TIGGER: Scalable Generative Modelling for Temporal Interaction Graphs

Shubham Gupta, Sahil Manchanda, Srikanta Bedathur, Sayan Ranu
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi
{shubham.gupta,sahil.manchanda,srikanta,sayanranu}@cse.iitd.ac.in

Abstract
There has been a recent surge in learning generative models for graphs. While impressive progress has been made on static graphs, work on generative modeling of temporal graphs is at a nascent stage with significant scope for improvement. First, existing generative models do not scale with either the time horizon or the number of nodes. Second, existing techniques are transductive in nature and thus do not facilitate knowledge transfer. Finally, due to their reliance on one-to-one node mapping from source to the generated graph, existing models leak node identity information and do not allow up-scaling/down-scaling the source graph size. In this paper, we bridge these gaps with a novel generative model called TIGGER. TIGGER derives its power through a combination of temporal point processes with auto-regressive modeling enabling both transductive and inductive variants. Through extensive experiments on real datasets, we establish TIGGER generates graphs of superior fidelity, while also being up to 3 orders of magnitude faster than the state-of-the-art.

Introduction and Related Work
Modelling and generating graphs find applications in various domains such as drug discovery (Hrinchuk, Popova, and Ginsburg 2020; Li, Zhang, and Liu 2018), anomaly detection (Ranu and Singh 2009), data augmentation (Bojchevski et al. 2018), and data privacy (Casas-Roma, Herrera-Joancomartí, and Torra 2017). Initial works on graph generative modelling relied on making prior assumptions about the graph structure. Examples include Erdős-Rényi (Karonski and Rucinski 1997) graphs, small-world models (Watts DJ 1998), and scale-free graphs (Albert and Barabasi 2002). Recently, learning-based algorithms have been developed that circumvent this limitation (You et al. 2018; Goyal, Jain, and Ranu 2020; Hrinchuk, Popova, and Ginsburg 2020; De Cao and Kipf 2018; Liao et al. 2019). Specifically, these algorithms directly learn the underlying hidden distribution of graph structures from training data.

Unfortunately, most of the learning-based generative models are limited to static graphs. In today’s world, there is an abundance of graphs that are temporal in nature. Examples include financial transactions (Kumar et al. 2016; Dal Pozzolo et al. 2018), online shopping (He and McAuley 2016), community interaction graphs like Reddit (Liu, Benson, and Charikar 2019), and user behaviour networks (Yang et al. 2013). The interactions (edges) between nodes in a temporal graph are timestamped and the structure of these graphs change with time. The key challenge in generative modelling is therefore to learn the rules that govern their evolution over the time horizon (Michail 2015).

TAGGEN (Zhou et al. 2020) models temporal graphs by converting them into equivalent static graphs by combining node-ids with each of their interaction edge timestamps, and connecting only those nodes in the resulting static graph that satisfy a specified temporal neighbourhood constraint. They perform random walks on this transformed graph, which are then modified using heuristic local operations to generate many synthetic random walks. Finally, the synthetic random walks that are classified by a discriminator as real random walks are collected and combined to construct the generated temporal graph. More recently, DYMOND (Zeno, La Fond, and Neville 2021) presented a non-neural, 3-node motif based approach for the same problem. They assume that each type of motif follows a time-independently exponentially distributed arrival rate and learn the parameters to fit the observed arrival rate.

These approaches suffer from the following limitations:

• **Weak Temporal Modelling:** DYMOND makes two key assumptions: first, the arrival rate of motifs is exponential; and second, the structural configuration of a motif remains the same throughout the time horizon being modeled on. Both these assumptions do not hold in practice – motifs themselves may evolve with time and could arrive with time-dependent rates. This leads to poor fidelity of structural and temporal properties of the generated graph. TAGGEN, on the other hand, does not model the graph evolution rate explicitly. It assumes that the timestamps in the input graph are discrete random variables prohibiting TAGGEN from generating new(unseen in source graph) timestamps. More critically, the generated graph duplicates a large portion of edges from the source graph – our experiments found up to 80% edge overlap between the generated and the source graph. While the design choices of TAGGEN generate graphs that exhibit high fidelity of graph structural and temporal interaction properties, unfortunately it achieves them by generating graphs that are largely indistinguishable from the source graph due to their poor modelling of interaction times.
• **Poor Scalability to Large Graphs:** Both TAGEN and DYMOND are limited to graphs where the number of nodes are less than $\approx 10000$ and the number of unique timestamps are below $\approx 200$. However, real graphs are not only of much larger size, but also grow with significantly high interaction frequency (Paranjape, Benson, and Leskovec 2017). In such scenarios, the key design choice of TAGEN to convert the temporal graph into a static graph, fails to scale to long time horizons since the number of nodes in the resulting static graph multiplies linearly with the number of timestamps. Further, TAGEN also requires the computation of the inverse of an $N' \times N'$ matrix, where $N'$ is the number of nodes in the equivalent static graph to impute node-node similarity. This leads to the quadratic increase in memory consumption and even higher cost of matrix inversion, thus making TAGEN not scalable. On the other hand, DYMOND has an $O(N^3T)$ complexity, where $N$ is the number of nodes and $T$ is the number of timestamps. In contrast, the complexity of the algorithm we propose is in $O(NM)$ for a graph with $N$ nodes and $M$ timestamped edges, and is independent of the time horizon length.

• **Lack of Inductive Modelling:** Inductivity allows transfer of knowledge to unseen graphs (Hamilton, Ying, and Leskovec 2017). In the context of graph generative modelling, inductive modelling is required to (1) upscale or downscale the source graph to a generated graph of a different size, and (2) prevent leakage of node-identity from the source graph. Both TAGEN and DYMOND rely on one-to-one mapping from source graph node ids to the generated graph and hence are non-inductive.

**Contributions:** The proposed generative model, TIGGER (Temporal Interaction Graph GEnerator), addresses the above mentioned gaps in existing literature through the following novel contributions:

• **Assumption-free Modelling:** We utilize intensity-free temporal point processes (TPP) (Shchur, Bilots, and Günneemann 2020) to jointly model the underlying distribution of node interactions and their timestamps through temporal random walks. Our modelling of time is assumption-free as we fit a continuous distribution over time. This allows TIGGER to generate timestamps that were not even present in the input graph. Moreover, this empowers TIGGER to sample interaction graphs for future-timestamps. Thus, TIGGER is capable of upsampling/down-sampling in the temporal dimension.

• **Inductive Modelling:** TIGGER supports inductive modelling through a novel multi-mode decoder that learns the distribution over node embeddings instead of learning distribution over node IDs. In addition, through the usage of a WGAN (Arjovsky, Chintala, and Bottou 2017), we support up-sampling/down-sampling of generated graph size. Thus, in contrast to DYMOND and TAGEN, TIGGER is capable of generating graphs of arbitrary sizes without leaking information from the source graph – potentially useful in many privacy-sensitive applications.

• **Large-scale Empirical Evaluation:** Extensive evaluation over five large, real temporal graphs with up to 10 millions of timestamps comprehensively establishes that TIGGER breaks new ground in terms of its scalability, while also ensuring superior fidelity of structural and temporal properties of the generated graph.

**Problem Formulation**

**Definition 1** (Temporal Interaction Graph). A temporal interaction graph is defined as $G = (V, E)$ where $V$ is a set of $N$ nodes and $E$ is a set of $M$ temporal edges $\{ (u, v, t) \mid u, v \in V, t \in [0, T] \}$. $T$ is the maximum time of interaction.

**Problem 1** (Temporal Interaction Graph Generator).

**Input:** A temporal interaction graph $G$.

**Output:** Let there be a hidden joint distribution of structural and temporal properties from which given $G$ has been sampled. Our goal is to learn this hidden distribution. Towards that end, we want to learn a generative model $p(G)$ that maximizes the likelihood of generating $G$. This generative model, in turn, can be used to generate new graphs that come from the same distribution as $G$, but not $G$ itself.

The above problem formulation is motivated by the one-shot generative modelling paradigm i.e., it only requires one temporal graph $G$ to learn the hidden joint distribution of structural and temporal interaction graph properties. Defining the joint distribution of temporal and structural properties is hard. In general, these properties are characterized by inter-interaction time distribution and evolution of static graph properties like degree distribution, power law exponent, no. of connected components, largest connected component, distribution of pair wise shortest distances, closeness centrality etc. Typically, a generative model optimizes over one of these properties under the assumption that the remaining properties are correlated and hence would be implicitly modeled. For example, DYMOND uses small structural motifs and TAGEN uses random walks over the transformed static graph. In our work, we perform temporal random walks, which are then modeled using point processes.

**Training Data: Temporal Random Walks**

**Definition 2** (Temporal Neighborhood). The temporal neighbourhood of a node $v$ at time $t$ contains all edges with a higher time stamp. Formally,

\[ N_t(v) = \{ e \mid e = (v, u, t') \in E \land t < t' \} \]

**Definition 3** (Temporal Random Walk). Given a node $v$ and time $t$, an $\ell$-length temporal random walk starts from $v$ and takes $\ell$ jumps through an edge in the temporal neighborhood of the current node. More formally, it is a sequence of tuples $S = \{(s_1, \ldots, s_{\ell})\}$, where each tuple $s \in S$ is a $(node, time)$ pair such that, $s_1 = (v, t)$ and for every $i \in [2, \ell]$ the edge $(s_{i-1}.v, s_{i}.v, s_{i}.t) \in N_{s_{i-1}.t}(s_{i-1}.v)$. A walk ends after taking $\ell$ jumps or if $N_{s_{\ell-1}.t}(s_{\ell-1}.v) = \emptyset$.
Since each jump is constrained to edges within the temporal neighborhood, it is guaranteed that $s_i.t > s_{i−1}.t$. To capture the temporal characteristics, the probability of jumping through edge $e \in \mathcal{N}_{s_i}(v)$ decreases exponentially with time gap from $(s_i.t)$. More formally, 

$$p(e = (s_i.v, u, t) | s_i) = \frac{\exp(s_i.t − t)}{\sum_{e'=(s_i, v', u', t') \in \mathcal{N}_{s_i}(v, t')} \exp(s_i.t − t')}$$

Note that $s_i.t < t$ and hence a smaller gap leads to increased chances of being sampled. While we exponentiate the time gap, other functions, such as linear, may also be used. We use exponentiation due to superior empirical results. A random walk starts from an edge chosen uniformly at random. Examples of temporal random walks are shown in the Data Generation component of Fig. 1. In Table 3 in appendix, we summarise all the notations used in our work. As per convention, we use boldface symbols to denote learnable vectors and weight matrices.

**Modelling Temporal Random Walks**

We train a generative model $p(S)$ on a set $S$ of temporal random walks. Formally, 

$$p(S) = \prod_{S \in \mathcal{S}} p(S)$$

where, $p(S) = p(s_1, \ldots, s_t)$

(1)

Owing to the auto-regressive nature of a sequence, we express $p(S)$ as the product of the conditionals.

$$p(S) = p(s_1) \prod_{i=2}^{t} p(s_i | (s_1, \ldots, s_{i−1}))$$

We simplify this conditional by decomposing as follows.

$$p(S) = p(s_1) \prod_{i=2}^{t} p(s_i | (s_1, \ldots, s_{i−1})) \times p(s_i, t | (s_1, v, (s_1, \ldots, s_{i−1})))$$

(2)

(3)

To learn the above conditional distribution, we utilize a recurrent neural network (RNN) based generator. Formally, 

$$h_1 = r_{\text{hidden}} \cdot h_1 \cdot (s_1, \ldots, s_{i−1}) = r_{\text{hidden}} \cdot h_1 \cdot (s_1, s_{i−1})$$

$$o_i = r_{\text{output}} \cdot o_i \cdot (h_1, s_1, \ldots, s_{i−1}) = r_{\text{output}} \cdot o_i \cdot (h_1, s_{i−1})$$

Here, $r_{\text{output}} \cdot o_i \cdot (h_1, x)$ is the output of RNN cell and $r_{\text{output}} \cdot o_i \cdot (h_1, x)$ is the updated hidden state. Both $h_1$ and $o_i$ are vectors and we initialize $h_1 = 0$. Semantically, $o_i$ captures the prior to predict the next node $v_i$ in the temporal random walk. More formally, $p(S)$ in Eq. (3) is re-written as:

$$p(S) = p(s_1) \prod_{i=2}^{t} p(s_i | o_i) \ast p(s_i, t | s_i, v, o_i)$$

(4)

In the following sections, we discuss the internals of the RNN, and formulate how exactly Eq. (4) is learned. We develop two procedures: first is a transductive learning algorithm, and the second is an inductive model.

**Transductive Recurrent Generative Model:** Given a sequence of node and time pairs $s_1, \ldots, s_t$, we transform $s_i.v$ and $s_i.t$ to vector representations $f_i(s_i.v, s_i.t) \in \mathcal{R}^{d_v} \ast \mathcal{R}^{d_t}$ respectively.

First, we transform the node ids to a vector using $f(v) = W_v v \in \mathcal{V}$ where $W_v \in \mathcal{R}^{d_v} \ast \mathcal{R}^{d_v}$ is a learnable weight matrix and $v \in \mathcal{R}^{d_v}$ is one-hot encoding of the node ID of $v$. Next, to learn vector representation of time $t \in \mathcal{R}$, we use following Time2Vec (Kazemi et al. 2019) transformation.

$$f_i(t)[v] = \begin{cases} \omega_r \cdot t + \zeta_r, & \text{if } r = 0 \\ \sin(\omega_r \cdot t + \zeta_r), & -1 \leq r < \mathcal{d}_r \end{cases}$$

(5)

where $\omega_1, \omega_2, \ldots, \omega_{d_{fr}}, \zeta_2, \ldots, \zeta_{d_r}$ are trainable weights and shared across each pair of the input sequence. $r$ is the index of $f_i(t)$. After embedding both $s_{i−1}.v$ and $s_{i−1}.t$, we concatenate them resulting in a vector of $\mathcal{R}^{d_v + d_t}$ dimension. This vector is fed into the RNN cell along with $h_{i−1}$ which outputs $o_i$ and $h_i$. We represent $p(s_i.v | o_i)$ in Eq. (4) as multinomial distribution over $v \in \mathcal{V}$ parameterized by $\theta_v$.

$$p(s_i.v | v, o_i) = \theta_v(o_i)$$

$$\begin{align*}
&= \theta_v(r_{\text{output}} \cdot o_i \cdot (h_{i−1}, s_{i−1}.v, s_{i−1}.t))) \\
&= \theta_v(r_{\text{output}} \cdot o_i \cdot (h_{i−1}, f_i(s_{i−1}.v) \cdot f_i(s_{i−1}.v))) \\
&= \frac{\exp(W_{o_i} \cdot o_i)}{\sum_{v \in \mathcal{V}} \exp(W_{o_i} \cdot o_i)}
\end{align*}$$

(6)
where $W_v^O \in \mathbb{R}^{1 \times d_0}$, $\forall v \in V$ is a node-specific learnable weight vector. $d_0$ is the dimension of $o_v$.

Temporal point processes (TPP) are de-facto models for modeling distributions of continuous, inter-event time over discrete events in event sequences $\{(c_0, t_0), (c_1, t_1) \ldots (c_n, t_n)\}$. TPPs are generally defined using conditional intensity function $\lambda(t)$.

$$
\lambda(t) = \frac{p(t \mid H_n)}{1 - F(t \mid H_n)}
$$

Here, $p(t \mid H_n)$ is the probability distribution of next event time $t$ after observing events till time $t_n$. $F(t)$ is the cumulative probability distribution corresponding to $p$. $H_n$ is the summary of events till time $t_n$. $\lambda(t)$ is the expected number of events around infinitesimal interval $[t, t + dt]$ given the history before $t$. It results in following probability distribution $p$ for next event time.

$$
p(t \mid H_n) = \lambda(t) \exp(-\int_{t_n}^{t} \lambda(x) \, dx)
$$

Resulting log likelihood contains integral due to $p(t)$ which needs to be estimated using Monte-Carlo sampling (Mei and Eisner 2017) leading to high variance, unstable updates during training and high computation cost (Omi, Ueda, and Ahara 2019). Motivated by strong performance on event prediction task by (Shchur, Bilò, and Günemann 2020), we adopt their TPP formulation, which directly defines $p(t)$ as mixture of log normal distribution instead of deriving it from $\lambda(t)$. From Eq. 4

$$
p(s_i, t \mid s_i, v, o_i) = p(s_i, t - s_{i-1}, t \mid s_i, v, o_i) = \theta_i(\Delta t \mid s_i, v, o_i)
$$

$$
= \sum_{c=1}^{C} \phi_c^C \frac{1}{\Delta t \sigma_c^C} \exp\left(-\frac{(\log \Delta t - \mu_c^C)^2}{2(\sigma_c^C)^2}\right)
$$

(7)

where $\Delta t$ is time difference between $s_i$ and $s_{i-1}$, $p(t)$ is parameterized by $\theta_i$ and $\mu_c^C, \sigma_c^C, \phi_c^C$ are parameters of $\theta_i$.

$$
\mu_c^C = W_c^C(f^{c}(s_i, v) \parallel o_i), \quad \sigma_c^C = \exp(W_c^C(f^{c}(s_i, v) \parallel o_i))
$$

$$
\phi_c^C = \frac{\exp(W_c^{\phi C}(f^{c}(s_i, v) \parallel o_i))}{\sum_{j=1}^{C} \exp(W_c^{\phi C}(f^{c}(s_i, v) \parallel o_i))}
$$

Moreover, $C$ is no. of components in the log normal mixture distribution and $W_c^C, W_c^{\phi C} \in \mathbb{R}^{d_v + d_k}$, $\forall c \in \{1, C\}$. Note that every components’ learnable weights are shared across each time-stamp in the sequence.

**Training loss:** The loss over the set $S$ of temporal random walks is derived from Eqs. [3][4][5] and [7]. Specifically,

$$
\mathcal{L} = -\log(p(S)) = -\sum_{S \in \mathcal{S}} \log(p(S))
$$

$$
= -\sum_{S \in \mathcal{S}} \log p(s_1) \sum_{i=2}^{\mathcal{L}} \log(p(s_i \mid s_{i-1}, v, o_i))
$$

+ \log(p(s_1 \mid s_i, v, o_i))

In the above loss function, $p(s_1) = 1/|E|$ since the first edge is chosen uniformly at random. $p(s_i \mid s_{i-1}, v, o_i)$ and $p(s_i \mid s_i, v, o_i)$ are computed using Eq. 6 and Eq. 7 respectively. A pictorial summary of the training process is available in the training component of Fig. 1.

**Inductive Recurrent Generative Model:** The primary distinction between transductive and inductive generative models are the construction of node representation and the procedure of learning next node distribution given the past information in the sequence. In the transductive model, a node is represented by its ID $\{1, \ldots, N\}$ in the form of a one-hot vector. In the inductive model, we use a Graph Convolution Network (GCN) to embed nodes.

**Node Representations:** We first transform the input temporal graph $G = (V, E)$ to a static graph $G^{static} = (V, E^{static})$, where $E^{static} = \{(u, v) \mid \exists (u, v, t) \in E\}$. On $G^{static}$, we utilize GraphSAGE (Hamilton, Ying, and Leskovec 2017) to learn unsupervised structural node representations. Details can be found in the appendix.

We denote embedding of node $v$ as $v \in \mathbb{R}^{d_v}$. Given a temporal walk sequence $S = (s_1, \ldots, s_t)$, we replace $s_i, v$ with $s_i, v$ $\forall i \in \{1, \ldots, \ell\}$. Similar to the transductive variant, in order to learn $p(s_i \mid s_{i-1}, \ldots, s_1)$, each $s_1$, $s_i$ and $s_{i-1}$, $t$ is transformed using $f_c(s_i, v)$ where $W_c$ $\in \mathbb{R}^{d_v \times d_v}$ and $f_c$ using Eq. 5. Both $f_c(s_i, v)$ and $f_c(s_{i-1}, t)$ are concatenated, which is fed into the RNN cell along with the previous hidden state $h_{i-1}$. The RNN outputs $o_i \in \mathbb{R}^{d_o}$ and $h_i \in \mathbb{R}^{d_h}$. These steps are the same as in the transductive variant.

**Multi-mode node embedding decoder:** Owing to working with node embeddings, the objective of the RNN is to predict the next node embedding instead of a node ID (in addition to the timestamp). Towards that end, we develop a multi-mode node embedding decoder. Fig. 2 presents the internals. The decoder has three distinct semantic phases. We explain them below.

We first note that node embeddings of a graph may not follow a uni-modal distribution since real-world graphs are known to have communities. The presence of communities would create a multi-modal distribution (Hamilton, Ying, and Leskovec 2017). To model this distribution, we perform $K$-means clustering on the node embeddings; each cluster would correspond to a community. The appropriate value of $K$ may be learned using any of the established mechanisms (Han, Pei, and Kamber 2011). Next, we design a multi-mode decoder that operates in two steps: first, it predicts the next node embedding instead of a node ID given the past embeddings. Then, it predicts the node embedding from that cluster.

Formally, we would like to the learn probability distribution $p(k_i = k \mid o_i)$, where $k_i$ denotes the cluster membership of next node $s_i, v$ and $k \in \{1, \ldots, K\}$. From this distribution, $k_i$ is sampled. Given cluster $k_i$, we next sample a vector $z$ from $p(z \mid o_i, k_i)$. Then, $s_i, v$ is sampled from $p(s_i \mid v, z)$. Since, we need to learn the distribution of $s_i, v$ given $o_i$ and $k_i$, we introduce a latent random variable $z$ in the multi-mode decoder. Mathematically,

$$
p(s_i, v \mid o_i) = p(k_i \mid o_i) \int p(z \mid o_i, k_i)p(s_i \mid v, z)dz
$$

$$
= p(k_i \mid o_i)E_{z \sim p(z,o_i, k_i)}[p(s_i \mid v, z)]
$$

where,
Training loss is derived from Eqs. 1, 4, 7 and 11 by substituting the regularizer on\footnote{Motivated by (Kingma and Welling 2014), which shows state-of-the-art empirical results on image generation tasks using L=1.} \(W_s\) used to learn the model parameters \(W\). Note that adding the regularizer on the temporal dimension.

We approximate the \(E\) term in Eq. 8 using the reparameterization trick from the auto-encoding variational bayes approach (Kingma and Welling 2014) by defining a deterministic function \(g\) to represent \(z \sim p(z | o_i, k_i)\) as follows:

\[
\begin{align*}
    z &= g(\mu_k, \sigma_k^2) = \mu_k^z + \varepsilon \sigma_k^z, \quad \varepsilon \sim \mathcal{N}(0, 1) \\
    \mathbb{E}_{z \sim p(z | o_i, k_i)}[p(s_i, v | z)] &= \mathbb{E}_{z \sim \mathcal{N}(0, 1)}[p(s_i, v | g(\mu_k, \sigma_k^2))] \\
    &\approx \frac{1}{T} \sum_{j=1}^{T} p(s_i, v | g(\mu_k, \sigma_k^2, \varepsilon_j)) \quad \varepsilon_j \sim \mathcal{N}(0, 1)
\end{align*}
\]

Taking the logarithm of Eq. 8, substituting the expectation term using Eq. 10 and assuming \(L=1\) we get the following:

\[
\begin{align*}
    \log p(s_i, v | o_i) &\gtrapprox \log p(k_i | o_i) + \log p(s_i, v | g(\mu_k, \sigma_k^2, \varepsilon)) \\
    p(s_i, t | s_i, v, o_i) &\text{is modelled the same as Eq. 8 except } f_t(s_i, v) \text{ which is replaced by } f_t(s_i, v) \text{ where } s_i, v \text{ is the vector representation of node } s_i, v.
\end{align*}
\]

Training loss is derived from Eqs. 4, 7 and 11 by substituting:

\[
\begin{align*}
    \mathcal{L} &\gtrapprox \sum_{s \in S} \log p(s_i) - \beta \mathcal{D}_{kl}(p(z | o_i, k_i) || \mathcal{N}(0, 1)),
\end{align*}
\]

where \(\mathcal{D}_{kl}\) is KL-Divergence. Empirical observations indicate that adding the regularizer on \(p(z)\) helps in reducing over-fitting. Thus, we have added the KL distance regularization on \(p(z)\) to restrict its sample space near to the distribution \(\mathcal{N}(0, 1)\). Here \(\beta \in (0, 1)\) is a hyper parameter, which decides the weightage of the regularizer term. \(\mathcal{L}\) is then used to learn the model parameters \(W, W^{K}_k, W^{Z}_k\) \(\forall k \in \{1 \ldots K\}, W^{Z}_{\phi}, W^{\sigma Z}\) and parameters of \(f_i, f_v, \text{rnn}_o\).

### Generating Interaction Graphs

Once the recurrent generative model is trained over the collection \(S\), we sample synthetic temporal random walks \(S'\) from the trained model. This synthetic collection is then assembled to form the synthetic temporal graph \(G'\). Similar to (Zhou et al. 2020), from each sequence \(S \in S\), we store the first item \(s_1\) and denote \(S_1\) as the collection of \(s_1\).

**Transductive model:** Alg. 2 explains a method to sample synthetic temporal random walks using transductive variant of TIGGER. Specifically, in Alg. 2, \(\Delta t\) in line 7 is sampled using below equation (Shchur, Biloš, and Gümenn̈2020).

\[
\phi \sim \text{Categorical}([\phi_1^C \ldots \phi_C^C])
\]

\[
\Delta t = \exp((\sigma^C)^T \phi + (\mu^C)^T \phi) \quad \varepsilon \sim \mathcal{N}(0, 1)
\]

Where \(\mu^C = (\mu_1^C \ldots \mu_C^C)\) and \(\sigma^C = (\sigma_1^C \ldots \sigma_C^C)\) and \(\phi\) is one-hot vector of size \(C\).

After collecting synthetic temporal random walks \(S'\), we assemble them by maintaining the same edge density as in the original graph within time range \(t \in [1, T]\). First, we count the frequency of each temporal occurrence in the synthetic random walks. We denote this as \(\alpha(v_i, v_j, t)\), i.e. the frequency of occurrence of node pair \((v_i, v_j)\) at time \(t\) in \(S'\).

We denote the set of edges present at time \(t\) in \(S'\) as \(E_t\). Now, for each uniquely sampled time stamp \(t \in [1, T]\), we define the distribution of occurrence on node pairs present at time \(t\) in synthetic temporal random walks \(S'\) as follows:

\[
p_{t|v_i, v_j} = \frac{\alpha(v_i, v_j, t)}{\sum_{(u_i, u_j) \in E_t} \alpha(u_i, u_j, t)}
\]
and hence allows control over the generated graph size. The pseudocode is provided in Alg. 2 in appendix. First, we train a WGAN (Arjovsky, Chintala, and Bottou 2017) generative model on node embeddings obtained from $G^{\text{static}}$ (Ji et al. 2021). From the trained WGAN model, we sample $N'$ node embeddings to construct $V'$. Finally, we match each embedding in $S'$ to its closest node in $V'$ using cosine similarity.

**Theorem 1.** The computation complexities of generating a temporal interaction graph $G_t = (V', E')$ through the transductive and inductive versions are $O(M \times |E| \times (N + C))$ and $O(M \times |E| \times (K + C + N'))$ respectively.

**Proof.** Provided in Appendix.

### Experiments

In this section, we benchmark TIGGER against DYMOND and TAGGEN and establish that it (1) it is up to 2000 times faster, (2) breaks new ground on scalability against number of timestamps, and (3) generates graphs of high fidelity. Our codebase and datasets are available at [https://github.com/data-iitd/tigger](https://github.com/data-iitd/tigger).

#### Experimental Setup

**Datasets:** For our empirical evaluation, we use the publicly available datasets listed in Table 1. Columns 2 to 4 of Table 1 summarize the sizes of the temporal interaction graphs. Our datasets span various domains including message exchange platform (UC Irvine) (Kunegis 2013a), financial network (Bitcoin) (Kumar et al. 2016), communication forum (Reddit) (Leskovec and Krevl 2014), shopping (Ta-feng) (Bai et al. 2018), and Wikipedia edits (Wiki) (Leskovec and Krevl 2014). Further details are provided in Table 5 in appendix. Since DYMOND and TAGGEN do not scale to graphs with large number of timestamps, we sample a smaller subset of Wiki by considering only the first 50 hours. This dataset is denoted as Wiki-Small.

**Baselines and Training:** We benchmark the performance of TIGGER against DYMOND and TAGGEN. For TIGGER, we denote the inductive version as TIGGER-I. To allow uniform comparison, in TIGGER-I, we generate graph of the same size as the source. For both TAGGEN and DYMOND, we use the code shared by authors. For all algorithms, the entire input graph is used for training and a single synthetic graph is generated. Parameter details along with machine configuration are provided in the appendix.

**Evaluation metrics:** The performance of a generative model is satisfactory if (1) it runs fast, (2) generates graphs with similar properties as in the source, (3) but without duplicating the source itself. To quantify these three objectives, we utilize the following metrics.

- **Efficiency:** Efficiency is measured through running time of the graph generation component.
- **Fidelity:** To quantify preservation of original graph properties, we compare various graph statistics of the snapshots of original graph $G_t$ and synthetic graph $G'_t$ for each unique timestamp $t \in \{1 \ldots T\}$. We use the following graph statistics (Kunegis 2013b): (i) mean degree, (ii) wedge count, (iii) triangle count, (iv) power law exponent of degree distribution (PLE), (v) relative edge distribution entropy, (vi) largest connected component size (LCC), (vii) number of components (NC), (viii) global clustering coefficient (CF), (ix) mean betweenness centrality (BC), (x) mean closeness centrality (CC). We explain these metrics in Table 6 in appendix. The error with respect to a given graph statistic $P$ is quantified as the median absolute error, that is, $\text{Median}_{t \in \{1 \ldots T\}}|P(G_t) - P(G'_t)|$. We use median instead of mean to reduce the impact of outliers. Nonetheless, the mean absolute errors (MAE) are also reported in the appendix.

- **Duplication:** To capture the level of duplication, we compute the percentage of overlapping edges, i.e., $\frac{|E \cap E'|}{|E|} \times 100$. Measuring duplication is important since an algorithm that duplicates the source graph would obtain perfect scores with respect to property preservation, although the generated graph is of limited use. We note that duplication has not been studied by TAGGEN or DYMOND.

#### Transductive: Comparison against Baselines

Table 1 presents the performance of all transductive algorithms across all metrics. We summarize the key observations below.

- **Efficiency and Scalability:** TIGGER is by far the most efficient of all models, while DYMOND is the slowest due to its $O(N^3 T)$ time complexity. TAGGEN is nearly 2 orders slower than TIGGER. In the inference phase, TAGGEN samples paths from the original graph, uses heuristics to modify these paths and then employs a discriminator to select from the generated paths. This process is prohibitively slow. Additionally, TAGGEN performs an expensive inversion of $N' \times N'$ matrix where $N'$ is number of unique pairs of nodes and their interaction timestamps in $G$. Consequently, TAGGEN fails to scale on Wiki and Reddit with millions of timestamps, and on Ta-feng which has much larger node and edge sets (see Table 1). Note that DYMOND fails to complete in all but Wiki-Small, the smallest dataset. TIGGER on the other hand is orders of magnitude faster, and can scale to large datasets, since it simply uses the trained RNN to sample paths, and generates the graph using these paths. In Fig. 3, we plot the growth of running time against the number of timestamps and graph size. As visible, TIGGER is not only faster, but also have a slower growth rate. For this experiment, we sample the desired number of timestamps/temporal edges from the Wiki dataset.

- **Duplication:** TAGGEN consistently duplicates ~80% of the original graph. Hence, the utility of TAGGEN as a graph generator is questionable. Both DYMOND and TIGGER do not suffer from this limitation.

**Fidelity:** From Table 1, we observe that TIGGER and TAGGEN achieve the best results in majority of graph statis-
Table 1: TIGGER’s performance against TAGGEN and DYMOND in terms of graph generation time (Col 6), edge duplication percent (Col 7), and median error across various graph statistics (Cols 8-17). For all performance metrics, lower values are better. For each statistic, we also list the Median value over original graph snapshots to better contextualize the error values. The best result in each dataset is in boldface. We do not report the results for an algorithm if it does not complete within 24 hours. Errors smaller than five decimal places are approximated to 0.

Table 2: Median errors across various graph statistics for inductive version. Each entry, row 3 onwards, denotes (median absolute error/median value of the corresponding original graph property across snapshots).

Scalability: TIGGER-I, is orders of magnitude faster than TAGGEN and DYMOND. However, TIGGER-I is 5–8 times slower than transductive TIGGER. This is unsurprising since TIGGER-I needs to perform nearest neighbor search in the inference phase on node embeddings. Additionally, TIGGER-I is challenging to train on large graphs due to its reliance on WGAN, which often fails to converge on large graphs. Hence, we have not reported results on full Wiki, Reddit and Bitcoin. Fig. 3 reveals that the growth rate of running time in TIGGER-I is similar to TIGGER.

Duplication: The edge-overlap of TIGGER-I is 0 across all benchmarked datasets, which is the ideal score.

Fidelity: The modelling task in inductive mode is inherently more difficult due to not having access to node IDs. Despite this challenge, we observe that the errors are low when compared to the median values of graph statistics in the original graph (Table 2). More importantly, despite being inductive, the errors are significantly better than DYMOND and comparable to TAGGEN and TIGGER (compare Tables 2 and 1).

This trend is also visible in Fig. 4.

Conclusion

The success of a temporal graph generative model rests on two key properties: (1) Scalability to large temporal graphs since real-world graphs are large, and (2) the ability to learn the underlying distribution of rules governing graph evolution rather than duplicating the training graph. Existing techniques fail to show the above desired behaviour. As established in our empirical evaluation, the proposed method, TIGGER, achieves the above desiderata. TIGGER derives its power through an innovative use of intensity-free temporal point processes to jointly model the node interaction times and the structural properties of the source graph. Additionally, we introduce an inductive version called TIGGER-I, which directly learns the distribution over node embeddings instead of node IDs. Future Work: The scalability of the inductive model is limited by its graph embeddings and the use of WGAN. Hence, we plan to explore mechanisms that address this limitation and eventually move towards a model that is inductive, scalable and accurate in terms of fidelity.
References

Albert, R.; and Barabási, A.-L. 2002. Statistical mechanics of complex networks. Reviews of modern physics, 74(1): 47.
Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In Precup, D.; and Teh, Y. W., eds., Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, 214–223. PMLR.
Bai, T.; Nie, J.-Y.; Zhao, W. X.; Zhu, Y.; Du, P.; and Wen, J.-R. 2018. An Attribute-Aware Neural Attentive Model for Next Basket Recommendation, 1201–1204. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356572.
Bojchevski, A.; Shchur, O.; Zügner, D.; and Günnemann, S. 2018. NetGAN: Generating Graphs via Random Walks. In Bojchevski, A.; Shchur, O.; Zügner, D.; and Günnemann, S., eds., Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, 610–619. PMLR.
Casas-Roma, J.; Herrera-Joancomartí, J.; and Torra, V. 2017. A Survey of Graph-Modification Techniques for Privacy-Preserving on Networks. Artif. Intell. Rev., 47(3): 341–366.
Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; and Bonetempi, G. 2018. Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8): 3784–3797.
De Cao, N.; and Kipf, T. 2018. MolGAN: An implicit generative model for small molecular graphs. ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models.
Goyal, N.; Jain, H. V.; and Ranu, S. 2020. GraphGen: a scalable approach to domain-agnostic labeled graph generation. In Proceedings of The Web Conference 2020, 1253–1263.
Guo, R.; Sun, P.; Lindgren, E.; Geng, Q.; Simcha, D.; Chern, F.; and Kumar, S. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In III, H. D.; and Singh, A., eds., Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, 3887–3896. PMLR.
Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, 1025–1035. Red Hook, NY, USA: Curran Associates Inc. ISBN 97815108060964.
Han, J.; Pei, J.; and Kamber, M. 2011. Data mining: concepts and techniques. Elsevier.
He, R.; and McAuley, J. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In Proceedings of the 25th International Conference on World Wide Web, WWW ’16, 507–517. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.
Hrinchuk, O.; Popova, M.; and Ginsburg, B. 2020. Correction of Automatic Speech Recognition with Transformer Sequence-To-Sequence Model. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, 7074–7078. IEEE.
Ji, Y.; Huang, R.; Chen, J.; and Xi, Y. 2021. Generating a Doppelganger Graph: Resembling but Distinct. ArXiv, abs/2101.09593.
Karoliński, M.; and Ruciński, A. 1997. The Origins of the Theory of Random Graphs, 311–336. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-60408-9.
Kazemi, S. M.; Goel, R.; Eghbali, S.; Ramanan, J.; Sahota, J.; Thakur, S.; Wu, S.; Smyth, C.; Poupard, P.; and Brubaker, M. A. 2019. Time2Vec: Learning a Vector Representation of Time. ArXiv, abs/1907.05321.
Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings.
Kumar, S.; Spezzano, F.; Subrahmanian, V.; and Faloutsos, C. 2016. Edge weight prediction in weighted signed networks. In Data Mining (ICDM), 2016 IEEE 16th International Conference on, 221–230. IEEE.
Kunegis, J. 2013a. KONECT: The Koblenz Network Collection. In Proceedings of the 22nd International Conference on World Wide Web, WWW ’13 Companion, 1343–1350. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320382.
Kunegis, J. 2013b. KONECT: The Koblenz Network Collection. In Proceedings of the 22nd International Conference on World Wide Web, WWW ’13 Companion, 1343–1350. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320382.
Leskovec, J.; and Krevl, A. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
Li, Y.; Zhang, L.; and Liu, Z. 2018. Multi-objective de novo drug design with conditional graph generative model. Journal of cheminformatics, 10(1): 1–24.
Liao, R.; Li, Y.; Song, Y.; Wang, S.; Hamilton, W. L.; Duvenaud, D.; Urtasun, R.; and Zemel, R. S. 2019. Efficient Graph Generation with Graph Recurrent Attention Networks. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 4257–4267.
Liu, P.; Benson, A. R.; and Charikar, M. 2019. Sampling methods for counting temporal motifs. In Proceedings of the ACM International Conference on Web Search and Data Mining.
Mei, H.; and Eisner, J. 2017. The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, 6757–6767. Red Hook, NY, USA: Curran Associates Inc. ISBN 97815108060964.
Appendices

Code optimizations

To increase the sampling efficiency during training, we restrict the $\mathcal{N}_t(v)$ by considering only the future $W$ edges. Moreover, we implement alias-table based sampling procedure [Walker 1977] which enables $O(1)$ sampling.

Attention based $p(S)$

$p(S)$ can be rewritten using attention [Vaswani et al. 2017] based generator too. In this case, the individual conditionals $p(s_i, v)$ and $p(s_i, t)$ will depend upon the whole sequence $\{s_1 \ldots s_{i-1}, s_{i+1} \ldots s_t\}$ and not only the hidden state. In order to perform future pair prediction with this model, future pairs need to be masked during training. Note that this is the same issue with bidirectional RNNs. Moreover, sequence length is less than 50 where LSTMs are known to perform well empirically. Hence, we are using LSTM as the $rnn_{th}$.

Computational Complexity

**Transductive:**

For constructing a temporal graph, we need to sample $O(M)$ temporal random walks. Each sampled temporal random walk is of length $\ell'$. Then, for each step we first pass the current node through a node embedding layer which is an $O(1)$ operation. Further, the time embedding layer of TIME2VEC takes $O(d_f)$ time. The concatenated node and time embedding is passed through an LSTM with hidden layer size $d_H$ which takes $O(d_H \times (d_V + d_T))$ time. The mixture model has $C$ components, and computing $\mu_c^S$, $\sigma_c^S$ and $\phi_c^S$ for each component $c$ requires a dot product operation which is $O(d_V + d_O)$. Predicting the next node is done by first computing a multinomial distribution over all nodes $v \in V$ and requires dot product operation of $O(d_O)$ which takes total time $O(N \times d_O)$. Combining all terms, the time complexity of the transductive model is $O(M \times \ell' \times (N \times d_O + C \times (d_V + d_O) + d_C \times (d_T + d_V)))$. Ignoring the dimension terms of weight matrices, the above expression simplifies as $O(M \times \ell' \times (N + C))$.

**Inductive:** Similar to transductive variant, we need to sample $O(M)$ temporal random walks for constructing a temporal graph. Each sampled temporal random walk is of length $\ell'$. Then, for each step, we first pass the current node through an MLP layer of hidden size $d_V$ which is $O(d_V \times d_V)$ operation. The time embedding layer of TIME2VEC takes $O(d_T)$ time. The concatenated node and time embedding is passed through an LSTM with hidden layer size $d_H$ which takes $O(d_H \times (d_V + d_T))$ time. Sampling the next cluster requires computing multinomial distribution over $K$ clusters which takes total time $O(K \times d_O)$). Predicting $\mu^K_c$ and $\sigma^K_c$ for each cluster $k \in K$ takes total time $O(K \times (d_O + d_Z))$. The mixture model has $C$ components, and computing $\mu^K_c$, $\sigma^K_c$ and $\phi^K_c$ for each component $c$ requires $O(d_V + d_O)$ time. The number of nodes in the generated graph is $N'$. The time required to perform nearest neighbor search at each step is $O(\ell')$.

Table 3: Notations used in the paper
\(O(N' \times d_V)\) where \(d_V\) is the dimension of the sampled node embeddings from WGAN.

Combining all terms above, the time complexity of the inductive model is \(O(M \times t' \times (K \times (d_O \times d_Z) + C \times (d_V + d_O) + (d_V)^2 + d_H \times (d_Y + d_V)) + N' \times d_V)\).

As an optimization for the nearest neighbour search, we use ScANN \(\text{Guo et al.}[2020]\) to perform faster approximate nearest neighbour search. The running time complexity of ScANN is \(O(qd + N')\) where \(q\) is the size of each quantization codebook, \(d\) is the dimension of vectors and \(N'\) is the number of vectors.

Simplifying the dimension terms as done earlier in transductive model, the total time complexity of inductive model \(\text{TIGGER-I}\) is \(O(M \times t' \times (K + C + N')).\)

### Datasets and Pre-processing

The semantics of the datasets are as follows:
- **UC Irvine messages**: It is a homogeneous graph of messages exchange between students of UC Irvine \(\text{Kunegis}[2013a]\).
- **Bitcoin alpha network**: It is a homogeneous financial transaction graph of bitcoin trading between users of bitcoin-alpha trading platform \(\text{Kumar et al.}[2016]\).
- **Reddit Interaction network**: Its a bipartite graph of users’ post on subreddits \(\text{Leskovec and Krevl}[2014]\). In the table \(\text{I}\)
- **Wiki Edit**: It a bipartite graph between human editors and Wikipedia pages \(\text{Leskovec and Krevl}[2014]\). Additionally, we curate a small Wiki-Small which corresponds to first 50 hours of wiki edit.
- **Ta-feng grocery shopping dataset** \(\text{Bai et al.}[2018]\): It is a bipartite graph of grocery shopping dataset spanning from November 2000 to February 2001.

**Data Pre-processing**: Apart from removing the duplicate interactions at same timestamps, we don’t perform any pre-processing on the dataset cited from the source.

| Dataset                                      | Source                                                      |
|----------------------------------------------|-------------------------------------------------------------|
| UC Irvine messages                           | http://konect.cc/networks/opsahl-ucsocial/                 |
| Bitcoin-alpha                                | http://snap.stanford.edu/data/social-sign-bitcoin-alpha.html|
| Reddit Interaction network                   | http://snap.stanford.edu/ca/w/                             |
| Wiki Edit network                            | http://snap.stanford.edu/ca/w/                             |
| Ta-feng grocery shopping network             | https://www.kaggle.com/chiranjivdas09/ta-feng-grocery-net/  |

Table 5: Sources of datasets

**Leskovec[2017]**. Please note that this method can produce similar embeddings for multiple nodes even having no edges between them. Hence, we follow boosting training approach as suggested \(\text{Ji et al.}[2021]\). After 1 round of training, we increase of weight of nodes in \(P_u(v)\) which contain false positive edge with node \(v\). We repeat this process, until the number of false positive edges comes down below to certain threshold.

### WGAN

We follow the similar training procedure as described in \(\text{Ji et al.}[2021]\). Given node embedding \(v \ni v \in G^{static}\), we initially remove the duplicate embeddings. Following this, we define a generator and critic based on 3 layer MLP. Finally, we optimize the WGAN value function by training generator for 1 epoch and critic for 4 epochs. We repeat this process until the convergence of loss. In order to avoid vanishing/explosion of gradients, we use WGAN along with gradient clipping. For training WGAN on GRAPHSAGE embeddings, we have used the code shared by \(\text{Ji et al.}[2021]\).

### Training and Parameter details

All experiments are performed on a machine running Intel Xeon E5-2698v4 processor with 64 cores, having 1 Nvidia 1080 Ti GPU card with 11GB GPU memory, and 376 GB RAM running Ubuntu 16.04.

We set the length of a temporal random walk \((\ell)\) to 20 during training. We note that during training, we expanded the node set \(V\) by adding an additional node \(end\_node\) to represent an empty temporal neighbourhood. We stop the generation of a temporal random walk if an \(end\_node\) is sampled as the next node or max length is reached during sampling procedure. We use 2 layer LSTM cell for \(run\_g\) and select \(d_V = 100\), \(d_T = 64\), \(d_O = 200\), \(C = 128\) and \(K = 300\). In TIGGER-I, we additionally set \(d_V = 128\) and \(d_Z = 128\). Both \(d_V\) and \(d_Z\) are constrained by the GRAPHSAGE embedding dimensions. To train both variants, we sample a single temporal random walk from every temporal edge of \(G\) thus collecting \(M\) temporal random walks. We assume 1

| Baseline         | Source                                           |
|------------------|--------------------------------------------------|
| TagGen           | https://github.com/davidchouzdw/TagGen            |
| DYMOND           | https://github.com/zeno129/DYMOND                |

Table 4: Sources of baseline implementation

### Node representation using GRAPHSAGE

For each node \(v\) in the network \(G^{static}\), a representation \(v\) is learnt by concatenating self information with information received from 1-hop neighbourhood by mean message passing. We utilize the following unsupervised loss on output representation \(v\) to learn the message-passing parameters.

\[
L = - \log(\sigma(v^T v_j)) - QE_{v_k \sim P_n(v)}( - v^T v_k)
\]

where \(v_j \ni \{ u \ni d(u, v) = 1\}\) and \(Q\) is number of negative samples and \(P_n(v)\) is probability distribution of negative nodes \(v_k \ni \{ u \ni d(u, v) \neq 1\}\). (Hamilton, Ying, and
walks. We re-sample $M$ training epoch as training over these $M$ mean $\pm$ represent the scale of the graph. Each value is in form of original graph statistic across timestamps. This is shown to in the method column represent the mean of corresponding each succeeding round of epoch. We set $\text{Mean}$ $\text{ness}$, we also report in Table 7 and 8 the Mean absolute error terms of median absolute error. For the sake of complete-ness, we set in the main paper, we have reported the performance in $\ell$ set divergence term as

\[ \text{Mean } (S_{t}) \approx ...T \]

\[ G_0 \] . During graph generation, we use the implementation provided by authors to $\text{Mean }$ $\text{Fidelity- Mean Errors}$

In the main paper, we have reported the performance in terms of median absolute error. For the sake of completeness, we also report in Table 7 and 8 the Mean absolute error i.e. $Mean_{e \in [1...T]} [P(G_t) - P(G_{t-1})]$ for all 5 datasets. Mean in the method column represent the mean of corresponding original graph statistic across timestamps. This is shown to represent the scale of the graph. Each value is in form of $\text{mean } \pm \text{ std. deviation.}$

| Metric                                      | Description                      |
|---------------------------------------------|----------------------------------|
| Mean degree                                 | Average of node degrees          |
| Wedge count                                 | Number of two hops path          |
| Triangle count                              | Number of triangles in the network |
| Power law exponent (PLE)                    | Exponent of power law distribution on the node degrees |
| Relative edge distribution entropy (RED entropy) | It measures the skewness of node degrees |
| Largest connected component size (LCC)      | Size of largest connected component in the network |
| Number of components (NC)                   | Number of connected component in the network |
| Global clustering coefficient (Global CF)   | It is computed as the fraction of number of closed triplets and number of all triplets. |
| Mean betweenness (BC)                       | Mean of each node’s betweenness centrality. Betweenness centrality of node $v$ is the fraction of all shortest paths which pass through $v$. |
| Mean Closeness centrality (CC)              | Mean of each node’s closeness centrality. Closeness centrality of node is the reciprocal of average shortest path distance to other reachable nodes. |

Table 6: Description of undirected graph properties

Algorithm 2: Sampling synthetic temporal random walks from a trained inductive recurrent generative model

Require: $S_t$, $f_\ell$, $f_i$, $\rho_\mu$, $\theta_k$, $W_k^{\omega}$, $W^{\omega}$ $\forall k \in \{1...K\}$, $\theta_i$, $W^{\omega}$ $\forall \omega$, $t'$
Ensure: $S'$
1: $S' = \{\}$
2: for $s_1 \in S_1$ do
3: $S' \leftarrow \{\}$, $(v_1, t_1) \leftarrow s_1$
4: $v_2 \leftarrow f_i(v_1)$, $t_2 \leftarrow f_\ell(t_1)$
5: $h_1 \leftarrow 0$
6: for $i \in \{2, 3, \ldots, t'\}$ do
7: \( o_i, h_i \leftarrow \text{rnn}_\mu(h_{i-1}, (v_{i-1} || t_{i-1})) \)
8: \( k_i \sim \text{Multinomial}(\theta_{k_1}(o_i), \theta_{k_2}(o_i) \ldots \theta_{j_K}(o_i)) \) \{Sample next cluster\}
9: \( z \sim \mathcal{N}(W_k^{\omega} o_i, \exp(W_k^{\omega} o_i)) \)
10: \( v_i \sim \mathcal{N}(W^{\omega} z, \exp(W^{\omega} z)) \) \{Sample next node embedding\}
11: \( \Delta t \sim \theta_t(t - t_{i-1} | v_i, o_i) \) \{Sample next time using eq 12\}
12: \( t_i = t_{i-1} + \Delta t \)
13: \( S' = S' + (v_{i-1}, v_i, t_i) \)
14: end for
15: \( S'' = S' + S' \)
16: end for
17: Return $S''$

Table 7: Mean absolute errors across various graph statistics for inductive version. Each entry denotes (Mean absolute error $\pm$ std. dev across snapshots).

| Metric                                      | Wiki-Small | UC Irvine | Bitcoin |
|---------------------------------------------|------------|-----------|---------|
| Mean degree                                 | 0.0701 ± 0.0063 | 0.2599 ± 0.2734 | 0.4392 ± 0.2607 |
| Wedge Count                                 | 12.2449 ± 15.3377 | 321.1081 ± 1099.1924 | 147.2024 ± 354.5897 |
| Triangle Count                              | 0.0816 ± 0.3403 | 2.5351 ± 7.6297 | 4.2362 ± 11.1031 |
| PLE                                         | 7.0495 ± 6.0917 | 2.0852 ± 3.8017 | 3.6262 ± 5.9867 |
| Edge Entropy                                | 0.0062 ± 0.0003 | 0.0019 ± 0.0243 | 0.0222 ± 0.0191 |
| LCC                                         | 3.0816 ± 4.5256 | 14.4054 ± 18.8709 | 21.9167 ± 19.9414 |
| Global CF                                   | 0.0066 ± 0.0027 | 0.0038 ± 0.2473 | 0.0273 ± 0.0577 |
| Mean BC                                     | 0.0001 ± 0.0003 | 0.0051 ± 0.0007 | 0.0366 ± 0.0188 |
| Mean CC                                     | 0.0025 ± 0.0028 | 0.0088 ± 0.0437 | 0.0718 ± 0.0556 |
Table 8: Tigger’s performance against TAGGEN and DYMOND in terms of mean absolute error ± std. dev. across various graph statistics. For all performance metrics, lower values are better. For each statistic, we also list the mean value over original graph snapshots to better contextualize the error values. We do not report the results for an algorithm if it does not complete within 24 hours. Errors smaller than five decimal places are approximated to 0.