Sandia National Laboratories

Exceptional service in the national interest

GRAPH ATTENTION EMBEDDINGS AS A CAUSAL LENS IN TEMPORAL LINK PREDICTION

Presented by: Sarah Simpson

Dan Krofcheck, Kate Cauthen, Michael Xi, Marco Campos, Matt Sweitzer, Asmeret Naugle, Casey Doyle

MLDL Workshop 2024

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

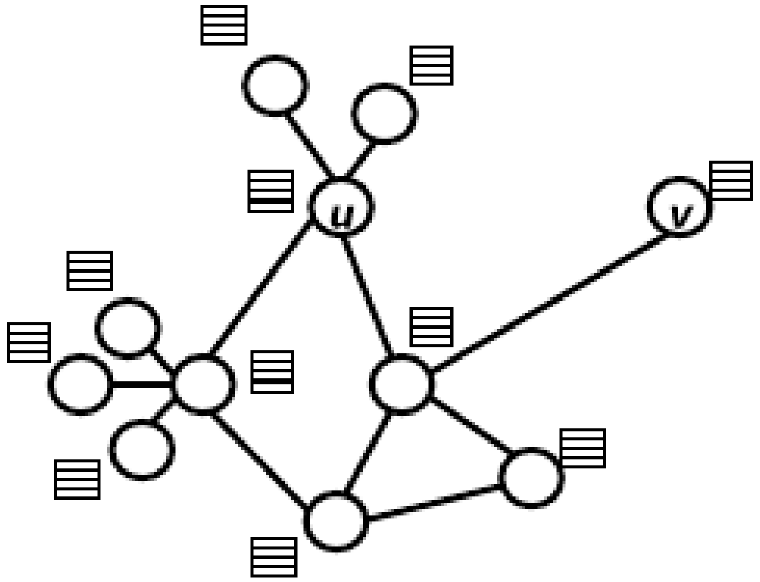
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-09855C

Background and Motivation

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

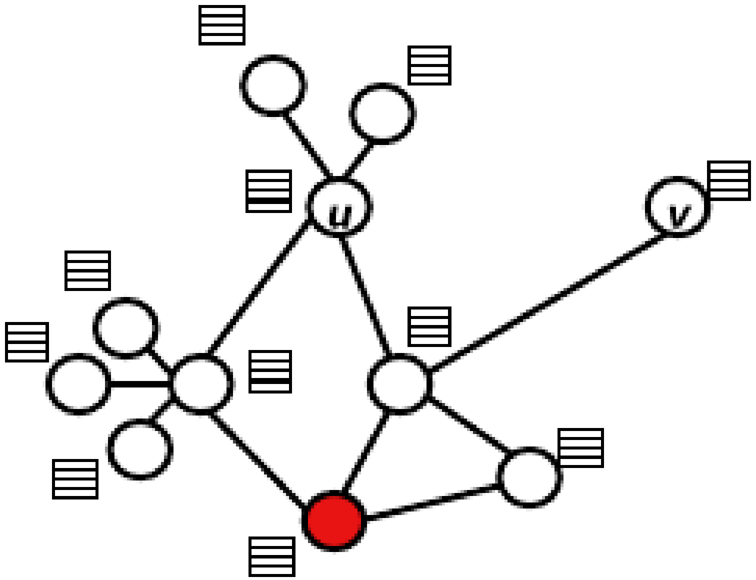


TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- **Node classification**
- Community detection
- Link prediction



Example applications

Identify bad actors in social systems

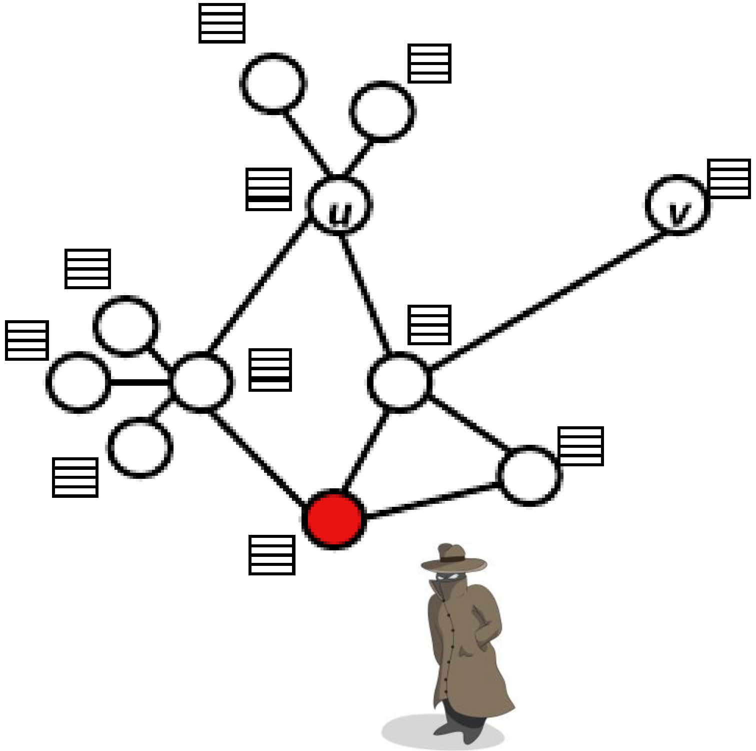
Find vulnerable components in an electric grid

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- **Node classification**
- Community detection
- Link prediction



<u>Example applications</u>

Identify bad actors in social systems

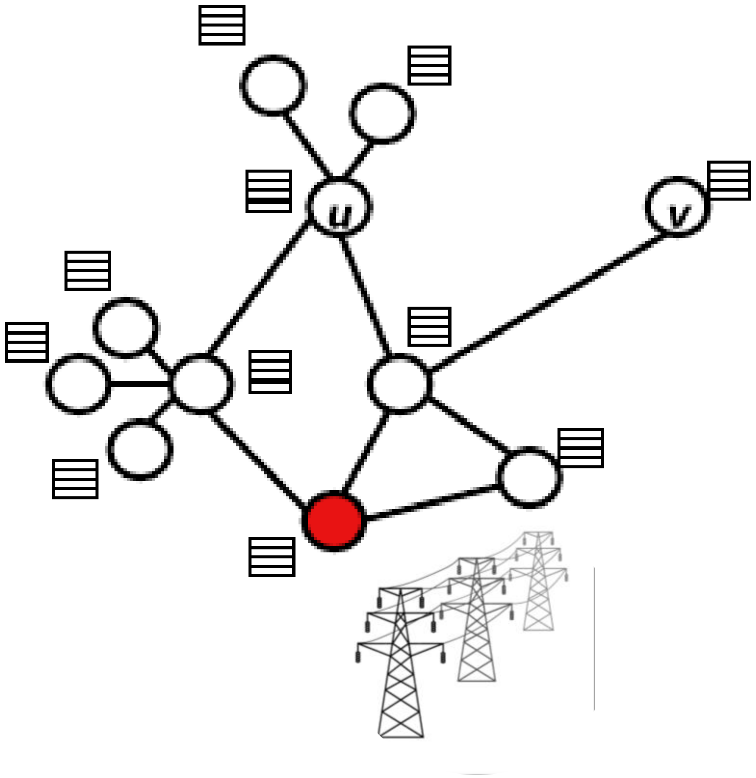
Find vulnerable components in an electric grid

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- **Node classification**
- Community detection
- Link prediction



Example applications

Identify bad actors in social systems

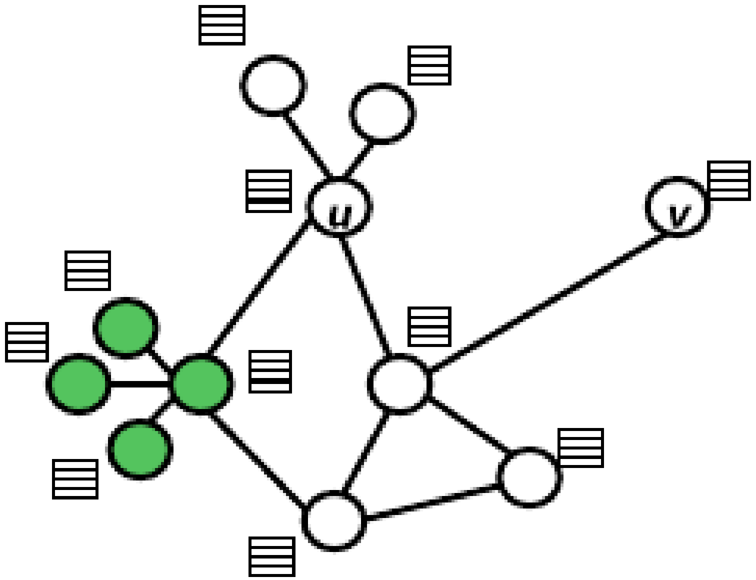
Find vulnerable components in an electric grid

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- Node classification
- **Community detection**
- Link prediction



Example applications

Identify vulnerable populations for pandemic intervention and countermeasure delivery

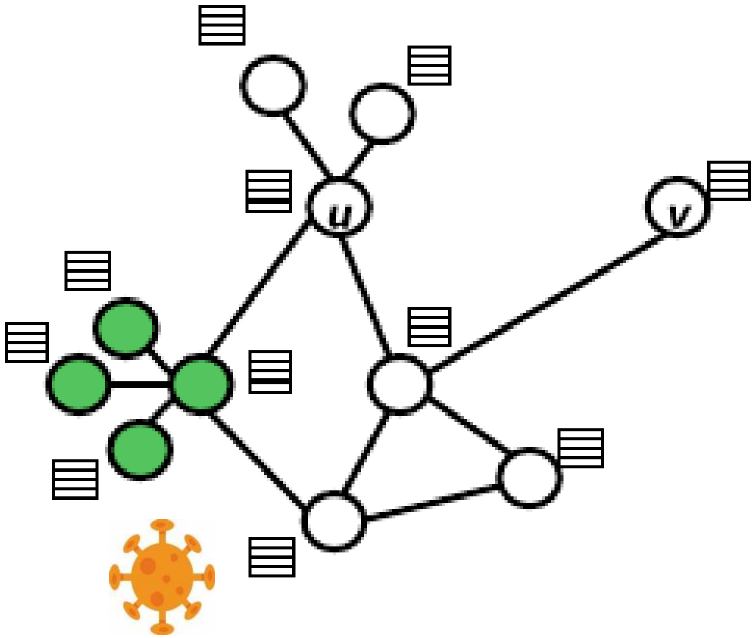
Disinformation and influence campaign identification

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- Node classification
- **Community detection**
- Link prediction



Example applications

Identify vulnerable populations for pandemic intervention and countermeasure delivery

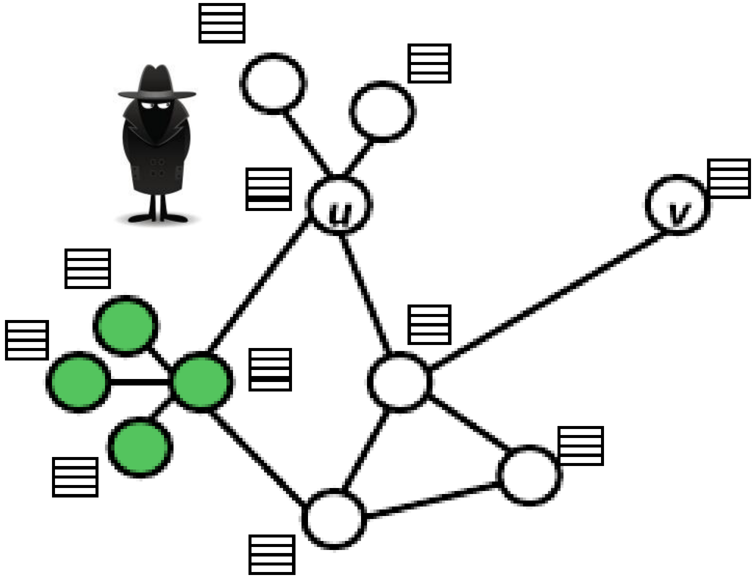
Disinformation and influence campaign identification

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- Node classification
- **Community detection**
- Link prediction



Example applications

Identify vulnerable populations for pandemic intervention and countermeasure delivery

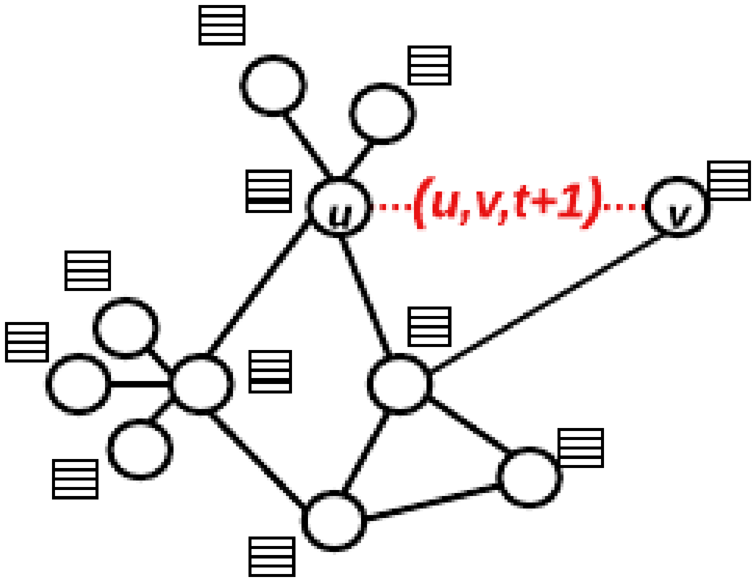
Disinformation and influence campaign identification

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- Node classification
- Community detection
- **Link prediction**



Example applications

Predict cascading failures in electric grid systems

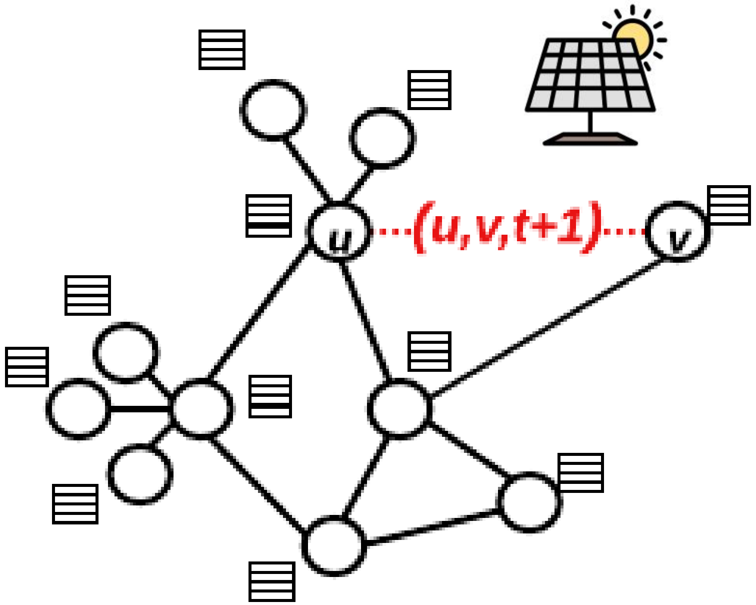
Anticipate high impact scientific collaborations

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- Node classification
- Community detection
- **Link prediction**



Example applications

Predict cascading failures in electric grid systems

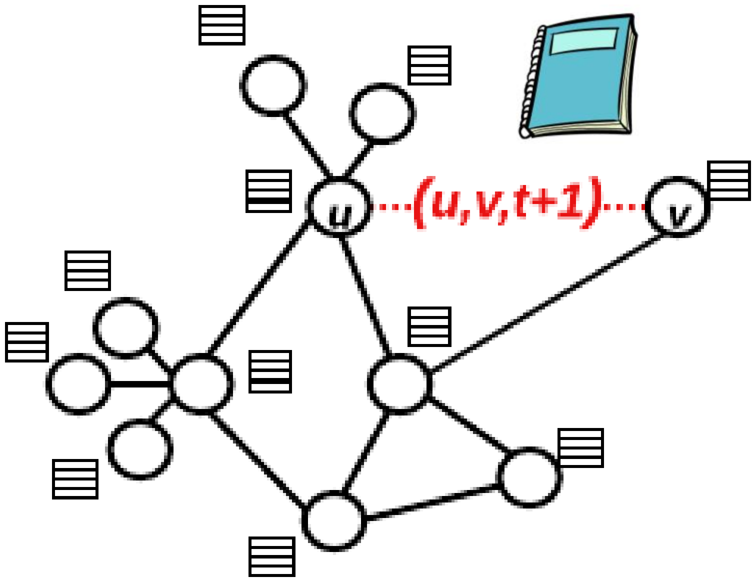
Anticipate high impact scientific collaborations

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

Temporal graphs are an extensible structure that is heavily used to represent complex dynamical systems across many high impact domains.

Common tasks on these systems include

- Node classification
- Community detection
- **Link prediction**



Example applications

Predict cascading failures in electric grid systems

Anticipate high impact scientific collaborations

TEMPORAL ATTRIBUTED GRAPHS AND CRITICAL SYSTEMS

These representations are used for far more than advertising, and marketing, social media. They are frequently used in high consequence modeling applications.

Cyber networks

Global trade

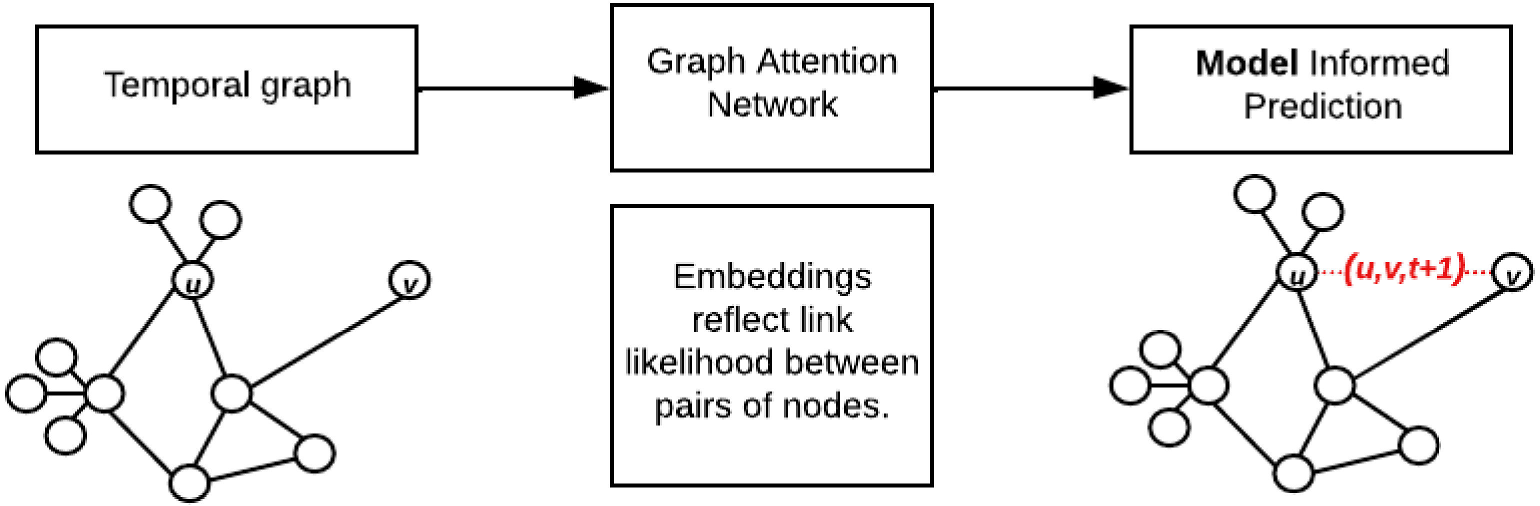
Transportation and logistics

Social networks

CURRENT MODELING STATE OF THE ART

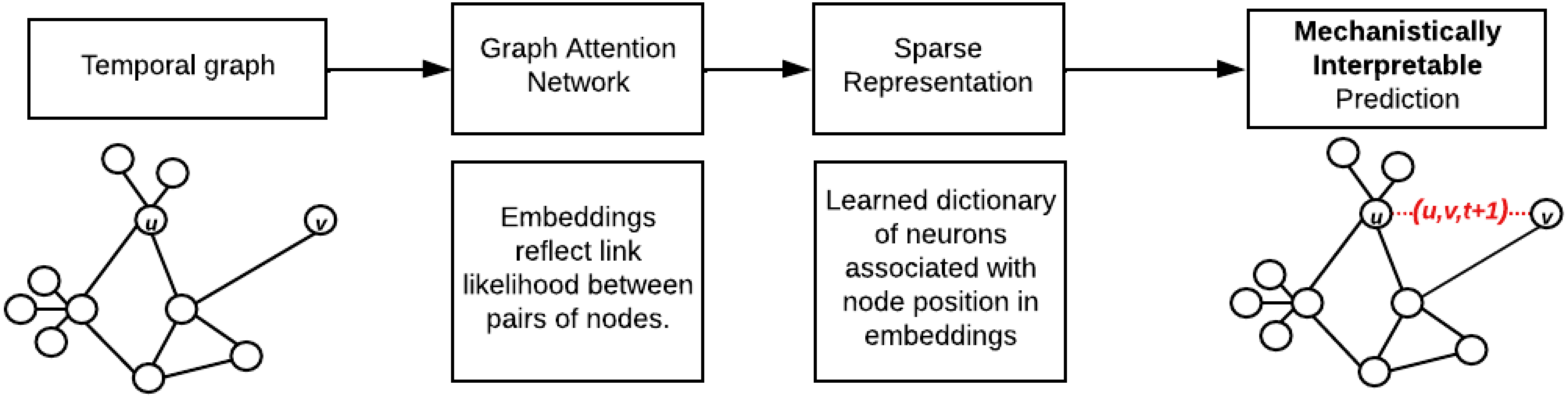
Current state-of-the-art algorithms are performant but lack causal structure and explainability.

This can preclude their use in higher consequence modeling and decision making.


Temporal graph
Graph Attention Network
Model Informed Prediction
Embeddings reflect link likelihood between pairs of nodes.
u
v
u
(u,v,t+1)
v

INCORPORATING MECHANISTIC INTERPRETABILITY

We seek to develop **algorithm agnostic** methods to incorporate interpretable, causal explanations into graph prediction models using the latent representations that underpin model inference.



Temporal graph

Graph Attention Network

Embeddings reflect link likelihood between pairs of nodes.

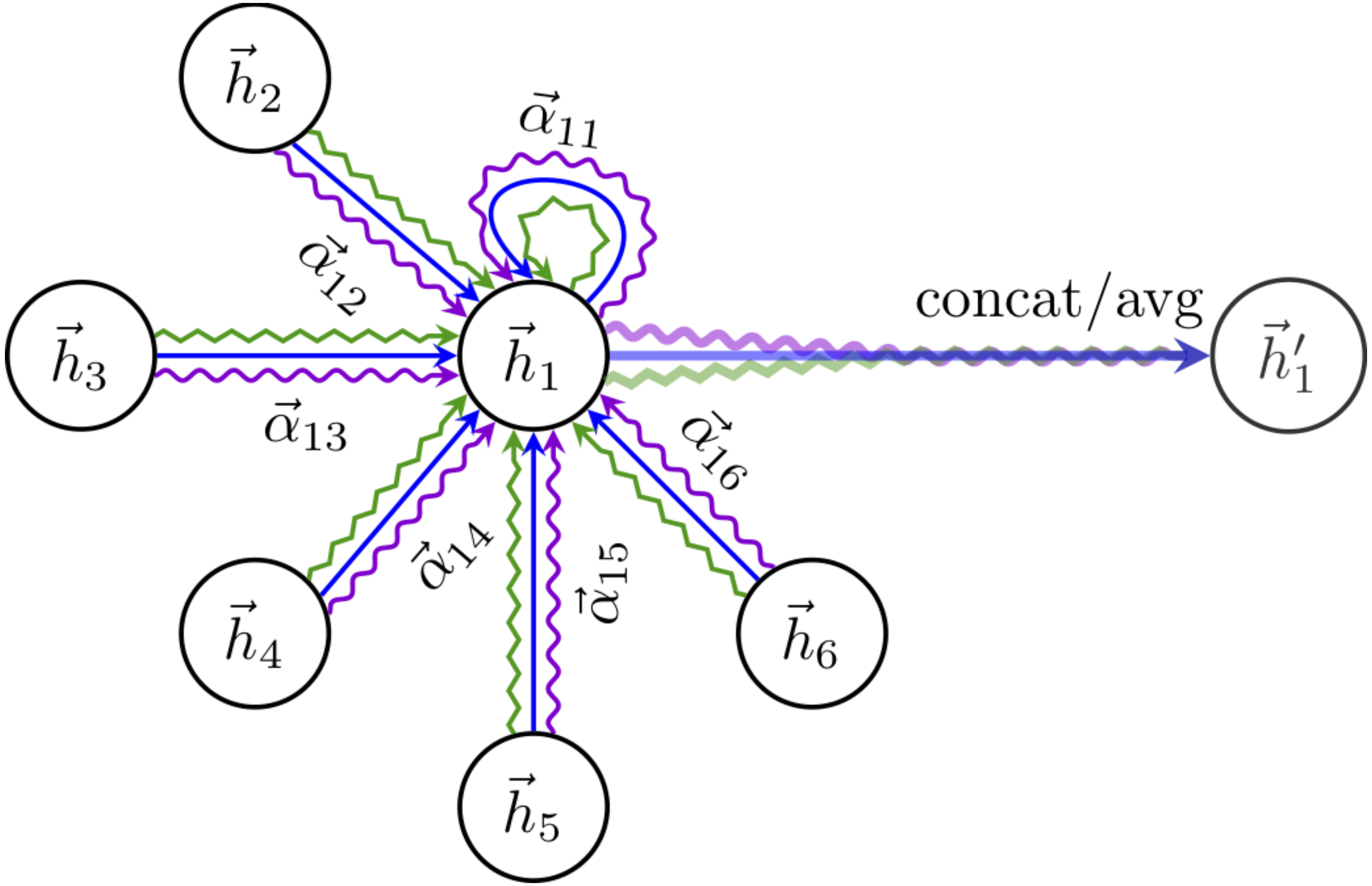
Sparse Representation

Learned dictionary of neurons associated with node position in embeddings

Mechanistically Interpretable Prediction

LATENT REPRESENTATIONS: EMBEDDINGS AND NODE SELF ATTENTION

Contextually rich relationships between nodes are learned through the attention mechanism.

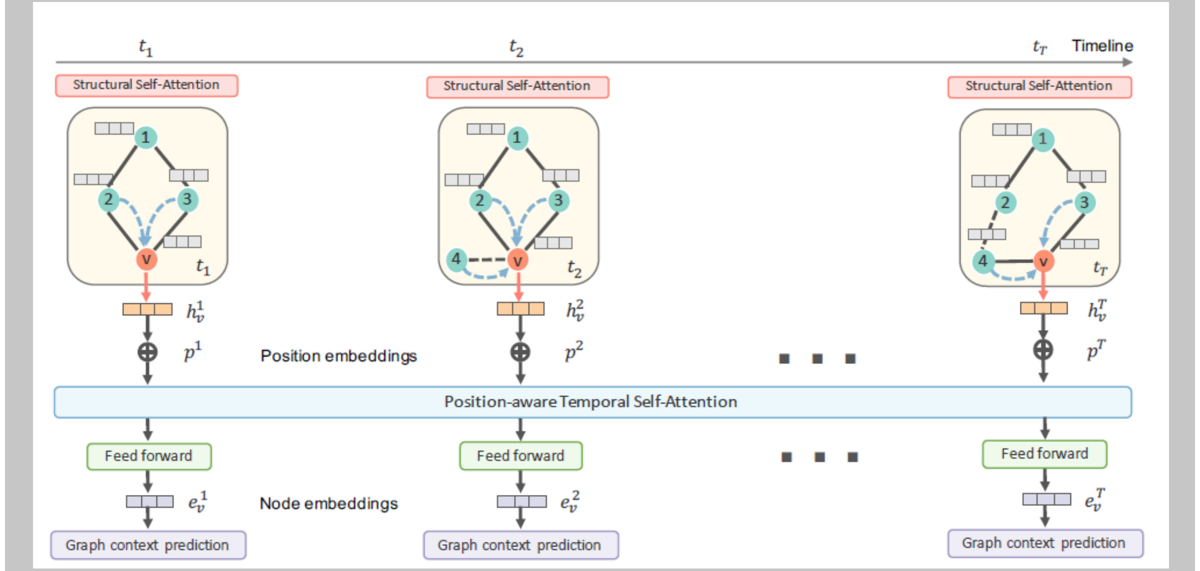


- Technique originally introduced in large language modeling
- Each node and edge embedding is associated with query, key, and value learned vectors.
- Similarity between the query of each element and the key of all other elements is used by the model to learn relationships between sentences, paragraphs, etc

EXEMPLAR MODELS

DySAT (Dynamic Self-Attention Network)

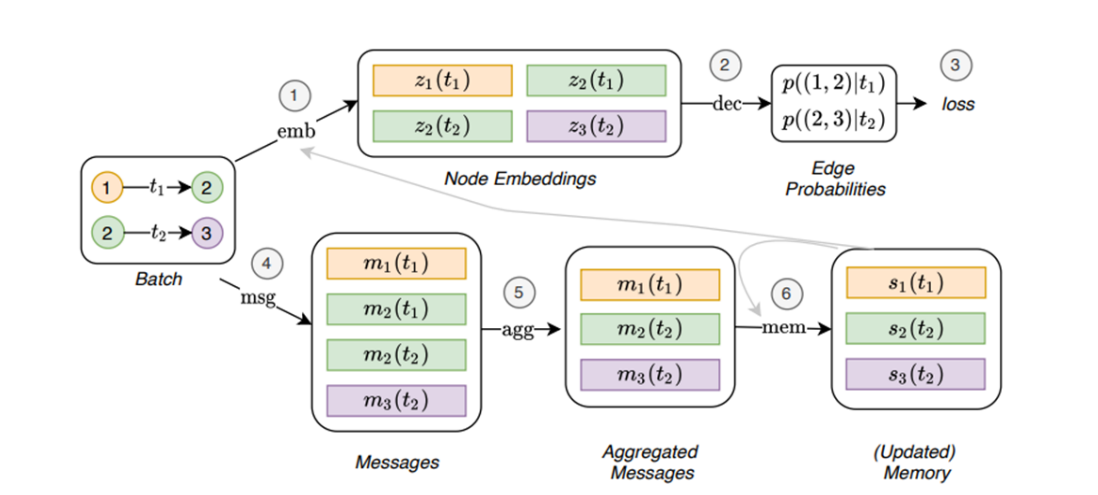
- Learns local and global dependencies through hierarchical self-attention layers
- Discrete-time dynamics



Sankar, Aravind, et al. "Dysat: Deep neural representation learning on dynamic graphs via self-attention networks." *Proceedings of the 13th international conference on web search and data mining*. 2020.

TGN (Temporal Graph Networks)

- Adopts a message-passing framework to handle dynamic graphs
- Continuous-time dynamics

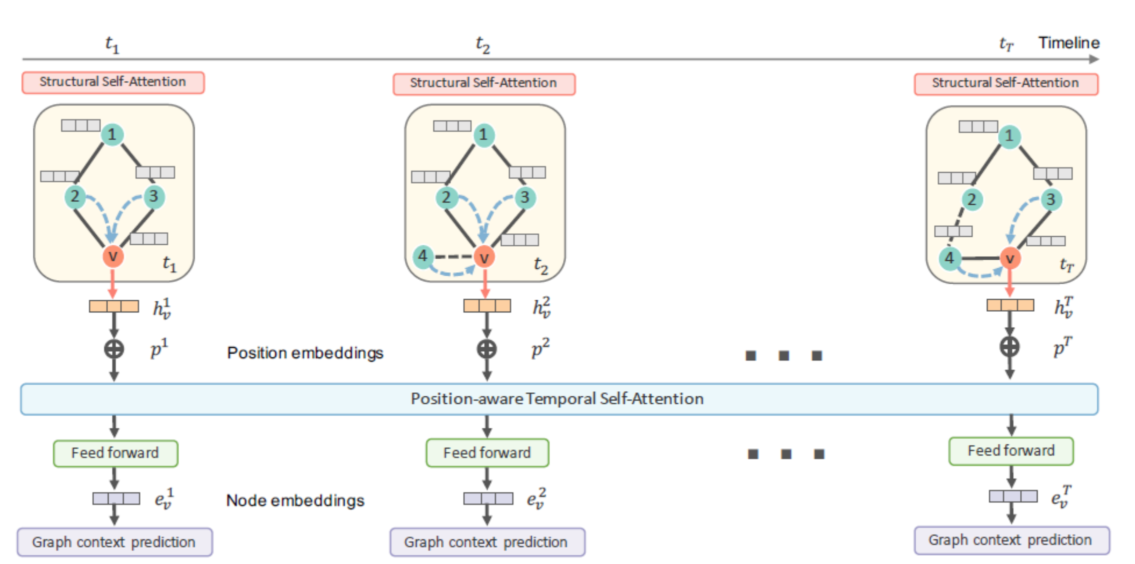


Rossi, Emanuele, et al. "Temporal graph networks for deep learning on dynamic graphs." *arXiv preprint arXiv:2006.10637* (2020).

EXEMPLAR MODELS

DySAT (Dynamic Self-Attention Network)

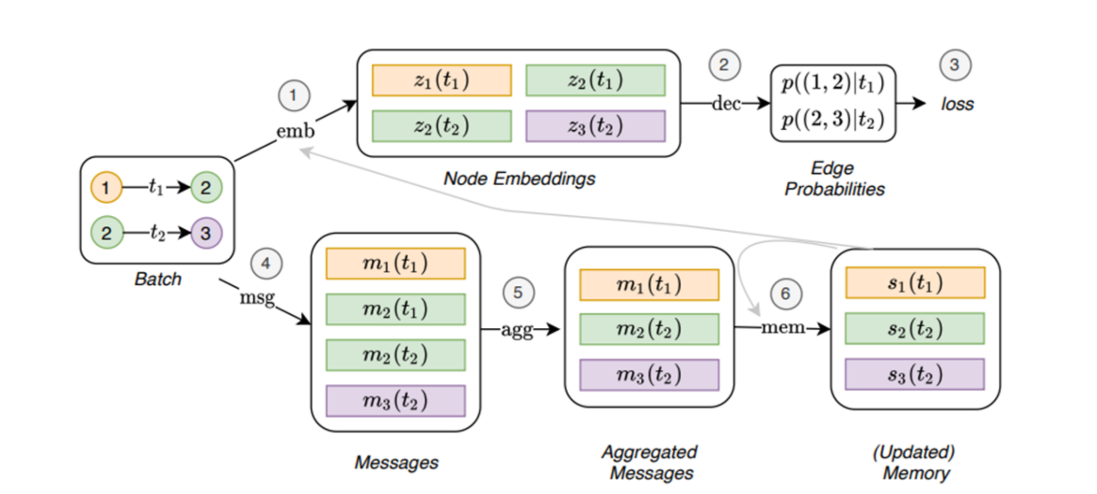
- Learns local and global dependencies through hierarchical self-attention layers
- Discrete-time dynamics



Sankar, Aravind, et al. "Dysat: Deep neural representation learning on dynamic graphs via self-attention networks." *Proceedings of the 13th international conference on web search and data mining*. 2020.

TGN (Temporal Graph Networks)

- Adopts a message-passing framework to handle dynamic graphs
- Continuous-time dynamics



Rossi, Emanuele, et al. "Temporal graph networks for deep learning on dynamic graphs." *arXiv preprint arXiv:2006.10637* (2020).

DYSAT ARCHITECTURE

Input: Adjacency matrix, node attributes

Step 1: Structural self-attention

Step 2: Encode temporal position

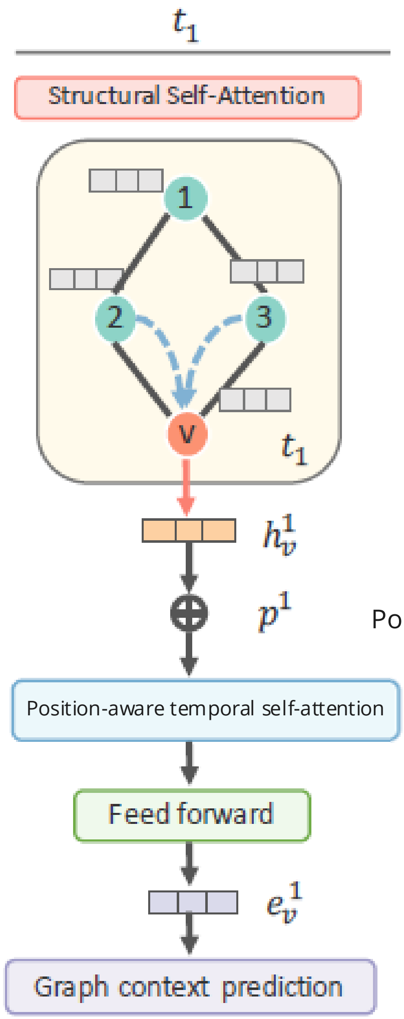
Step 3: Temporal self-attention

Output: 1 node embedding vector per node per snapshot

Input: $\boldsymbol{x}_u$, adjacency matrix

$$\boldsymbol{z}_v = \sigma\left(\sum_{u \in N_v} \alpha_{uv} \boldsymbol{W}^s \boldsymbol{x}_u\right)$$

$$\boldsymbol{Z}_v = \boldsymbol{\beta}_v(\boldsymbol{X}_v \boldsymbol{W}_v)$$

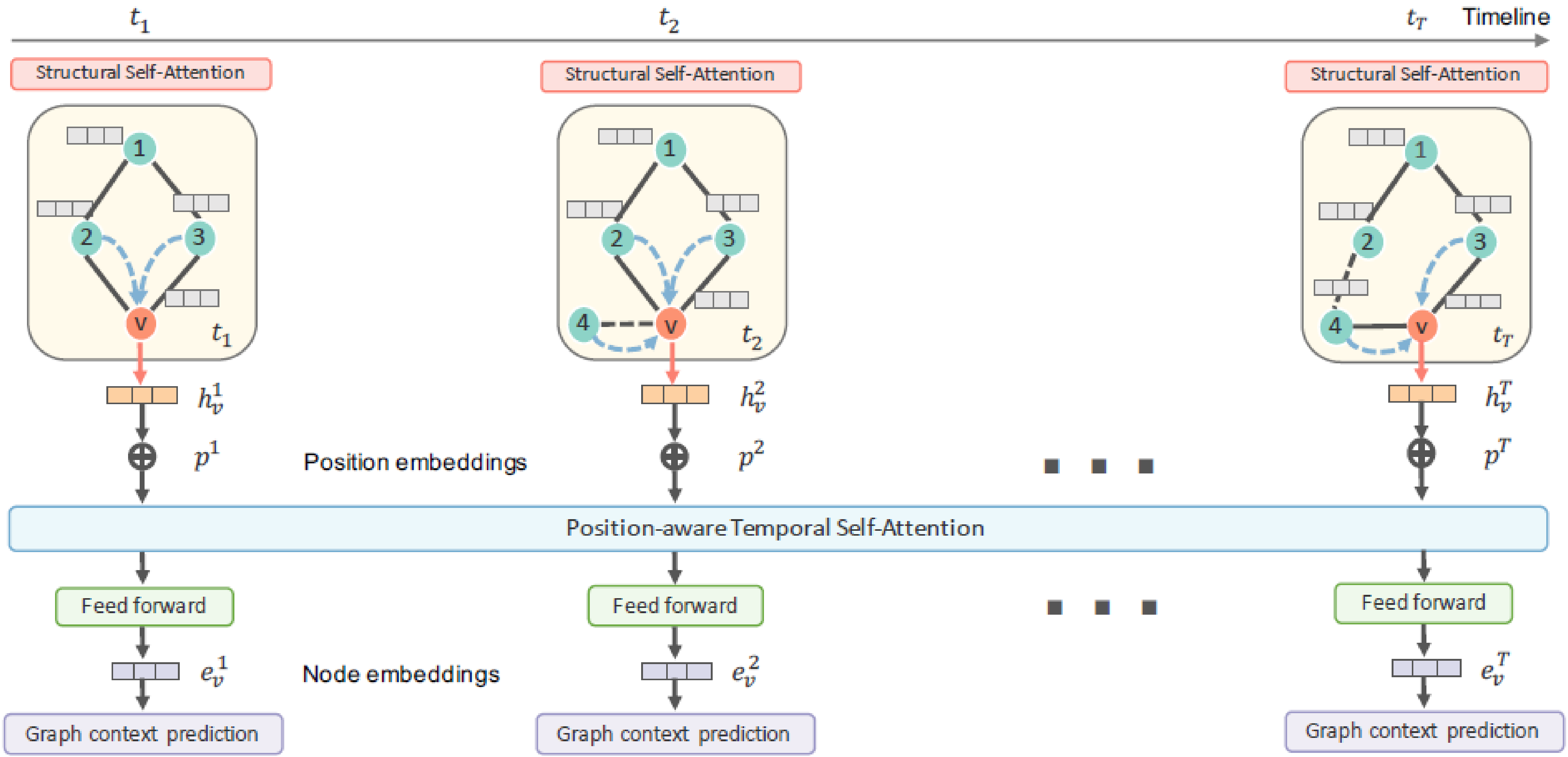


Position embeddings

Node embeddings

Sankar, A., Wu, Y., Gou, L., Zhang, W., & Yang, H. (2020). **"Dynamic Graph Representation Learning via Self-Attention Networks."** In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 6039-6046).

DYSAT ARCHITECTURE

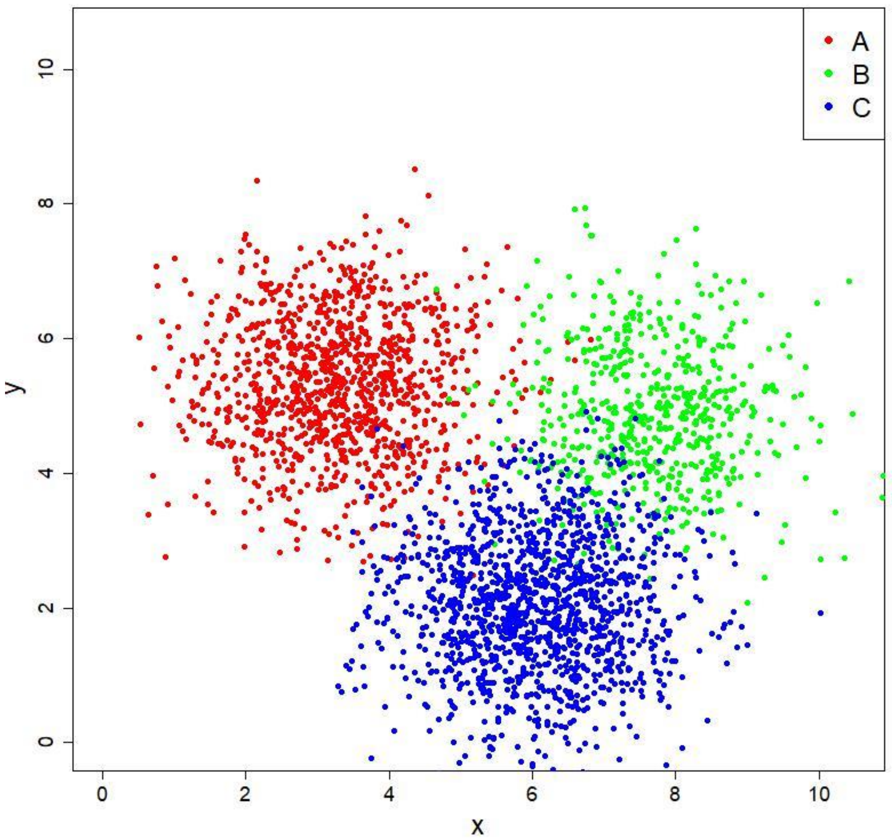


Sankar, A., Wu, Y., Gou, L., Zhang, W., & Yang, H. (2020). **"Dynamic Graph Representation Learning via Self-Attention Networks."** In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 6039-6046).

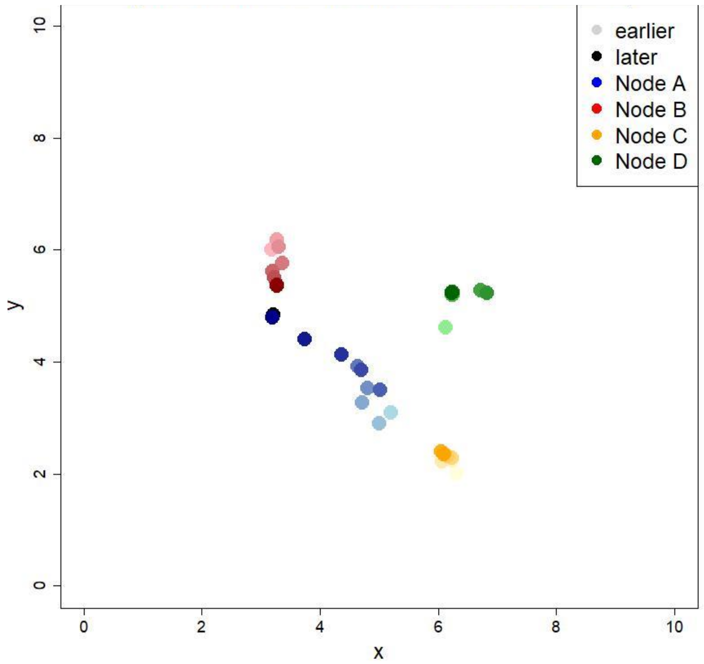
SIMULATED SOCIAL INTERACTION EXPERIMENT

- 3 communities with locations sampled in a distribution around their center
- 30 days, broken into 10 ESSD snapshots
- 100 nodes, each belonging to 1 or more communities
  - Nodes with shared communities have positive probability of check-in
- Probability of edges varies by
  - Number of communities nodes share
  - Average location proximity

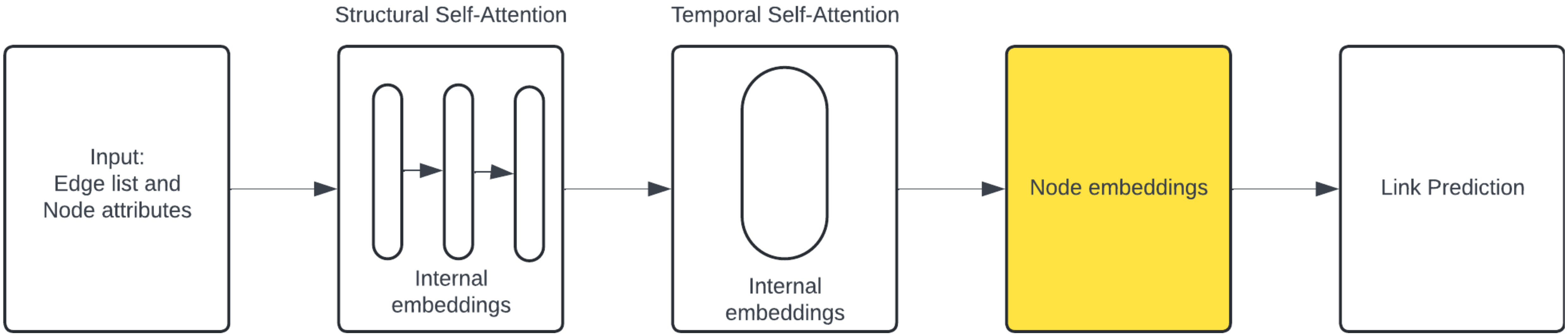
Social Check-Ins by Community



Moving Average Location for Ten Preceding Check-Ins



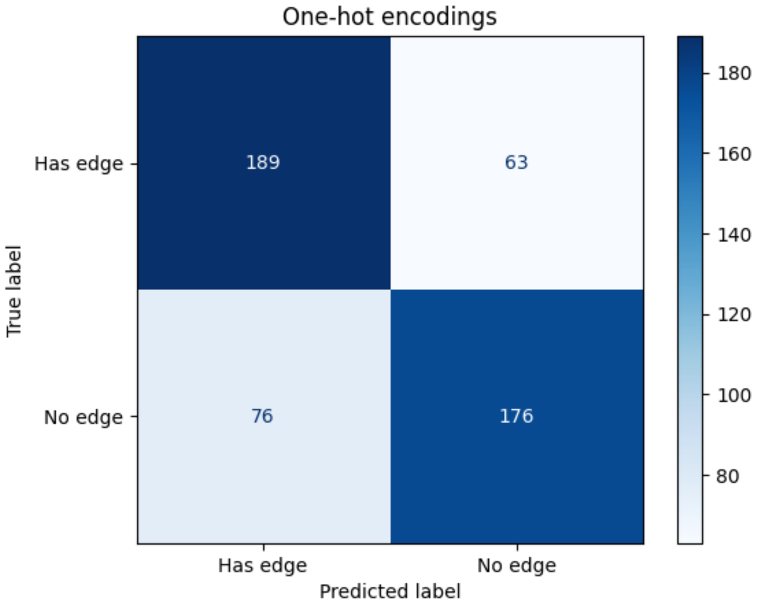
SIMULATED SOCIAL INTERACTION EXPERIMENT

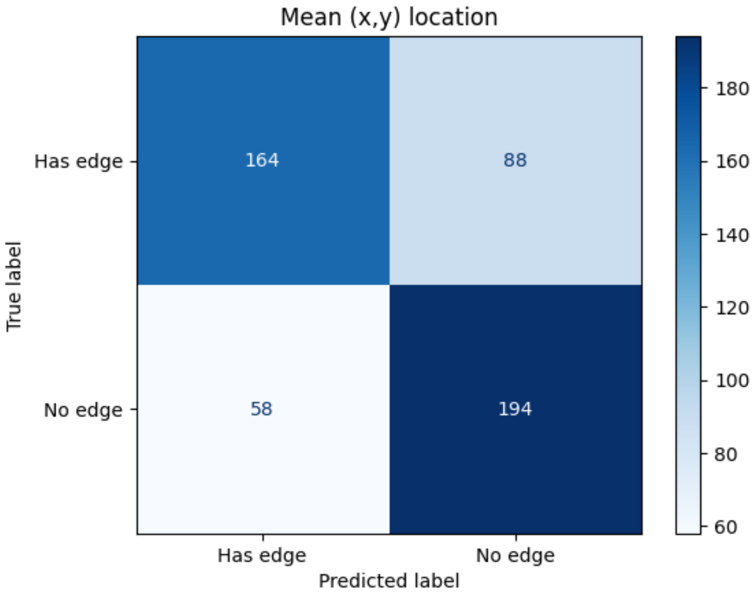


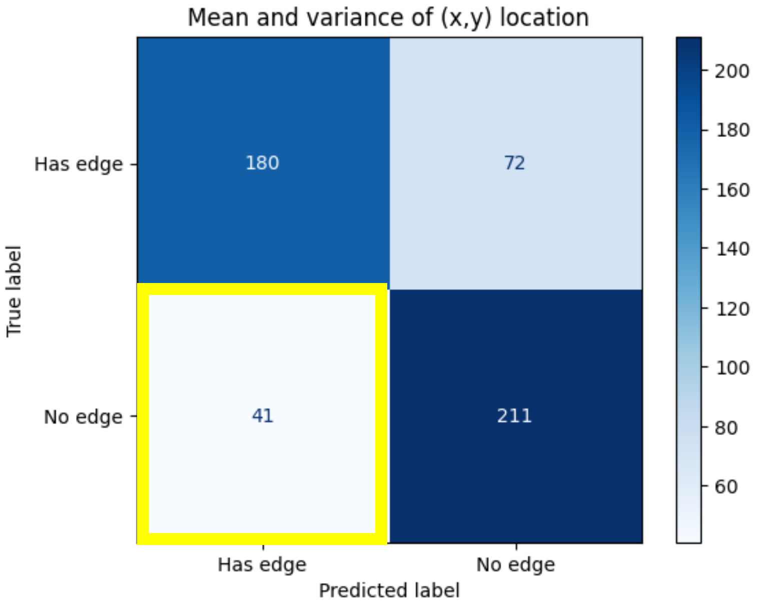
Node Attribute Modification Experiments

NODE ATTRIBUTE VARIATION AND MODEL PERFORMANCE

- How does modifying attributes impact model performance?
  - Number of false positive predictions decreases with addition of features which enhance underlying dataset dynamics

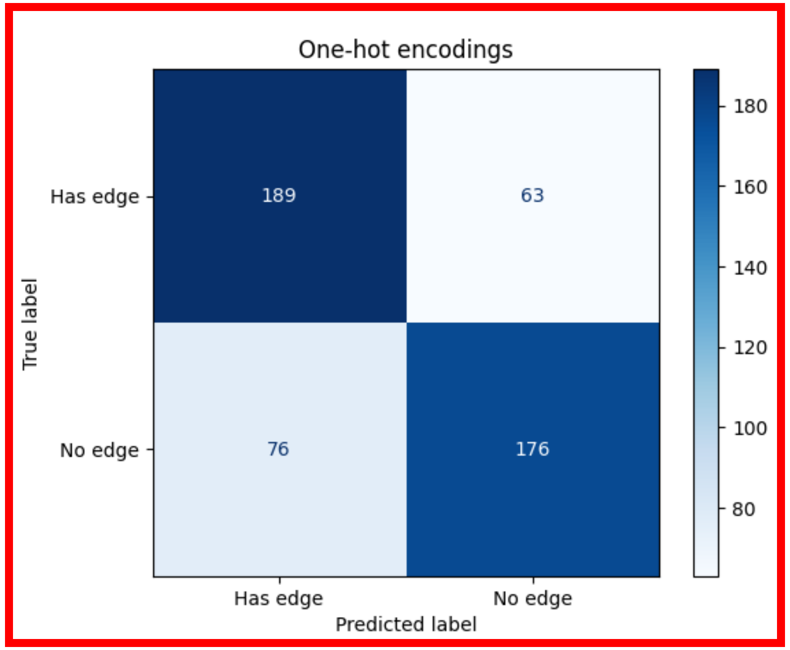

One-hot encodings


Mean (x,y) location

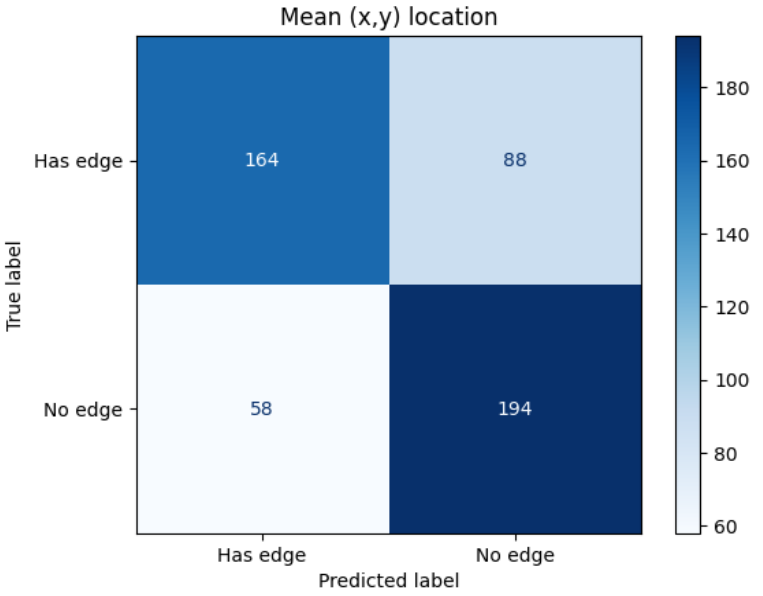

Mean and variance of (x,y) location

NODE ATTRIBUTE VARIATION AND MODEL PERFORMANCE

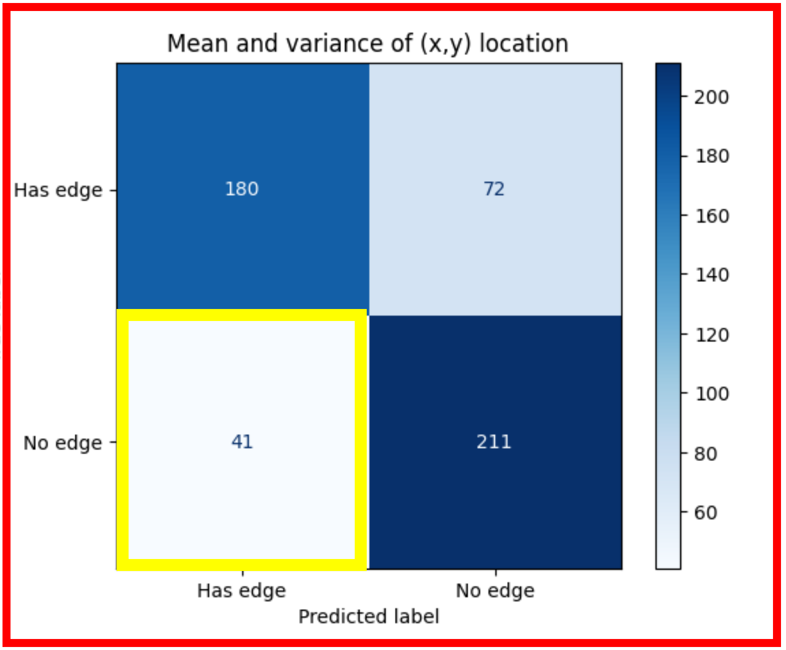
Number of false positive predictions decreases with addition of features which enhance underlying dataset dynamics



One-hot encodings



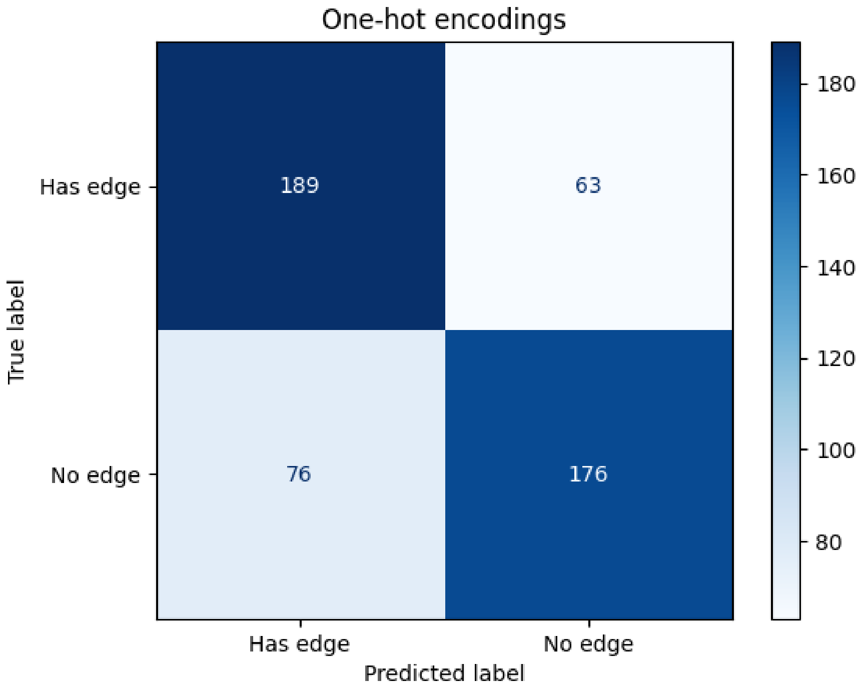
Mean (x,y) location

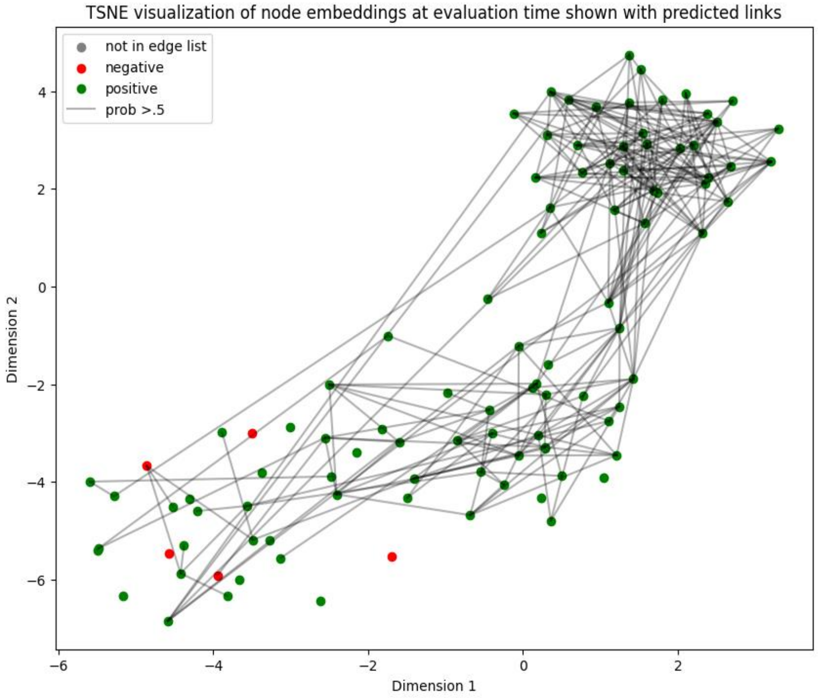


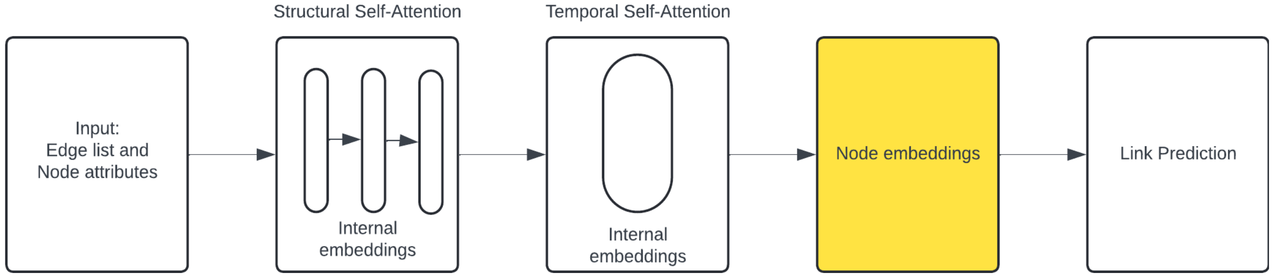
Mean and variance of (x,y) location

NODE ATTRIBUTE VARIATION AND LATENT REPRESENTATIONS

These embeddings are used by the model to predict link likelihood


One-hot encodings


TSNE visualization of node embeddings at evaluation time shown with predicted links


Structural Self-Attention

Temporal Self-Attention

Input:
Edge list and
Node attributes

Internal embeddings

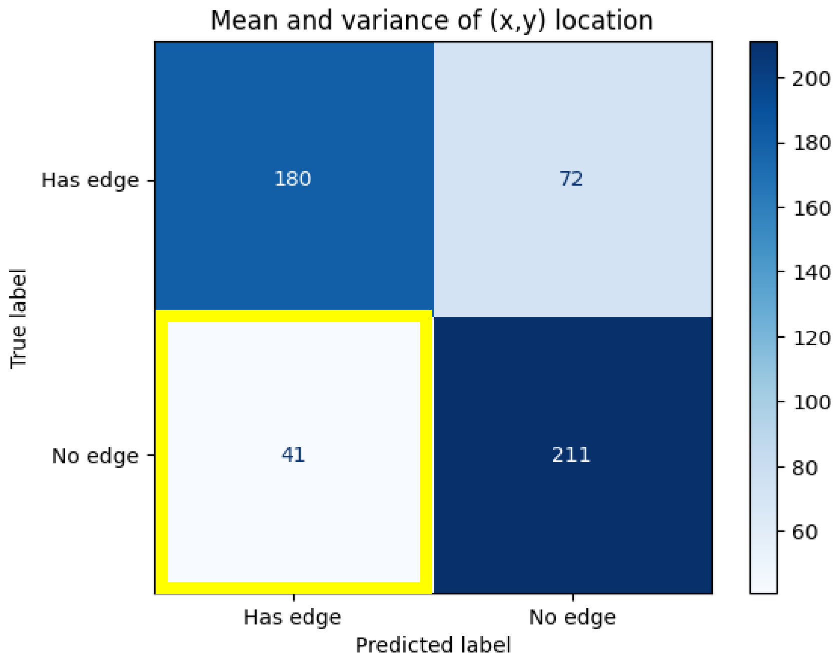
Internal embeddings

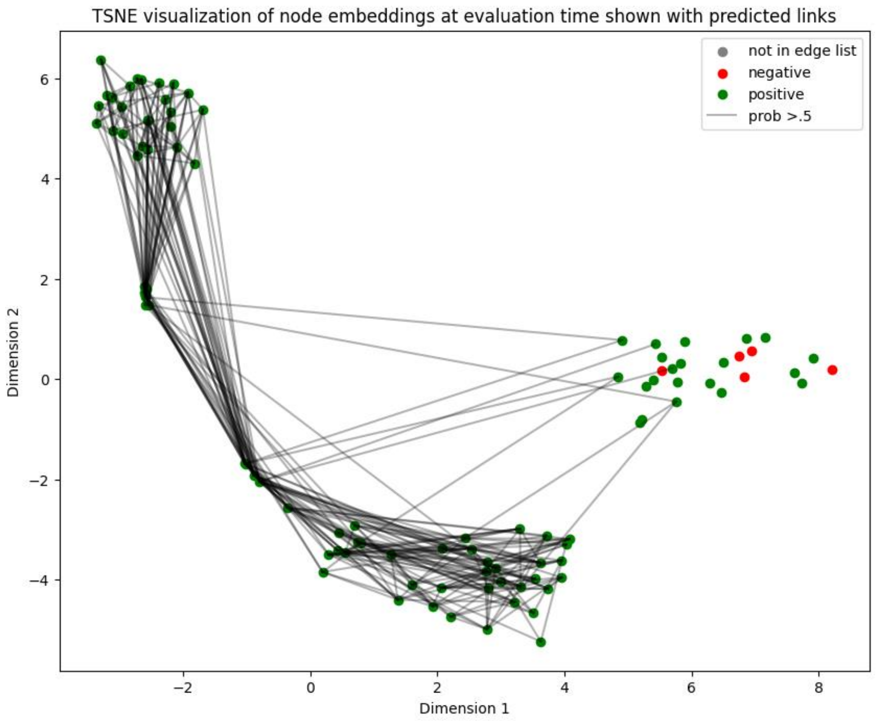
Node embeddings

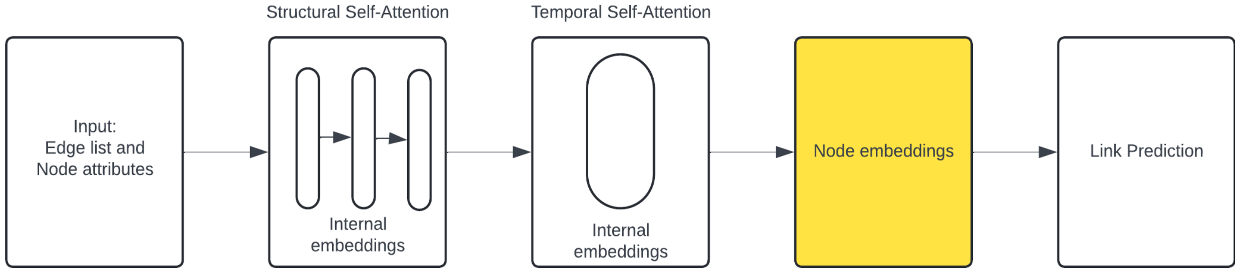
Link Prediction

NODE ATTRIBUTE VARIATION AND LATENT REPRESENTATIONS

The increased attribute richness decreases false positives, and creates a more complex and separable node embedding

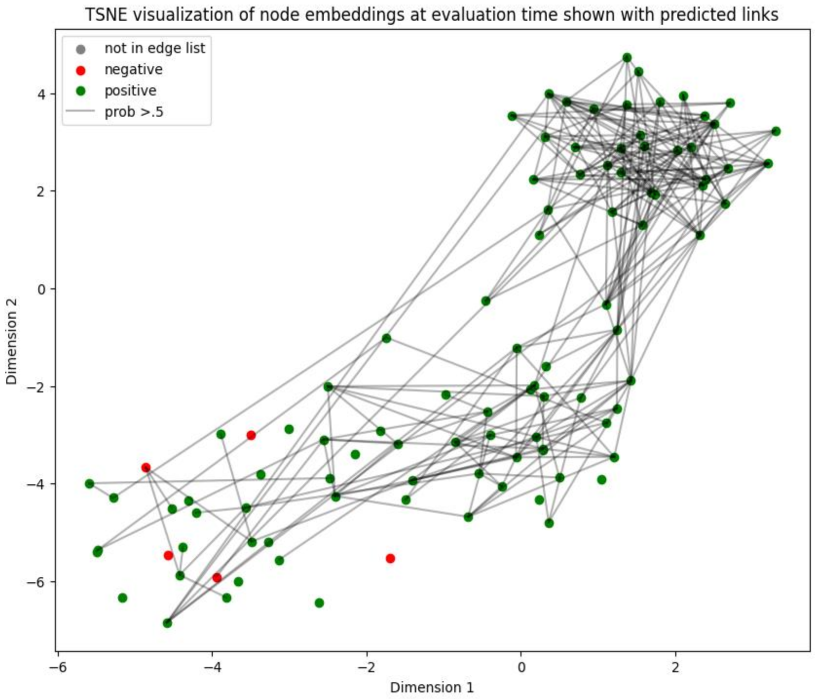


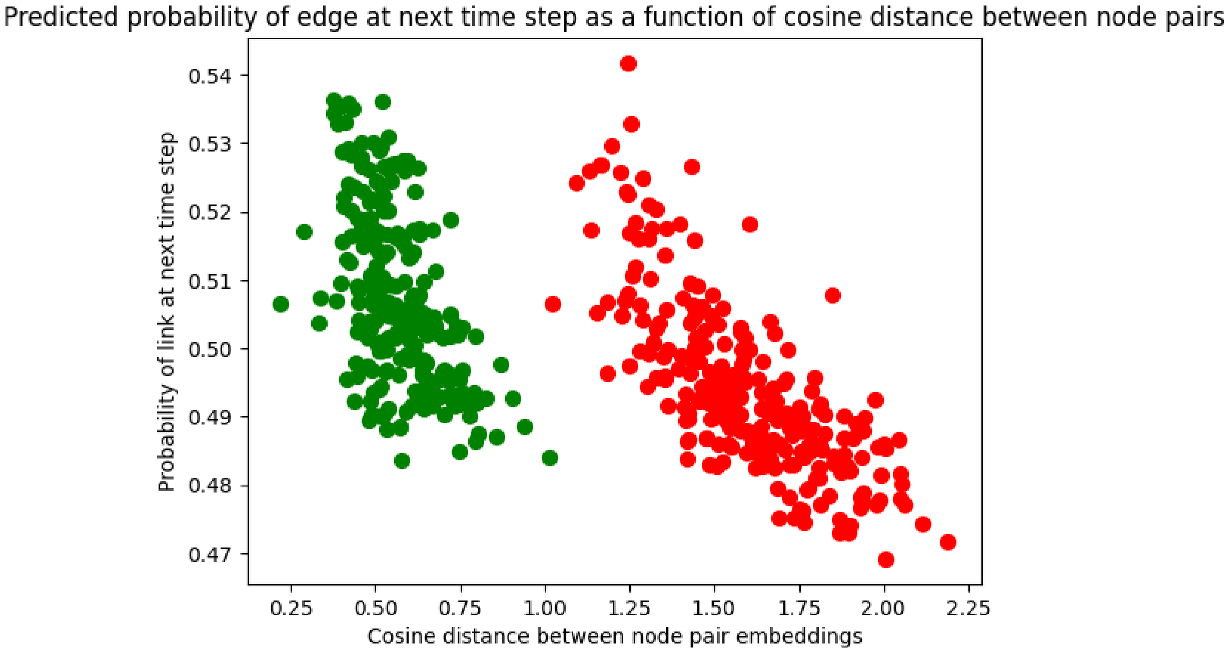


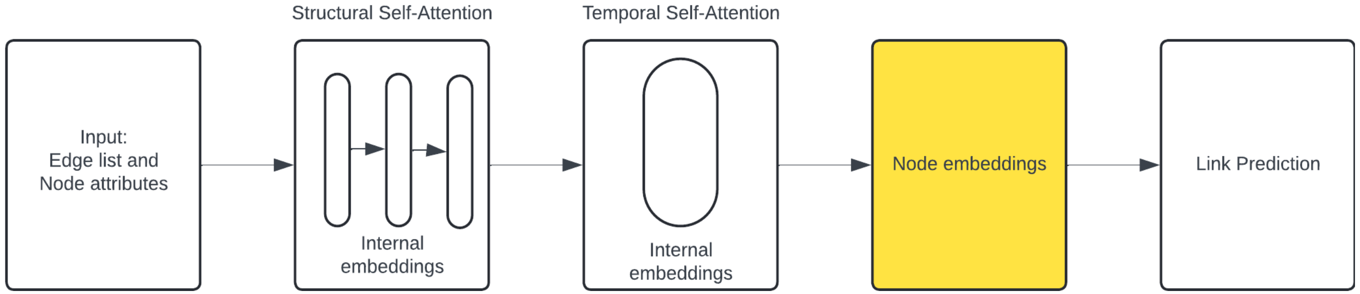


NODE ATTRIBUTE VARIATION AND LATENT REPRESENTATIONS

The mechanism that underlies link prediction in this case is heavily influenced by node proximity in the embeddings

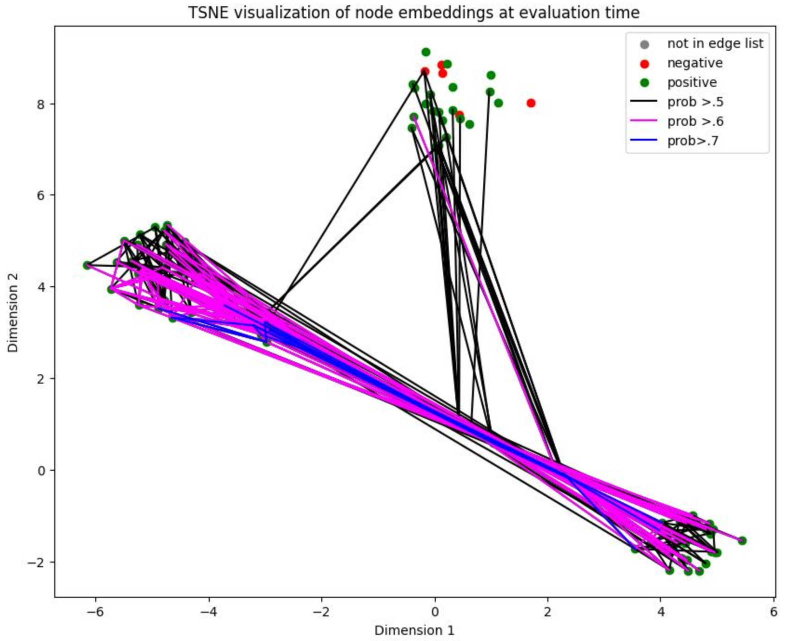


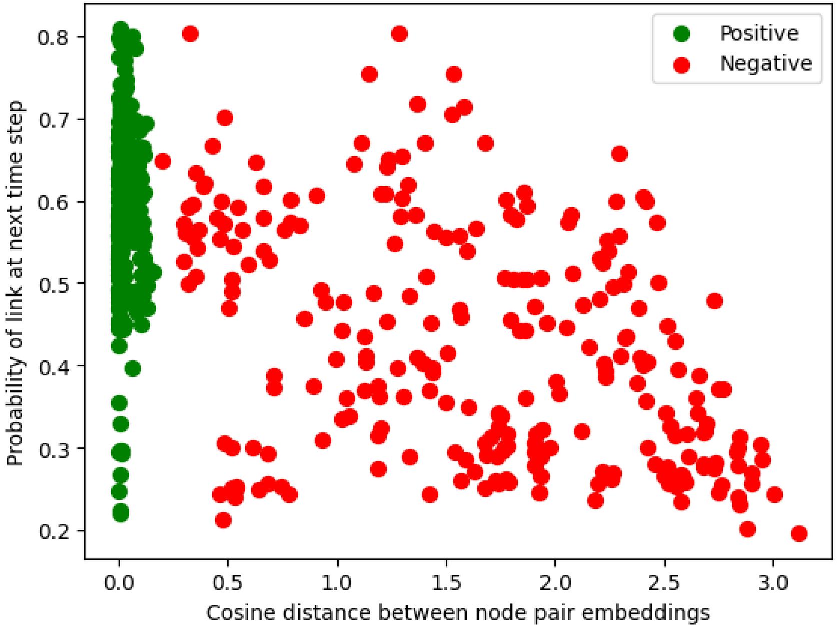


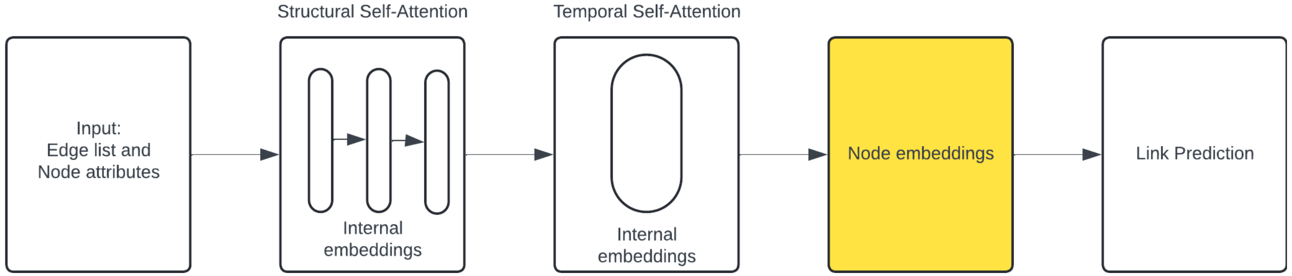


LATENT REPRESENTATIONS AND MODEL PREDICTIONS

Probability thresholds illustrate the complexity of this relationship






Input:
Edge list and
Node attributes

Structural Self-Attention

Internal
embeddings

Temporal Self-Attention

Internal
embeddings

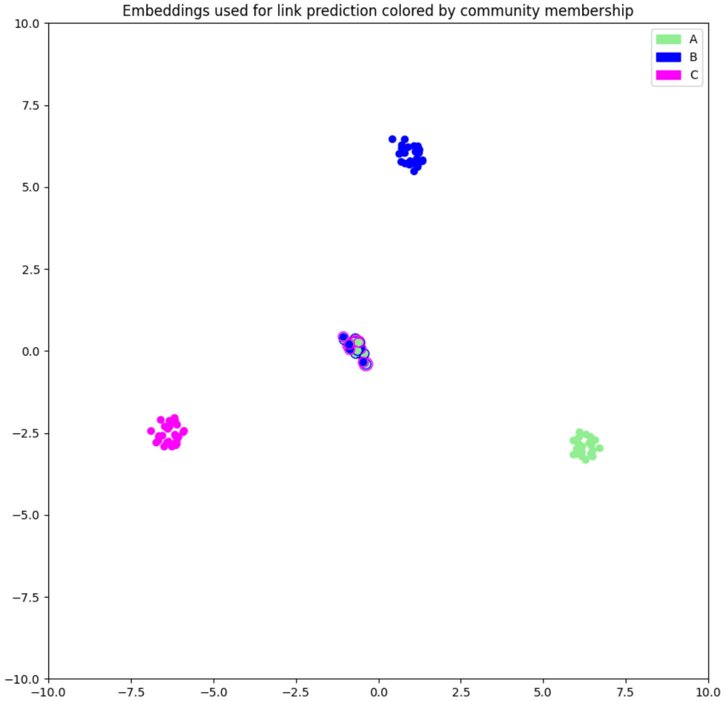
Node embeddings

Link Prediction

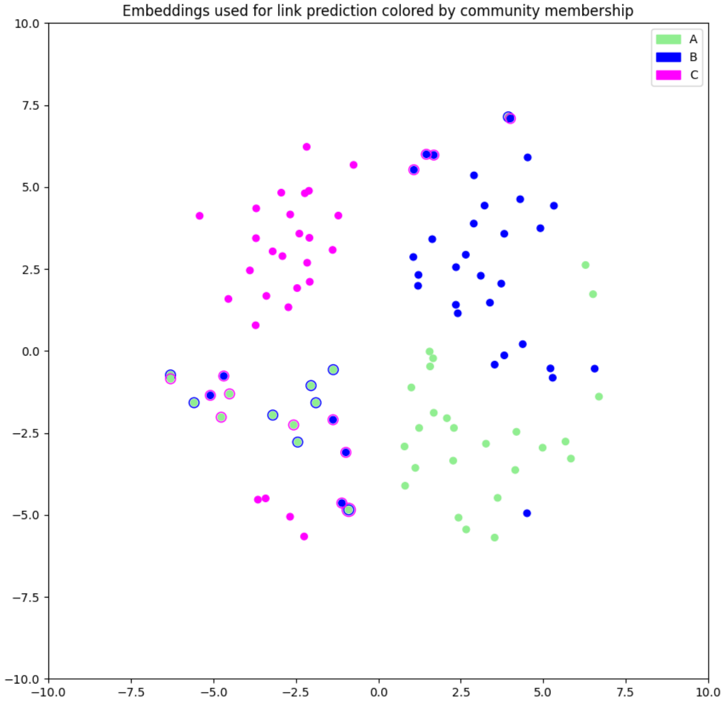
LESSONS LEARNED

- Addition of meaningful features increases separability in latent representations
- More features isn't always better
- Even in simple models it is difficult to communicate the mechanism of link prediction in terms of model inputs

Node community membership features, trained for **100 epochs**



Node community membership features, trained for **300 epochs**



Embeddings as a Lens for Causal Analysis

TWO PROPOSED METHODS FOR CAUSAL EXPLANATIONS FROM GRAPH EMBEDDINGS

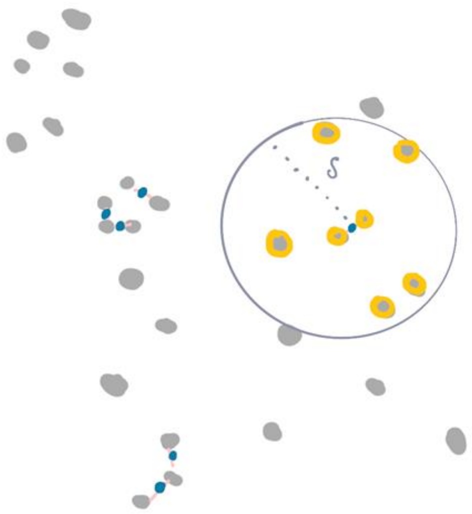
Define causal neighborhoods based on proximity in embedding space

$$D_{ij} = \|\mathbf{e}_i - \mathbf{e}_j\|_2$$

$$\text{Link}_{ij} = \begin{cases} 1 & \text{if } D_{ij} < \tau \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{m}_{ij} = \frac{\mathbf{e}_i + \mathbf{e}_j}{2}$$

$$\mathcal{N}_{ij} = \{k \mid \|\mathbf{e}_k - \mathbf{m}_{ij}\|_2 < \delta\}$$



- Goal: Provide analyst with a mechanistic metric that can be used to inform the context of predictions in terms of familiar model inputs

TWO PROPOSED METHODS FOR CAUSAL EXPLANATIONS FROM GRAPH EMBEDDINGS

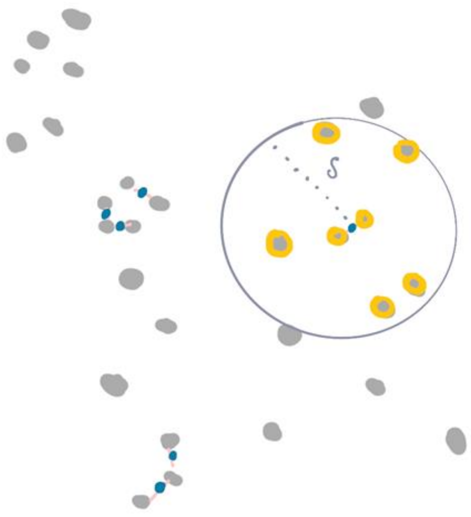
Define causal neighborhoods based on proximity in embedding space

$$D_{ij} = \|\mathbf{e}_i - \mathbf{e}_j\|_2$$

$$\text{Link}_{ij} = \begin{cases} 1 & \text{if } D_{ij} < \tau \\ 0 & \text{otherwise} \end{cases}$$

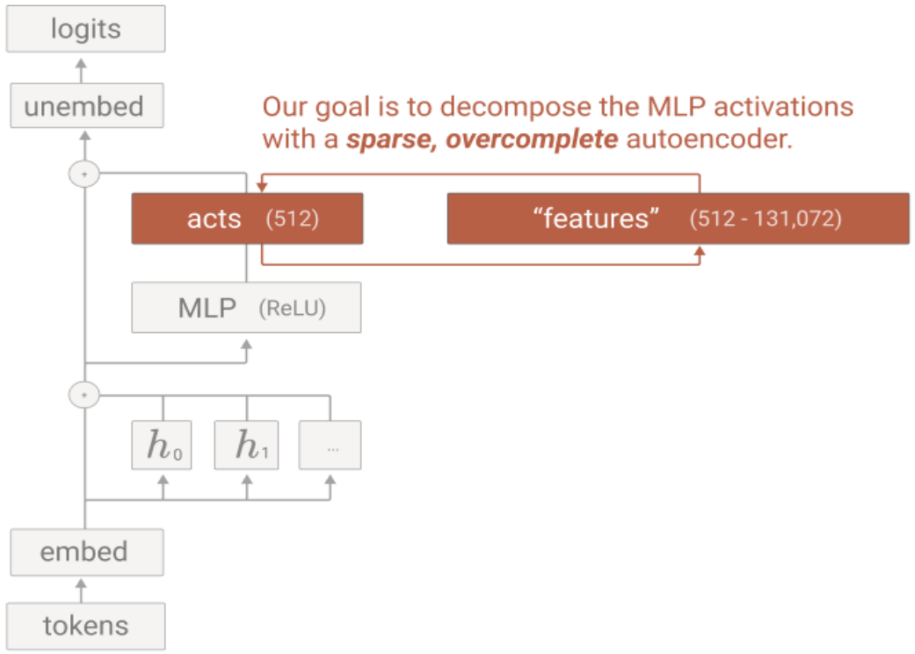
$$\mathbf{m}_{ij} = \frac{\mathbf{e}_i + \mathbf{e}_j}{2}$$

$$\mathcal{N}_{ij} = \{k \mid \|\mathbf{e}_k - \mathbf{m}_{ij}\|_2 < \delta\}$$



Generate causal diagrams via utilizing sparse autoencoder to highlight key model features influencing predictions

- Utility demonstrated by recent advances in language modeling



logits

unembed

Our goal is to decompose the MLP activations with a ***sparse, overcomplete*** autoencoder.

acts (512)

"features" (512 - 131,072)

MLP (ReLU)

h_0 h_1 ...

embed

tokens

https://transformer-circuits.pub/2023/monosemantic-features/index.html

TWO PROPOSED METHODS FOR CAUSAL EXPLANATIONS FROM GRAPH EMBEDDINGS

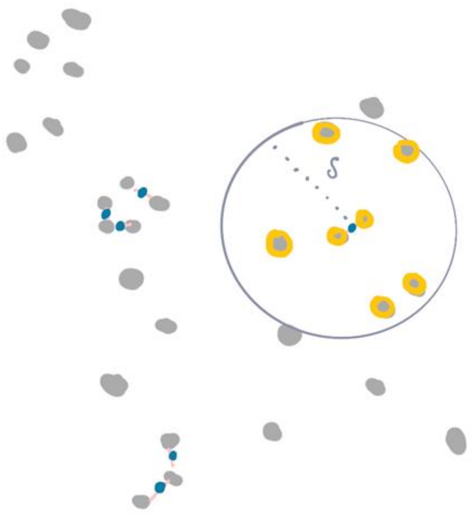
Define causal neighborhoods based on proximity in embedding space

$$D_{ij} = \|\mathbf{e}_i - \mathbf{e}_j\|_2$$

$$\text{Link}_{ij} = \begin{cases} 1 & \text{if } D_{ij} < \tau \\ 0 & \text{otherwise} \end{cases}$$

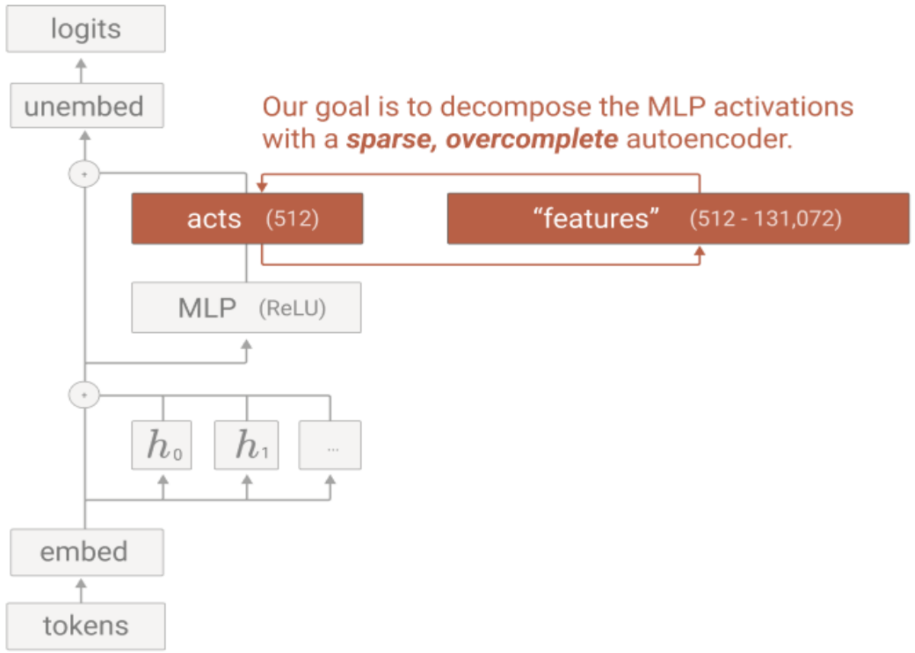
$$\mathbf{m}_{ij} = \frac{\mathbf{e}_i + \mathbf{e}_j}{2}$$

$$\mathcal{N}_{ij} = \{k \mid \|\mathbf{e}_k - \mathbf{m}_{ij}\|_2 < \delta\}$$



Generate causal diagrams via utilizing sparse autoencoder to highlight key model features influencing predictions

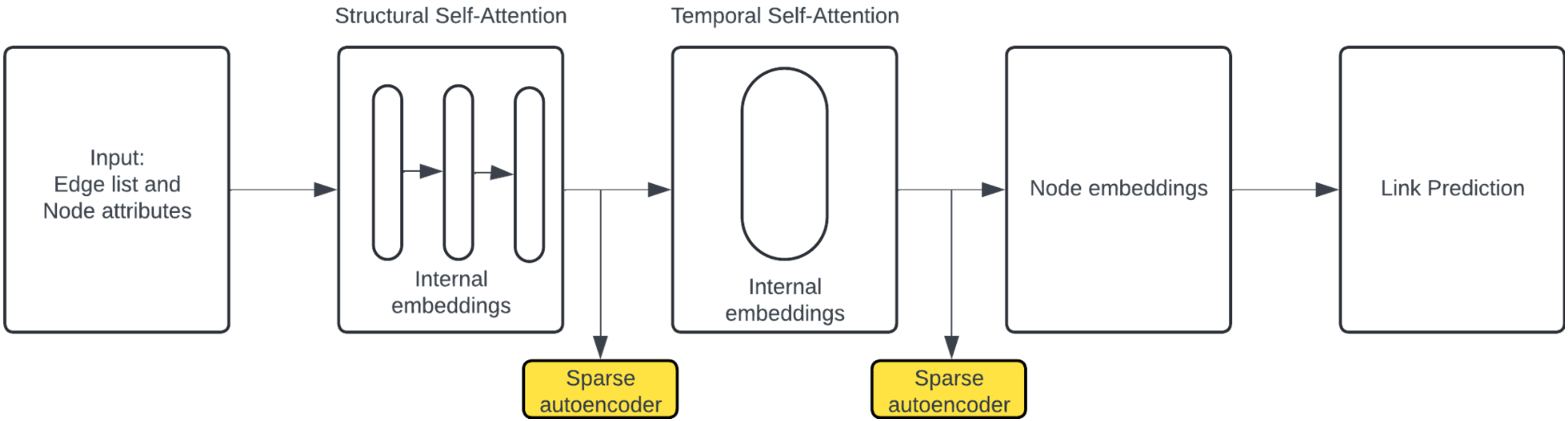
- Utility demonstrated by recent advances in language modeling



Our goal is to decompose the MLP activations with a ***sparse, overcomplete*** autoencoder.

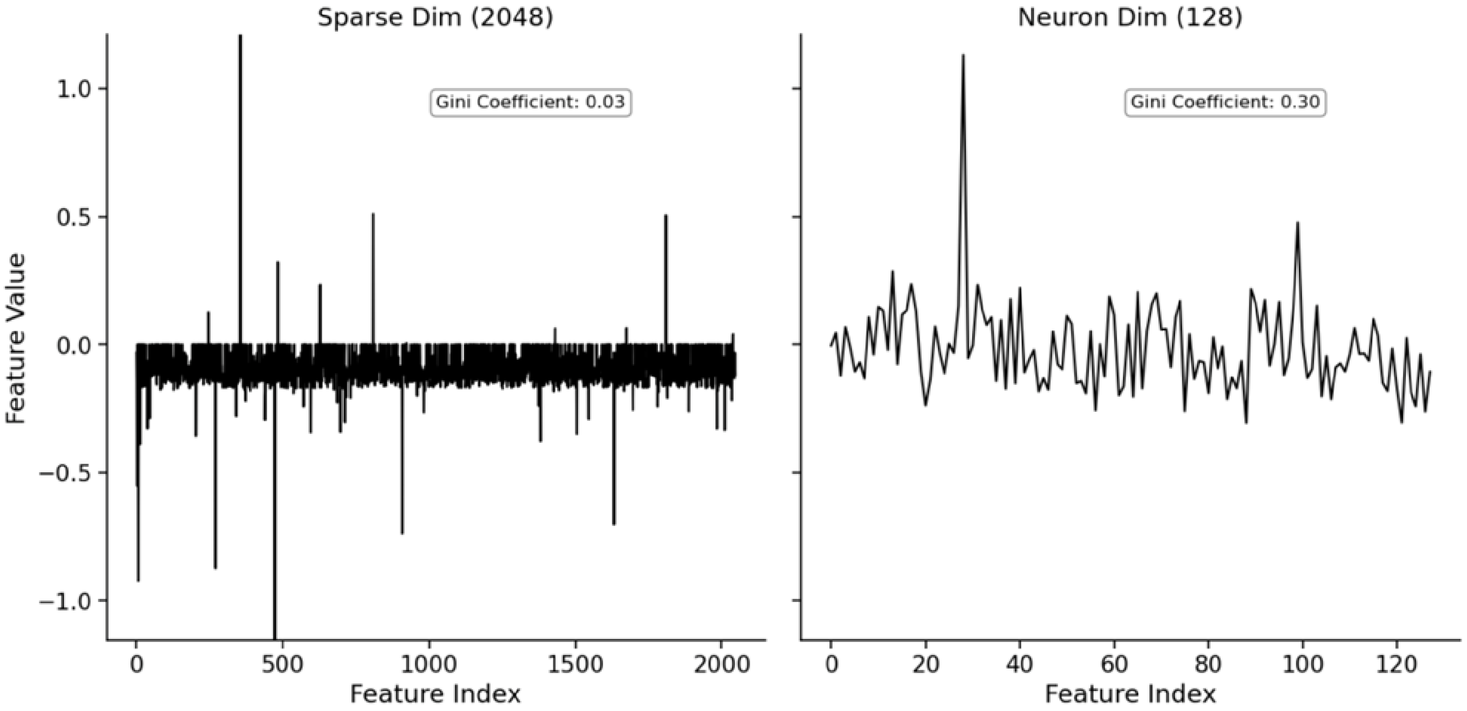
https://transformer-circuits.pub/2023/monosemantic-features/index.html

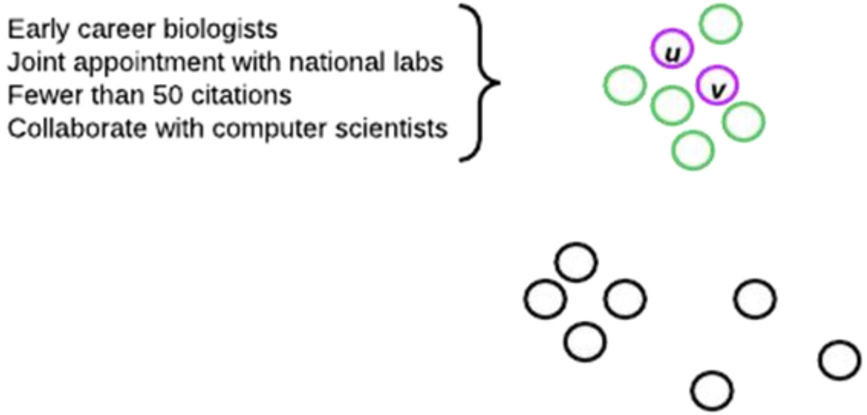
SPARSE AUTOENCODER APPROACH



LINK ATTRIBUTION

- Sparse and neuron dimensions provide different information
- Aim to generate more sparse representations to reduce the number of neurons responsible for placement in a given cluster




Early career biologists
Joint appointment with national labs
Fewer than 50 citations
Collaborate with computer scientists

Marks, Samuel, et al. "Sparse feature circuits: Discovering and editing interpretable causal graphs in language models." *arXiv preprint arXiv:2403.19647* (2024)

SUMMARY

- Graph representations are ubiquitous in modeling complex systems but lack causal explanation
- Graph attention networks achieve state-of-the-art performance while providing latent representations that can be leveraged to extract mechanistic interpretability
- Promising recent results in large language modeling provide a good starting point for generating causal explanations
- Causal explanations are only as good as the representations they came from

Questions?