than EULER on any prediction tests
- EULER is better with significance on new FB test, and equivalent elsewhere

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

The Importance of Statistical Significance

TABLE III: Comparison of EULER to related work on dynamic link prediction

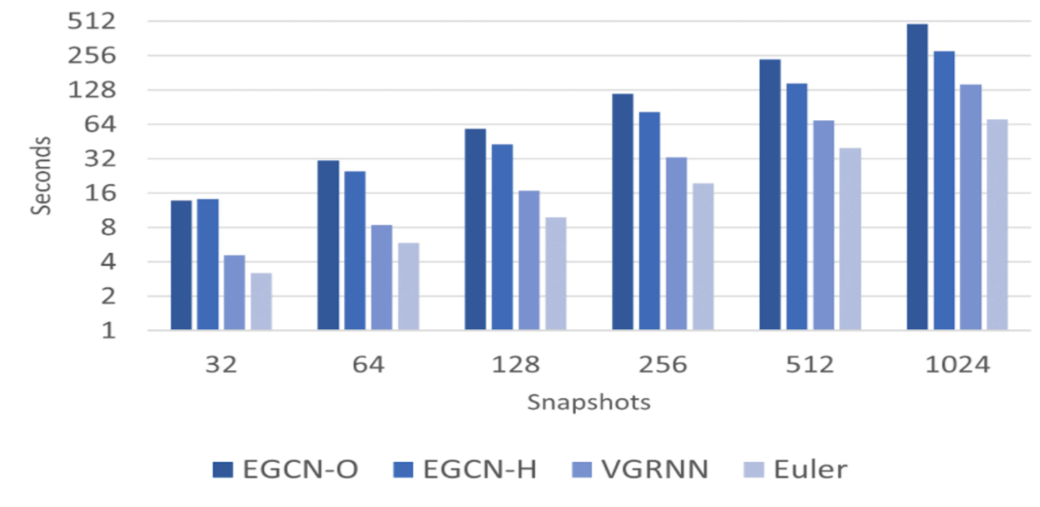
Metrics	Methods	Enron	COLAB	Facebook
AUC	DynAE	74.22 ± 0.74	63.14 ± 1.30	56.06 ± 0.29
	DynRNN	86.41 ± 1.36	75.7 ± 1.09	73.18 ± 0.60
	DynAERNN	87.43 ± 1.19	76.06 ± 1.08	76.02 ± 0.88
	EGCN-O	84.28 ± 0.87	78.63 ± 2.14	77.31 ± 0.58
	EGCN-H	88.29 ± 0.87	80.80 ± 0.95	75.88 ± 0.32
	VGRNN	93.10 ± 0.57	**85.95 ± 0.49**	89.47 ± 0.37
	SI-VGRNN	**93.93 ± 1.03**	85.45 ± 0.91	**90.94 ± 0.37**
	EULER	**93.15 ± 0.42**	**86.54 ± 0.20**	**90.88 ± 0.12**
AP	DynAE	76.00 ± 0.77	64.02 ± 1.08	56.04 ± 0.37
	DynRNN	85.61 ± 1.46	78.95 ± 1.55	75.88 ± 0.42
	DynAERNN	89.37 ± 1.17	81.84 ± 0.89	78.55 ± 0.73
	EGCN-O	86.55 ± 1.57	81.43 ± 1.69	76.13 ± 0.52
	EGCN-H	89.33 ± 1.25	83.87 ± 0.83	74.34 ± 0.53
	VGRNN	93.29 ± 0.69	87.77 ± 0.79	89.04 ± 0.33
	SI-VGRNN	**94.44 ± 0.85**	**88.36 ± 0.73**	**90.19 ± 0.27**
	EULER	**94.10 ± 0.32**	**89.03 ± 0.08**	**89.98 ± 0.19**

- When are models essentially the same?
- Similar avg. AUC/AP lower stderr
- Use hypothesis testing:

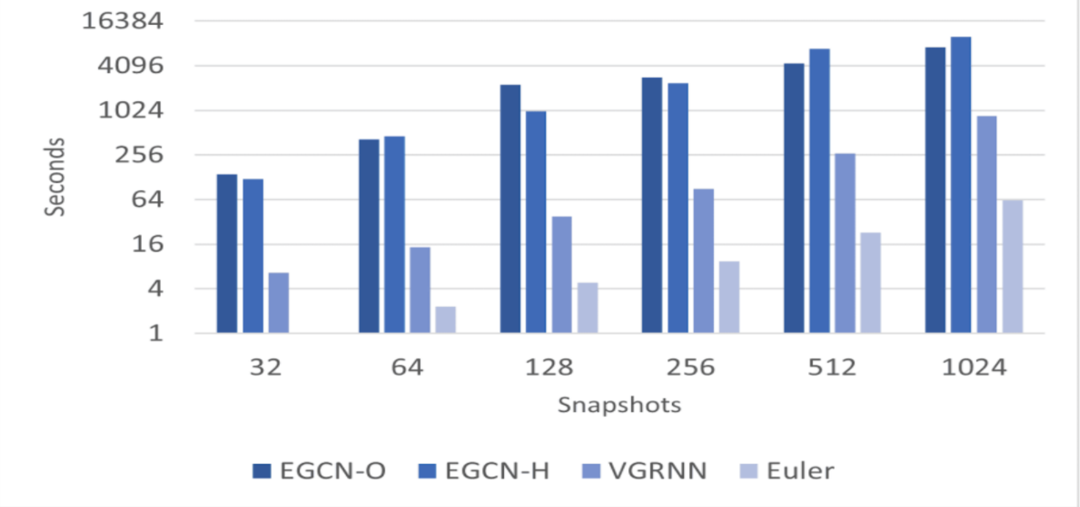
$$t = \frac{0 - (\mu(B) - \mu(A))}{\sqrt{\frac{\mathrm{Var}(B-A)}{N}}} = \frac{\mu(A) - \mu(B)}{\sqrt{\sigma_M(A)^2 + \sigma_M(B)^2}}$$

- t < 2.228 means not significantly different (p-value > 0.05)

Performance Comparison



Forward Time



Backward Time

Euler uses 16 workers; prior works use 16 inter-op threads for fair comparison

- Euler is consistently faster than prior works
- Forward time is about 2x faster
- Backward time is 16x better (showing near-perfect scaling)

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Real-world data sets

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

The LANL Dataset

TABLE V: LANL Data Set Metadata

Nodes	17,685
Events	45,871,390
Anomalous Edges	750
Duration (Days)	58

- 58 Days of log files in a real-world system
- Attack campaigns sporadically
- Redlog identifies 750 authorization events “involved in compromise”
- Nodes: Users, Computers, System
- Edges: Authorizations, weighted according to frequency:

$$W((u,v) \in \mathcal{E}) = \sigma\left(\frac{C(u,v) - \mu_{\mathcal{E}}}{\Sigma_{\mathcal{E}}}\right)$$

- Features: 1-hot ID, and 1-hot vector of node’s role

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

LANL Tests

- Tested 3 Encoders
  - GCN
  - GraphSAGE (Maxpool aggr.)
  - GAT (3 attn. heads)
- Tested 3 RNNs
  - GRU
  - LSTM
  - None (ablation study)
- Compared to 4 prior works
  - GL-LV, GL-GV are static, graph-based
  - UA is a simple rules-based method
  - VGRNN is SoTA temporal LP method

Tests:

- Link Detection
  - Real world use: forensic audit
- Link Prediction
  - Real world use: live detector

Results

- Link Detection:
  - Best precision was GCN-GRU
  - Surprisingly, ablation study had best AUC (with GRU). RNN may not be necessary
  - SAGE also performed well

- Link Prediction
  - SAGE had best precision this time
  - AUC not as good as GCN

- Overall
  - Regression metrics are better than all prior works
  - Higher TPR and lower FPR on classification metrics than prior works

Link Detection

Encoder	RNN	AUC	AP	TPR	FPR
GCN	GRU	0.9912	**0.05230**	86.10	0.5698
	LSTM	0.9913	0.01692	89.65	0.5723
	None	**0.9916**	0.01163	88.57	0.4798
SAGE	GRU	0.9872	0.03065	84.71	0.6874
	LSTM	0.9887	0.03892	83.55	0.6591
	None	0.8652	0.00515	79.58	24.5669
GAT	GRU	0.9094	0.00762	85.21	21.533
	LSTM	0.8713	0.00219	96.83	19.873
	None	0.9867	0.00787	99.88	23.174
GL-LV [9]		–	–	67.00	1.200
GL-GV [9]		–	–	85.00	0.900
UA		–	–	72.00	4.400
VGRNN		0.9315	0.0000	59.69	4.938

Link Prediction

Encoder	RNN	AUC	AP	TPR	FPR
GCN	GRU	**0.9906**	0.0155	85.49	0.6088
	LSTM	0.9885	0.0166	78.91	0.5987
	None	0.9902	0.0092	86.42	0.5425
SAGE	GRU	0.9847	0.0200	86.30	1.6542
	LSTM	0.9865	**0.0228**	85.29	0.8037
	None	0.9284	0.0020	86.23	16.525
GAT	GRU	0.8826	0.0020	87.82	21.971
	LSTM	0.8383	0.0002	83.42	29.297
	None	0.9352	0.0079	88.83	20.093
VGRNN		0.9503	0.0004	70.00	0.280

A More Detailed Data Set: OpTC

- With LANL it's unclear how "anomalous events" are defined

- OpTC has entire redlog—more informative labels
- Edges are `FLOW-START` events
- Weighted and directed the same way as LANL
- No node features, just 1-Hot IDs
- Edges Anomalous if
    - SRC or DST IP in redteam event
    - PID in redteam and time >= ts
    - Edges to/from compromised IPs remain anomalous until the end of the day

TABLE VIII: OpTC Data Set Metadata

Nodes	1,114
Events	7,773,514
Anomalous Edges	21,872
Duration (Days)	7

Results

- Fewer hosts allows us to use softmax anomaly detector
- Boosts scores significantly
- With easier to interpret results, Euler has low enough FPR for IDS

TABLE VI: Effectiveness of link prediction models on the OpTC Data Set

Detection						
Model	δ (h)	F1	AUC	AP	TPR (%)	FPR (%)
EGCN-O	5	0.005	0.554	0.003	67.5	58.7
EGCN-H	3.5	0.004	0.484	0.002	83.9	85.4
VGRNN	5	0.048	0.988	0.367	**99.3**	15.0
EULER GRU	2.5	0.140	0.888	0.088	17.8	0.473
EULER LSTM	2.5	0.189	0.882	0.118	17.8	0.168
EULER-SM GRU	0.125	0.937	**0.995**	0.973	97.0	0.021
EULER-SM LSTM	0.125	**0.955**	**0.995**	**0.984**	96.7	**0.012**

Prediction						
Model	δ (h)	F1	AUC	AP	TPR (%)	FPR (%)
EGCN-O	5	0.005	0.563	0.003	72.7	63.2
EGCN-H	3.5	0.004	0.507	0.003	80.0	80.2
VGRNN	0.125	0.014	0.692	0.008	73.1	42.1
EULER GRU	3	0.167	0.785	0.180	37.6	10.4
EULER LSTM	3	0.207	0.779	0.243	42.7	6.75
EULER-SM GRU	0.125	0.931	**0.995**	0.969	93.8	0.017
EULER-SM LSTM	0.5	**0.944**	0.994	**0.986**	**94.9**	**0.013**

Conclusion

Euler accomplished the following:

- Consistently as powerful or better than prior work
- Parallelized temporal link prediction
- First use of graph temporal link prediction for IDS
- Achieved high scores on OpTC; good scores on LANL

Discussion

- Why do so few ML papers make use of t-tests?
- Why don’t results on small data sets apply to real world ones?
- How valuable is LANL v. OpTC for evaluating IDS models?
- How to integrate speed into evaluation? What is a fair comparison?

Thank You

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC