Temporal Motifs in Heterogeneous Information Networks

Yuchen Li* Zhengzhi Lou* Yu Shi Jiawei Han

University of Illinois at Urbana-Champaign, Urbana, IL USA

{li215, zlou4, yushi2, hanj}@illinois.edu

ABSTRACT

Network motifs are crucial building blocks of understanding and modeling complex networks for their capacity in characterizing higher-order interactions. Meanwhile, heterogeneous information networks (HINs) are ubiquitous in real-world applications, which often come with rich temporal information. We are hence motivated to study temporal motifs in the context of heterogeneous information networks. With examples from real-world datasets, we demonstrate HIN motifs can be armed with substantially more discriminability by incorporating temporal information. Furthermore, counting temporal HIN motif instances in large-scale networks is time consuming. We therefore develop efficient counting algorithm for the HIN motifs that are of the most interests in the literature. Empirical observations in the experiment have shown that interesting motif instances can be identified from large-scale HINs thanks to the improved discriminability of temporal HIN motifs, and the proposed efficient counting algorithm enjoys linear complexity that is multiple orders of magnitude faster than the baseline method in three real-world HINs.

CCS CONCEPTS

•Information systems → *Data mining;* **•Theory of computation** → **Graph algorithms analysis;**

KEYWORDS

Heterogeneous information networks, network motifs, algorithms, graph mining.

1 INTRODUCTION

Networks in real world applications are complex in nature, wherein higher-order interactions reflect certain mechanisms that can not be revealed merely by analyzing nodes and edges [2, 12, 14, 17, 23, 24]. As a result, network motifs have been studied to characterizing such higher-order interactions and have been proved to be useful in various classic network mining tasks such as clustering [2, 25] and representation learning [15, 27]. Meanwhile, the complex, real world networks are often heterogeneous due to the ubiquitous heterogeneity in the human society and the physical world [16, 19]. Therefore, we are motivated to study network motifs in the context of heterogeneous information networks (HINs) in the hope of developing motif-based network mining techniques that can better model heterogeneous network data.

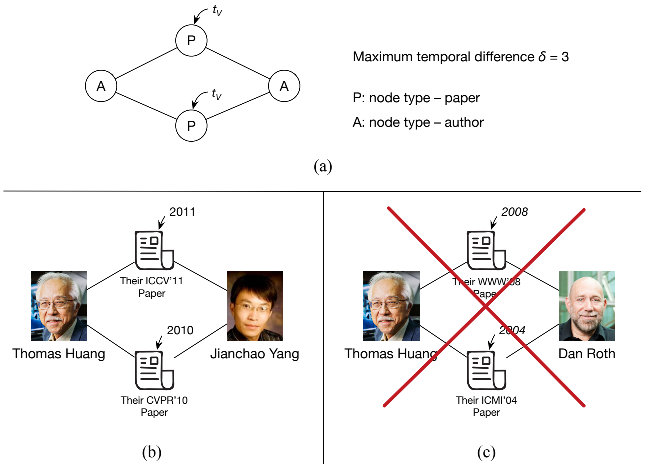


Figure 1: (a) An example temporal motif. (b) A motif instance under the example temporal motif. (c) An instance that does not match the example temporal motif for violating the time constraint.

A large portion of real world HINs are abstractions of the dynamic world, and their accompanying temporal information hence becomes an indispensable aspect of these networks. For instance, the heavily studied bibliographic network, DBLP, centers on the nodes that represent research papers, and each paper has its own publication time. Two authors jointly publishing two papers in consecutive years implies a higher likelihood of close collaboration than their coauthoring another paper five years after the first joint paper. Likewise, in the Twitter social network, each post is tweeted at a certain timestamp. A user using a tag in multiple posts that are tweeted within a few days may be a good indication that the user is engaged in an ongoing event. Hence, it is of interest to study the problem of incorporating temporal information into HIN motifs.

However, the time space is continuous. One method to deal with the continuity is to discretize it into several intervals, each represented by a node in the network. An event node is connected to a time node if the event has a time stamp that belongs to the time interval. The problem with this approach is that trivially mapping the temporal information into nodes of an HIN would inevitably result in information loss in the discretization process. Also, in order to be of pragmatic use, the attempt to incorporate time into HIN motifs should be accompanied by practical algorithms that could count such motifs in large-scale datasets. Therefore, we approach this problem by defining temporal motifs in HINs with time constraints. We use an example from a real-world bibliographic newtork, DBLP, to illustrate the idea. As shown in Figure 1, Jianchao Yang was an advisee of Thomas Huang who used to frequently co-author paper during Yang's Ph.D. study; while Dan Roth and Thomas Huang are

*These authors contributed equally to this work.

—, —

DOI: 10.475/123_4

both renowned professors who have published many papers and have co-authored papers for multiple times over the years. If the temporal information is neglected, both the instances in Figure 1 (b) and Figure 1 (c) would be considered as legit motif instance under the example motif in Figure 1 (a). However, by considering the time constraint, the instance for Roth and Huang in Figure 1 (c) no longer matches the example motif. In other words, an HIN motif can be armed with more discriminability by the temporal information, so that it would identify the close collaboration between Yang and Huang from that between Roth and Huang. Despite the utility of temporal information, it can be challenging to count temporal HIN motif instances in large networks. To tackle this challenge, we propose an efficient algorithm for a class of temporal HIN motifs, which is of the most interest according to our survey into the related work.

Lastly, we summarize our contributions as follows:

(1) We propose to study the problem of temporal motifs in the context of heterogeneous information networks, which capture the pervasive and informative temporal higher-order interactions in HINs.
(2) We identify a class of temporal HIN motifs that are of the most interests by surveying related work, and develop efficient algorithms to count instances of this motif class.
(3) Experiments with three real-world large-scale datasets demonstrate the utility of temporal HIN motifs and the superior efficiency of the proposed algorithms.

2 RELATED WORK

Network motif. Higher-order structures are crucial building blocks of complex networks [2, 12, 24], the understanding of which has been shown to be instrumental in many research domains such as neuroscience [17], biological networks [14], and social networks [23]. Such high-order structures are often referred to as network motifs or graphlets. One line of data mining research on network motifs has centered on efficiently counting motif instances including triangles and more complex motifs [1, 3, 8, 18]. Researchers have also found applications of motifs in clustering and network partitioning [2, 11, 22, 25]. One recent work [13] discusses the utility of temporal information in network motifs, which has been proven instrumental in various mining tasks for homogeneous networks.

Motifs in heterogeneous information network. Heterogeneous information network (HIN) has been extensively studied as a powerful and effective paradigm to model networked data with rich and informative type information [16, 19], upon which a great many methods have been proposed for applications such as classification, clustering, recommendation, and outlier detection [16, 19, 20, 26, 30].

Motifs in the context of HINs have been studied very recently [5–7, 9, 15, 27–29], where meta graphs and meta structures are sometimes used as synonyms to HIN motifs. Many of these works focus their scope on pairwise relationship such as relevance or similarity [5–7, 28, 29], while others tackle the problem of label propagation [9] or representation learning [15, 27]. Note that some of these prior works define meta graphs or meta structures to be directed acyclic graphs [7, 27–29], while in general the definition of HIN motifs, meta graphs, or meta structures is not restricted as such. One recent work [15] proposes a convolutional neural network method via motif and its application in HINs. The type of motif discussed in this work is defined to always have a target node, a context node, and multiple auxiliary nodes, which is a subset of motifs that are shown to be useful in heterogeneous networks.

Co-Author	TCount	TRank	Count	Rank	Percentage
Yun Fu	586	1	861	1	0.681
Shuicheng Yan	277	2	406	2	0.682
Ming Liu	240	3	406	2	0.591
Hao Tang	207	4	406	2	0.510
Xi Zhou	170	5	300	6	0.567
Mark H-J	158	6	406	2	0.389

Table 1: Co-authors with Thomas S. Huang and the temporal motif counts and non-temporal motif counts between Huang and each of them. The last co-author's full last name is Hasegawa-Johnson.

3 PRELIMINARIES

In this section, we define related concepts and notations.

Definition 3.1 (Temporal Heterogeneous Information Network). An information network is a directed graph $G = (\mathcal{V}, \mathcal{E})$ with a node type mapping $\varphi : \mathcal{V} \rightarrow \mathcal{T}$ and an edge type mapping $\psi : \mathcal{E} \rightarrow \mathcal{R}$. When the number of node types $|\mathcal{T}| > 1$ or the number of edge types $|\mathcal{R}| > 1$, the network is referred to as a heterogeneous information network (HIN). Particularly, an HIN is a **temporal heterogeneous information network** (temporal HIN) if it has a node time mapping t_V that maps a subset of nodes $\mathcal{V}_T \subseteq \mathcal{V}$ to their timestamps and an edge time mapping t_E that maps a subset of edge $\mathcal{E}_T \subseteq \mathcal{V}$ to their timestamps.

Note that we refer to $\mathcal{V}_T$ as the set of all temporal nodes and $\mathcal{E}_T$ at the set of all temporal edges, and $\mathcal{V}_T$ and $\mathcal{E}_T$ can be empty sets for a specific temporal HIN.

Definition 3.2 (Temporal Motifs in HINs). In an HIN $G = (\mathcal{V}, \mathcal{E})$, an **HIN motif** $m = (\{\varphi_i\}, \{\psi_j\})$ is a graph on the type level, described by a set of node types, $\{\varphi_i\} \subseteq \mathcal{T}$, and a set of edge types, $\{\psi_j\} \subseteq \mathcal{R}$. An **HIN motif instance** under motif m is a subgraph of the HIN $G^{(m)} = (\{v_i\}, \{e_j\}) \subseteq G$ such that $\varphi(v_i) = \varphi_i$ for all i and $\psi(v_j) = \psi_j$ for all j. Moreover, an HIN motif m is a **temporal HIN motif** if it has a further maximum temporal difference requirement, δ, such that each of its instances $G^{(m)} = (\{v_i\}, \{e_j\}) \subseteq G$ satisfies $\max(t_V(\{v_i\} \cap \mathcal{V}_T) \cup t_E(\{e_i\} \cap \mathcal{E}_T)) < \min(t_V(\{v_i\} \cap \mathcal{V}_T) \cup t_E(\{e_i\} \cap \mathcal{E}_T)) + \delta$.

We note that the motifs in HINs are sometime referred to as the meta graphs or meta structures as discussed in Section 2.

4 TEMPORAL HIN MOTIFS

Temporal information is integral to many heterogeneous information networks, and incorporating it into HIN motifs can largely boost their representational power. For example, we compare the effects of temporal motifs versus their non-temporal counterparts in the task of finding close authors in the DBLP dataset as shown in Table 1. If we count the total number of motifs between two authors without considering time information, we will find that Shuicheng Yan, Ming Liu, and Mark Hasegawa-Johnson all have the same

Rank with Temporal Motif	Rank with Non-temporal Motif
Illinois	Illinois
Massachusetts	New York
New York	Massachusetts

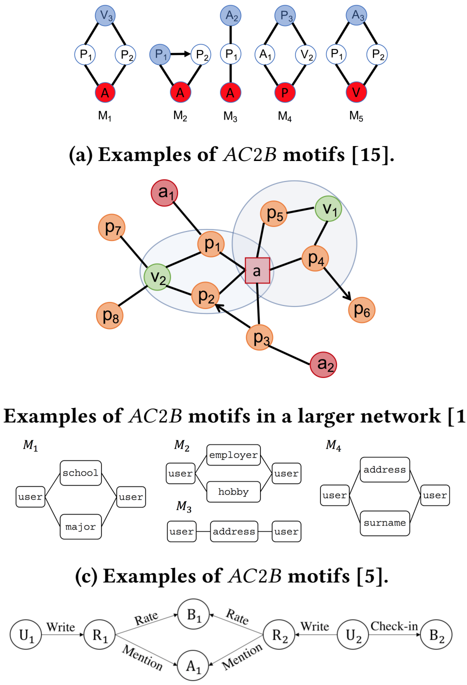
Table 2: Comparison showing the effect of temporal information on motifs: in which states is hockey the most popular sport?

number of co-authored papers with Thomas S. Huang. However, in terms of temporal motifs, Huang has nearly 50% more motifs with Liu and Yan than with Hasegawa-Johnson. One explanation for this observation arises when we examine the relationship between these authors. Among the list, Shuicheng Yan and Ming Liu are Huang's students, whereas Hasegawa-Johnson is a colleague of Huang within the same department.

Another example comes from the News dataset. We would like to briefly introduce the dataset first, with a more detailed description in Section 6.1. The dataset is a collection of news from 2013. Each news document is labeled with a timestamp, a location, and a topic. Moreover, the hierarchical information about locations and topics is also included. For locations, the "is a part of" relation is included: "Chicago" is part of "Illinois", which is part of the "United States". Topic-wise, the "is a subtopic of" relation is there: "Internet" is a sub-topic of "Technology". To facilitate information extraction, we construct an HIN based on this dataset. After that, we attempt to find the relationship between topics and locations, based on the information from the documents.

For a given topic, we compare the results on finding the most relevant locations using temporal and non-temporal methods. An example for the topic "hockey" is given in Table 2. For the ease of comparison, we restrict the location to be the states in the USA. The table shows the top three states mined using the temporal and the non-temporal methods. Clearly, the main difference lies in the ranking of Massachusetts and New York. According to an article from the New York Times [10] , Massachusetts has an estimated average of 66.9 hockey players per 10,000 population, whereas for New York the statistic is only 23.8. This tells us that hockey plays a more important role in Massachusetts than in New York. After a closer inspection, we find that compared with Massachusetts, New York receives more attention in the News dataset in general. Therefore, its superior amount of appearances may confuse the methods that do not consider temporal information. In contrast, by adding temporal information into consideration, a method can do better in filtering out the superfluous relations, and thus make the true information more visible.

In the context of heterogeneous information networks, researchers have studied network motifs under the name HIN motifs, meta graphs, or meta structures [5–7, 9, 15, 27–29]. By surveying these prior arts, we identify that many HIN motifs that the researchers use in practice can be generalized to a class, which we refer to as the *fusiform motifs*. A fusiform motif has the topology where two node types are connected by multiple intermediate node types in the middle as shown in Figure 3. For example, we lay out the HIN motifs used in the previous studies [5, 15, 28] in Figure 2a, Figure 2b, Figure 2c, and Figure 2d. It can been seen that these HIN motifs can all be characterized into fusiform motifs either directly or after collapsing certain internal nodes and edges. Specifically,



(a) Examples of $AC2B$ motifs [15].

(b) Examples of $AC2B$ motifs in a larger network [15].

(c) Examples of $AC2B$ motifs [5].

(d) Example of motifs that can be collapsed into $AC2B$ motifs [28].

Figure 2: Examples of $AC2B$ motifs in various contexts.

motifs M1, M4, and M5 from Figure 2a, the two motifs in circle in Figure 2b, motifs M1, M2, and M4 from Figure 2c are of the type of fusiform motifs, while the motif in Figure 2d would be of this type by collapsing a few nodes into edges. Specifically, in Figure 2d, node R1, as well as the three edges connected to it, is collapsed into two edges, one connecting U1 with A1, and the other connecting U1 with B1. Similarly, nodes R2 and U2, as well as the four edges connected to them, are collapsed into two edges, one connecting B2 with A1, and the other connecting B2 with B1.

Formally, the fusiform motif can be represented by $S = (N, M)$ where N is a set of node types, and M is a set of edge types. Inspired by the observation of real-world scenarios, we delve deep into several subclasses of fusiform motifs with $n = 4$, as shown in Figure 4. In the case of $2A2B$ motifs, where $\varphi_1 = \varphi_2 = A$ and $\varphi_3 = \varphi_4 = B$, N consists of four nodes $\{A_1, A_2, B_1, B_2\}$, $\varphi(N)$ contains two node types $\{A, B\}$, M includes four edges $\{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$, and $\psi(M)$ consists of only one edge type with one of the endpoints being type A, and the other endpoint being type B. On the other hand, in the case of the $AC2B$ motif, where $\varphi_1 = A$, $\varphi_2 = C$ and $\varphi_3 = \varphi_4 = B$, N consists of four nodes $\{A_1, C_1, B_1, B_2\}$, $\varphi(N)$ contains three node types $\{A, B, C\}$, and M includes four edges $\{A_1B_1, C_1B_1, A_1B_2, C_1B_2\}$, and $\psi(M)$ consists of two edge types: $A - B$-edge and $C - B$-edge. In the example of Figure 2c, the $AC2B$ motif $M1$ can be represented by (N, M) where N, the set of vertex types, can be represented as $\{User, School, Major\}$ and M, the set of edge types, can be represented as $\{User - School - edge, User - Major - edge\}$. For M2, $N = \{User, Employer, Hobby\}$ and $M = \{User - Employer - edge, User - Hobby - edge\}$.

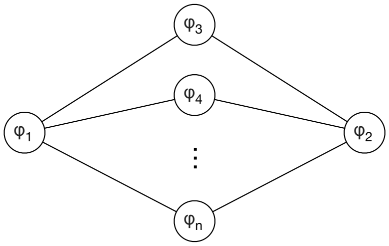


Figure 3: Fusiform motifs

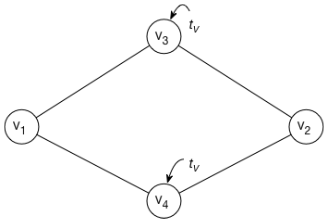


Figure 4: Temporal motifs of our interest. Here, $\varphi(v_3) = \varphi(v_4)$, but $\varphi(v_1)$ and $\varphi(v_2)$ can either be the same or be different. If $\varphi(v_1) = \varphi(v_2)$, then we refer to it as "the $2A2B$ motif". Otherwise, we refer to it as "the $AC2B$ motif".

Furthermore, temporal information naturally exists in the HIN motifs found in the previously mentioned cases. For example, in Figure 2a, note that the *AC2B* motif M_1 can be converted to a temporal motif by adding timestamps on the nodes P_1 and P_2, representing the time that the two papers are published. By enforcing a time constraint on the temporal motif, stronger relations between the nodes A and V_3 can be mined. Specifically, if we require that the timestamps on P_1 and P_2 be less than δ years away, then each instance of the motif represents that the author A published twice on venue V_3 within δ years (here $\delta = 3$ is a nice valuation that fits well into the context). This would establish a strong indication that the author's research interest aligns with the topic of the venue. In contrast, without the time information, such a motif would contain significantly less semantics. For example, even if an author published twice on a venue, if the the second paper was decades after the first one, then this pair of publications does not imply such a strong tie between the author and the venue as the case in which $\delta = 3$. With additional time constraint incorporated, we use *popular temporal HIN motifs* to refer to the above-mentioned *2A2B* and *AC2B* motifs with time information on nodes. In both cases, the time stamps are on the two type-*B* nodes. Figure 4 illustrates these two types of motifs. Formally, in the spirit of Definition 3.2, a temporal *2A2B* motif S is a graph (N, M) with parameter δ, where $N = \{A_1, A_2, B_1, B_2\}$, and $M = \{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$. Let $N_T = \{B_1, B_2\}$ and and $M_T = \emptyset$ be the sets of temporal nodes and temporal edges in S, respectively. According to the time constraints, $max(T) - min(T) < \delta$ where $T = \{\varphi(v)|v \in N_T\} \cup \{\psi(e)|e \in M_T\} = \{\varphi(B_1), \varphi(B_2)\}$. In a similar way, the *AC2B* motif can be defined, with $N = \{A, C, B_1, B_2\}$, $M = \{AB_1, AB_2, CB_1, CB_2\}$, $N_T = \{B_1, B_2\}$, and $M_T = \emptyset$.

In the next section, we show that efficient counting algorithms can be developed for fusiform temporal HIN motifs.

```
Algorithm 1: 2A2B Temporal Motif Counting
  Input  : as and bs are arrays of nodes of type-A and type-B
           respectively, sorted by time. Adjacency relations in
           HIN are also included.
  Output : Count of temporal motifs between type A nodes
  begin
      initialize $count\_total$ and $count\_window$
      $start \leftarrow 1$
      for $end \leftarrow 1..m$ do
          while $t_{start} + \delta \leqslant t_{end}$ do
              decrement_count_window($bs[start]$)
              $start+ = 1$
          increment_count_window_and_count_total($bs[end]$)
  Function decrement_count_window($b$):
      // $N_a(b)$ is the neighbor of node $b$ in the array of nodes $as$
      for $a_1, a_2 \in N_a(b), a_1 \neq a_2$ do
          $count\_window[a_1, a_2]- = 1$
  Function increment_count_window_and_count_total($b$):
      for $a_1, a_2 \in N_a(b), a_1 \neq a_2$ do
          $count\_total[a_1, a_2]+ = count\_window[a_1, a_2]$
          $count\_window[a_1, a_2]+ = 1$
```

5 COUNTING ALGORITHM

To begin with, we try to count the number of *2A2B* motif instances as we have discussed above and in Figure 4.

Without loss of generality, we only care about the relationship between two type-*A* entities. So we aggregate the count for each pair of type-*A* nodes. We utilize the schemes of sliding window and dynamic programming. First, we order the type-*B* nodes by their time stamps. Then, we loop through all type-*B* nodes based on the order. For each node, we first update the current window. This involves removing all nodes that fall out of the current window and then adding the node we are concerning about into the current window. Along the way, we use an array *count_window* to count the meta-path $A - B - A$ instances that exist in the current window. We also keep an array *count_total* for each pair of type-*A* entities. When we add a new node into the time window, we can form a motif as desired by combining the current type-*B* node with the previous nodes that *count_window* keeps track off. In this way, we can efficiently count the number of motifs. The pseudocode for this algorithm is shown in Algorithm 1.

We realize that correlating two entities of a single type might be too restrictive; it would be much more beneficial if we can come up with a measure of relationship between two nodes of different types. Therefore, instead of using two type-*A* nodes, we correlate one type-*A* entity with one type-*C* entity using two type-*B* entities, i.e. *AC2B* motifs, as discussed above and in Figure 4. The modified algorithm for this motif is shown in Algorithm 2.

6 EXPERIMENTS

In this section, we present the empirical observations made in real world HINs by leveraging temporal motifs, and evaluate the performance boosted by the proposed efficient counting algorithm.

Algorithm 2: AC2B Temporal Motif Counting

```
Input  : Nodes of types $A$, $B$, or $C$, and edges between them in HIN
Output : Count of temporal motifs between one type A node and one type C node
Function decrement_count_window($b$):
    for $a \in N_a(b), c \in N_c(b)$ do
        $count\_window[a, c] -= 1$
Function increment_count_window_and_count_total($b$):
    for $a \in N_a(b), c \in N_c(b)$ do
        $count\_total[a, c] += count\_window[a, c]$
        $count\_window[a, c] += 1$
```

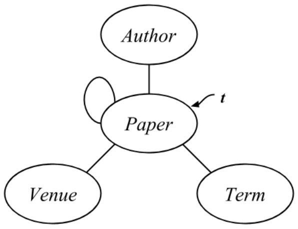
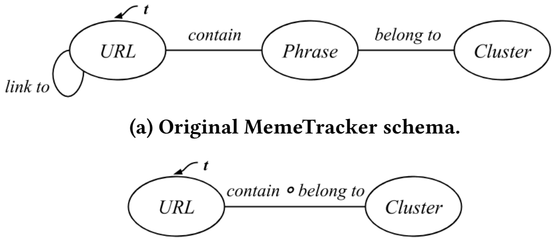


Figure 5: DBLP schema.

Node Type	Count
Paper	1595783
Author	1003836
Term	402687
Venue	7528
Year	62
Total	3009896

Edge Type	Count
Paper-Author	4650898
Paper-Paper	6510282
Paper-Term	12773973
Paper-Venue	1595783
Paper-Year	1595783
Total	27126719

Table 3: DBLP statistics.



(a) Original MemeTracker schema.

(b) Collapsed MemeTracker schema.

Figure 6: Nodes of type "Phrase" are removed in the collapsing of schema. As a result, the relation between a "URL" node and a "Cluster" node is a composition of "contain" and "belong-to".

Node Type	Count
URL	4455215
Phrase	310457
Cluster	71568
Total	4837240

Edge Type	Count
Phrase-Cluster	310457
Phrase-URL	210999824
URL-URL	418237269
Total	629547550

Table 4: MemeTracker statistics.

6.1 Data Description

Datasets. We use three publicly available real-world temporal HIN datasets: DBLP, MemeTracker and News.

- **DBLP** is a bibliographical network in the computer science domain [21]. There are four types of nodes in the network: author, paper, key term, and venue. The key terms are extracted and released by Chen et al. [4]. The edge types include authorship (aut.), term usage (term) and publishing venue (ven.) of a paper, and the reference relationship from a paper to another (ref.). It has a nonempty temporal node set, $\mathcal{V}_T = \{u \in \mathcal{V} \mid \varphi(u) = \text{paper}\}$, where a paper is timestamped by its publishing time. The corresponding network schema is depicted in Figure 5, with statistics shown in Table 3.
- **MemeTracker** aims at finding frequent quotes and phrases from a large collection of online texts including news and blogs. A temporal HIN can be constructed from this dataset. There are three types of nodes in this HIN: uniform resource locator (URL), phrase, and cluster. Specifically, each URL represents a document online. Phrases are extracted from these documents. Moreover, similar phrases are grouped into clusters. The edge types include the mentioning relationship between documents and phrases, and the belong-to relationship between phrases and clusters. Naturally, the time information is on each URL-typed node, representing the time stamp of the document. Formally the temporal node set of this dataset is $\mathcal{V}_T = \{u \in \mathcal{V} \mid \varphi(u) = \text{URL}\}$. The corresponding network schema is depicted in Figure 6a, with statistics shown in Table 4. In this project, we are interested in a collapsed version of this HIN. Specifically, the nodes for "phrases" are collapsed, allowing clusters to connect with documents directly. The network schema for the collapsed HIN is depicted in Figure 6b.
- **News** is a collection of news articles from year 2013. Each piece of news comes with a date, a location label, and a topic label. Moreover, the hierarchies of locations and topics are provided. For example, the dataset contains the information that Illinois is part of the USA, and "Internet" belongs to the topic of "Technology". A temporal HIN can be constructed from this dataset. There are three types of nodes in this HIN: document, location, and topic. Specifically, each document represents a news article, and its location and topic labels are provided. The edge types include the "talks about" relationship between documents and locations, and between documents and topics. Moreover, directed edges exist between location nodes, representing the hierarchy of locations. Likewise, the edges representing the hierarchy of topics are also included. Naturally, the time information is on the document-typed nodes, representing the time stamp of the document. Formally the temporal node set of this dataset is $\mathcal{V}_T = \{u \in \mathcal{V} \mid \varphi(u) = \text{document}\}$. The corresponding network schema is depicted in Figure 7, with statistics shown in Table 5.

6.2 Empirical Observations

By applying our algorithm on the data we have described in 6.1, we found several different observations based on the particular scenario that each dataset describes.

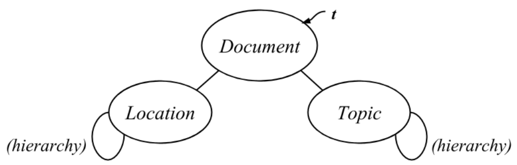


Figure 7: News schema. Note the hierarchical relations between different nodes of locations and between different nodes of topics.

Node Type	Count
Document	41959
Location	354
Topic	60
Total	42738

Edge Type	Count
Document-Location	41959
Document-Topic	41959
Location-Location	644
Topic-Topic	69
Total	84631

Table 5: News statistics. The data range over 365 days.

Co-Author	TCount	TRank	Count	Rank	Percentage
Mladen Kolar	11	8	28	5	0.393
Fei-Fei Li	22	4	28	5	0.786
Noah A. Smith	15	6	21	7	0.714
Jacob Eisenstein	19	5	21	7	0.905
Gunhee Kim	11	8	21	7	0.524

Table 6: Co-authors with Eric P. Xing based on temporal motif counts and non-temporal motif counts.

DBLP. In this dataset, we investigate the relationship between authors based on the temporal motifs of their co-authored papers. The time stamp in this data is given only as the year when the paper is published, and we set the time window $\delta = 3$. A list of selected co-authors with Eric P. Xing is shown in Table 6, ranked by their non-temporal motif counts.

MemeTracker. In this dataset, we try to find clusters of phrases that often appear together in a short period of time based on the common URLs that they appear in. The experimental result is shown in Table 7 and Table 8. In fact, it can be seen that pairs found by both methods are meaningful. Those picked by the temporal approach have a clearer clue: they are all from Obama's inaugural address. In contrast, those picked up by the non-temporal approach include the pairs formed by an entity without strong evidence of time, and a related remark about that entity. Therefore, compared with the non-temporal motifs, the temporal ones are better at capturing the strong ties in time, not just semantically relatedness.

News. In this dataset, we are trying to correlate information of two different types: topic and location. We find the following interesting cases, as shown in Table 9 and Table 10. We notice from the result that not all locations are actually located within the Asia Pacific region; rather, these locations correlate strongly with Asia Pacific in the news. Beijing and Hong Kong rank high in the list because of their important roles in the politics and economics of the Asia Pacific region. However, New York and San Francisco Peninsula are also in the list, probably owing to their strong economic ties with the Asia Pacific region. Similarly, motif counts correctly identify the critical locations in the business sector, from Boston, New York, and the Northeast corridor, to Michigan in the Midwest, and California on the west coast.

hope over fear unity of purpose over conflict and discord	to the muslim world we seek a new way forward based on mutual interest and mutual respect
hope over fear unity of purpose over conflict and discord	starting today we must pick ourselves up dust ourselves off and begin again the work of remaking america
hope over fear unity of purpose over conflict and discord	what is required of us now is a new era of responsibility a recognition on the part of every american that we have duties to ourselves our nation and the world duties that we do not grudgingly accept but rather seize gladly

Table 7: Pairs of phrases that have high temporal motif counts.

joe the plumber	i think when you spread the wealth around it's good for everybody
hope over fear unity of purpose over conflict and discord	what is required of us now is a new era of responsibility a recognition on the part of every american that we have duties to ourselves our nation and the world duties that we do not grudgingly accept but rather seize gladly
the daily show with jon stewart's	everybody did saturday night live the colbert report they did the jon stewart show by showing they want to be closer to people politicians are showing they want to be more like us

Table 8: Pairs of phrases that have high non-temporal motif counts.

Location	Temporal Motif Count
Beijing	2380
Hong Kong	790
New York	153
San Francisco Peninsula	102
Shandong	90

Table 9: Temporal motif count with "Asia Pacific" as the topic.

Location	Temporal Motif Count
New York	1496
Massachusetts	69
California	56
Northeast megalopolis	33
Michigan	20

Table 10: Temporal motif count with "business sectors" as the topic.

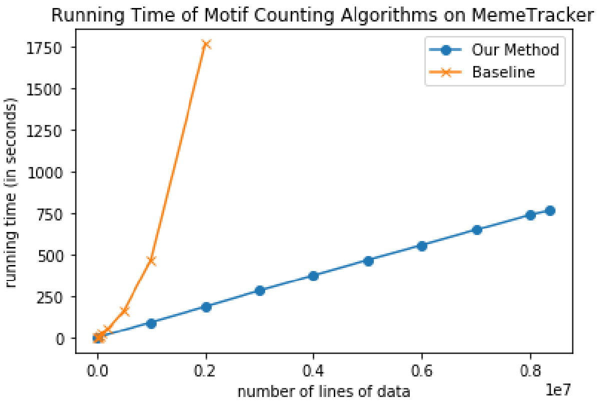


Figure 8: Comparison of the running time of our algorithm versus that of baseline method. Our algorithm runs in time linear to the size of the data.

6.3 Efficiency Study

Compared to the baseline method, which generates all pairs of possible motifs and filters them by their time stamp differences, our algorithm runs in linear to the input size as illustrated in Figure 8. Utilizing our algorithm, we efficiently aggregate the count of edges at the first place, instead of enumerating all the possible combinations. Whereas the running time of the baseline is superlinear to the data size, our method runs in linear time.

7 CONCLUSION AND FUTURE WORKS

In response to the growing demand of finding time-sensitive frequent patterns in heterogeneous information networks, we proposed a series of algorithms that address this problem in linear time. Throughout the paper, we have encountered several case studies on how these newly proposed algorithms can help understand the temporal information on graphs in a way that is not efficiently achieved before. This paper in no way sets a definitive tone in discovering interesting temporal motifs in heterogeneous networks; rather, it is a proposal to bring a richer mix of information into the analysis. Specifically, one can extend our algorithms and the way counts are interpreted to apply these algorithms in a wider range of applications. We believe that the incorporation of temporal information in network analysis can bring a new set of discoveries in the near future.

REFERENCES

[1] Nesreen K Ahmed, Jennifer Neville, Ryan A Rossi, and Nick Duffield. 2015. Efficient graphlet counting for large networks. In *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 1–10.

[2] Austin R Benson, David F Gleich, and Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* 353, 6295 (2016), 163–166.

[3] Marco Bressan, Flavio Chierichetti, Ravi Kumar, Stefano Leucci, and Alessandro Panconesi. 2017. Counting graphlets: Space vs time. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 557–566.

[4] Ting Chen and Yizhou Sun. 2017. Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification. In *WSDM*. ACM.

[5] Yuan Fang, Wenqing Lin, Vincent W Zheng, Min Wu, Kevin Chen-Chuan Chang, and Xiao-Li Li. 2016. Semantic Proximity Search on Graphs with Metagraph-based Learning. In *ICDE*. IEEE.

[6] Valeria Fionda and Giuseppe Pirrò. 2017. Meta Structures in Knowledge Graphs. In *International Semantic Web Conference*. Springer, 296–312.

[7] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. 2016. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. In *KDD*. ACM.

[8] Madhav Jha, C Seshadhri, and Ali Pinar. 2015. Path sampling: A fast and provable method for estimating 4-vertex subgraph counts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 495–505.

[9] He Jiang, Yangqiu Song, Chenguang Wang, Ming Zhang, and Yizhou Sun. 2017. Semi-supervised learning over heterogeneous information networks by ensemble of meta-graph guided random walks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1944–1950.

[10] Jeff Klein. 2011. Hockeyfis Heartland, State by State. https://slapshot.blogs.nytimes.com/2011/02/20/hockeys-heartland-state-by-state/. (Feb. 2011). Accessed: 2018-05-15.

[11] Christine Klymko, David Gleich, and Tamara G Kolda. 2014. Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874* (2014).

[12] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[13] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. 2017. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 601–610.

[14] Nataša Pržulj. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* 23, 2 (2007), e177–e183.

[15] Aravind Sankar, Xinyang Zhang, and Kevin Chen-Chuan Chang. 2017. Motif-based Convolutional Neural Network on Graphs. *arXiv preprint arXiv:1711.05697* (2017).

[16] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2017. A survey of heterogeneous information network analysis. *TKDE* 29, 1 (2017), 17–37.

[17] Olaf Sporns and Rolf Kötter. 2004. Motifs in brain networks. *PLoS biology* 2, 11 (2004), e369.

[18] Lorenzo De Stefani, Alessandro Epasto, Matteo Riondato, and Eli Upfal. 2017. Triest: Counting local and global triangles in fully dynamic streams with fixed memory size. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 4 (2017), 43.

[19] Yizhou Sun and Jiawei Han. 2013. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explorations* 14, 2 (2013), 20–28.

[20] Yizhou Sun, Yintao Yu, and Jiawei Han. 2009. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*. ACM, 797–806.

[21] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 990–998.

[22] Charalampos E Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. 2017. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1451–1460.

[23] Johan Ugander, Lars Backstrom, and Jon Kleinberg. 2013. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1307–1318.

[24] Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. 2014. Revealing the hidden language of complex networks. *Scientific reports* 4 (2014), 4547.

[25] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. 2017. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 555–564.

[26] Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. 2014. Personalized entity recommendation: A heterogeneous information network approach. In *WSDM*. ACM.

[27] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. Meta-Graph2Vec: Complex Semantic Path Augmented Heterogeneous Network Embedding. *arXiv preprint arXiv:1803.02533* (2018).

[28] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Dik Lun Lee. 2017. Meta-graph based recommendation fusion over heterogeneous information networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 635–644.

[29] Yu Zhou, Jianbin Huang, Heli Sun, Yizhou Sun, and Hong Chong. 2017. DMSS: A Robust Deep Meta Structure Based Similarity Measure in Heterogeneous Information Networks. *arXiv preprint arXiv:1712.09008* (2017).

[30] Honglei Zhuang, Jing Zhang, George Brova, Jie Tang, Hasan Cam, Xifeng Yan, and Jiawei Han. 2014. Mining query-based subnetwork outliers in heterogeneous information networks. In *ICDM*. IEEE.