Analytical Models for Motifs in Temporal Networks: Discovering Trends and Anomalies

ABSTRACT

Dynamic evolving networks capture temporal relations in domains such as social networks, communication networks, and financial transaction networks. In such networks, temporal motifs, which are repeated sequences of time-stamped edges/transactions, offer valuable information about the networks' evolution and function. However, to spot trends and anomalies statistically significant temporal motifs have to be identified, and using existing approaches this is infeasible due to high computational complexity. Here, we develop the *Temporal Activity State Block Model (TASBM)*, to model temporal motifs in temporal graph. We develop efficient model fitting methods, and derive closed-form expressions for the expected motif frequencies and their variances in a given temporal network, thus enabling the discovery of statistically significant temporal motifs. Our TASMB framework can accurately track the changes in the expected motif frequencies over time, and also scales well to networks with tens of millions of edges/transactions as it does not require time-consuming generation of many random temporal networks and then computing motif counts for each one of them. Applications of our model include anomaly and trend detection as well as fraud detection in temporal networks. We show that in a network of financial transactions our framework can successfully identify the motif anomalies associated with the financial crises by looking at the significance profile of temporal motifs. Moreover, we are able to identify trends in motifs in a phone call network at multiple time scales.

1 INTRODUCTION

Networks are ubiquitous models for real world systems, with applications ranging from social interactions to protein relationships [14]. Many such systems are not static, but the edges are active only at certain points in time. The networks in which *temporal edges* appear and disappear over time are called time-varying or *temporal* networks. Examples of temporal networks include communication and transaction networks where each link is relatively short or instantaneous, such as phone calls or financial exchanges. Another example is networks of physical proximity, capturing times at which individuals encounter each other. Time dependent and temporal properties can be analyzed on time-varying networks. Extracting recurring and persistent patterns of interaction in temporal networks is of particular interest, as it provides higher order information about the network transformation and functionality [3]. For example, an abundance of triangles in financial transaction networks is associated with anomalies and is identified as the signature of financial crisis [18]. Similarly, phone calls have been shown to significantly increase in volume and change patterns during major events such as earthquakes [23].

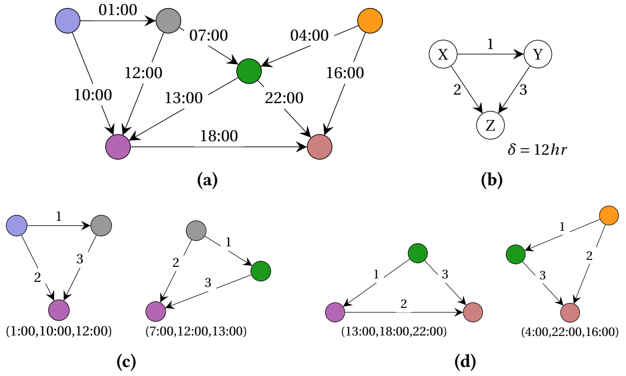


Figure 1: (a) A temporal network with edges appearing over a day, (b) a motif M with a temporal window of 12 hours, (c) examples of M in the network, and d) triangles in the network which are not δ-instances of M, either due to edge order or time between first and last edge.

Repeated patterns of interconnections between nodes occurring at a significantly higher frequency than those in randomized networks are called *motifs*. Formally, a *temporal motif* is a subgraph and an ordering on the temporal occurrences of its edges. δ-instances of a temporal motif are instances in which all the temporal edges appear according to the ordering specified by the temporal motif within a time window of length δ [15]. Figure 1 illustrates a small temporal network with δ-instances of a temporal motif M.

A critical task when analyzing temporal networks is to identify significant temporal motifs. To do so, one must compute both the expected number of occurrences of each possible motif as well as its variance. Current approaches require generating thousands of randomized networks (e.g., by rewiring or reshuffling the time-order of the edges [8]), counting the motif frequencies in each of them, and then computing averages and standard deviations of those counts. While even creating randomized networks is computationally expensive, counting motifs is a task with exponential complexity in the motif size [15]. Thus, identifying significant temporal motifs in large real-world networks is impossible using current approaches.

Here we propose the *Temporal Activity State Block Model (TASBM)*, which models different activity levels of groups of nodes in a temporal network and can effectively capture intermittent activation between individual nodes. More precisely, our proposed Temporal Activity State Block Model first partitions the nodes into different groups based on their activity level, *i.e.*, the rate of temporal edges they are likely to send or receive. We then model rate of out-links and in-links between every pair of groups using Poisson processes. Every pair of groups has distinct rate for sending and receiving temporal edges. The nodes' activity level, and hence their group assignment, may change over time. In real temporal networks, streams of edges often arrive in "bursts", resulting in sharp rises in the nodes' activity levels. The TASBM allows us to robustly and efficiently model the bursty arrival of temporal edges that is observed in a real temporal network and thus allows for accurate identification of temporal network properties.

More importantly, our proposed model (TASBM) allows for efficient identification of temporal patterns of interaction that are statistically significant in the network. To identify statistically significant temporal motifs we first calculate the number of occurrences of a temporal pattern in the given network. We then compare its frequency with the expected number provided by the TASBM. Crucially, we develop an efficient TASBM parameter fitting technique and also derive closed-form solutions for the expected δ-instances (as well as the variance) of a temporal motif in every time interval T. Conceptually, the expected motif frequencies calculated by our framework are equivalent to the expected frequencies obtained by averaging motif instances over a large ensemble of re-wired networks with shuffled timestamps. However, we develop an analytical framework for calculating the expected motif frequencies. This brings a crucial benefit of our approach as we do not rely on enumerating motif instances on large random ensemble of re-wired networks. As a result, our framework allows for accurate modeling of motif frequencies and scales large real-world temporal networks with tens of millions of transactions.

We conduct experiments on both synthetic and real-world temporal networks. Results on synthetic networks demonstrate that our analytical framework can closely track the changes in the frequencies of δ-instances of temporal motifs over time. When applying our framework to analyze synthetic networks containing planted anomalous motifs, our results show that although the planted motif anomalies cannot be identified based on motif counts alone, they can be discovered through comparison to the expected frequencies provided by our analytical model.

In our real-world experiments, we apply our analytical model to discover anomalies in a financial transaction network with 118.7 thousand nodes and 2.9 million temporal edges. In addition, we apply our framework to find trends in a phone call network with 1.2 million nodes and 21.9 million temporal edges. In the financial network, our method can successfully localize the anomalies caused by a financial crisis when it hits the network. We also observe that certain motifs can be interpreted as signals of improvement of the economy when the network starts to recover. In the phone call network, we observe that our analytical framework can identify daily and weekly phone call trends such as peak hours, weekends, and a widely observed holiday.

2 RELATED WORK

In this section, we review the related work to dynamic network and activity state models, as well as temporal motifs. We note that existing methods for calculating expected motif frequencies rely on randomization and shuffling timestamps, and thus do not scale to large temporal networks. In contrast, we provide the first analytical model for temporal motifs that does not require random network ensembles and hence scales well to real-world temporal networks.

Dynamic network models. There is a body of work on modeling dynamic networks, mostly by extending a static model to the dynamic setting [1, 4, 7, 20–22]. Among the existing methods, temporal extensions of stochastic block models (SBMs) are the most relevant to our work. In particular, Yang et al. [22] proposed a dynamic SBM, in which a transition matrix specifies the probability of nodes to switch classes over time. Ho et al. [7] proposed a temporal extension of a mixed-membership SBM (MMSBM) using linear state-space models for the class membership vectors of node clusters. However, the above works both assume that edge probabilities do not change over time. More recently, Xu et al. [21] proposed a model in which the network snapshots are modeled using SBM [4], and the state evolution od the dynamic network is modeled by a stochastic dynamic system. Our work here differs from the above in that in our proposed TASBM model, a particular block is not a community with vertices more likely to be connected to each other, but a set of vertices with similar local behavior, in terms of out- and in- edge arrival rates.

Temporal motifs. Paranjape et. al. [15] extended the concept of static motifs to temporal networks and proposed a framework for counting the exact number of relatively small temporal motifs. Prior to [15], existing methods either do not account for ordering of the temporal edges [24], or require temporal edges in a motif to arrive consecutively to a node [12]. Among other examples are the algorithm of [6] that uses ideas from sub-sequence mining to identify patterns in temporal graphs, and methods of [17] for finding temporal isomorphic subgraphs. While this line of work aims to accurately count or approximate the actual numbers of motif instances in any graph, the goal of our work is to determine the expected number and variance of motif instances in a graph, given its underlying statistical model.

Finally, there have been attempts to identify the significance of various motifs in temporal graphs. Initial approaches to randomization of temporal networks have included shuffling and reversing timestamps. Bajardi et al. [2] showed the inherent ordering of motifs due to their disappearance in shuffled or reverse-ordered time series. Donker et al. [5] used time reversal in their analysis of a hospital network. Kovanen et al. [12] compared temporal network data to randomly permuted timestamps as well as permuted timestamps with bias toward shorter inter-event times. Holme et al. [9] and Li et al. [13] presented models of temporal randomization with additional constraints to preserve the lifetimes of edges. Unlike these methods, our approach does not require the simulation of network ensembles, and hence scales well to large temporal networks.

3 NETWORK ACTIVITY MODEL

In this section we describe our Temporal Activity State Block Model (TASBM). Our goal is to construct a temporal network model which can be updated efficiently and thus maintained online, so that it can be used to describe the network before the full edge set is known.

Formally, a *temporal graph* can be viewed as a sequence of static directed graphs over the same (static) set V of nodes and the set E of $m = |E|$ *temporal edges*. Each *temporal edge* is a timestamped ordered pair of nodes $(e_i = (u, v), t_i), i \in [m]$, where $u, v \in V$ and $t_i \in \mathbb{R}$ is the timestamp at which the edge arrives. Multiple temporal edges between the same pair of nodes u and v and different timestamps can exist. We assume that the timestamps t_i are unique so that the temporal edges may be strictly ordered. However, our methods do not rely on this assumption and can easily be adapted to the case where timestamps are not unique.

The *stochastic block model* [4] on static networks is defined as dividing the nodes into communities, or blocks, such that a higher proportion of the possibly edges within a block occur, compared to those between blocks. The model then represents the graphs as a set of blocks with edge probabilities within and between each block. Taking this approach on temporal networks is not realistic because communities evolve over time and determining them at frequent time intervals is prohibitively expensive. We propose a temporal variant of the stochastic block model in which we partition the nodes of the network into groups based on their activity levels, which we define as the rates at which in- and out- edges arrive to nodes. This means that a particular block is not a *community* with vertices more likely to be connected to each other, but a set of vertices with *similar local behavior*. Specifically, nodes within each block will all have similar rates of out- and in- edge arrivals.

3.1 Temporal Activity State Block Model (TASBM)

In our Temporal Activity State Block Model (TASBM) we consider two sets of groups or activity states $G^{in} = \{1, \cdots, C^{in}\}, G^{out} = \{1, \cdots, C^{out}\}$. Every node u in the network belongs to a group $a_u^{in} = i \in G^{in}$ based on its activity level for receiving in-links and a group $a_u^{out} = j \in G^{out}$ based on its activity level for sending out-links. Nodes in the same group $i \in G^{in}$ have similar rate of receiving temporal edges. Similarly, nodes in the same group $j \in G^{out}$ have similar rate of sending temporal edges. We model group assignments a_u^{in}, a_u^{out} for $u \in V$ as independent draws from multinomial distributions parameterized by π^{in}, π^{out}. Thus, $a_u^{in} \sim \text{Multinomial}(\pi^{in})$, and $a_u^{out} \sim \text{Multinomial}(\pi^{out})$.

We consider a $C^{out} \times C^{in}$ matrix $\boldsymbol{\theta}$ such that θ_{ij} denotes the rate of temporal edges from nodes in group $i \in G^{out}$ to the nodes in group $j \in G^{in}$. After assigning nodes to different activity states, we model the temporal edges between every pair (u, v) of nodes with $a_u^{out} = i, a_v^{in} = j$ as independent Poisson draws, where the means of these Poisson draws are specified by θ_{ij}. More formally, every temporal edge $(e_r = (u, v), t_r)$ between the node u in out-link activity state $a_u^{out} = i$ to the node v in in-link activity state $a_v^{in} = j$ is an independent Poisson draw. I.e.,

$$e_r = (u, v)|a_u^{in} = i, a_v^{out} = j \sim \text{Poisson}(\theta_{ij}).$$

For the ease of notation, instead of $\theta_{a_u^{out} a_u^{in}}$, we subsequently use $\theta_{a_u a_u}$ to denote the rate of out-links from nodes in activity state a_u to the nodes in activity state a_v.

3.2 Parameter Inference

It has been observed that in real temporal networks, stream of edges usually arrive in bursts, resulting in sharp rises in the nodes' activity levels [10, 11]. We base our model on the hierarchical activity state model of [11], which uses an infinite state automaton to represent activity states. At each event instance, the model decides whether the entity's activity classification should move up or down a state via the automata transitions, or remain the same. By balancing transitions costs with rewards of being the state closest to the observed activity level, this method can capture bursts in activity of various sizes or with nested structure. We apply this principle in our model on each node individually, where events are arriving edges involving the node. However, allowing all nodes to change states at any time would make it difficult to maintain the parameter set θ trained on the latest data, since the average rates between node groups change when the composition of the groups changes because the transition cost prevents the model from fitting every single change in activity level.

Another approach to inferring the model parameters is to assume that while edge arrival rates can vary over time, node group assignments do not change over intervals of length T. Then we model arrivals of temporal edges between each pair of nodes as a Poisson process with a constant parameter on time intervals of length T. Specifically, for all pairs of vertices (u, v) such that $a_u^{out} = i$ and $a_v^{in} = j$ for some $i \in G^{out}$ and $j \in G^{in}$, the Poisson process modeling temporal edges between u and v will be parameterized by a constant $\theta_{ij} = \theta_{a_u a_v}$ in every time interval of length T. Across intervals, we calculate the posterior model parameters α and θ as:

$$\hat{\alpha} = \frac{n_r}{n}, \hat{\theta}_{rs} = \frac{m_{rs}}{n_r n_s},$$

where we denote by n_r the number of nodes in group r, and by m_{rs} the number of edges connecting group r to group s in a time interval of length T. Model inference can be done in at most two passes over the edges and in practice, can be well-approximated in one pass (see Section 5.1 for details). Thus, despite its simplicity, this method is extremely scalable.

4 ANALYTICAL MODEL FOR TEMPORAL MOTIFS

In this section, we first provide formal definitions of temporal motifs and δ-instances of temporal motifs. We then introduce our analytical framework for calculating expectation and variance of the number of motif instances, without the need for generating many randomized networks by re-wiring or shuffling edge timestamps. Finally, we provide the computational complexity analysis of our method and show that it can easily scale to large real-world temporal networks with millions of temporal edges.

4.1 Temporal Network Motifs

Formally, a temporal motif M defines a particular sequence of interactions between a set of nodes over time.

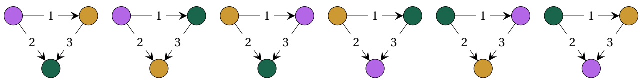


Figure 2: $3! = 6$ unique bijections between a 3-node 3-edge subgraph and the motif in Figure 1b.

Definition 1 (temporal motif). *A k-node z-edge temporal motif $M = (G_M, \prec_M)$ consists of a graph $G_M = (V_M, E_M)$, such that $|V_M| = k$ and $|E_M| = z$, and a strict total ordering $\prec_M$ on the edges E_M. We index $E_M = \{e'_1, \cdots, e'_z\}$, such that $e'_1 \prec_M e'_2 \prec_M \cdots \prec_M e'_z$.*

Note that multiple interactions between the same pair of nodes may occur in the sequence defined by M, but each edge is indexed and ordered uniquely. In a dynamic network, any subgraph of a temporal graph is a δ-instance of a temporal motif M if it is isomorphic to G_M, the set of its temporal edges follows the same ordering imposed by M, and it occurs within a time window of δ.

Definition 2 (δ-Instance of a temporal motif). *A temporal subgraph $G_s = (V_s, E_s)$, with $E_s = \{(e_1, t_1), \cdots, (e_z, t_z)\}$ is a δ-instance of a temporal motif M if 1) isomorphism: there exist an edge-preserving bijection $f : V_s \longrightarrow V_M$ between nodes of the subgraph and nodes of the motif such that $\forall e = (u, v) \in E_s \Leftrightarrow f(e = (u, v)) \in E_M$; 2) temporal ordering: the edges of the temporal motif occur according to the ordering $\prec_M$, i.e., for the ordered sequence $f(e_h) \prec_M f(e_i) \prec_M \cdots \prec_M f(e_j)$ we get a set of strictly increasing timestamps $t_h < t_i < \cdots < t_j$; and 3) temporal window: all the edges in E_s occur within δ time, i.e. $t_j - t_h \leq \delta$.*

Here, our goal is to calculate the expected number (and the variance) of δ-instances of a given temporal motif in a time-varying network. More precisely, given the number of nodes and degree distribution of a temporal network at each point in time, we are interested in calculating the expected frequency of ordered subsets of edges from the temporal network that are δ-instances of a particular temporal motif. In the following, we present a general analytic approach, and show how to apply this approach to the Temporal Activity State Block Model (TASBM).

4.2 Expected Motif Frequencies in TASBM

In this section, we present our analytical framework for calculating the expected number of δ-instances of temporal motifs in a temporal network. We provide a closed form solution for the expected δ-instances of temporal motifs in TASBM networks. However, our analysis is not limited to TASBM and can be easily generalized to other temporal network models.

To calculate the expected number of δ-instances of a temporal motif over a time window of length T, we need to calculate the expected number of subgraphs $G_s = (V_s, E_s)$ satisfying the conditions specified in Definition 2: 1) G_s and G_M are isomorphic under an edge-preserving bijection f; 2) the edges of the subgraph appear according to ordering specified by $\prec_M$, i.e., for $f(e_h) \prec_M f(e_i) \prec_M \cdots \prec_M f(e_j)$ we have $t_h < t_i < \cdots < t_j$; and 3) all the edges occur within δ time, i.e., $t_j - t_h \leq \delta$.

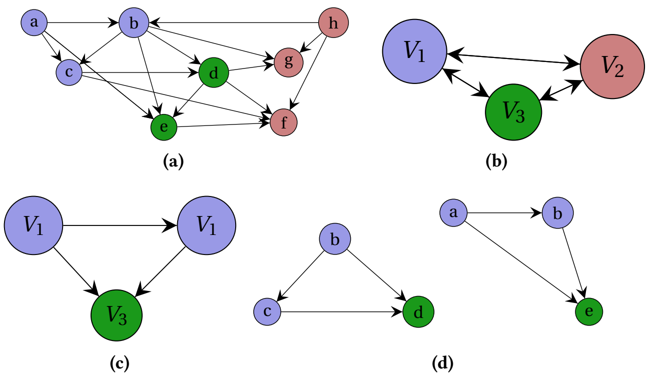


Figure 3: (a) Example graph, (b) assignment of partition $V = \{V_1, V_2, V_3\}$, (c) activity state assignment to a motif, and (d) two instances of motifs with the assigned activity states.

Isomorphic Subgraphs. We start by counting the number of ways that subgraphs isomorphic to the motif graph G_M may occur in the temporal network. In a static network, the number of subgraphs isomorphic to G_M is equal to the number of unique edge-preserving bijections between nodes of the network and nodes of the motif M. However, in a temporal graph multiple edges may appear between every pair of nodes in a time interval of length T, and therefore each bijection may result in multiple isomorphic subgraphs to G_M. To count the number of isomorphic subgraphs, we first count the number of unique bijections $f : V_s \longrightarrow V_M$ between the nodes of every k-node z-edge subgraphs $G_s = (V_s, E_s)$ and $G_M = (V_M, E_M)$. Then we count the expected number of isomorphic subgraphs that may result from each bijection f in a time interval of length T.

In general, in a graph with n nodes, there are $\binom{n}{k}$ ways to choose nodes to form a k-node subgraph. Furthermore, there are $k!$ permutations of a set of k nodes, each corresponding to a unique bijection from V_M to V_s. Figure 2 shows an example of the $3! = 6$ vertex bijections between a 3-node 3-edge subgraph and the motif in Figure 1b, each of which correspond to a unique potential motif instance. As shown in Figure 2, some of these bijections correspond to the same static graph, but the ordering on the temporal motif edges are different. Therefore, the number of bijections $\mathcal{B}_{k,V}$ between nodes of the temporal network and nodes of the motif is

$$|\mathcal{B}_{k,V}| = \binom{n}{k} k! = P(n, k).$$

We now count the number of ways that subgraphs isomorphic to the motif graph G_M may occur in the Temporal Activity State Block Model (TASBM). Here, we first count the number of ways nodes in different activity states can form a k-node subgraph. Any node in the subgraph can be selected from at most C activity states. Figure 3c shows two different activity state assignment to the nodes of a 3-node subgraph. Hence, the number of activity state assignments $\mathcal{A}_{C,k}$ for k nodes in TASBM is[1]

$$|\mathcal{A}_{C,k}| \leq C^k$$

To count the number of bijections, assume that nodes of the temporal graph are partitioned into C activity states as $V = \{V^1, \cdots, V^C\}$, and let $n^c = |V^c|$ be the number of nodes in activity state $c \in [C]$.

[1] If some groups have fewer than k nodes, we have $\mathcal{A}_{C,k} < C^k$. For example, let $k = 3$ and s_i be the number groups of size i for $i \in \{0, 1, 2\}$, then $\mathcal{A}_{C,k} = (C - s_0)^k - s_1\binom{k}{2}(C - 1 - s_0) - s_2$. If each node is in a separate group with $C = |V|$ we get $|\mathcal{B}_{k,V,n}| = |\mathcal{B}_{k,V}| = P(n, k)$.

Consider an activity state assignment $A = \{a_1, \cdots, a_k\} \in \mathcal{A}_{C,k}$, where $a_i \in [C]$ is the activity state of the i-th node. We now count the number of subsets V_s of k nodes in the network that are consistent with the activity state assignment A. Let n^c_A be the number of nodes in A that are assigned to the activity state c. There are $\binom{n^c}{n^c_A}$ ways to select n^c_A nodes from V^c. Therefore, there are $\binom{n^1}{n^1_A} \cdots \binom{n^C}{n^C_A}$ ways of forming a k-node subgraph in the network that are consistent with the activity state assignment A. Figures 3c shows an example. Finally, as shown in Figure 3d, for each permutation of the n^c_A nodes in activity state c we get a bijection. Hence, the number of unique bijections in TASBM is

$$|\mathcal{B}_{k,V,C}| = \sum_{A \in \mathcal{A}_{C,k}} \prod_{c \in [C]} P(n^c, n^c_s), \tag{1}$$

where $P(n^c,n^c_A)=\binom{n^c}{n^c_A} n^c_A!$ is the number of n^c_A-permutations of n^c. Note that $\prod_{c \in [C]} P(n^c, n^c_A)$ is constant for every activity state assignment $A \in \mathcal{A}_{C,k}$, and hence the cost of calculating Eq. 1 only depends on the number of possible activity state assignments $O(C^k)$.

Next, we calculate the expected number of isomorphic subgraphs that can result from each bijection in a time interval of length T. As we discussed in Section 4.1, in the TASBM the arrival rate of temporal edges between any pair of nodes (u, v) depends on their activity states. The expected number of temporal edges that occur from node u in activity state a_u to node v in activity state a_v in an interval $[t_0, t_0 + T)$ is

$$\mathbb{E}[N_{e_{uv}} | t \in [t_0, t_0 + T)] = \int_{t_0}^{t_0+T} \theta_{a_u a_v}(t) dt,$$

where $\theta_{a_u a_v}(t)$ is the temporal edge arrival rate from activity state a_u to activity a_v at time t.

Let $f : V_s \rightarrow V_M$ be a bijection, where V_s is a k-node subgraph that is consistent with the activity state assignment $A \in \mathcal{A}_{C,k}$. In other words, $V_s = \{v_1, \cdots, v_k | a_{v_i} = A[i]\}$. We now calculate the expected number of k-node z-edge isomorphic subgraphs $S^{k,z}_{V,C,f}$ that can result from f in a TASBM with nodes V partitioned into C activity states, in a time interval of $[t_0, t_0 + T)$.

$$\mathbb{E}[N_{S^{k,z}_{V,C,f}} | t \in [t_0, t_0 + T)] = \prod_{\substack{u,v \in V_s, \\ (f(u),f(v)) \in E_M}} \int_{t_0}^{t_0+T} \theta_{a_u a_v}(t) dt. \tag{2}$$

Note that for each activity state assignment $A \in \mathcal{A}_{C,k}$, the expected number of subgraphs for all the $\prod_{c \in [C]} P(n^c, n^c_s)$ bijections corresponding to A is the same. Therefore, from Eq. 1 and 2 we get the following lemma.

Lemma 1. *The expected number of k-node subgraphs isomorphic to the motif subgraph G_M in a Temporal Activity State Block Model (TASBM) during a time interval $[t_0, t_0 + T)$ for $T \leq \delta$ is*

$$\mathbb{E}[N_{S^{k,z}_{V,C}} | t \in [t_0, t_0 + \delta)] =$$

$$\sum_{A \in \mathcal{A}_{C,k}} \prod_{c \in [C]} P(n^c, n^c_s) \prod_{\substack{u,v \in V_s | R(V_s)=A, \\ (f(u),f(v)) \in E_M}} \int_{t_0}^{t_0+T} \theta_{a_u a_v}(t) dt,$$

where $R(V_s) = A$ is the set of all k-node subgraphs consistent with activity state assignment A.

Temporal Restrictions. We now calculate the probability that for an isomorphic subgraph $G_s = (V_s, E_s)$, the timestamps of the mapped edges ordered by $\prec_M$ are strictly increasing and within a time window of δ. For ease of notation, we subsequently assume that for each subgraph G_S and corresponding bijection $f : V_s \rightarrow V_M$, we have $f(e_1) \prec_M f(e_2) \prec_M \cdots \prec_M f(e_z)$. Therefore, we need to calculate the probability that $t_0 \leq t_1 < t_2 < \cdots < t_z < t_0 + \delta$.

The marginal probability for a temporal edge $e = (u, v)$ from activity state a_u to activity state a_v to occur at time t in the time window $[t_0, t_0 + T)$ is

$$\Theta^{[t_0, t_0+T)}_{e=(u,v)}(t) = \frac{\theta_{a_u a_v}(t)}{\int_{t_0}^{t_0+T} \theta_{a_u a_v}(t') dt'}.$$

Lemma 2. *The probability that temporal edges of a subgraph $G(V_s, E_s)$ occur in the order $\prec_M$ specified by motif M in an interval $[t_0, t_0 + T)$ is*

$$Pr(t_0 \leq t_1 < t_2 < \ldots < t_z < t_0 + T) =$$

$$\int_{t_0}^{t_0+T} \Theta^{[t_0, t_0+T)}_{e_1}(t_1) \int_{t_1}^{t_0+T} \Theta^{[t_0, t_0+T)}_{e_2}(t_2) \cdots$$

$$\int_{t_{z-1}}^{t_0+T} \Theta^{[t_0, t_0+T)}_{e_z}(t_z) dt_z dt_{z-1} dt_{z-2} \ldots dt_1.$$

Note that in the above equation, if the edge arrival rates θ_{ij} between activity states do not change in interval $[t_0, t_0 + T)$, the marginal probability for every temporal edge to happen is $1/T$. Hence, every ordering of the temporal edges in the subgraph has a probability of $1/z!$. Therefore, for constant edge arrival rates we have $Pr(t_0 \leq t_1 < t_2 < \cdots < t_z < t_0 + T) = 1/z!$. On the other hand, varying edge arrival rates in $[t_0, t_0 + T)$ can be modeled by integrable functions, for which we can calculate the value of the nested integrals to get the probability of the correct ordering $Pr(t_0 \leq t_1 < t_2 < \cdots < t_z < t_0 + T)$.

Finally, from Lemma 1 and Lemma 2, we get the following theorem for the expected number of δ instances of a temporal motif M in a time interval of length $T \leq \delta$.

Theorem 1. *The expected number of δ instances of a temporal motif M in a Temporal Activity State Block Model (TASBM) during a time interval $[t_0, t_0 + T)$ for $T \leq \delta$ is*

$$\mathbb{E}[N_M | T \leq \delta] = \mathbb{E}[N_{S^{k,z}_{V,C}}] \cdot Pr(t_0 \leq t_1 < t_2 < \ldots < t_z < t_0 + T).$$

The pseudocode of our method is given in Algorithm 1.

4.3 Expected Motif Frequency for $T > \delta$

Next, we consider the case where $T > \delta$. Here, the δ-instances of a temporal motif may have at least one edge occurring in $[t_0, t_0{+}T{-}\delta)$ or they may have all edges occurring in $[t_0{+}T{-}\delta, t_0{+}T)$. Hence, to calculate the expected number of instances in T, we take a sum over the instances with at least one edge in $[t_0, t_0 + T - \delta)$ and instances fully appearing in $[t_0 + T - \delta, t_0 + T)$.

To count the number of motif instances with the first edge occurring in $[t_0, t_0{+}T{-}\ \delta)$, assume that $e'_1 = (u', v') \in E_M$ is the edge of the motif that comes first in the ordering $\prec_M$. We first count the number of isomorphic 2-nodes subgraphs in the network to $\{u', v'\}$. Then, for each isomorphic subgraph, we count the number of subgraphs in the remaining network isomorphic to $V_M \setminus \{u', v'\}$.

Algorithm 1 Calculate Expected Motif Frequency

```
Input: Set of nodes $V$, set of $C$ activity states, edge arrival rates
    between activity states $\theta_{ij}$, time interval $[t_0, t_0 + T)$.
Output: Expected number of $\delta$-instances of motif $M$ within time
    interval $[t_0, t_0 + T)$.
 1: for $(i, j) \in ([C^{out}] \times [C^{in}])$ do
 2:     $E[N_{e_{ij}}] = \textsc{Integrate}(\theta_{ij}, t_0, t_0 + T)$
 3: $\mathbb{E}[N_M] = 0$
 4: $\mathcal{A}_{C,k}$ = set of $O(C^k)$ activity state assignments to $k$ nodes.
 5: for $A \in \mathcal{A}_{C,k}$ do
 6:     $|\mathcal{B}_{k,V,C}| = \prod_{c \in [C]} P(n^c, n^c_s)$                ▷ Eq. 1
 7:     $V_s = \{v_1, \cdots, v_k | a_{v_i} = A[i]\}$
 8:     $f$ : a bijection from $V_s$ to $V_M$
 9:     $\mathbb{E}[N_S] = 1$
10:     for $(u, v) \in V_s$ such that $(f(u), f(v)) \in E_M$ do      ▷ Eq. 2
11:         $\mathbb{E}[N_S] = \mathbb{E}[N_S] \times E[N_{e_{a_u a_v}}]$
12:     $P_{\text{order}} = Pr(t_0 \le t_1 < t_2 < \ldots < t_z < t_0 + T)$    ▷ Lemma 2
13:     $\mathbb{E}[N_M] = \mathbb{E}[N_M] + |\mathcal{B}_{k,V,C}| \times \mathbb{E}[N_S] \times P_{\text{order}}$
    return $\mathbb{E}[N_M]$
```

Lemma 3. *The expected number of k-node subgraphs isomorphic to the motif subgraph G_M in a Temporal Activity State Block Model (TASBM) with the first edge occurring in $[t_0, t_0 + T - \delta)$ for $T > \delta$ is*

$$\mathbb{E}[N_{S^{k,z}_{V,C}} \mid t_1 \in [t_0, t_0 + T - \delta)] =$$

$$\mathbb{E}[N_{S^{2,1}_{V,C}} | t \in [t_0, t_0 + T - \delta)] \cdot \mathbb{E}[N_{S^{k-2,z-1}_{V \setminus V_{e_1}, C}} | t \in [t_1, t_1 + \delta)],$$

where t_1 is the time of appearance of the first edge, and V_{e_1} is the set of two nodes in subgraph $S^{2,1}_{V,C}$.

We can now use marginal edge probabilities to calculate expected frequency on intervals with $T > \delta$. For subgraph G_S isomorphic to G_M, we compute the probability that the first edge occurs in $[t_0, t_0 + T - \delta]$ and the remaining $z - 1$ edges occur sequentially within a time window of length δ starting at t_1.

Lemma 4. *The probability that temporal edges of a subgraph $G(V_s, E_s)$ occur in the order $\prec_M$ specified by motif M in an interval of length $T > \delta$ with the first edge appearing in $[t_0, t_0 + T - \delta)$ is*

$$Pr(t_2 < t_3 < \ldots < t_z < t_1 + \delta | t_1 < t_2, t_1 \in [t_0, t_0 + T - \delta)) =$$

$$\int_{t_0}^{t_0 + T - \delta} \Theta_{e_1}^{[t_0, t_0 + T - \delta]}(t_1) \int_{t_1}^{t_1 + \delta} \Theta_{e_2}^{[t_1, t_1 + \delta]}(t) \int_{t_2}^{t_1 + \delta} \Theta_{e_3}^{[t_1, t_1 + \delta]}(t_3)$$

$$\cdots \int_{t_{z-1}}^{t_1 + \delta} \Theta_{e_z}^{[t_1, t_1 + \delta]}(t_z) dt_z dt_{z-1} dt_{z-2} \ldots dt dt_1.$$

The expected frequency of δ-instances of temporal motif M in a time window of length $T > \delta$, conditional on at least one edge occurring in $[t_0, t_0 + T - \delta)$, can be calculated using Lemmas 3 and 4.

$$\mathbb{E}[N_M | T > \delta, t_1 < t_0 + T - \delta] = \mathbb{E}[N_{S^{k,z}_{V,C}} | t_1 \in [t_0, t_0 + T - \delta)] \cdot \quad (3)$$

$$Pr(t_2 < t_3 < \ldots < t_z < t_1 + \delta | t_1 < t_2, t_1 \in [t_0, t_0 + T - \delta))$$

Finally, The expected number of instances fully appearing in $[t_0 + T - \delta, t_0 + T)$ can be calculated from Theorem 1. Hence, we can compute the expected number of δ-instances of temporal motif M in a time interval $T > \delta$ using the following Theorem.

Theorem 2. *The expected number of δ instances of a temporal motif M in a Temporal Activity State Block Model (TASBM) during a time interval $[t_0, t_0 + T)$ for $T > \delta$ is*

$$\mathbb{E}[N_M | T > \delta] = \mathbb{E}[N_M | T = \delta] + \mathbb{E}[N_M | T > \delta, t_1 < t_0 + T - \delta].$$

4.4 Variance

We next discuss how we can use our framework for computing the variance of the number of motif instances $\mathbb{V}[N_M]$. I.e.,

$$\mathbb{V}[N_M] = \mathbb{E}[N_M^2] - \mathbb{E}[N_M]^2.$$

While we can simply calculate $\mathbb{E}[N_M]^2$ using Algorithm 1, computing $\mathbb{E}[N_M^2]$ involves calculating the expected number of *pairs* of δ-instances of motif M. A pair of instances can overlap in up to k vertices and up to z temporal edges. Examples of two overlapping motif instances sharing vertices and/or edges are show in Figure 4.

In order to calculate $\mathbb{E}[N_M^2]$, we need to consider both independent and dependent pairs of δ-instances of M. Motifs instances which do not share an edge, such as the examples in Figures 4a and 4b, are conditionally independent. On the other hand, pairs of instances which share at least one edge are not independent. For dependent instances, we must consider all possible total orderings of the edges of the two pairs of δ-instances of M. For example in Figure 4c, we have that $t_1 < t_2 < t_3 = t_{3'}$ from the first δ-instance and $t_{1'} < t_{2'} < t_{3'} = t_3$ from the second δ-instance, but the ordering of the edges $\{t_1, t_2, t_{1'}, t_{2'}\}$ is not fully specified by the two instances and thus we need to count all the possibilities for the remaining 4 edges, including $t_1 < t_2 < t_{1'} < t_{2'}$, $t_1 < t_{1'} < t_2 < t_{2'}$, $t_{1'} < t_{2'} < t_1 < t_2$, etc.

Let S_1, S_2 be a pair of of δ-instances of M. The time interval that the edges $E_{S_1} \cup E_{S_2}$ of both instances may occur is within an interval $[t_0 + \delta, t_0 + 2\delta)$, depending on which temporal edges, if any, are shared. We denote by $\mathbb{E}_\delta$ and $\mathbb{E}_{\delta'}$ expected δ- and δ'-instances of motif M, respectively. Then by linearity of expectation, we get

$$\mathbb{E}[N_M^2] = \sum_{\substack{(S_1, S_2): \\ E_{S_1} \cap E_{S_2} = \emptyset}} \mathbb{E}_\delta[N_{S_1} | t \in [t_0, t_0 + T])\mathbb{E}_\delta[N_{S_2} | t \in [t_0, t_0 + T)] +$$

$$\sum_{\substack{(S_1, S_2): \\ E_{S_1} \cap E S_2 \neq \emptyset}} \sum_{\delta' \in [t_0 + \delta, t_0 + 2\delta)} \mathbb{E}_{\delta'}[N_{S_1 \cup S_2} | t \in [t_0, t_0 + T)].$$

4.5 Computational Complexity

We describe the computational complexity of our method to compute the expected number of δ-instances of motif $M = (G_M, \prec_M)$ on time window of length T. First consider the case where $T \le \delta$. Here, for every pair (i, j) of activity state assignments, we need to calculate an integral to get the expected number of edges from nodes in state i to nodes in state j (line 2 of Algorithm 1). Then, for each of the $|\mathcal{A}_{C,k}| = O(C^k)$ activity state assignment to k nodes, we need to calculate the expected number of isomorphic subgraphs to G_M that can be formed on V_s. This involves iterating over z edges (line 10 of Algorithm 1). We then need to calculate the probability of the correct temporal ordering for the edges of the subgraph. If edge arrival rates between activity states do not change within the time interval, the probability of correct ordering is $1/z!$. Otherwise we need to calculate an additional set of z integrals as specified in Lemma 2. Assuming the cost of computing each

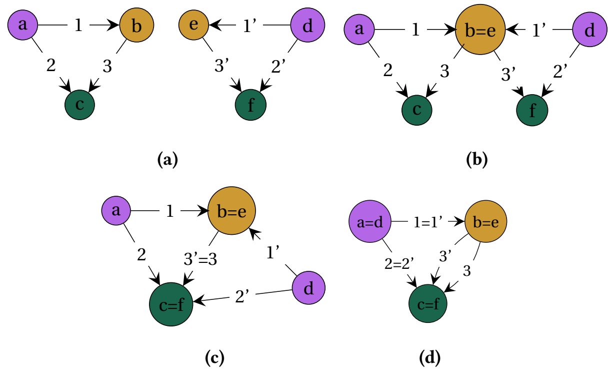


Figure 4: Examples of joint instances of motif M: S_1 on vertices (a, b, c) with edges labeled $1, 2,$ and 3 at times $t_1 < t_2 < t_3$ and S_2 on vertices (d, e, f) with edges labeled $1', 2',$ and $3'$ at times $t_{1'} < t_{2'} < t_{3'}$.

integral is $O(1)$, the total computational complexity of Algorithm 1 is $O(C^k)$. For the case where $t > \delta$, the additional computation of $\mathbb{E}[N_M|T > \delta, t_1 < t_0 + T - \delta]$ for $T > \delta$ (Eq. 3) follows the same structure as in the case where $T \leq \delta$ with different integral bounds. Therefore, the computational complexity for calculating the expected motif frequencies is $O(C^k)$. If the number of blocks, C, is constant relative to the size of the network, we get a computational complexity of $O(1)$. Our method and analysis are easily generalized to any temporal network model by considering $C = n$ activity states, one for each node. However, without benefiting from TASBM, we have $C = |V|$, and the computational complexity of calculating the expected motif frequencies is $O(n^k)$.

Furthermore, note that in edge re-wiring and timestamp shuffling approaches, for an ensemble of size r, we need to generate r (in the order of thousands) randomized networks and enumerate the motif instances on each of them to calculate the average and standard deviation for the number of instances of each motif. While even creating thousands of randomized networks is computationally expensive, each enumeration of motifs on an instance takes at least $O(|E|)$ steps (see [15] for algorithms) and thus the $O(r|E|)$ cost of motif enumeration in an ensemble is much greater than cost for analytical method using TASBM.

5 EXPERIMENTS

In this section, we present the results of applying our analytical framework to an inferred TASBM to calculate the expected motif frequencies in synthetic and real-world temporal graphs. We compare the expected number of motif instances to the number of observed motif instances counted by the method of [15]. To demonstrate our TASBM model and analytical method, we focus on expected counts for motifs with 2 or 3 nodes and 3 edges (Figure 5). While our framework can be used for larger motifs, including those representing joint motif instance (e.g., Figure 4.4), we do not compute these in practice due to the computational cost. Thus we do not include experiments computing variance, which require such counts described in Section 4.4.

We first discuss how we fit the TASBM model to observed temporal network data. We then present synthetic experiments, in which we investigate the effect of average degree of the temporal network, length of the time window T, and motif window δ on the accuracy of calculating the expected motif frequencies. Finally, we show real-world experiments, where we apply our analytical model to localize anomalies in a financial transaction network and identify daily and weekly phone-call trends in a phone call network.

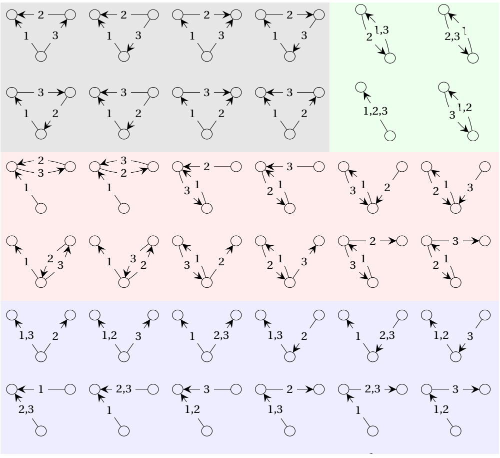


Figure 5: All 2- and 3- node motifs with 3 edges, shaded by type: triangle (gray), reciprocated edge (red), double edge (blue), and two-node (green).

5.1 Fitting the Model

Our framework first computes the average out-edge and in-edge arrival rates for all nodes on every window of T time unit $[iT, (i{+}1)T)$, $i \geq 0$. Out-edge rates are partitioned into C^{out} groups and in-edge rates are partitioned into C^{in} groups. For a total of $C{=}C^{in} \cdot C^{out}$ groups corresponding to the out-and in-edge rate combinations. In a single pass over the edges of time interval $[0, T]$, we can determine the out- and in-degree of each node. To assign our group partitions in a single pass over the nodes at the end of the time interval, we take the following approach. We start with a large number of groups, i.e. a large value of C^{in} and C^{out} and assign each group to one of the exponentially spaced *target rates*. Then nodes are assigned to the group which has the target rate closest to their observed rate, for out- and in-rates independently. Empty groups can be dropped from subsequent computations, so starting with a large initial set does not incur significant computational cost.

We then use our group assignments to approximate $\boldsymbol{\theta}$ without making another pass over the edges. During assignment of nodes to groups, we compute the observed average out- and in- degrees of each group in $[C^{out}]$ and $[C^{in}]$, respectively. These total out and in rates correspond to the column and row sums of $\boldsymbol{\theta}$. Rather than iterating over all edges once group assignments have been made to determine the exact breakdown of each sum, we make the assumption that the out-edges of a group $i \in [C^{out}]$ will be distributed among all nodes according to their in-edge rate. Thus we need only the set of average in-rates for each group in $[C^{in}]$ to determine how the total rate for $i \in [C^{out}]$ is divided up over a row of $\boldsymbol{\theta}$. In practice, this method results in an accurate approximation of $\boldsymbol{\theta}$, so we use it for all the following experiments.

5.2 Synthetic Networks

Our synthetic networks are generated according to the TASBM with $C = C^{out} \cdot C^{in}$ groups with C^{out} out-edge and C^{in} in-edge states, respectively. Each node is assigned to a randomly chosen group in $[C]$. We select a distinct edge arrival rate between every pair of groups. For each pair of nodes (u, v), temporal edges are then sampled according to a Poisson process with the rate corresponding to the out-edge group of u and in-edge group of v.

Accuracy of Model Inference. We first show that we can accurately infer the TASBM parameters used to generate the synthetic network and thus accurately use our analytical approach to determine expected motifs. We measure accuracy of our model over a set of r synthetic networks, using Mean Squared Relative Error,

$$MSRE = \frac{1}{r}\sum_{i=1}^{r}\left(\frac{N_M^i - N_M}{N_M^i}\right)^2,$$

where N_M^i is the actual number of motif instances counted using the method of [15] in the i-th generated network, and N_M is the expected motif frequency calculated by our framework.

Table 1 shows MSRE for 30 networks with 300 nodes generated using TASBM with $C^{out} = C^{in} = 5$. For generating out-edges, we partition the nodes to groups of size 10, 30, 60, 80, and 120, with out-rates of 1e-7, 1e-6, 1e-5, 1e-4, and 1e-3. For generating in-edges, we have all the nodes in one group, hence having in-rate of 11111e-3. We got 384,580 edges on average in the generated networks. We varied the number of groups in our framework for calculating the expected motif frequencies from $C^{out} = C^{in} = 1$ to $C^{out} = C^{in} = 5$ and used $T = 10K$ and $\delta = 5K$ time units. It can be seen that the error quickly vanishes when the model is allowed to use a higher number of groups. However, the improvements from increasing the number of groups quickly diminish. It can be observed that using only $C^{out} = C^{in} = 2$ groups to calculate the expected frequencies, we get almost the same accuracy as using $C^{out} = C^{in} = 5$ groups. This indicates that while real data is likely to have a large variety of node activity levels, a relatively small value of C can be used to calculate accurate expected motif counts.

Robustness of Model to Hyper-Parameter Choices. Next we investigate robustness of our method to choices of hyper-parameters. Figure 6 compares MSRE for 30 networks with 100 nodes generated using TASBM with $C^{out} = C^{in} = 3$. Here, for generating out-links we divide the nodes into groups of sizes 10, 30, and 60 and use the initial out-rates of 5e-6, 1e-4, and 1e-3, respectively. The rates were chosen to be sufficiently distinct and to generate sufficiently large edge volumes without motif counts exceeding the maximum capacity of the motif counter from [15].

Table 1: MSRE for varying number of groups C used by TASBM using $T = 10K$, $\delta = 5K$. The values are calculated over 30 networks generated with 300 nodes and $C^{out} = C^{in} = 5$.

C	Triangles	Two Vertex	Reciprocated	Double Edge
1	0.229	0.381	0.147	0.381
4	1.99e-05	4.35e-05	2.84e-05	1.69e-05
9	1.89e-05	4.26e-05	2.78e-05	1.60e-05
16	1.04e-05	3.59e-05	2.25e-05	7.90e-06
25	1.04e-05	3.59e-05	2.25e-05	7.91e-06

Figure 6a shows the accuracy of our model for networks with increasing average degree, generated by scaling the entire set of arrival rates exponentially, both within and between groups. The peaks in error as the average degree increases occur at approximately the point when motif counts become non-zero, but have high variance due to low edge rates. For two-node motifs, this peak happens at a larger edge volume since the frequency of such motifs are lower than that of three-node motifs.

Figure 6b shows how MSRE converges to zero as the motif window δ increases. The convergence is due to the same reason as above; as the motif window increases there is a greater volume of randomly generated edges over which each computation is made, and thus the average motif counts are closer to our expected counts. Again, as there is a smaller number of two-node motifs in a temporal network generated by stochastic block model, the MSRE converges more slowly for two-node motifs.

Performance for Dynamic Edge Arrival Rates. In our next set of synthetic experiments, we demonstrate the performance of our framework when edge arrival rates change over time. We generated a network using TASBM with 100 nodes and $C^{out} = C^{in} = 2$. We set baseline out-and in- arrival rates of 1e-4 for 50 nodes and 1e-3 for the other 50. The baseline rate is then scaled by a random rate scale for all edges generated on a given time-subinterval. Figure 7 shows the result of the process over 3 and 64 distinct time intervals with 3 and 64 corresponding rate scales. Figure 7a shows the average edge arrival rate over each of the 3 distinct generation periods and Figure 7b shows the number of observed and expected instances of an example motif over each intervals. Figures 7c and 7d repeat this process for 64 intervals.

Identifying Synthetic Planted Anomalies. We next show how our framework can be used to identify motif anomalies. We generate a 100 node network with the same baseline rates as in the previous experiment, this time over 32 distinct intervals of 1000 time steps, with randomly chosen rate scales. In addition, on the 10-th time interval (labeled T_1 in Figures 8 and 9) we plant anomalous reciprocated edges and on interval 25 (labeled T_2) we plant anomalous repeated edges. We introduce both anomalies by the following process: after an edge is drawn from the Poisson process, we generate a corresponding anomalous edge with probability 0.25, and place it at a random time within the next 10 and 100 steps.

We show that observed motif counts alone are not sufficient to identify planted motif anomalies. Figure 8 shows the number of each motif observed over time. We see that although each anomaly specifically target a subset of the motifs, they are not notable features on any of the individual motif plots. Figure 9 shows the log of the ratio of observed and expected motifs. Here, we immediately see which motifs are most affected by the anomalies at T_1 and T_2. The reason for some of the motifs being more affected is the pattern of introducing anomalies. In particular, since we add either reciprocate or repeated anomalous edges within the next 10 and 1000 steps, among the motifs with reciprocated edges, the ones that have sequential reciprocated edges are more affected. Similarly, among the motifs with a double edge, the ones that have sequential repeated edges are more affected.

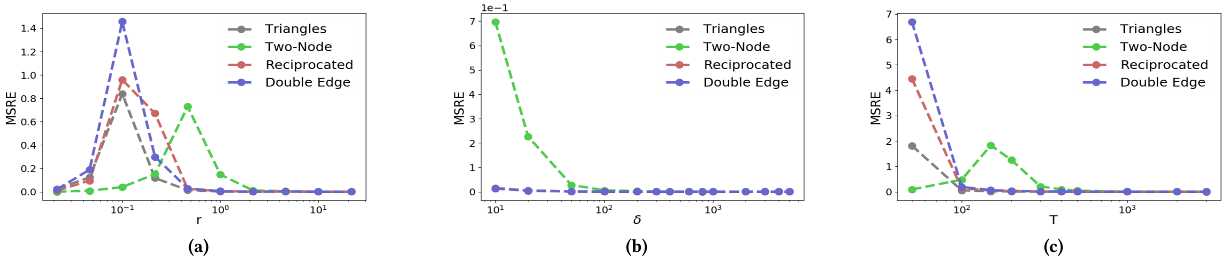


Figure 6: MSRE for the expected motif frequencies over 30 synthetic networks with 100 nodes for varying (a) number of edges, (b) motif window δ, and (c) time window T.

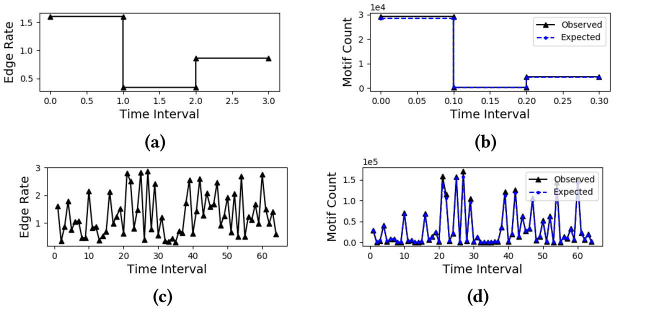


Figure 7: Varying edge rates and accuracy of motif counts for a network of 100 nodes generated over (a) 3 distinct edge-rate intervals and (b) 64 distinct edge-rate intervals.

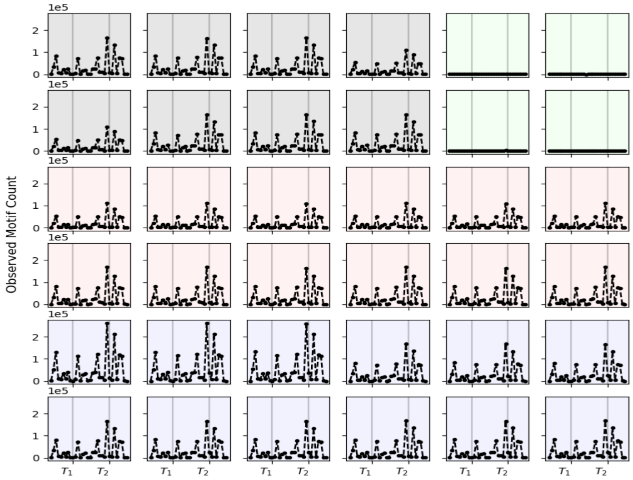


Figure 8: Observed motif counts on synthetic network with anomalies at T_1 and T_2. Plot at a given x-y position corresponds to the motif at the same position in Figure 5.

5.3 Real-world Temporal Networks

In our real-world experiments, we use our analytical model to calculate the expected motif frequencies in a financial transaction network and a phone call network. We show that our framework can localize anomalies in a financial transaction network and identify daily and weekly phone-call trends.

Financial transaction network. In our first real-world experiment, we applied our framework to calculate the expected motif counts in a small European country's financial transaction network. The data is collected from the entire country's transaction log for all transactions larger than 50K Euros over 10 years from 2008 to 2018, and includes 118,739 nodes and 2,982,049 temporal edges. The number of temporal edges from June 2008 to April 2008 is shown in Figure 10b. As edges do not occur on weekends, we do not count them toward values of T and δ or in computing edge rates.

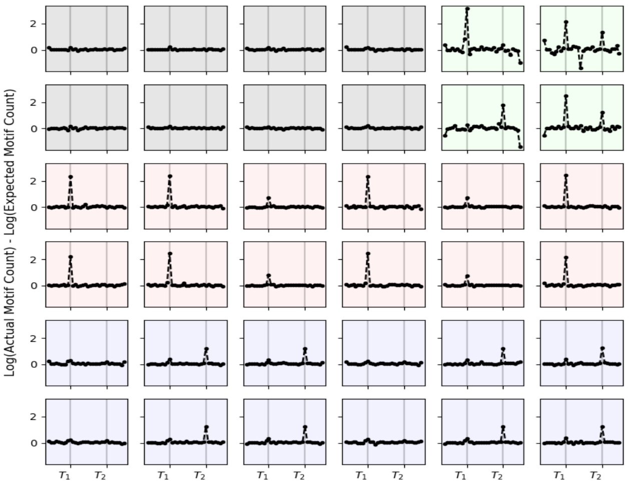
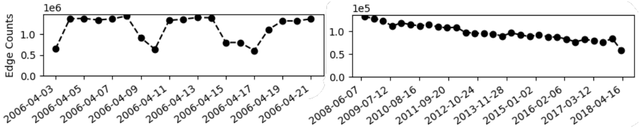


Figure 9: Ratio of observed and expected motifs on synthetic network with anomalies at T_1 and T_2.

Figure 11 shows the difference in the log of observed and expected motif frequencies for each motif in Figure 5, $\delta = T = 90$ weekdays and $C_1 = C_2 = 6$. It can be seen that although the number of edges is decreasing over time, perhaps surprisingly, we can localize the time the financial crisis hits the country around September 2011, from the difference in the actual vs. expected motif frequencies. While the network has a small number of triangles in general, relative frequency of the triangles increases significantly during the crisis, which is an indicator of anomalies in financial networks [19]. On the other hand, although all the other motifs are almost equally represented in the network, two-node motifs became much more frequent during the crisis. This is potentially due to the collapse of the larger connected components, and increase in the number of transactions between immediate partners. It can also be observed that the network starts to recover around March 2017, with blocking reciprocates (row 3) and blocking double-edge motifs (row 6, columns 1-3) showing signs of improvements over time between 2011 and 2017. This could potentially be a result of returning loans and debts as the economy improves.



(a) Phone call network **(b) Financial transactions**

Figure 10: Number of temporal edges in real networks.

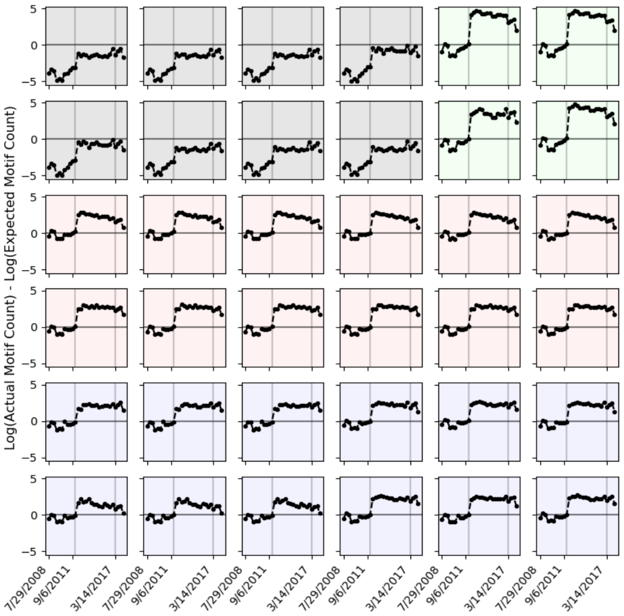
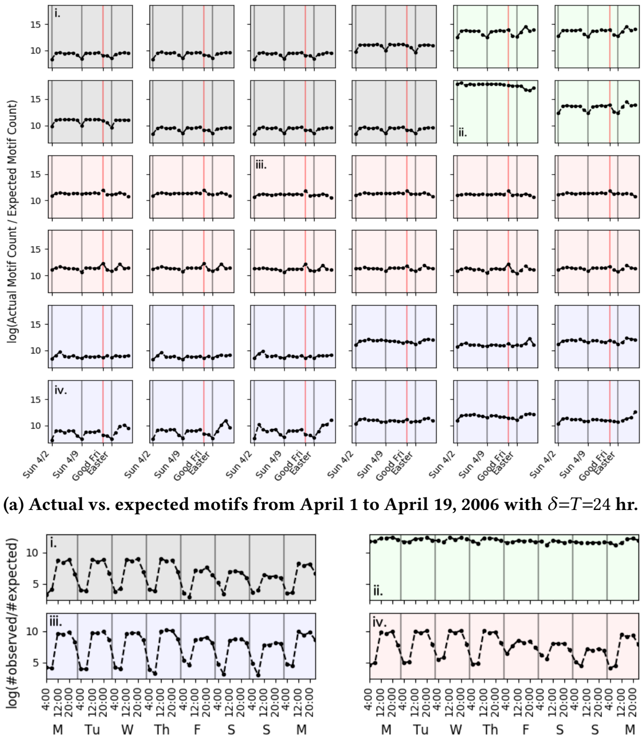


Figure 11: Financial transaction network, $\delta = T = 90$ days.

Phone Call Network. Our second dataset is a temporal network of phone calls made in April 2006 within a European country. The data includes 1,218,293 nodes and 21,907,608 temporal edges over a 19 day period. We computed the expected frequencies of 2− and 3− node motif with 3 edges at time scales of $\delta = T = 4$ and 24 hours with 48 and 19 intervals, respectively, and $C_1 = C_2 = 4$. Figure 12a and 12b show the differences between the log of the expected and observed motif frequencies for each motif in Figure 5.

Figure 12a shows the differences in motif counts from April 1, 2006 to April 19, 2006 with $\delta = T = 24$ hours. A weekly cycle can be observed, with dips on the weekends for triangles (1-2 rows, 1-4 columns), blocking double-edge motifs (4th row, column 1-3), and non-blocking reciprocates (row 6). This could be due to the fact that institutions either only receive phone calls from different people (customer services), or they return or forward every phone call during weekdays. On the other hand, blocking and non-blocking reciprocated motifs (row 5, 6) are significantly more frequent than expected on April 14th, demonstrating that people exchange a lot more phone calls on Good Friday.

The general difference between the number of expected and observed motifs in the phone call network is due to the existence of multiple disconnected communities in the underlying static graph. This is in contrast to our assumption that nodes in the same activity state have similar rate of out-links to all the other groups. Despite this difference, our method is able to accurately capture the trends in the frequency of temporal motifs in the network.



(a) Actual vs. expected motifs from April 1 to April 19, 2006 with $\delta = T = 24$ hr.

(b) Actual vs. expected motifs from April 10 to 17, 2006 with $\delta = T = 4$ hr.

Figure 12: Phone call network.

Figure 12b shows the differences in motif counts from April 10, 2006 through April 17, 2006 with $\delta = T = 4$ hours for the marked motifs inf Figure 12a. The most prominent feature is the daily cycle of motif frequencies with the most motif over-expression occurring during the day. In particular, the peaks correspond to approximately the hour intervals 12:00-16:00 and 16:00-20:00. Over the eight day period, which begins on a Monday, we can also see that weekends have different motif patterns than weekdays for most of the motifs. Friday, April 14th, also has relative motif counts similar to Saturday and Sunday, perhaps because of the holiday. We can also see that the two-node motif with 3 edges in one direction appears to be less effected by time of day or day of week.

6 DISCUSSION

We have developed an analytical model to determine the expected number as well as the variance of motifs in a temporal network. We developed an efficient parameter inference technique as well as provided closed form solutions for the expected motif frequencies in the general case where temporal edges appear with distinct rates between different pairs of nodes, and the arrival rate of temporal edges between every pair of nodes may change over time. We demonstrated the effectiveness of our Temporal Activity State Block Model combined with our analytical model of temporal motifs for discovering trends and anomalies in temporal networks. Applied to a financial transaction network, our framework can successfully localize anomalies caused by a financial crisis. Moreover, we identify trends such as weekends and an observed holiday by looking at the significance profile of temporal motifs in a phone call network.

REFERENCES

[1] A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

[2] P. Bajardi, A. Barrat, F. Natale, L. Savini, and V. Colizza. Dynamical patterns of cattle trade movements. *PloS one*, 6(5):e19869, 2011.

[3] A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[4] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[5] T. Donker, J. Wallinga, and H. Grundmann. Dispersal of antibiotic-resistant high-risk clones by hospital networks: changing the patient direction can make all the difference. *Journal of Hospital Infection*, 86(1):34–41, 2014.

[6] S. Gurukar, S. Ranu, and B. Ravindran. Commit: A scalable approach to mining communication motifs from dynamic networks. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 475–489. ACM, 2015.

[7] Q. Ho, L. Song, and E. Xing. Evolving cluster mixed-membership blockmodel for time-evolving networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 342–350, 2011.

[8] P. Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234, 2015.

[9] P. Holme and F. Liljeros. Birth and death of links control disease spreading in empirical contact networks. *Scientific reports*, 4:4999, 2014.

[10] P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[11] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[12] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005, 2011.

[13] M. Li, V. D. Rao, T. Gernat, and H. Dankowicz. Lifetime-preserving reference models for characterizing spreading dynamics on temporal networks. *Scientific reports*, 8(1):709, 2018.

[14] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[15] A. Paranjape, A. R. Benson, and J. Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610. ACM, 2017.

[16] F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, and S. Robin. Assessing the exceptionality of network motifs. *Journal of Computational Biology*, 15(1):1–20, 2008.

[17] U. Redmond and P. Cunningham. Temporal subgraph isomorphism. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 1451–1452. IEEE, 2013.

[18] T. Squartini, I. Van Lelyveld, and D. Garlaschelli. Early-warning signals of topological collapse in interbank networks. *Scientific reports*, 3:3357, 2013.

[19] B. M. Tabak, M. Takami, J. M. Rocha, D. O. Cajueiro, and S. R. Souza. Directed clustering coefficient as a measure of systemic risk in complex banking networks. *Physica A: Statistical Mechanics and its Applications*, 394:211–216, 2014.

[20] A. H. Westveld, P. D. Hoff, et al. A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, 5(2A):843–872, 2011.

[21] K. S. Xu and A. O. Hero. Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 201–210. Springer, 2013.

[22] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. Detecting communities and their evolutions in dynamic social networksâĂŤa bayesian approach. *Machine learning*, 82(2):157–189, 2011.

[23] X. Yu, T. Pei, K. Gai, and L. Guo. Analysis on urban collective call behavior to earthquake. In *High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on*, pages 1302–1307. IEEE, 2015.

[24] Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, and W.-C. Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1645–1648. ACM, 2010.