

# Heterogeneous Graphlets

RYAN A. ROSSI, Adobe Research, USA  
 NESREEN K. AHMED, Intel Labs, USA  
 ALDO CARRANZA, Stanford University, USA  
 DAVID ARBOUR, Adobe Research, USA  
 ANUP RAO, Adobe Research, USA  
 SUNGCHUL KIM, Adobe Research, USA  
 EUNYEE KOH, Adobe Research, USA

In this paper, we introduce a generalization of graphlets to heterogeneous networks called *typed graphlets*. Informally, typed graphlets are small typed induced subgraphs. Typed graphlets generalize graphlets to rich heterogeneous networks as they explicitly capture the higher-order typed connectivity patterns in such networks. To address this problem, we describe a general framework for counting the occurrences of such typed graphlets. The proposed algorithms leverage a number of combinatorial relationships for different typed graphlets. For each edge, we count a few typed graphlets, and with these counts along with the combinatorial relationships, we obtain the exact counts of the other typed graphlets in  $o(1)$  constant time. Notably, the worst-case time complexity of the proposed approach matches the time complexity of the best known untyped algorithm. In addition, the approach lends itself to an efficient lock-free and asynchronous parallel implementation. While there are no existing methods for typed graphlets, there has been some work that focused on computing a different and much simpler notion called colored graphlet. The experiments confirm that our proposed approach is orders of magnitude faster *and* more space-efficient than methods for computing the simpler notion of colored graphlet. Unlike these methods that take hours on small networks, the proposed approach takes only seconds on large networks with millions of edges. Notably, since typed graphlet is more general than colored graphlet (and untyped graphlets), the counts of various typed graphlets can be combined to obtain the counts of the much simpler notion of colored graphlets. The proposed methods give rise to new opportunities and applications for typed graphlets.

Additional Key Words and Phrases: Heterogeneous graphlets, typed graphlets, position-aware typed graphlets, labeled graphlets, heterogeneous network motifs, heterogeneous networks, attributed graphs, large networks

## 1 INTRODUCTION

Higher-order connectivity patterns such as small induced subgraphs called graphlets<sup>1</sup> are known to be the fundamental building blocks of homogeneous networks [Milo et al. 2002] and are essential for modeling and understanding the fundamental components of these networks [Ahmed et al. 2015, 2016; Benson et al. 2016]. Furthermore, graphlets are also important for many predictive and descriptive modeling application tasks [Ahmed et al. 2017a; Hayes et al. 2013; Lichtenwalter and Chawla 2012; Milenković and Pržulj 2008; Milo et al. 2002; Pržulj et al. 2004; Shervashidze et al. 2009; Vishwanathan et al. 2010; Zhang et al. 2016] such as image processing and computer vision [Zhang et al. 2016, 2013], network alignment [Crawford and Milenković 2015; Koyutürk

<sup>1</sup>The terms graphlet and induced subgraph are used interchangeably.

---

Authors' addresses: Ryan A. Rossi, Adobe Research, 345 Park Ave, San Jose, CA, USA, rrossi@adobe.com; Nesreen K. Ahmed, Intel Labs, 3065 Bowers Avenue, Santa Clara, CA, USA, nesreen.k.ahmed@intel.com; Aldo Carranza, Stanford University, Huang Building 475 Via Ortega, Stanford, CA, USA, aldogael@stanford.edu; David Arbour, Adobe Research, 345 Park Ave, San Jose, CA, USA, arbour@adobe.com; Anup Rao, Adobe Research, 345 Park Ave, San Jose, CA, USA, anuprao@adobe.com; Sungchul Kim, Adobe Research, 345 Park Ave, San Jose, CA, USA, sukim@adobe.com; Eunyee Koh, Adobe Research, 345 Park Ave, San Jose, CA, USA, eunyee@adobe.com.

---

2020. 1556-4681/2020/10-ART9 \$15.00  
<https://doi.org/10.1145/3418773>

et al. 2006; Milenković and Pržulj 2008; Pržulj 2007], classification [Shervashidze et al. 2009; Vishwanathan et al. 2010], visualization and sensemaking [Ahmed et al. 2015, 2016], dynamic network analysis [Hulovatyy et al. 2015; Kovanen et al. 2011], community detection [Benson et al. 2016; Palla et al. 2005; Radicchi et al. 2004; Solava et al. 2012], role discovery [Ahmed et al. 2017b, 2018], anomaly detection [Akoglu et al. 2015; Noble and Cook 2003], and link prediction [Rossi et al. 2018].

However, such (untyped) graphlets are *unable* to capture the rich typed connectivity patterns in more complex networks such as those that are heterogeneous, which includes bipartite, k-partite, k-star, and attributed graphs as special cases, among others. In heterogeneous networks, nodes and edges can be of different types and explicitly modeling such types is crucial [Acar et al. 2011; Banerjee et al. 2007; Carranza et al. 2018; Gu et al. 2018]. Such heterogeneous networks arise ubiquitously in the natural world where nodes and edges of multiple types are observed, e.g., between humans [Kong et al. 2013], neurons [Bassett and Bullmore 2006; Bullmore and Sporns 2009], routers and autonomous systems (ASes) [Rossi et al. 2013], web pages [Yin et al. 2009], devices & sensors [Eagle and Pentland 2006], infrastructure (roads, airports, power stations) [Wang and Rong 2009], vehicles (cars, satellites, UAVs) [Hung et al. 2008], and information in general [Rossi and Zhou 2016; Sun et al. 2011; Yu et al. 2014].

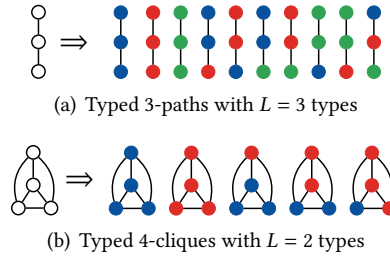


Fig. 1. Examples of typed (heterogeneous) graphlets

In this work, we introduce the notion of a *typed graphlet* that naturally generalizes the notion of graphlet to heterogeneous networks.<sup>2</sup> Typed graphlets generalize the notion of graphlets to rich heterogeneous networks as they capture both the induced subgraph of interest and the types associated with the nodes in the induced subgraph (Figure 1). These small induced typed subgraphs are the fundamental *building blocks of rich heterogeneous networks*. Typed graphlets naturally capture the higher-order typed connectivity patterns in bipartite, k-partite, signed, k-star, attributed graphs, and more generally heterogeneous networks. As such, typed graphlets are useful for a wide variety of predictive and descriptive modeling applications in these rich complex networks. Closest work related to our own has focused on colored graphlets [Gu et al. 2018; Ribeiro and Silva 2014], which is a different problem. See Figure 2 for an intuitive illustration of the difference between the proposed notion of typed graphlets and recent work that focuses on colored graphlets.

Despite their fundamental and practical importance, counting typed graphlets remains a challenging and unsolved problem. To address this problem, we propose a fast, parallel, and space-efficient framework for counting typed graphlets in large networks. The time complexity is provably optimal and matches the time complexity of the best known untyped graphlet counting algorithm, i.e., PGD [Ahmed et al. 2015] and variants based on it [Dave et al. 2017; Pinar et al. 2017]. Using non-trivial combinatorial relationships between lower-order  $(k-1)$ -node typed graphlets, we derive equations that allow us to compute many of the  $k$ -node typed graphlet counts in  $o(1)$  constant

<sup>2</sup>The terms heterogeneous and typed graphlet are used interchangeably.

Table 1. Summary of notation. Matrices are bold upright letters; vectors are bold lowercase letters.

$G$	graph
$H, F$	graphlet of $G$
$N, M$	number of nodes $N =  V $ and edges $M =  E $ in the graph
$K$	size of a graphlet (# nodes)
$L$	number of types ( <i>i.e.</i> , labels)
$\mathcal{H}$	set of all untyped graphlets in $G$
$\mathcal{H}_T$	set of all typed graphlets in $G$
$T$	# of typed graphlets $T =  \mathcal{H}_T $ observed in $G$ with $L$ types
$T_{\max}$	# of possible typed graphlets with $L$ types, hence $T \leq T_{\max}$
$T_H$	# of different typed graphlets for a particular graphlet $H \in \mathcal{H}$
$\mathcal{T}_V$	set of node types in $G$
$\mathcal{T}_E$	set of edge types in $G$
$\phi$	type function $\phi : V \rightarrow \mathcal{T}_V$
$\xi$	type function $\xi : E \rightarrow \mathcal{T}_E$
$\mathbf{t}$	$K$ -dimensional type vector $\mathbf{t} = [\phi_{w_1} \cdots \phi_{w_K}]$
$f_{ij}(H, \mathbf{t})$	# of instances of graphlet $H$ that contain nodes $i$ and $j$ with type vector $\mathbf{t}$
$\mathbb{F}$	an arbitrary typed graphlet hash function (Section 4.5)
$\Delta$	maximum degree of a node in $G$
$\Gamma_i^t$	set of neighbors of node $i$ with type $t$
$d_i^t$	degree of node $i$ with type $t$ , $d_i^t =  \Gamma_i^t $
$T_{ij}^t$	set of nodes of type $t$ that form typed triangles with $i$ and $j$
$S_i^t, S_j^t$	set of nodes of type $t$ that form typed 3-node stars centered at $i$ (or $j$ )
$\mathcal{M}_{ij}$	set of typed graphlets for a given pair of nodes $(i, j)$
$\mathcal{X}_{ij}$	nonzero (typed-graphlet, count) pairs for edge $(i, j) \in E$
$\Psi$	hash table for checking whether a node is connected to $i$ or $j$ and its "relationship" ( <i>e.g.</i> , $\lambda_1, \lambda_2, \lambda_3$ ) in constant time

time. Thus, we avoid explicit enumeration of many typed graphlets by simply computing the exact count directly in constant time using the discovered combinatorial relationships. For every edge, we count a few typed graphlets and obtain the exact counts of the remaining typed graphlets in  $o(1)$  constant time. Furthermore, we store only the nonzero typed graphlet counts for every edge. To better handle large-scale heterogeneous networks with an arbitrary number of types, we propose an efficient parallel algorithm for typed graphlets that scales almost linearly as the number of processing units increases. As an aside, this paper focuses on counting typed graphlets with up to four nodes. Typed graphlets of a larger size are outside the scope of this paper and left for future work. However, the ideas and theoretical foundations formalized in this work naturally extend to typed graphlets of larger sizes (See Section 4.8 for further discussion).

Theoretically, we show that typed graphlets are more powerful and encode more information than untyped graphlets. In addition, we theoretically show the worst-case time and space complexity of the proposed framework. Notably, the time complexity of the proposed approach is shown to be equivalent to the best untyped graphlet counting algorithm. Furthermore, we derive many of the typed graphlets directly in  $o(1)$  constant time using counts of lower-order  $(k-1)$ -node typed graphlets.

Empirically, the proposed approach is shown to be orders of magnitude faster than state-of-the-art methods for the simpler colored graphlet counting problem. In particular, we observe between 89 and 10,981 times speedup in runtime performance compared to the best method. Notably, on graphs of even moderate size (thousands of nodes/edges), these approaches fail to finish in a reasonable amount of time (24 hours). In terms of space, the proposed approach uses between 42x and 776x

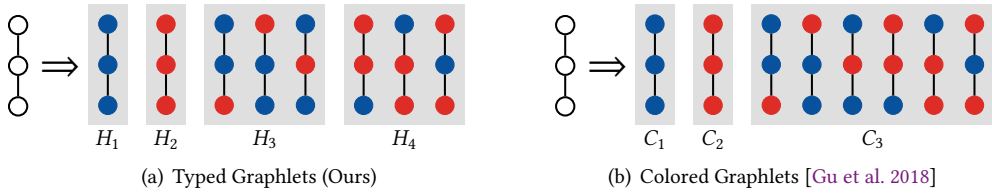


Fig. 2. **Typed graphlets vs. colored graphlets.** The intuitive example shows the difference between typed graphlets that are formally defined in this paper and colored graphlets from [Gu et al. 2018]. In particular, (a) shows the typed 3-paths with  $L = 2$  types whereas (b) shows the “colored 3-paths”. In the above example, there are three colored graphlets, that is, the last 4 typed graphlets are considered a single colored graphlet. Note given  $L$  colors, there are  $2^L - 1$  colored graphlets. However, for  $K$  nodes and  $L$  types, there are  $\binom{L+K-1}{K}$  typed graphlets.

less space than these methods. We also demonstrate the parallel scaling of the parallel algorithm and observe nearly linear speedups as the number of processing units increases. In addition to real-world graphs from a wide range of domains, we also show results on a number of synthetically generated graphs from a variety of graph models. Finally, we demonstrate the utility of typed graphlets for exploratory network analysis using a variety of well-known networks.

Compared to the untyped/homogeneous graphlet counting problem (which has found many important applications [Ahmed et al. 2017b, 2018; Akoglu et al. 2015; Benson et al. 2016; Koyutürk et al. 2006; Noble and Cook 2003; Pržulj 2007; Rossi et al. 2018; Shervashidze et al. 2009; Solava et al. 2012; Vishwanathan et al. 2010]), typed graphlets are more powerful containing a significant amount of additional information. We show this formally using information theory (Section 7) and demonstrate the importance of typed graphlets empirically using real-world graphs for exploratory analysis (Section 8.7) and graph-based predictive modeling (Section 8.8). Importantly, we find that only a handful of the possible typed graphlets actually occur in the real-world graphs studied in this work (Table 7). Furthermore, among the typed graphlets with nonzero counts (*i.e.*, the typed graphlets that actually occur in  $G$ ), we find that a few of those typed graphlets occur very frequently while the vast majority have very few occurrences (see Figure 12 and Figure 14 for a few examples). This observation indicates a power-law relationship between the counts of the different typed graphlets. The rare typed graphlets (*i.e.*, typed graphlets that rarely occur in the graph) also contain useful information as the appearance of these typed graphlets may indicate anomalies/outliers or simply unique structural behaviors that are fundamentally important but extremely difficult to identify using traditional methods. Moreover, the typed graphlets found to be important are easily interpretable and provide key insights into the structure and underlying phenomena governing the formation of the complex network that would otherwise be hidden using traditional untyped methods, see Section 8.7 for further details. Finally, we also demonstrate the effectiveness of typed graphlets in Section 8.8 for improving a predictive modeling task.

This work introduces and formally defines a generalization of the notion of graphlet to heterogeneous networks called *typed graphlets*. We describe a general framework for counting the proposed formalization of typed graphlets. The proposed framework has the following desired properties:

- **Fast:** The approach is fast for large graphs by leveraging novel *non-trivial combinatorial relationships* to derive many of the typed graphlets in  $o(1)$  constant time. Theoretically, the worst-case time complexity is shown to match the best untyped graphlet algorithm (Section 7.1). As shown in Table 5-6, the approach is orders of magnitude faster than recent methods proposed for the simpler colored graphlet problem.

- **Space-Efficient:** The approach is space-efficient by hashing and storing only the typed graphlet counts that appear on a given edge.
- **Scalable for Large Networks:** The proposed approach is scalable for large heterogeneous networks. In particular, the approach scales nearly linearly as the size of the graph increases.
- **Parallel:** The typed graphlet approach lends itself to an efficient lock-free & asynchronous parallel implementation. We observe near-linear parallel scaling results in Section 8.3.
- **Effectiveness:** We demonstrate the utility of typed graphlets for graph mining/exploratory analysis (Section 8.7) and predictive modeling (Section 8.8) where leveraging typed graphlets significantly improves predictive performance. This work brings new opportunities to leverage typed graphlets for many other real-world applications.

## 2 RELATED WORK

Closest work related to our own is that of colored graphlets [Gu et al. 2018; Ribeiro and Silva 2014]. However, the notion of colored graphlet is different from the notion of typed graphlets (and position-aware typed graphlets (Def. 8)) that are formally defined and investigated in this paper. For an intuitive illustration of the difference between the proposed notion of typed graphlets and the colored graphlet counting problem studied in prior work, see Figure 2. Besides the difference in problem, all of the prior work has focused on counting colored graphlets for *nodes* whereas this paper focuses on the problem of counting typed graphlets for *edges* (or more generally, between a pair of nodes  $i$  and  $j$ ). It is straightforward to see that the definition of colored graphlets from Gu et al. [2018] is only able to cover a subset of the typed graphlets given by our definition. Thus, the notion of typed graphlet described in our work is more general than the notion of colored graphlet. Besides the fundamental difference in problem as shown in Figure 2, that work also focused mainly on the application to network alignment (using very small networks) and not on the approach for computing colored graphlets. Nevertheless, the method GC used in that work and the other methods for colored graphlets are only able to handle extremely small graphs as shown in Table 5.

Despite the difference between colored and typed graphlets (Figure 2), the approach proposed in this work also differs from the colored graphlet methods in three fundamental ways. First, while we leverage new combinatorial relationships to derive a number of typed graphlets in  $o(1)$  constant time, GC and other colored graphlet methods must enumerate all graphlets in order to obtain their color configuration. Therefore, our approach is significantly faster (even though we count typed graphlets, a more complex and representationally powerful notion) than these methods as they require a lot of extra work to compute the colored graphlets that our approach can derive in constant time.<sup>3</sup> For instance, the small citeseer graph with only 3.3k nodes and 4.5k edges takes 46.27 seconds using the best method (for colored graphlets) whereas our approach for typed graphlets takes only a fraction of a second, notably,  $2/100$  seconds. In addition, while the methods for colored graphlets (a simpler relaxation of typed graphlet, see Figure 2) are only able to handle small networks, our approach naturally scales to large networks with millions of nodes and edges (Section 8). Second, our approach is significantly more space-efficient and stores only the nonzero counts of the typed graphlets discovered at each edge (Section 8.2). Third, our approach lends itself to an efficient, lock-free, and asynchronous parallelization. As an aside, unlike the methods for colored graphlets, our approach enumerates only a few typed graphlets and derives the remaining typed graphlets in  $o(1)$  constant time using new non-trivial combinatorial relationships that involve counts of lower-order typed graphlets. These lower-order typed graphlet counts are used as building blocks

<sup>3</sup>Notice from Figure 2 that any method for counting typed graphlets can by definition be used to count colored graphlets, but not vice-versa.

to directly derive many of the higher-order typed graphlet counts directly without any enumeration or knowledge of the explicit node types. Therefore, the worst-case time complexity of the proposed approach is equivalent to the best known untyped/homogeneous graphlet algorithm (as shown formally in Section 7).

### 3 HETEROGENEOUS GRAPHLETS

This section introduces a generalization of graphlets called *heterogeneous graphlets* (or simply *typed graphlets*). See Table 1 for a summary of key notation.

#### 3.1 Heterogeneous Graph Model

We use the following heterogeneous graph formulation:

**DEFINITION 1 (HETEROGENEOUS NETWORK).** *A heterogeneous network is defined as  $G = (V, E)$  consisting of a set of node objects  $V$  and a set of edges  $E$  connecting the nodes in  $V$ . A heterogeneous network also has a node type mapping function  $\phi : V \rightarrow \mathcal{T}_V$  and an edge type mapping function defined as  $\xi : E \rightarrow \mathcal{T}_E$  where  $\mathcal{T}_V$  and  $\mathcal{T}_E$  denote the set of node object types and edge types, respectively. The type of node  $i$  is denoted as  $\phi_i$  whereas the type of edge  $e = (i, j) \in E$  is denoted as  $\xi_{ij} = \xi_e$ .*

A few special cases of heterogeneous networks are shown in Figure 3.

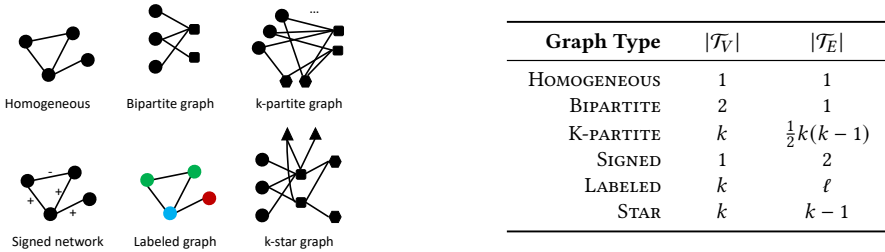


Fig. 3. *Typed graphlets* are useful for a wide variety of graphs. These graphs are only a few examples that are naturally supported by the proposed framework.

#### 3.2 Graphlet Generalization

In this section, we introduce a more general notion of graphlet called *typed graphlet* that naturally extends to both homogeneous and general heterogeneous networks. We use  $G$  to represent a graph and  $H$  or  $F$  to represent graphlets.

**3.2.1 Untyped Graphlets.** We begin by defining untyped graphlets for graphs with a single type.

**DEFINITION 2 (UNTYPED GRAPHLET).** *An untyped graphlet  $H$  is a connected induced subgraph of  $G$ .*

Given a graphlet in some graph, it may be the case that we can find other topologically identical “appearances” of this structure in that graph. We call these “appearances” *graphlet instances*.

**DEFINITION 3 (UNTYPED GRAPHLET INSTANCE).** *An instance of an untyped graphlet  $H$  in graph  $G$  is an untyped graphlet  $F$  in  $G$  that is isomorphic to  $H$ .*

**3.2.2 Typed Graphlets.** In heterogeneous graphs, nodes/edges can be of many different types and so explicitly and jointly modeling such types is essential. In this work, we introduce the notion of a *typed graphlet* that explicitly captures both the connectivity pattern of interest and the types. Notice that typed graphlets are a generalization of graphlets to heterogeneous networks.

**DEFINITION 4 (TYPED GRAPHLET).** *A typed graphlet of a graph  $G = (V, E, \phi, \xi)$  is a connected induced heterogeneous subgraph  $H = (V', E', \phi', \xi')$  of  $G$  such that*

- (1)  $(V', E')$  is a graphlet of  $(V, E)$ ,
- (2)  $\phi' = \phi|_{V'}$ , that is,  $\phi'$  is the restriction of  $\phi$  to  $V'$ ,
- (3)  $\xi' = \xi|_{E'}$ , that is,  $\xi'$  is the restriction of  $\xi$  to  $E'$ .

The terms typed graphlet and heterogeneous graphlet are used interchangeably. See Figure 1 for examples of typed graphlets and untyped graphlets (in which the type structure is ignored). We can consider the presence of topologically identical “appearances” of a typed graphlet in a graph.

**DEFINITION 5 (TYPED GRAPHLET INSTANCE).** *An instance of a typed graphlet  $H = (V', E', \phi', \xi')$  of graph  $G$  is a typed graphlet  $F = (V'', E'', \phi'', \xi'')$  of  $G$  such that*

- (1)  $(V'', E'')$  is isomorphic to  $(V', E')$ ,
- (2)  $\mathcal{T}_{V''} = \mathcal{T}_{V'}$  and  $\mathcal{T}_{E''} = \mathcal{T}_{E'}$ , that is, the multisets of node and edge types are correspondingly equal. The set of typed graphlet instances of  $H$  in  $G$  is denoted as  $I_G(H)$ .

Comparing the above definitions of graphlet and typed graphlet, we see at first glance that typed graphlets are nontrivial extensions of their homogeneous counterparts. The “position” of an edge (node) in a typed graphlet is often topologically important, e.g., an edge at the end of the 4-path vs. an edge at the center of a 4-path. These topological differences of a typed graphlet are called (automorphism) *typed orbits* since they take into account “symmetries” between edges (nodes) of a graphlet. Typed graphlet orbits are a generalization of (homogeneous) graphlet orbits [Pržulj 2007].

### 3.3 Number of Typed Graphlets

For a single  $K$ -node untyped graphlet (e.g.,  $K$ -clique), the number of *typed graphlets* with  $L$  types is:

$$\binom{L}{K} = \binom{L+K-1}{K} \tag{1}$$

where  $L$  = number of types and  $K$  = size of the graphlet (# of nodes). Table 2 shows the number of *typed graphlets* that arise from a single graphlet  $H \in \mathcal{H}$  of size  $K \in \{2, \dots, 4\}$  nodes as the number of types varies from  $L = 1, 2, \dots, 9$ . For instance, the total number of typed graphlet orbits with 4 nodes that arise from 7 types is  $10 \cdot 210 = 2100$  since there are 10 connected 4-node (untyped) graphlet orbits. See Figure 1 for other examples. Unlike homogeneous graphlets, it is obviously impossible to show all the heterogeneous graphlets counted by the proposed approach since it works for general heterogeneous graphs with any arbitrary number of types  $L$  and structure.

Table 2. Number of *typed graphlets* (for a single untyped graphlet) as the size  $K$  (i.e., # of nodes in the typed graphlet) and number of types  $L$  varies.

	Types $L$								
	1	2	3	4	5	6	7	8	9
<b>K=2</b>	1	3	6	10	15	21	28	36	45
<b>K=3</b>	1	4	10	20	35	56	84	120	165
<b>K=4</b>	1	5	15	35	70	126	210	330	495

### 3.4 Generalization to Other Graphs

The proposed notion of *typed graphlets* can be used for applications on bipartite, k-partite, signed, attributed, and more generally heterogeneous networks. A few examples of such graphs are shown in Figure 3. The proposed framework naturally handles general heterogeneous networks with arbitrary structure and an arbitrary number of types. It is straightforward to see that homogeneous, bipartite, k-partite, signed, star, and attributed networks are all special cases of heterogeneous graphs (Figure 3). Therefore, the framework for deriving typed graphlets can easily support such networks. For attributed graphs with more than one attribute/feature, the attributes of a node or edge can be mapped to types using any arbitrary approach such as role2vec [Ahmed et al. 2018] or WL [Shervashidze et al. 2011].

## 4 FRAMEWORK

This section describes the general framework for counting typed graphlets. The typed graphlet framework can be used for counting typed graphlets locally for every edge in  $G$  as well as global typed graphlet counting (Problem 2) that focuses on computing the total frequency of all typed graphlets. This paper mainly focuses on the harder local typed graphlet counting problem:

**PROBLEM 1 (LOCAL TYPED GRAPHLET COUNTING).** *Given a graph  $G$  and an edge  $(i, j) \in E$ , the local typed graphlet counting problem is to find the set of all typed graphlets that contain nodes  $i$  and  $j$  and their corresponding frequencies. This work focuses on computing all  $\{2, 3, 4\}$ -node typed graphlet counts for every edge in  $G$ .*

Algorithm 1 shows the general approach for counting all typed graphlets with up to four nodes. Note that we do not make any restriction or assumption on the number of node or edge types. The algorithm naturally handles heterogeneous graphs with arbitrary number of types and structure. See Table 1 for a summary of notation.

---

### Algorithm 1 Typed Graphlets

---

**Input:** a graph  $G$

**Output:** nonzero typed graphlet counts  $X_{ij}$  for each edge  $(i, j) \in E$

```

1  parallel for each  $(i, j) \in E$  do
2       $T_{ij}^t = \Gamma_i^t \cap \Gamma_j^t$ , for  $t = 1, \dots, L$  ▷ typed triangles
3       $S_i^t = \Gamma_i^t \setminus T_{ij}^t$ , for  $t = 1, \dots, L$  ▷ typed 3-paths centered at  $i$ 
4       $S_j^t = \Gamma_j^t \setminus T_{ij}^t$ , for  $t = 1, \dots, L$  ▷ typed 3-paths centered at  $j$ 
5       $|S_{ij}^t| = |S_i^t| + |S_j^t|$ , for  $t = 1, \dots, L$  ▷ typed 3-path count
6      Store nonzero counts of the 3-node typed graphlets derived above
7      Given  $S_i$  and  $S_j$ , use Algorithm 2 to derive a few typed path-based graphlets
8      Given  $T_{ij}$ , use Algorithm 3 to derive a few typed triangle-based graphlets
9      for  $t, t' \in \{1, \dots, L\}$  such that  $t \leq t'$  do
10         Derive remaining typed graphlet orbits in constant time via Eq. 19-30 and update counts  $x$  and set
            of typed graphlets  $M_{ij}$  (with nonzero count)
11     for  $c \in M_{ij}$  do  $X_{ij} = X_{ij} \cup \{(c, x_c)\}$  ▷ store nonzero typed graphlet counts
12 end parallel

```

---



#### 4.1 Counting 3-Node Typed Graphlets

We begin by introducing the notion of a typed neighborhood *and* typed degree of a node. These are then used as a basis for deriving all typed 3-node graphlet counts in worst-case  $O(\Delta)$  time (Theorem 2).

**DEFINITION 6 (TYPED NEIGHBORHOOD).** *Given an arbitrary node  $i$  in  $G$ , the typed neighborhood  $\Gamma_i^t$  is the set of nodes with type  $t$  that are reachable by following edges originating from  $i$  within 1-hop distance. More formally,*

$$\Gamma_i^t = \{j \in V \mid (i, j) \in E \wedge \phi_j = t\} \quad (2)$$

*Intuitively, a node  $j \in \Gamma_i^t$  iff there exists an edge  $(i, j) \in E$  between  $i$  and  $j$  and the type of node  $j$  denoted as  $\phi_j$  is  $t$ .*

**DEFINITION 7 (TYPED DEGREE).** *The typed-degree  $d_i^t$  of node  $i$  with type  $t$  is defined as  $d_i^t = |\Gamma_i^t|$  where  $d_i^t$  is the number of nodes connected to node  $i$  with type  $t$ .*

Using these notions as a basis, we can define  $S_i^t$ ,  $S_j^t$ , and  $T_{ij}^t$  for  $t = 1, \dots, L$  (Figure 4). Obtaining these sets is equivalent to computing all 3-node typed graphlet counts. These sets are all defined with respect to a given edge  $(i, j) \in E$  between node  $i$  and  $j$  with types  $\phi_i$  and  $\phi_j$ . Since typed graphlets are counted for each edge  $(i, j) \in E$ , the types  $\phi_i$  and  $\phi_j$  are fixed ahead of time. Thus, there is only one remaining type to select for 3-node typed graphlets.

**COROLLARY 1.** *Given an edge  $(i, j) \in E$  between node  $i$  and  $j$  with types  $\phi_i$  and  $\phi_j$ , let  $T_{ij}^t$  denote the set of nodes of type  $t$  that complete a typed triangle with node  $i$  and  $j$  defined as:*

$$T_{ij}^t = \Gamma_i^t \cap \Gamma_j^t \quad (3)$$

*where  $|T_{ij}^t|$  denotes the number of nodes that form triangles with node  $i$  and  $j$  of type  $t$ .*

Let  $\Gamma_i^t$  denote the set of neighbors of  $i$  with type  $t$ . If  $k \in \Gamma_i^t$  and  $k \in \Gamma_j^t$ , then since  $(i, j) \in E$ , node  $k$  must form a typed triangle with  $i$  and  $j$  (i.e.,  $k \in T_{ij}^t$ ). Hence,  $(i, j) \in E$  closes a triangle with node  $k$  of type  $t$ . This is straightforward to see since  $k \in \Gamma_i^t$  implies  $(i, k) \in E$ ,  $k \in \Gamma_j^t$  implies  $(j, k) \in E$ , and  $(i, j) \in E$ . Since typed triangles are counted for each edge  $(i, j) \in E$ , the types  $\phi_i$  and  $\phi_j$  are fixed ahead of time. Therefore, there is only one remaining type to select. Let  $t$  denote the remaining node type, then  $T_{ij}^t = \Gamma_i^t \cap \Gamma_j^t$ . Furthermore, since every node  $k \in T_{ij}^t$  is of type  $t$  and thus completes a typed triangle with node  $i$  and  $j$  consisting of types  $\phi_i$ ,  $\phi_j$ , and  $\phi_k = t$ .

**COROLLARY 2.** *Given an edge  $(i, j) \in E$  between node  $i$  and  $j$  with types  $\phi_i$  and  $\phi_j$ . Let  $S_i^t$  denote the set of nodes of type  $t$  that form 3-node stars (or equivalently 3-node paths) centered at node  $i$  (and not including  $j$ ). More formally,*

$$S_i^t = \{k \in (\Gamma_i^t \setminus \{j\}) \mid k \notin \Gamma_j^t\} \quad (4)$$

$$= \Gamma_i^t \setminus (\Gamma_j^t \cup \{j\}) = \Gamma_i^t \setminus T_{ij}^t \quad (5)$$

*where  $|S_i^t|$  denotes the number of nodes of type  $t$  that form 3-stars centered at node  $i$  (not including  $j$ ).*

Let  $\Gamma_i^t$  denote the set of neighbors of  $i$  with type  $t$ . Let  $k \in \Gamma_i^t$  be a node that forms a typed 3-star centered at  $i$  with type  $t$ , then  $k \notin \Gamma_j^t$ . Otherwise if  $k \in \Gamma_j^t$ , then  $k \in T_{ij}^t$ , which implies  $k \notin S_i^t$ . Similarly, it is straightforward to define the set  $S_j^t$  of typed 3-star/path nodes of type  $t$  centered at  $j$  in a similar fashion:

$$S_j^t = \{k \in (\Gamma_j^t \setminus \{i\}) \mid k \notin \Gamma_i^t\} \quad (6)$$

$$= \Gamma_j^t \setminus (\Gamma_i^t \cup \{i\}) = \Gamma_j^t \setminus T_{ij}^t \quad (7)$$

where  $|S_j^t|$  denotes the number of nodes of type  $t$  that form 3-stars centered at node  $j$  (not including  $i$ ). This follows from Corollary 2.

PROPERTY 1.

$$T_{ij} = \bigcup_{t=1}^L T_{ij}^t, \quad S_i = \bigcup_{t=1}^L S_i^t, \quad S_j = \bigcup_{t=1}^L S_j^t \quad (8)$$

This property follows directly from Corollary 1-2 and is shown in Figure 4. These lower-order 3-node typed graphlet counts are used to derive many higher-order typed graphlet counts in  $o(1)$  constant time (Section 4.3).

COROLLARY 3 (TYPED 3-STARS). *Given an edge  $(i, j) \in E$  between node  $i$  and  $j$  with types  $\phi_i$  and  $\phi_j$ , the number of typed 3-node stars that contain  $(i, j) \in E$  with types  $\phi_i, \phi_j, t$  is:*

$$|S_{ij}^t| = |S_i^t| + |S_j^t| \quad (9)$$

where  $|S_{ij}^t|$  denotes the number of typed 3-stars that contain nodes  $i$  and  $j$  with types  $\phi_i, \phi_j, t$ .

Moreover, the number of typed triangles centered at  $(i, j) \in E$  with types  $\phi_i, \phi_j, t$  is simply  $|T_{ij}^t|$  (Corollary 1) whereas the number of typed 3-node stars that contain  $(i, j) \in E$  with types  $\phi_i, \phi_j, t$  is  $|S_{ij}^t| = |S_i^t| + |S_j^t|$  (Corollary 3). We do not need to actually store the sets  $S_i^t, S_j^t$ , and  $T_{ij}^t$  for every type  $t = 1, \dots, L$ . We only need to store the *size/cardinality* of the sets (as shown in Algorithm 1) since these are the counts of all possible 3-node typed graphlets. For convenience, we denote the size of those sets as  $|S_i^t|, |S_j^t|$ , and  $|T_{ij}^t|$  for all  $t = 1, \dots, L$ , respectively. At this point, all typed 3-node graphlets with nonzero counts have been computed for edge  $(i, j) \in E$  in  $O(|\Gamma_i| + |\Gamma_j|) = O(\Delta)$  time where  $\Delta$  is max degree (See Section 7.1 for proof). Note  $|\Gamma_i| = \sum_t |\Gamma_i^t|$ .

---

### Algorithm 2 Typed Path-based Graphlets

---

**Input:** a graph  $G = (V, E, \Phi, \xi)$ , an edge  $(i, j)$ , sets of nodes  $S_i$  and  $S_j$  that form 3-paths centered at  $i$  and  $j$ , respectively, a typed graphlet count vector  $\mathbf{x}$  for  $(i, j)$ , and set  $\mathcal{M}_{ij}$  of unique typed graphlets for  $i$  and  $j$ .

```

1  for each  $w_k \in S_i$  do
2    for  $w_r \in \Gamma_{w_k} \setminus \{i, j\}$  do
3      if  $w_r \notin (\Gamma_i \cup \Gamma_j)$  then ▷ typed 4-path-edge orbit
4         $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_3, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
5      else if  $w_r \in S_i \wedge w_r \leq w_k$  then ▷ typed tailed-tri (tail orbit)
6         $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_7, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
7  for each  $w_k \in S_j$  do
8    for  $w_r \in \Gamma_{w_k} \setminus \{i, j\}$  do
9      if  $w_r \notin (\Gamma_i \cup \Gamma_j)$  then ▷ typed 4-path-edge orbit
10        $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_3, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
11      else if  $w_r \in S_j \wedge w_r \leq w_k$  then ▷ typed tailed-tri (tail orbit)
12        $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_7, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
13      else if  $w_r \in S_i$  then ▷ typed 4-cycle
14        $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_6, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
15  return set of typed graphlets  $\mathcal{M}_{ij}$  between  $i$  and  $j$  and counts  $\mathbf{x}$ 

```

---

**Algorithm 3** Typed *Triangle*-based Graphlets

**Input:** a graph  $G = (V, E, \Phi, \xi)$ , an edge  $(i, j)$ , set of nodes  $T_{ij}$  that form triangles with  $i$  and  $j$ , sets of nodes  $S_i$  and  $S_j$  that form 3-paths centered at  $i$  and  $j$ , respectively, a typed graphlet count vector  $\mathbf{x}$  for  $(i, j)$ , and set  $\mathcal{M}_{ij}$  of unique typed graphlets for  $i$  and  $j$ .

```

1 for each  $w_k \in T_{ij}$  do
2   for  $w_r \in \Gamma_{w_k} \setminus \{i, j\}$  do
3     if  $w_r \in T_{ij} \wedge w_r \leq w_k$  then ▷ typed 4-clique
4        $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_{12}, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
5     else if  $w_r \in (S_i \cup S_j)$  then ▷ typed chord-cycle-edge orbit
6        $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_{10}, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
7     else if  $w_r \notin (\Gamma_i \cup \Gamma_j)$  then ▷ typed tailed-tri-center orbit
8        $\langle \mathbf{x}, \mathcal{M}_{ij} \rangle = \text{UPDATE}(\mathbf{x}, \mathcal{M}_{ij}, \mathbb{F}(g_8, \Phi_i, \Phi_j, \Phi_{w_k}, \Phi_{w_r}))$ 
9 return set of typed graphlets  $\mathcal{M}_{ij}$  between  $i$  and  $j$  and counts  $\mathbf{x}$ 

```

**4.2 Counting 4-Node Typed Graphlets**

To derive  $k$ -node typed graphlets, the framework leverages the lower-order  $(k-1)$ -node *typed graphlets*. Therefore, 4-node typed graphlets are derived by leveraging the *typed* sets  $T_{ij}^t = \Gamma_i^t \cup \Gamma_j^t$ ,  $S_i^t = \Gamma_i^t \setminus T_{ij}^t$ , and  $S_j^t = \Gamma_j^t \setminus T_{ij}^t$  (for  $t \in \{1, \dots, L\}$ ) computed from the lower-order 3-node typed graphlets along with the set  $I^t$  of non-adjacent nodes of type  $t$  w.r.t.  $(i, j) \in E$  defined formally as follows:

$$\begin{aligned} I^t &= V^t \setminus (\Gamma_i^t \cup \Gamma_j^t) \\ &= V^t \setminus (T_{ij}^t \cup S_i^t \cup S_j^t \cup \{i, j\}). \end{aligned} \quad (10)$$

where  $V^t \subseteq V$  is the set of nodes in  $V$  of type  $t$ .

PROPERTY 2.

$$|V^t| = |I^t| + |\Gamma_i^t| + |\Gamma_j^t| \quad (11)$$

The proof is straightforward by Eq. 10 and applying the principle of inclusion-exclusion.

**4.2.1 A General Principle for Typed Graphlet Counting.** We now introduce a general typed graphlet formulation. Let  $N_{P,Q}^{e,t}$  denote the number of distinct typed 4-node graphlets of  $H$  with the type vector  $\mathbf{t}$  that contain edge  $(i, j) \in E$  and have properties  $P \in \{S_i^t, S_j^t, T_{ij}^t, I^t\}$  and  $Q \in \{S_i^{t'}, S_j^{t'}, T_{ij}^{t'}, I^{t'}\}$  for any  $t, t' \in \{1, \dots, L\}$  defined as:

$$\begin{aligned} N_{P,Q}^{e,t} &= \left| \left\{ \{i, j, w_k, w_r\} \mid w_k \in P \wedge w_r \in Q \wedge \right. \right. \\ &\quad \left. \left. w_r \neq w_k \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right| \end{aligned} \quad (12)$$

Now let  $e'$  denote the event  $(w_k, w_r) \in E$ , and let  $[e']$  be the Iverson bracket that is 1 when  $(w_k, w_r) \in E$  and 0 otherwise. Then,  $N_{P,Q,[e']}^e$  denotes the number of all possible typed 4-node graphlets conditional on  $(w_k, w_r) \in E$ .

$$\begin{aligned} N_{P,Q,[e']}^{e,t} &= \left| \left\{ \{i, j, w_k, w_r\} \mid w_k \in P \wedge w_r \in Q \wedge \right. \right. \\ &\quad \left. \left. [e'] \wedge w_r \neq w_k \wedge \right. \right. \\ &\quad \left. \left. \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right| \end{aligned} \quad (13)$$

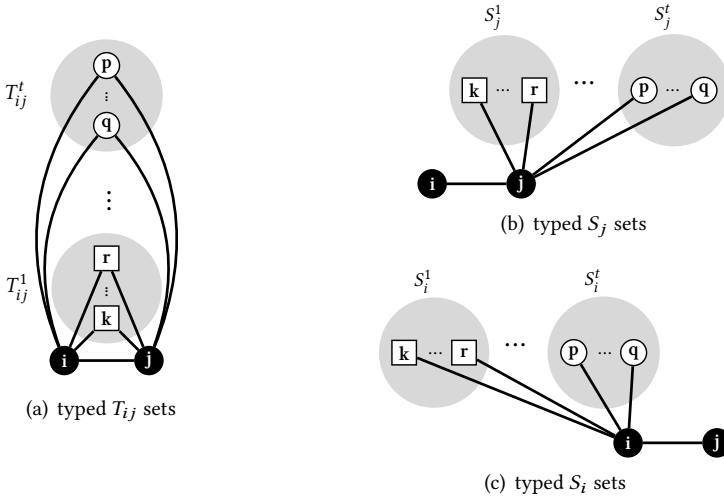


Fig. 4. Typed lower-order sets used to derive many higher-order graphlets in constant time. Note node  $i$  and  $j$  can be of arbitrary types.

**THEOREM 1 (GENERAL PRINCIPLE FOR TYPED GRAPHLET COUNTING).** *Given a graph  $G$ , for any edge  $e = (i, j)$  in  $G$ , for some type vector  $\mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}]$ , where  $w_k \in P$ ,  $w_r \in Q$ ,  $\phi_{w_k} = t$ , and  $\phi_{w_r} = t'$ , then the number of all 4-node typed graphlets  $\{i, j, w_k, w_r\}$  satisfies the general principle,*

$$N_{P,Q,0}^{e,\mathbf{t}} = N_{P,Q}^{e,\mathbf{t}} - N_{P,Q,1}^{e,\mathbf{t}} \quad (14)$$

**PROOF.** Assume there is a typed induced subgraph  $J \subset V$  such that  $J = \{i, j, w_k, w_r\}$  is incident to the edge of interest  $e = (i, j)$  and  $J$  is associated with a type vector  $\mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}]$  where  $\phi_{w_k} = t$  and  $\phi_{w_r} = t'$ . Suppose  $(w_k, w_r) \in E$ . Then by definition, we have  $J$  being counted once in the term  $N_{P,Q,1}^{e,\mathbf{t}}$  and once in the term  $N_{P,Q}^{e,\mathbf{t}}$  of Eq. 14. By the principle of inclusion-exclusion [Stanley 1986], the total contribution of the typed subgraph  $J$  with type vector  $\mathbf{t}$  to  $N_{P,Q,0}^{e,\mathbf{t}}$  is zero. ■

Notice that  $N_{P,Q}^{e,\mathbf{t}}$  does not indicate whether  $t = t'$  or  $t \neq t'$ . For clarity, we often use  $N_{P,Q}^{e,t=t'}$  and  $N_{P,Q}^{e,t \neq t'}$  to denote this explicitly. Theorem 1 shows it is sufficient to compute only two of the three quantities  $\{N_{P,Q}^{e,t=t'}, N_{P,Q,0}^{e,t=t'}, N_{P,Q,1}^{e,t=t'}\}$ . For instance, it is enough to compute  $N_{P,Q}^{e,t=t'}$  and  $N_{P,Q,1}^{e,t=t'}$ , and then  $N_{P,Q,0}^{e,t=t'}$  can be derived in  $o(1)$  constant time

$$N_{P,Q,0}^{e,t=t'} = N_{P,Q}^{e,t=t'} - N_{P,Q,1}^{e,t=t'} \quad (15)$$

It is straightforward to see that for  $t \neq t'$  it also holds:

$$N_{P,Q,0}^{e,t \neq t'} = N_{P,Q}^{e,t \neq t'} - N_{P,Q,1}^{e,t \neq t'} \quad (16)$$

This implies that selecting the two least computationally expensive quantities offers an obvious computational advantage.

We now show two fundamental properties that simplify the theory and discussion in Section 4.3. Property 3 applies to  $N_{P,Q}^{e,t=t'}$  only, whereas Property 4 applies to either  $N_{P,Q}^{e,t=t'}$  or  $N_{P,Q}^{e,t \neq t'}$ .

**PROPERTY 3.** *Let  $e = (i, j) \in E$  and  $N_{P,Q}^{e,t=t'}$  denote the number of 4-node typed graphlet orbits  $\{i, j, w_k, w_r\}$  such that  $w_k$  and  $w_r$  satisfy property  $P \in \{S_i^t, S_j^t, T_{ij}^t, I^t\}$  and  $Q \in \{S_i^{t'}, S_j^{t'}, T_{ij}^{t'}, I^{t'}\}$  for any*

$t = t' \in \{1, \dots, L\}$ , respectively. We say  $N_{P,Q}^{e,t=t'}$  is an unrestricted count since  $N_{P,Q}^{e,t=t'} = N_{P,Q,0}^{e,t=t'} + N_{P,Q,1}^{e,t=t'}$ . If  $P = Q$ , then  $t = t'$ . Therefore,

$$N_{P,Q}^{e,t=t'} = \binom{|P|}{2} = \frac{|P|(|P| - 1)}{2} = \frac{|Q|(|Q| - 1)}{2} \quad (17)$$

Clearly, Property 3 holds iff  $t = t'$  and  $P = Q$ . Suppose  $t \neq t'$ , then  $P \cap Q = \emptyset$  by definition, hence,  $P \neq Q$ . In other words,  $P = Q$  implies  $t = t'$ . Assuming  $t = t'$ , this property is useful for deriving the count of typed 4-stars and typed chordal-cycles (center orbit) in  $o(1)$  time as shown later in Section 4.3. For instance, suppose  $t = t'$ , then  $P = S_i^t$  and  $Q = S_j^t$ , and therefore the number of typed 4-stars  $f_{ij}(g_5, \mathbf{t})$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t]$  that occur between node  $i$  and  $j$  is

$$f_{ij}(g_5, \mathbf{t}) = \frac{1}{2} \left[ |S_i^t|(|S_i^t| - 1) + |S_j^t|(|S_j^t| - 1) \right] - f_{ij}(g_7, \mathbf{t})$$

where  $f_{ij}(g_7, \mathbf{t})$  is the typed tailed-triangle (tail-edge orbit) count.

PROPERTY 4. If  $P \neq Q$ , then  $P \cap Q = \emptyset$ . Hence,  $P$  and  $Q$  are mutually exclusive. This implies

$$N_{P,Q}^{e,t=t'} = |P| \cdot |Q| \quad \text{and} \quad N_{P,Q}^{e,t \neq t'} = |P| \cdot |Q| \quad (18)$$

Hence, since  $P \neq Q$ , then the above clearly holds for both  $t = t'$  and  $t \neq t'$ . Notice that in the untyped case, if  $P = S_i$  and  $Q = S_i$ , then  $P \cap Q = P$ . However, if we consider types and set  $P = S_i^t$  and  $Q = S_i^{t'}$ , then  $P \cap Q = \emptyset$  iff  $t \neq t'$ .

PROPERTY 5.  $\forall t, t', \text{ s.t. } t \neq t' \Rightarrow P \neq Q$

The above is straightforward. The converse is not true, that is, if  $P \neq Q$ , then  $t \neq t'$  does not necessarily hold. Assume  $t = t'$ , and let  $P = T_{ij}^t$  and  $Q = S_i^{t'}$ , then clearly  $P \neq Q$  despite  $t = t'$ . Suppose  $P = T_{ij}^{t'}$  and  $Q = T_{ij}^{t'}$ . If  $P = Q$ , then  $T_{ij}^{t'} = T_{ij}^{t'}$  and therefore  $t = t'$  must hold. Otherwise, if  $t \neq t'$  then  $P \neq Q$ .

The equations for deriving every typed graphlet orbit of size 4 are provided in Table 3. Notice that all typed graphlets with  $k$ -nodes are formulated with respect to the typed node sets  $\{S_i^t, S_j^t, T_{ij}^t, I^t\}$  derived from the typed graphlets with  $(k-1)$ -nodes. Hence, higher-order typed graphlets of order  $k$  are derived from lower-order  $(k-1)$ -node typed graphlets. We classify typed graphlets as path-based or triangle-based. Typed path-based graphlets are the typed 4-node graphlets derived from the sets  $S_i = \bigcup_i S_i^t$  and  $S_j = \bigcup_i S_j^t$  of nodes that form 3-node typed paths centered at node  $i$  and  $j$ , respectively (Algorithm 2). Conversely, typed triangle-based graphlets are the typed 4-node graphlets derived from the set  $T_{ij} = \bigcup_i T_{ij}^t$  of nodes that form typed triangles (typed 3-cliques) with node  $i$  and  $j$  (Algorithm 3). Naturally, typed path-based graphlets are the least dense (graphlets with fewest edges) whereas the typed triangle-based graphlets are the most dense.

The typed graphlet equations in Table 3 are mainly used to characterize the typed graphlets, and of course can be used to count them. However, using those equations to count all typed graphlets is still expensive since some non-negligible work is required to count every typed graphlet. Instead, we count only a few typed graphlets and use newly discovered combinatorial relationships (see Section 4.3) to derive the others directly in  $o(1)$  constant time. Notably, we make no assumptions about the number of types  $L$ , their distribution among the nodes and edges, or any other additional information. On the contrary, the framework is extremely general for arbitrary heterogeneous graphs (see Figure 3 for a number of popular special cases covered by the framework). In addition, we also avoid a lot of computations by symmetry breaking techniques, and other conditions to avoid unnecessary work.

Table 3. Typed graphlet orbit equations. All typed graphlet orbits with 4-nodes are formulated with respect to the typed node sets  $\{S_i^t, S_j^t, T_{ij}^t, I^t\}$  and  $\{S_i^{t'}, S_j^{t'}, T_{ij}^{t'}, I^{t'}\}$  for  $t, t' = 1, \dots, L$  derived from the typed 3-node graphlets. Recall  $T_{ij}^t = \Gamma_i^t \cap \Gamma_j^t$ ,  $S_j^t = \Gamma_j^t \setminus T_{ij}^t$ ,  $S_i^t = \Gamma_i^t \setminus T_{ij}^t$ , and  $I^t = V^t \setminus (\Gamma_i^t \cup \Gamma_j^t) = V^t \setminus (T_{ij}^t \cup S_i^t \cup S_j^t \cup \{i, j\})$  where  $V^t$  is the set of nodes in  $V$  of type  $t$ . In all cases,  $w_r \neq w_k$ .

TYPED GRAPHLET	ORBIT	$f_{ij}(H, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in P \wedge w_r \in Q \wedge \mathbb{I}\{(w_k, w_r) \in E\} \wedge w_r \neq w_k \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
4-path	edge	$f_{ij}(g_3, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid (w_k \in S_i^t \wedge w_r \in I^{t'}) \vee (w_k \in S_j^t \wedge w_r \in I^{t'}) \wedge (w_k, w_r) \in E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
	center	$f_{ij}(g_4, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in S_j^t \wedge w_r \in S_i^{t'} \wedge (w_k, w_r) \notin E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
4-star		$f_{ij}(g_5, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid (w_k \in S_i^t \wedge w_r \in S_i^{t'}) \vee (w_k \in S_j^t \wedge w_r \in S_j^{t'}) \wedge w_r \neq w_k \wedge (w_k, w_r) \notin E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
4-cycle		$f_{ij}(g_6, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in S_j^t \wedge w_r \in S_i^{t'} \wedge (w_k, w_r) \in E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
tailed-triangle	tail-edge	$f_{ij}(g_7, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid (w_k \in S_i^t \wedge w_r \in S_i^{t'}) \vee (w_k \in S_j^t \wedge w_r \in S_j^{t'}) \wedge w_r \neq w_k \wedge (w_k, w_r) \in E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
	center	$f_{ij}(g_8, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in T_{ij}^t \wedge w_r \in I^{t'} \wedge (w_k, w_r) \in E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
	tri-edge	$f_{ij}(g_9, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in T_{ij}^t \wedge w_r \in (S_i^{t'} \cup S_j^{t'}) \wedge (w_k, w_r) \notin E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
chordal-cycle	edge	$f_{ij}(g_{10}, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in T_{ij}^t \wedge w_r \in (S_i^{t'} \cup S_j^{t'}) \wedge (w_k, w_r) \in E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
	center	$f_{ij}(g_{11}, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in T_{ij}^t \wedge w_r \in T_{ij}^{t'} \wedge w_r \neq w_k \wedge (w_k, w_r) \notin E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$
4-clique		$f_{ij}(g_{12}, \mathbf{t}) = \left\{ \left\{ \{i, j, w_k, w_r\} \mid w_k \in T_{ij}^t \wedge w_r \in T_{ij}^{t'} \wedge w_r \neq w_k \wedge (w_k, w_r) \in E \wedge \mathbf{t} = [\phi_i \ \phi_j \ \phi_{w_k} \ \phi_{w_r}] \right\} \right\}$

Table 4. **Summary of Typed Graphlets and Position-aware Typed Graphlets.** Enumerative and combinatorial properties of typed graphlets and position-aware typed graphlets. With repetition allowed (in making the selection), the number of  $K$ -node *typed graphlets* and *position-aware typed graphlets* (for a single untyped graphlet) from  $L$  distinguishable labels/types is given below along with properties of each.

	TYPED GRAPHLETS (Definition 5)	POSITION-AWARE TYPED GRAPHLETS (Definition 8)
<b>With Repetition</b>	$\binom{L}{K} = \binom{L+K-1}{K}$	$L^K$
	Unordered Selections (Combinations)	Ordered Selections (Permutations)

### 4.3 Combinatorial Relationships for Typed Graphlets

Now, we show the existence of combinatorial relationships between the different *typed graphlets* and demonstrate how they can be leveraged to derive the counts of typed graphlets efficiently. These combinatorial relationships allow us to derive many *typed graphlets* in  $o(1)$  constant time and play a significant role in the speed/efficiency of the proposed approach (see Section 8.1). Using new combinatorial relationships between lower-order *typed graphlets*, we derive all remaining typed graphlet orbits in  $o(1)$  constant time via Eq. 19-30 (See Line 9-10 in Algorithm 1). Since we derive all typed graphlet counts for a given edge  $(i, j) \in E$  between node  $i$  and  $j$ , we already have two types  $\phi_i$  and  $\phi_j$ . Thus, these types are fixed ahead of time. In the case of 4-node typed

graphlets, there are two remaining types that need to be selected. Notice that for typed graphlet orbits, we must solve  $\frac{L(L-1)}{2} + L$  equations in the worst-case. The counts of all remaining typed graphlets are derived in  $o(1)$  constant time using the counts of the lower-order  $(k-1)$ -node typed graphlets and a few other counts from the  $k$ -node typed graphlets. After deriving the exact count of each remaining graphlet with types  $\phi_i, \phi_j, t$ , and  $t'$  for every  $t, t' \in \{1, \dots, L\}$  such that  $t \leq t'$  (Line 9-10), if such count is nonzero, we compute a graphlet hash  $c = \mathbb{F}(g, \phi_i, \phi_j, t, t')$  for graphlet orbit  $g$ , set  $\mathcal{M}_{ij} \leftarrow \mathcal{M}_{ij} \cup \{c\}$ , and then set the count of that typed graphlet in  $\mathbf{x}_c$  to the count derived in constant  $o(1)$  time.

We now demonstrate the relationship between different typed graphlets and prove the correctness of the equations used to derive a number of typed graphlet counts in  $o(1)$  constant time. See Figure 5 for intuition.

#### 4.3.1 Relationship between typed 4-cycles and 4-paths (center orbit).

**COROLLARY 4.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed 4-cycles containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{S_i^t S_j^t}^{e, t=t'}$  for  $t = t'$  and  $N_{S_i^t S_j^t}^{e, t \neq t'} + N_{S_i^{t'} S_j^{t'}}^{e, t \neq t'}$  otherwise.

**COROLLARY 5.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed 4-path center orbits containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{S_i^t S_j^0}^{e, t=t'}$  for  $t = t'$  and  $N_{S_i^t S_j^0}^{e, t \neq t'} + N_{S_i^{t'} S_j^0}^{e, t \neq t'}$  otherwise.

To count the typed 4-path center orbits for a given edge  $(i, j) \in E$  with types  $\phi_i$  and  $\phi_j$ , we simply select the remaining two types denoted as  $t$  and  $t'$  to obtain the 4-dimensional type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  and derive the count directly using Lemma 1.

**LEMMA 1.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$  and any type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$ , the relationship between the typed 4-cycle count  $f_{ij}(g_6, \mathbf{t})$  and the typed 4-path center orbit count  $f_{ij}(g_4, \mathbf{t})$  with type vector  $\mathbf{t}$  is

$$f_{ij}(g_4, \mathbf{t}) = \begin{cases} (|S_i^t| \cdot |S_j^t|) - f_{ij}(g_6, \mathbf{t}) & \text{if } t = t' \\ (|S_i^t| \cdot |S_j^t|) + (|S_i^{t'}| \cdot |S_j^{t'}|) - f_{ij}(g_6, \mathbf{t}) & \text{otherwise} \end{cases} \quad (19)$$

where  $f_{ij}(g_6, \mathbf{t})$  is the typed 4-cycle count for edge  $(i, j) \in E$  with type vector  $\mathbf{t}$ .

**PROOF.** Assume  $t = t'$ . From Theorem 1, we have

$$N_{S_i^t S_j^0}^{e, t=t'} = N_{S_i^t S_j^t}^{e, t=t'} - N_{S_i^t S_j^1}^{e, t=t'} \quad (20)$$

Since  $N_{S_i^t S_j^t}^{e, t=t'}$  is the number of typed 4-node induced subgraphs containing  $e = (i, j)$  such that  $w_k \in S_i^t$  and  $w_r \in S_j^t$ , then  $N_{S_i^t S_j^t}^{e, t=t'} = |S_i^t| \cdot |S_j^t|$  by Property 4. From Corollary 4, the number of typed 4-cycles that contain  $e = (i, j)$  is  $f_{ij}(g_6, \mathbf{t}) = N_{S_i^t S_j^1}^{e, t=t'}$ . From Corollary 5, the count of typed 4-paths centered at edge  $e = (i, j)$  is  $f_{ij}(g_4, \mathbf{t}) = N_{S_i^t S_j^0}^{e, t=t'}$ . Therefore, by direct substitution in Eq. 20, we obtain Eq. 19.

Assume  $t \neq t'$ . From Theorem 1, we have

$$N_{S_i^t S_j^0}^{e, t \neq t'} = N_{S_i^t S_j^{t'}}^{e, t \neq t'} - N_{S_i^t S_j^1}^{e, t \neq t'} \quad \text{and} \quad N_{S_i^{t'} S_j^0}^{e, t \neq t'} = N_{S_i^{t'} S_j^{t'}}^{e, t \neq t'} - N_{S_i^{t'} S_j^1}^{e, t \neq t'} \quad (21)$$

It is straightforward to rewrite this as

$$N_{S_i^t S_j^{t'}}^{e,t \neq t'} + N_{S_i^{t'} S_j^t}^{e,t \neq t'} = \left( N_{S_i^t S_j^t}^{e,t \neq t'} + N_{S_i^{t'} S_j^{t'}}^{e,t \neq t'} \right) - \left( N_{S_i^t S_j^{t'}}^{e,t \neq t'} + N_{S_i^{t'} S_j^t}^{e,t \neq t'} \right) \quad (22)$$

By Property 4, there are  $N_{S_i^t S_j^t}^{e,t \neq t'} + N_{S_i^{t'} S_j^{t'}}^{e,t \neq t'} = |S_i^t| \cdot |S_j^t| + |S_i^{t'}| \cdot |S_j^{t'}|$  typed 4-node induced subgraphs that contain edge  $e = (i, j)$  such that  $w_k \in S_i^t$  and  $w_r \in S_j^t$  or  $w_k \in S_i^{t'}$  and  $w_r \in S_j^{t'}$  where  $N_{S_i^t S_j^t}^{e,t \neq t'} = |S_i^t| \cdot |S_j^t|$  and  $N_{S_i^{t'} S_j^{t'}}^{e,t \neq t'} = |S_i^{t'}| \cdot |S_j^{t'}|$ . From Corollary 4, the typed 4-cycle count for edge  $e$  is  $f_{ij}(g_6, \mathbf{t}) = N_{S_i^t S_j^t, 1}^{e,t \neq t'} + N_{S_i^{t'} S_j^{t'}, 1}^{e,t \neq t'}$ . From Corollary 5, the count of typed 4-paths centered at edge  $e = (i, j)$  is  $f_{ij}(g_4, \mathbf{t}) = N_{S_i^t S_j^t, 0}^{e,t \neq t'} + N_{S_i^{t'} S_j^{t'}, 0}^{e,t \neq t'}$ . Therefore, by direct substitution in Eq. 20, we obtain Eq. 19. ■

The only difference between a typed 4-path centered at  $(i, j)$  (Corollary 5) and a typed 4-cycle (Corollary 4) is whether  $(w_k, w_r) \in E$  holds or not. Clearly, if  $(w_k, w_r) \in E$ , then we have a typed 4-cycle, otherwise  $(w_k, w_r) \notin E$  and it is a typed 4-path centered at  $(i, j)$  as shown in Figure 5(a).

#### 4.3.2 Relationship between typed 4-stars and tailed-triangles (tail-edge orbit).

**COROLLARY 6.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed 4-stars containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{S_i^t S_i^t, 0}^{e,t=t'} + N_{S_j^t S_j^t, 0}^{e,t=t'}$  for  $t = t'$  and  $N_{S_i^t S_i^t, 0}^{e,t \neq t'} + N_{S_j^t S_j^t, 0}^{e,t \neq t'}$  otherwise.

**COROLLARY 7.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed tailed-triangles (tail-edge orbit) containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{S_i^t S_i^t, 1}^{e,t=t'} + N_{S_j^t S_j^t, 1}^{e,t=t'}$  for  $t = t'$  and  $N_{S_i^t S_i^t, 1}^{e,t \neq t'} + N_{S_j^t S_j^t, 1}^{e,t \neq t'}$  otherwise.

To count the typed 4-stars for a given edge  $(i, j) \in E$  with types  $\phi_i$  and  $\phi_j$ , we simply select the remaining two types denoted as  $t$  and  $t'$  to obtain the 4-dimensional type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$ . We derive the typed 4-star counts with the type vector  $\mathbf{t}$  for edge  $(i, j) \in E$  in constant time using Lemma 2.

**LEMMA 2.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$  and any type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$ , the relationship between the typed 4-star count  $f_{ij}(g_5, \mathbf{t})$  and the typed tailed-triangle tail-edge orbit count  $f_{ij}(g_7, \mathbf{t})$  with type vector  $\mathbf{t}$  is

$$f_{ij}(g_5, \mathbf{t}) = \begin{cases} \left( \binom{|S_i^t|}{2} + \binom{|S_j^t|}{2} \right) - f_{ij}(g_7, \mathbf{t}) & \text{if } t = t' \\ \left( |S_i^t| \cdot |S_i^{t'}| \right) + \left( |S_j^t| \cdot |S_j^{t'}| \right) - f_{ij}(g_7, \mathbf{t}) & \text{otherwise} \end{cases} \quad (23)$$

where  $f_{ij}(g_7, \mathbf{t})$  is the tailed-triangle tail-edge orbit count for edge  $(i, j) \in E$  with type vector  $\mathbf{t}$ .

**PROOF.** Let  $S_i^t$  and  $S_i^{t'}$  be the nodes that form typed 3-node stars with  $(i, j) \in E$  of type  $t$  and  $t'$  where node  $i$  is the star-center node, respectively. Similarly,  $S_j^t$  and  $S_j^{t'}$  are the nodes that form typed 3-node stars with  $(i, j) \in E$  of type  $t$  and  $t'$  where node  $j$  is the star-center node, respectively.

Assume  $t = t'$ . From Theorem 1, we have

$$N_{S_i^t S_i^t, 0}^{e,t=t'} + N_{S_j^t S_j^t, 0}^{e,t=t'} = \left( N_{S_i^t S_i^t}^{e,t=t'} + N_{S_j^t S_j^t}^{e,t=t'} \right) - \left( N_{S_i^t S_i^t, 1}^{e,t=t'} + N_{S_j^t S_j^t, 1}^{e,t=t'} \right) \quad (24)$$



Therefore, by Property 3, there are  $N_{S_i^t S_i^{t'}}^{e,t=t'} + N_{S_j^t S_j^{t'}}^{e,t=t'} = \binom{|S_i^t|}{2} + \binom{|S_j^t|}{2}$  typed 4-node induced subgraphs that contain edge  $e = (i, j)$  such that  $w_k, w_r \in S_i^t$  or  $w_k, w_r \in S_j^t$  where  $N_{S_i^t S_i^{t'}}^{e,t=t'} = \binom{|S_i^t|}{2}$  and  $N_{S_j^t S_j^{t'}}^{e,t=t'} = \binom{|S_j^t|}{2}$ . From Corollary 6, the number of typed 4-stars that contain edge  $e = (i, j)$  is  $f_{ij}(g_5, \mathbf{t}) = N_{S_i^t S_i^{t'}, 0}^{e,t=t'} + N_{S_j^t S_j^{t'}, 0}^{e,t=t'}$ . Similarly, from Corollary 7, the typed tailed-triangle tail-edge orbit count for edge  $e$  is  $f_{ij}(g_7, \mathbf{t}) = N_{S_i^t S_i^{t'}, 1}^{e,t=t'} + N_{S_j^t S_j^{t'}, 1}^{e,t=t'}$ . Therefore, by direct substitution in Eq. 24, we obtain Eq. 23.

Assume  $t \neq t'$ . From Theorem 1, we have

$$N_{S_i^t S_i^{t'}, 0}^{e,t \neq t'} + N_{S_j^t S_j^{t'}, 0}^{e,t \neq t'} = \left( N_{S_i^t S_i^{t'}}^{e,t \neq t'} + N_{S_j^t S_j^{t'}}^{e,t \neq t'} \right) - \left( N_{S_i^t S_i^{t'}, 1}^{e,t \neq t'} + N_{S_j^t S_j^{t'}, 1}^{e,t \neq t'} \right) \quad (25)$$

Therefore, from Property 4, there are  $N_{S_i^t S_i^{t'}}^{e,t \neq t'} + N_{S_j^t S_j^{t'}}^{e,t \neq t'} = (|S_i^t| \cdot |S_i^{t'}|) + (|S_j^t| \cdot |S_j^{t'}|)$  typed 4-node induced subgraphs that contain edge  $e = (i, j)$  such that  $w_k \in S_i^t, w_r \in S_i^{t'}$  or  $w_k \in S_j^t, w_r \in S_j^{t'}$  where  $N_{S_i^t S_i^{t'}}^{e,t \neq t'} = |S_i^t| \cdot |S_i^{t'}|$  and  $N_{S_j^t S_j^{t'}}^{e,t \neq t'} = |S_j^t| \cdot |S_j^{t'}|$ . From Corollary 6, the number of typed 4-stars that contain edge  $e = (i, j)$  is  $f_{ij}(g_5, \mathbf{t}) = N_{S_i^t S_i^{t'}, 0}^{e,t \neq t'} + N_{S_j^t S_j^{t'}, 0}^{e,t \neq t'}$ . Similarly, from Corollary 7, the typed tailed-triangle tail-edge orbit count for edge  $e$  is  $f_{ij}(g_7, \mathbf{t}) = N_{S_i^t S_i^{t'}, 1}^{e,t \neq t'} + N_{S_j^t S_j^{t'}, 1}^{e,t \neq t'}$ . Therefore, by direct substitution in Eq. 25, we obtain Eq. 23. ■

The only path-based typed graphlet containing a triangle is the tailed-triangle tail-edge orbit. Observe that this is the only orbit needed to derive the typed 4-star counts in constant time.

#### 4.3.3 Relationship between typed tailed-triangles (tri-edge orbit) and chordal-cycles (edge orbit).

**COROLLARY 8.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed tailed-triangle (paw) tri-edge orbits containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t=t'}$  for  $t = t'$  and  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t \neq t'} + N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t \neq t'}$  otherwise.

**COROLLARY 9.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the typed chordal-cycle edge orbit count with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 1}^{e,t=t'}$  for  $t = t'$  and  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 1}^{e,t \neq t'} + N_{T_{ij}^t S_i^{t'} S_j^{t'}, 1}^{e,t \neq t'}$  otherwise.

**LEMMA 3.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$  and any type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$ , the relationship between the typed tailed-triangle tri-edge orbit count  $f_{ij}(g_9, \mathbf{t})$  and the typed chordal-cycle edge orbit count  $f_{ij}(g_{10}, \mathbf{t})$  with type vector  $\mathbf{t}$  is

$$f_{ij}(g_9, \mathbf{t}) = \begin{cases} (|T_{ij}^t| \cdot (|S_i^t| + |S_j^t|)) - f_{ij}(g_{10}, \mathbf{t}) & \text{if } t = t' \\ (|T_{ij}^t| \cdot (|S_i^{t'}| + |S_j^{t'}|)) + (|T_{ij}^t| \cdot (|S_i^t| + |S_j^t|)) - f_{ij}(g_{10}, \mathbf{t}) & \text{otherwise} \end{cases} \quad (26)$$

where  $f_{ij}(g_{10}, \mathbf{t})$  is the chordal-cycle edge orbit count for edge  $(i, j) \in E$  with type vector  $\mathbf{t}$ .

**PROOF.** Assume  $t = t'$ . From Theorem 1, we have

$$N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t=t'} = N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t=t'} - N_{T_{ij}^t S_i^{t'} S_j^{t'}, 1}^{e,t=t'} \quad (27)$$

Let  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t=t'} = N_{T_{ij}^t S_i^t}^{e,t=t'} + N_{T_{ij}^t S_j^t}^{e,t=t'}$ . Since  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t=t'}$  is the number of typed 4-node induced subgraphs containing  $e = (i, j)$  such that  $w_k \in T_{ij}^t$  and  $w_r \in S_i^t \cup S_j^t$ , then  $N_{T_{ij}^t S_i^{t'} S_j^{t'}, 0}^{e,t=t'} = |T_{ij}^t| \cdot (|S_i^t| + |S_j^t|)$  by

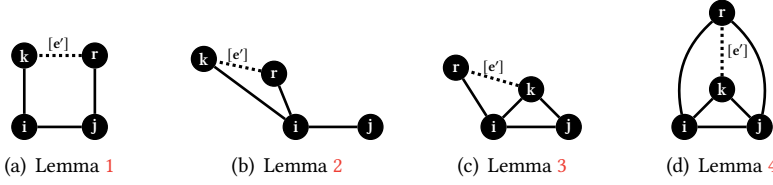


Fig. 5. Combinatorial relationships of typed graphlets. (a) Relationship between typed 4-paths and 4-cycles. (b) Relationship between typed 4-stars and tailed-triangles (tail-edge orbit). (c) Relationship between typed tailed-triangles (tri-edge orbit) and chordal-cycles (edge orbit). (d) Relationship between typed 4-cliques and chordal-cycles (center orbit).

**Property 4.** From Corollary 8, the number of typed tailed-triangles tri-edge orbits that contain  $e = (i, j)$  is  $f_{ij}(g_9, \mathbf{t}) = N_{T_{ij}^t, S_i^t \vee S_j^t, 0}^{e, t \neq t'}$ . From Corollary 9, the number of typed chordal-cycle edge orbits centered at edge  $e$  is  $f_{ij}(g_{10}, \mathbf{t}) = N_{T_{ij}^t, S_i^t \vee S_j^t, 1}^{e, t \neq t'}$ . Therefore, by direct substitution in Eq. 27, we obtain Eq. 26.

Assume  $t \neq t'$ . From Theorem 1, we have  $P = T_{ij}^t$  and  $Q = S_i^{t'} \cup S_j^{t'}$  or  $P = T_{ij}^{t'}$  and  $Q = S_i^t \cup S_j^t$ . Therefore,

$$N_{T_{ij}^t, S_i^t \vee S_j^t, 0}^{e, t \neq t'} = N_{T_{ij}^t, S_i^{t'} \vee S_j^{t'}}^{e, t \neq t'} - N_{T_{ij}^t, S_i^t \vee S_j^{t'}}^{e, t \neq t'} \quad \text{and} \quad N_{T_{ij}^t, S_i^t \vee S_j^t, 0}^{e, t \neq t'} = N_{T_{ij}^t, S_i^t \vee S_j^{t'}}^{e, t \neq t'} - N_{T_{ij}^t, S_i^{t'} \vee S_j^t}^{e, t \neq t'} \quad (28)$$

By rewriting the above,

$$N_{T_{ij}^t, S_i^t \vee S_j^t, 0}^{e, t \neq t'} + N_{T_{ij}^t, S_i^{t'} \vee S_j^{t'}, 0}^{e, t \neq t'} = \left( N_{T_{ij}^t, S_i^t \vee S_j^{t'}}^{e, t \neq t'} + N_{T_{ij}^t, S_i^{t'} \vee S_j^t}^{e, t \neq t'} \right) - \left( N_{T_{ij}^t, S_i^t \vee S_j^{t'}}^{e, t \neq t'} + N_{T_{ij}^t, S_i^{t'} \vee S_j^t}^{e, t \neq t'} \right) \quad (29)$$

By Property 4, there are  $N_{T_{ij}^t, S_i^t \vee S_j^t}^{e, t \neq t'} + N_{T_{ij}^t, S_i^{t'} \vee S_j^{t'}}^{e, t \neq t'} = |T_{ij}^t|(|S_i^t| + |S_j^t|) + |T_{ij}^t|(|S_i^{t'}| + |S_j^{t'}|)$  typed 4-node induced subgraphs that contain edge  $e = (i, j)$  such that  $w_k \in T_{ij}^t$  and  $w_r \in S_i^t \cup S_j^t$  or  $w_k \in T_{ij}^{t'}$  and  $w_r \in S_i^{t'} \cup S_j^{t'}$ . From Corollary 9, the typed chordal-cycle edge orbit count for edge  $e = (i, j)$  is  $f_{ij}(g_{10}, \mathbf{t}) = N_{T_{ij}^t, S_i^t \vee S_j^t, 1}^{e, t \neq t'} + N_{T_{ij}^t, S_i^{t'} \vee S_j^{t'}, 1}^{e, t \neq t'}$ . Similarly, from Corollary 8, the typed tailed-triangle tri-edge orbit count for edge  $e$  is  $f_{ij}(g_9, \mathbf{t}) = N_{T_{ij}^t, S_i^t \vee S_j^t, 0}^{e, t \neq t'} + N_{T_{ij}^t, S_i^{t'} \vee S_j^{t'}, 0}^{e, t \neq t'}$ . Therefore, by direct substitution in Eq. 29, we obtain Eq. 26. ■

#### 4.3.4 Relationship between typed 4-cliques and chordal-cycles (center orbit).

**COROLLARY 10.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed 4-cliques containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{T_{ij}^t, T_{ij}^t, 1}^{e, t=t'}$  for  $t = t'$  and  $N_{T_{ij}^t, T_{ij}^t, 1}^{e, t \neq t'}$  for  $t \neq t'$ .

**COROLLARY 11.** For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$ , the number of typed chordal-cycles (center orbit) containing edge  $(i, j)$  with type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$  is  $N_{T_{ij}^t, T_{ij}^t, 0}^{e, t=t'}$  for  $t = t'$  and  $N_{T_{ij}^t, T_{ij}^t, 0}^{e, t \neq t'}$  for  $t \neq t'$ .

Notice the only difference between Corollary 10 and 11 is that  $(w_k, w_r) \in E$  must hold for counting typed 4-cliques for a given edge whereas for counting chordal-cycle center orbits  $(w_k, w_r) \notin E$  must hold. Further, if  $t \neq t'$ , then  $w_r \neq w_k$  can be removed by Property 4 since by definition  $T_{ij}^t \cap T_{ij}^{t'} = \emptyset$ .

LEMMA 4. For any edge  $(i, j) \in E$  in  $G$  with types  $\phi_i$  and  $\phi_j$  and any type vector  $\mathbf{t} = [\phi_i \ \phi_j \ t \ t']$ , the relationship between the typed 4-clique count  $f_{ij}(g_{12}, \mathbf{t})$  and the typed chordal-cycle center orbit count  $f_{ij}(g_{11}, \mathbf{t})$  with type vector  $\mathbf{t}$  is

$$f_{ij}(g_{11}, \mathbf{t}) = \begin{cases} \binom{|T_{ij}^t|}{2} - f_{ij}(g_{12}, \mathbf{t}) & \text{if } t = t' \\ (|T_{ij}^t| \cdot |T_{ij}^{t'}|) - f_{ij}(g_{12}, \mathbf{t}) & \text{otherwise} \end{cases} \quad (30)$$

where  $f_{ij}(g_{12}, \mathbf{t})$  is the typed 4-clique count for edge  $(i, j) \in E$  with type vector  $\mathbf{t}$ .

PROOF. Let  $T_{ij}^t$  and  $T_{ij}^{t'}$  be the nodes that form typed triangles with  $(i, j) \in E$  of type  $t$  and  $t'$ , respectively. There are again two cases.

Assume  $t = t'$ . From Theorem 1, we have

$$N_{T_{ij}^t, T_{ij}^t, 0}^{e, t=t'} = N_{T_{ij}^t, T_{ij}^t}^{e, t=t'} - N_{T_{ij}^t, T_{ij}^t, 1}^{e, t=t'} \quad (31)$$

Since  $t = t'$ , then  $P = T_{ij}^t$  and  $Q = T_{ij}^t$ , hence  $P = Q$ . Therefore, by Property 3, there are  $N_{T_{ij}^t, T_{ij}^t}^{e, t=t'}$   $\binom{|T_{ij}^t|}{2}$  typed 4-node induced subgraphs that contain edge  $e = (i, j)$  such that  $w_k \in T_{ij}^t$  and  $w_r \in T_{ij}^t$ . From Corollary 10, the number of typed 4-cliques that contain edge  $e = (i, j)$  is  $f_{ij}(g_{12}, \mathbf{t}) = N_{T_{ij}^t, T_{ij}^t, 1}^{e, t=t'}$ . Similarly, from Corollary 11, the typed chordal-cycle center orbit count for edge  $e$  is  $f_{ij}(g_{11}, \mathbf{t}) = N_{T_{ij}^t, T_{ij}^t, 0}^{e, t=t'}$ . Therefore, by direct substitution in Eq. 31, we obtain Eq. 30.

Assume  $t \neq t'$ . From Theorem 1, we have  $P = T_{ij}^t$  and  $Q = T_{ij}^{t'}$ , therefore

$$N_{T_{ij}^t, T_{ij}^{t'}, 0}^{e, t \neq t'} = N_{T_{ij}^t, T_{ij}^{t'}}^{e, t \neq t'} - N_{T_{ij}^t, T_{ij}^{t'}, 1}^{e, t \neq t'} \quad (32)$$

By Property 4, there are  $N_{T_{ij}^t, T_{ij}^{t'}}^{e, t \neq t'} = |T_{ij}^t| \cdot |T_{ij}^{t'}|$  typed 4-node induced subgraphs that contain edge  $e = (i, j)$  such that  $w_k \in T_{ij}^t$  and  $w_r \in T_{ij}^{t'}$ . From Corollary 10, the number of typed 4-cliques that contain edge  $e = (i, j)$  is  $f_{ij}(g_{12}, \mathbf{t}) = N_{T_{ij}^t, T_{ij}^{t'}, 1}^{e, t \neq t'}$ . Similarly, from Corollary 11, the typed chordal-cycle center orbit count for edge  $e$  is  $f_{ij}(g_{11}, \mathbf{t}) = N_{T_{ij}^t, T_{ij}^{t'}, 0}^{e, t \neq t'}$ . Therefore, by direct substitution in Eq. 32 we obtain Eq. 30. ■

---

**Algorithm 4** Update Typed Graphlets. Add typed graphlet with hash  $c$  to  $\mathcal{M}_{ij}$  if  $c \notin \mathcal{M}_{ij}$  and increment  $\mathbf{x}_c$  (frequency of that typed graphlet for a given edge).

---

- 1 **procedure** UPDATE( $\mathbf{x}$ ,  $\mathcal{M}_{ij}$ ,  $c = \mathbb{F}(g, \Phi_i, \Phi_j, \Phi_k, \Phi_r)$ )
  - 2     **if**  $c \notin \mathcal{M}_{ij}$  **then**  $\mathcal{M}_{ij} \leftarrow \mathcal{M}_{ij} \cup \{c\}$  and set  $\mathbf{x}_c = 0$
  - 3      $\mathbf{x}_c = \mathbf{x}_c + 1$
  - 4     **return** updated set of typed graphlets  $\mathcal{M}_{ij}$  and counts  $\mathbf{x}$
-

#### 4.4 From Typed Orbits to Graphlets

Counts of the *typed graphlets* for each edge  $(i, j) \in E$  can be derived from the *typed graphlet orbits* using the following equations:

$$f_{ij}(h_3, \mathbf{t}) = f_{ij}(g_3, \mathbf{t}) + f_{ij}(g_4, \mathbf{t}) \quad (33)$$

$$f_{ij}(h_4, \mathbf{t}) = f_{ij}(g_5, \mathbf{t}) \quad (34)$$

$$f_{ij}(h_5, \mathbf{t}) = f_{ij}(g_6, \mathbf{t}) \quad (35)$$

$$f_{ij}(h_6, \mathbf{t}) = f_{ij}(g_7, \mathbf{t}) + f_{ij}(g_8, \mathbf{t}) + f_{ij}(g_9, \mathbf{t}) \quad (36)$$

$$f_{ij}(h_7, \mathbf{t}) = f_{ij}(g_{10}, \mathbf{t}) + f_{ij}(g_{11}, \mathbf{t}) \quad (37)$$

$$f_{ij}(h_8, \mathbf{t}) = f_{ij}(g_{12}, \mathbf{t}) \quad (38)$$

where  $h$  is the graphlet without considering the orbit (Table 3).

#### 4.5 Typed Graphlet Hash Functions

Given a general heterogeneous graph with  $L$  unique types such that  $L < 10$ , then a simple and efficient typed graphlet hash function  $\mathbb{F}$  is defined as follows:

$$\mathbb{F}(g, \mathbf{t}) = g10^4 + t_110^3 + t_210^2 + t_310^1 + t_4 \quad (39)$$

where  $g$  encodes the  $k$ -node graphlet orbit (e.g., 4-path center) and  $t_1, t_2, t_3, t_4$  encode the type of the nodes in  $H \in \mathcal{H}$  with type vector  $\mathbf{t} = [t_1 \ t_2 \ t_3 \ t_4]$ . Since the maximum hash value resulting from Eq. 39 is small (and fixed for any arbitrarily large graph  $G$ ), we can leverage a perfect hash table to allow for fast  $o(1)$  constant time lookups to determine if a typed graphlet was previously found or not as well as updating the typed graphlet count in  $o(1)$  constant time. For  $k$ -node graphlets where  $k < 4$ , we simply set the last  $4 - k$  types to 0. Note the simple typed graphlet hash function defined above can be extended trivially to handle graphs with  $L \geq 10$  types:

$$\mathbb{F}(g, \mathbf{t}) = g10^8 + t_110^6 + t_210^4 + t_310^2 + t_4 \quad (40)$$

In general, any non-cryptographic hash function  $\mathbb{F}$  can be used (see Chi and Zhu [2017] for some other possibilities). Thus, the approach is independent of  $\mathbb{F}$  and can always leverage the best known hash function. The only requirement of the hash function is that it is invertible  $\mathbb{F}^{-1}$ .

Thus far we have not made any assumption on the ordering of types in  $\mathbf{t}$ . As such, the hash function  $\mathbb{F}$  discussed above can be used directly in the framework for counting typed graphlets such that the type structure and position are preserved (See Section 5 for further discussion on position-aware typed graphlets). However, since we are interested in counting all typed graphlets *w.r.t.* Definition 5, then we map all such orderings of the types in  $\mathbf{t}$  to the same hash value using a precomputed hash table. This allows us to obtain the unique hash value in  $o(1)$  constant time for any ordering of the types in  $\mathbf{t}$ . In our implementation, we compute  $s = t_110^3 + t_210^2 + t_310^1 + t_4$  and then use  $s$  as an index into the precomputed hash table to obtain the unique hash value  $c$  in  $o(1)$  constant time.

#### 4.6 Sparse Typed Graphlet Format

This section describes a space-efficient representation for typed graphlets based on a key observation.

**PROPERTY 6.** *Let  $T$  denote the number of unique typed graphlets that appear in an arbitrary graph  $G$  with  $L$  types. Assuming the graph  $G$  has a skewed degree distribution, then most edges in  $G$  appear in only a small fraction of the  $T$  actual typed graphlets that can occur.*

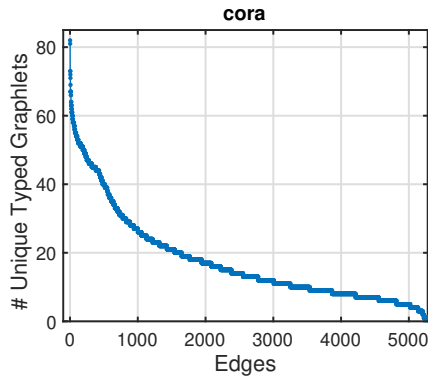


Fig. 6. Distribution of unique typed graphlets that occur on the edges. This experiment considers all typed graphlets of  $\{3, 4\}$ -nodes. Among the 1428 possible unique typed graphlets that could arise in  $G$ , there are only 876 unique typed graphlets that actually occur (at least once at an edge in  $G$ ). Even more striking, the maximum unique typed graphlets that occur on any edge in  $G$  (cora) is only 82. Overall, the mean number of unique typed graphlets over all edges in  $G$  is 17, *i.e.*, only about 1.1% of the possible typed graphlets. These results indicate the significance of only a few typed graphlets as the vast majority of the typed graphlet counts for any arbitrary edge is zero. Thus, the space required by the approach is nearly-optimal.

This property is shown empirically in Figure 6 and implies that using a  $M \times T$  matrix to store the typed graphlet counts is far from optimal in terms of the space required due to most of the  $T$  typed graphlet counts being zero for any given edge. Based on this observation, we ensure the proposed approach uses near-optimal space by storing only the typed graphlets with nonzero counts for each edge  $(i, j) \in E$  in the graph. Typed graphlet counts are stored in a sparse format since it would be impractical in large graphs to store all typed graphlets as there can easily be hundreds of thousands depending on the number of types in the input graph.

For each edge, we store only the nonzero typed graphlet counts along with the unique ids associated with them. The unique ids allow us to map the nonzero counts to the actual typed graphlets. We also developed a space-efficient format for storing the resulting typed graphlet counts to disk. Instead of using the typed graphlet hash as the unique id, we remap the typed graphlets to smaller consecutive ids (starting from 1) to reduce the space requirement even further. Finally, we store a typed graphlet lookup table that maps a given graphlet id to its description and is useful for looking up the meaning of the typed graphlets discovered.

#### 4.7 Parallelization

We now describe a parallelization strategy for the proposed typed graphlet counting approach. While our implementation uses shared memory, the parallelization is described generally such that it can be used with a distributed-memory architecture as well. As such, our discussion is on the general scheme. The parallel constructs we use are a worker task-queue and a global broadcast channel. Here, we assume that each worker has a copy of the graph and distribute edges to workers to find the typed graphlet counts that node  $i$  and  $j$  participate. At this point, we view the main while loop as a task generator and farm the current edge out to a worker to find the typed graphlet counts that co-occur between node  $i$  and node  $j$ . The approach is lock free since each worker uses the same graphlet hash function to obtain a unique hash value for every typed graphlet. Thus, each worker can simply maintain the typed graphlets identified and their counts for every edge assigned to it. In our own shared memory implementation, we avoid some of the communications by using

global arrays and avoiding locked updates to them by using a unique edge id. Counting typed graphlets on the edges as opposed to the nodes also has computational advantages with respect to parallelization and in particular load balancing. Let  $x_i$  and  $x_{ij}$  denote the node and edge count of an arbitrary graphlet  $H$ . Since  $|E| \gg |V|$  and  $\sum_{i \in V} x_i = \sum_{(i,j) \in E} x_{ij}$ , then  $\frac{1}{|V|} \sum_{i \in V} x_i < \frac{1}{|E|} \sum_{(i,j) \in E} x_{ij}$ . Hence, more work per vertex is required than per edge. Therefore, counting typed graphlets on the edges is guaranteed to have better load balancing than node-centric algorithms.

#### 4.8 Discussion

This work formalized the notion of typed graphlet and provided a time- and space-efficient framework for counting all  $\{2, 3, 4\}$ -node typed graphlets. Counting typed graphlets of a larger size is outside the scope of this paper and left for future work. However, the ideas introduced in this paper can be used to extend and derive equations for typed graphlets of 5-nodes and larger. In particular, Theorem 1 states the general principle of counting typed graphlets which is based on inclusion-exclusion, and therefore is straightforward to apply to typed graphlets of larger sizes. This would follow directly from recent work [Dave et al. 2017; Pinar et al. 2017] that extended the ideas of Ahmed et al. [2015, 2016] to 5-node untyped graphlets. For instance, just as we did in this work, the node sets used to derive the different 5-node untyped graphlet counts in [Dave et al. 2017; Pinar et al. 2017] are further partitioned into subsets where each subset represents nodes of the same type, e.g., just as a node  $w \in T_{ij}^t$  is a node of type  $t$  that forms a triangle with  $i$  and  $j$ . Afterwards, we can derive typed equations just as we did in this work to handle the different cases, i.e., when all types are the same vs. when they are different, and so on. Nevertheless, counting typed graphlets of 5 nodes and larger is outside the scope of this work and left for future work.

The approach is also straightforward to adapt for directed typed graphlets. In particular, we simply replace  $\Gamma_i^t$  with  $\Gamma_i^{t,+}$  and  $\Gamma_i^{t,-}$  for typed out-neighbors and typed in-neighbors, respectively. Thus, we also have  $T_{ij}^{t,+}$ ,  $T_{ij}^{t,-}$ ,  $S_j^{t,+}$ ,  $S_j^{t,-}$ ,  $S_i^{t,+}$ , and  $S_i^{t,-}$ . Now it is just a matter of enumerating all combinations of these sets with the out/in-neighbor sets as well. That is, we essentially have two additional versions of Algorithm 1 and Algorithm 2-3 for each in and out set (w.r.t. to the main for loop). The other trivial modification is to ensure each directed typed graphlet is assigned a unique id (this is the same modification required for typed orbits). The time and space complexity remains the same since all we did is split the set of neighbors (and the other sets) into two smaller sets by partitioning the nodes in  $\Gamma_i^t$  into  $\Gamma_i^{t,+}$  and  $\Gamma_i^{t,-}$ . Similarly, for  $T_{ij}^t$ ,  $S_j^t$ , and  $S_i^t$ .

### 5 POSITION-AWARE TYPED GRAPHLETS

#### 5.1 Formulation

We can consider the presence of topologically identical ‘‘appearances’’ of a typed graphlet in a graph such that the type structure is preserved as well. More formally, we define a *position-aware typed graphlet* that ensures node (edge) types coincide via the isomorphism:

**DEFINITION 8 (POSITION-AWARE TYPED GRAPHLET INSTANCE).** *An instance of a position-aware typed graphlet  $H = (V', E', \phi', \xi')$  of graph  $G$  is a typed graphlet  $F = (V'', E'', \phi'', \xi'')$  of  $G$  such that*

- (1)  $(V'', E'')$  is isomorphic to  $(V', E')$ ,
- (2)  $\mathcal{T}_{V''} = \mathcal{T}_{V'}$  and  $\mathcal{T}_{E''} = \mathcal{T}_{E'}$ , that is, the multisets of node and edge types are correspondingly equal.
- (3)  $\phi'' = \phi' \circ p$  and  $\xi'' = \xi' \circ q$  where  $q = p \times p$ , that is, the node and edge types coincide via the graph isomorphism  $p$ .

This formulation can be used to count position-aware typed graphlets that *preserve type structure*. Such typed graphlets are called *position-aware typed graphlets* to distinguish them from typed graphlets formalized in Definition 5. See Figure 7 for an example of position-aware typed graphlets

and Table 4 summarizes the combinatorial and enumerative properties. For a single  $K$ -node induced subgraph, the number of *position-aware typed graphlets* with  $L$  types is  $L^K$ .

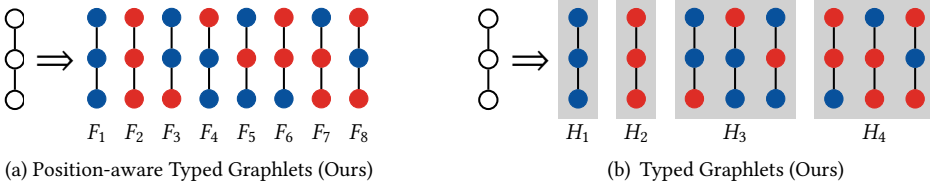


Fig. 7. **Position-aware Typed Graphlets (Def. 8) and Typed Graphlets (Def. 5).** This intuitive example shows the difference between position-aware typed graphlets *and* typed graphlets that do not impose the constraint that types coincide via the isomorphism.

## 5.2 Algorithm

To count position-aware typed graphlets, we only need a slight modification to the previous algorithm. Notice the algorithmic difference between Definition 5 and position-aware typed graphlets defined in Definition 8 is that to count typed graphlets, we only need to consider the types involved in the graphlet, and not the nodes/edges that correspond to those types (*i.e.*, types coincide via the isomorphism). In other words, typed graphlets formally defined in Definition 5 can be viewed as ignoring the “order” of the types with respect to their assignment to nodes in the induced subgraph. Recall that Section 4.5 used a function to map all possible orderings of a given multiset of  $L$  types to a single hash value (using a precomputed hash table). This allowed us to merge all such “position-aware typed graphlets” that have the same multiset of types (Figure 7(a)) to the appropriate typed graphlet (Figure 7(b)). For instance, the counts of the position-aware typed graphlets  $\{F_3, F_4, F_5\}$  in Figure 7(a) are all mapped to the same typed graphlet  $H_3$  shown in Figure 7(b). Therefore, the framework (Algorithm 1) actually counts position-aware typed graphlets and uses a mapping function to obtain a single hash value for all such type orderings of some multiset of types, which allows the position-aware typed graphlet counts of a typed graphlet to be merged into a single count. To count position-aware typed graphlets, we simply remove the lookup table that maps the appropriate position-aware typed graphlets to the corresponding typed graphlet. Interestingly, this equates to slightly less work required for computing position-aware typed graphlets. As such, the analysis in Section 7 and elsewhere clearly holds for both position-aware typed graphlets and typed graphlets.

**PROPERTY 7.** *Given an untyped induced subgraph  $H$ , let  $T_H$  denote the number of typed graphlets of  $H$  in  $G$ , and let  $P_H$  denote the number of position-aware typed graphlets of  $H$  in  $G$ . Then  $P_H \geq T_H$ .*

**PROOF.** If  $P_H = T_H$ , then for the position-aware typed graphlets that map to a specific typed graphlet (*e.g.*,  $\{F_3, F_4, F_5\}$  maps to  $H_3$  in Figure 7), only one position-aware typed graphlet must have nonzero count in  $G$ , and this must hold for all typed graphlets in  $G$ . Otherwise, if there exists a typed graphlet with more than one position-aware typed graphlet with nonzero count, then  $P_H > T_H$ . ■

## 6 GLOBAL TYPED GRAPHLET COUNTS

While Section 4 focused on counting typed graphlets locally for each edge  $(i, j) \in E$  in  $G$ , one may also be interested in the total counts of each typed graphlet in  $G$ . More formally,

**PROBLEM 2 (GLOBAL TYPED GRAPHLETS).** *Given a graph  $G$  with  $L$  types, the global typed graphlet counting problem is to find the set of all typed graphlets that occur in  $G$  along with their corresponding frequencies. This work focuses on computing all  $\{2, 3, 4\}$ -node typed graphlet counts for  $G$ .*

A general equation for solving the above problem for any arbitrary typed graphlet  $H$  is given below. Let  $H$  denote an arbitrary typed graphlet and  $\mathbf{x}$  be an  $M$ -dimensional vector of counts of  $H$  for every edge  $(i, j) \in E$ , then the frequency of  $H$  in  $G$  is:

$$C_H = \frac{1}{|E(H)|} \mathbf{x}^\top \mathbf{e} \quad (41)$$

where  $|E(H)|$  is the number of edges in the typed graphlet  $H$  and  $\mathbf{e} = [1 \ \cdots \ 1]$  is an  $M$ -dimensional vector of all 1's.

## 7 THEORETICAL ANALYSIS

First, we show the relationship between the count of an untyped graphlet  $H$  in  $G$  and the count of all typed graphlets in  $G$  with induced subgraph  $H$ .

**PROPOSITION 1.** *Let  $\mathbf{x}$  denote the vector of counts for any untyped graphlet  $H \in \mathcal{H}$  (e.g., 4-cycle). Further, let  $\mathbf{X}$  denote a  $M \times T_H$  matrix of typed graphlet counts for graphlet  $H$  where  $T_H$  denotes the number of typed graphlets that arise from  $L$  types. Then the following holds:*

$$C = \sum_{i=1}^M x_i = \sum_{i=1}^M \sum_{j=1}^{T_H} X_{ij} \quad (42)$$

Consider the counts of an untyped graphlet for a single edge. The above demonstrates that these counts are partitioned among the set of typed graphlets that arise from the untyped graphlet when types are considered. Let  $\mathbf{p} \in \mathbb{R}^{T_H}$  denote a typed graphlet probability distribution ( $\mathbf{p}^\top \mathbf{e} = 1$ ), the entropy (average information content) of  $\mathbf{p}$  is  $\mathbb{H}(\mathbf{p}) = -\sum_i p_i \log p_i$ . Hence,  $\mathbb{H}(\mathbf{p})$  quantifies the amount of information in the relative frequencies of the typed graphlets (of a given graphlet  $H \in \mathcal{H}$ ). In the case of untyped graphlets, the  $C = \sum_{i=1}^M x_i$  untyped graphlets are assumed to belong to a single homogeneous graphlet where all nodes are of the same type. This matches exactly the information we have if types are not considered.

**PROPOSITION 2.** *Assume  $\mathbf{p} \in \mathbb{R}^{T_H}$  is an arbitrary typed graphlet probability distribution such that  $p_i < 1, \forall i$  and  $\mathbf{q}$  is the untyped graphlet distribution where  $q_i = 1$  and  $q_j = 0, \forall j \neq i$ , then  $\mathbb{H}(\mathbf{p}) > \mathbb{H}(\mathbf{q})$ .*

This implies that typed graphlets contain more information than untyped graphlets. The proof is straightforward.

### 7.1 Time Complexity

We first introduce two properties that are useful to understand the complexity.

**PROPERTY 8.**

$$d_i + d_j = 2|T_{ij}| + |S_i| + |S_j| \quad (43)$$

where  $d_i = |\Gamma_i|$ ,  $d_j = |\Gamma_j|$ ,  $T_{ij} = \Gamma_i \cap \Gamma_j$ ,  $S_i = \Gamma_i \setminus T_{ij}$ , and  $S_j = \Gamma_j \setminus T_{ij}$ .

**PROPERTY 9.** *The space required to store  $T_{ij}$ ,  $S_i$ , and  $S_j$  is less than  $d_i + d_j$  iff  $|T_{ij}| > 0$ .*

This is straightforward to see since  $|S_i| + |S_j| = |S_i \cup S_j|$  always holds. However, if  $|T_{ij}| = 0$ , then  $|S_i| + |S_j| = d_i + d_j$ . Hence, triangles represent the smallest clique, and as shown in [Rossi and Zhou 2018] can be used to compress the graph. As the density of the graph increases, more triangles are



formed, and therefore less space is used. Notice that the worst case is also unlikely to occur because of this fact. For instance, suppose  $d_i = \Delta$ ,  $d_j = \Delta$ , and  $\Delta = n$  (worst case), then  $|T_{ij}| = d_i = d_j$ , and  $|S_i| = 0$ ,  $|S_j| = 0$ . Furthermore, if  $|S_i| = n$ , then  $|S_j| = 0$  and  $|T_{ij}| = 0$  must hold. Obviously, if node  $i$  is connected to all  $n$  nodes, then any node  $k \in \Gamma_j$  must form a triangle with  $i$  ( $k \in T_{ij}$ ). For any node with maximum degree  $\Delta$ , there is very low probability that  $|T_{ij}| = 0$ , which implies  $|T_{ij}| > 0$ ,  $|S_i| < \Delta$  and  $|S_j| < \Delta$ .

### 7.1.1 Typed 3-Node Graphlets.

**THEOREM 2.** *The worst-case time complexity for counting all 3-node typed graphlets for a given edge  $(i, j) \in E$  is:*

$$O(2|\Gamma_i| + |\Gamma_j|) = O(\Delta) \quad (44)$$

where  $|\Gamma_i|$  and  $|\Gamma_j|$  denote the number of nodes connected to node  $i$  and  $j$ , respectively. Further,  $\Delta$  is the maximum degree in  $G$ .

**PROOF.** It takes at most  $O(|\Gamma_i| + |\Gamma_j|)$  time to compute typed triangles (i.e.,  $T_{ij}^t$ , for all  $t = 1, \dots, L$ ) by hashing neighbors of  $i$  in  $O(|\Gamma_i|)$  time, and then checking if each node  $w \in \Gamma_j$  is hashed or not, taking  $O(|\Gamma_j|)$  time. Similarly, if  $w \in \Gamma_j$  is not hashed, then  $S_j^t \leftarrow S_j^t \cup \{w\}$  where  $t = \phi_w$ . Now all that remains is computing  $S_i^t$ , for all  $t$ . Notice  $|S_i^t| = |\Gamma_i^t| - |T_{ij}^t|$ , for all  $t = 1, \dots, L$ . ■

**7.1.2 Typed 4-Node Graphlets.** We first provide the time complexity of deriving path-based and triangle-based graphlet orbits in Lemma 5-6, and then give the total time complexity of all 3 and 4-node typed graphlets in Theorem 3 based on these results. Note that Lemma 5-6 includes the time required to derive all typed 3-node typed graphlets.

**LEMMA 5.** *For a single edge  $(i, j) \in E$ , the worst-case time complexity for deriving all typed path-based graphlet orbits is:*

$$O(\Delta(|S_i| + |S_j|)) \quad (45)$$

Note  $|S_i|\Delta \geq \sum_{k \in S_i} d_k$  and  $|S_j|\Delta \geq \sum_{k \in S_j} d_k$ .

**LEMMA 6.** *For a single edge  $(i, j) \in E$ , the worst-case time complexity for deriving all typed triangle-based graphlet orbits is:*

$$O(\Delta|T_{ij}|) \quad (46)$$

Notice  $|T_{ij}|\Delta \geq |T_{ij}|\Delta_T \geq \sum_{k \in T_{ij}} d_k$  where  $\Delta$  is the maximum degree of a node in  $G$  and  $\Delta_T$  is the maximum degree of a node in  $T_{ij}$ . Thus,  $|T_{ij}|\Delta$  only occurs iff  $\forall k \in T_{ij}, d_k = \Delta$  where  $\Delta =$  maximum degree of a node in  $G$ . In sparse real-world graphs,  $T_{ij}$  is likely to be smaller than  $S_i$  and  $S_j$  as triangles are typically more rare than 3-node paths. Conversely,  $T_{ij}$  is also more likely to contain high degree nodes, as nodes with larger degrees are obviously more likely to form triangles than those with small degrees.

From Lemma 5-6, we have the following:

**THEOREM 3.** *For a single edge  $(i, j) \in E$ , the worst-case time complexity for deriving all 3 and 4-node typed graphlet orbits is:*

$$O(\Delta(|S_i| + |S_j| + |T_{ij}|)) \quad (47)$$

**PROOF.** The time complexity of each step is provided below. Hashing all neighbors of node  $i$  takes  $O(|\Gamma_i|)$ . Recall from Lemma 2 that counting all 3-node typed graphlets takes  $O(2|\Gamma_i| + |\Gamma_j|) = O(\Delta)$  time for an edge  $(i, j) \in E$ . This includes the time required to derive the number of typed 3-node stars and typed triangles for all types  $t = 1, \dots, L$ . This information is needed to derive the remaining

typed graphlet orbit counts in constant time. Next, Algorithm 2 is used to derive a few path-based typed graphlet orbit counts taking  $O(\Delta(|S_i| + |S_j|))$  time in the worst-case. Similarly, Algorithm 3 is used to derive a few triangle-based typed graphlet orbit counts taking in the worst-case  $O(\Delta|T_{ij}|)$  time. As an aside, updating the count of a typed graphlet count is  $o(1)$  (Algorithm 4).

Now, we derive the remaining typed graphlet orbit counts in constant time (Line 9-10). Since each type pair leads to different typed graphlets, we must iterate over at most  $L(L-1)/2 + L$  type pairs. For each pair of types selected, we derive the typed graphlet orbit counts in  $o(1)$  constant time via Eq. 19-30 (See Line 9-10). Furthermore, the term involving  $L$  is for the worst-case when there is at least one node in all  $L$  sets (*i.e.*, at least one node of every type  $L$ ). Nevertheless, since  $L$  is a small constant,  $L(L-1)/2 + L$  is negligible. Therefore, for a single edge, the worst-case time complexity is  $O(\Delta(|S_i| + |S_j| + |T_{ij}|))$ .

Let  $\bar{T}$  and  $\bar{S}$  denote the average number of triangle and 3-node stars incident to an edge in  $G$ . More formally,  $\bar{T} = \frac{1}{M} \sum_{(ij) \in E} |T_{ij}|$  and  $\bar{S} = \frac{1}{M} \sum_{(ij) \in E} |S_i| + |S_j|$ . The total worst-case time complexity for all  $M$  edges is  $O(M\Delta(\bar{S} + \bar{T}))$ . Note that obviously  $\bar{S}M = \sum_{(ij) \in E} |S_i| + |S_j|$  and  $\bar{T}M = \sum_{(ij) \in E} |T_{ij}|$ . ■

**COROLLARY 12.** *The worst-case time complexity of counting typed graphlets using Algorithm 1 matches the worst-case time complexity of the best known untyped graphlet counting algorithm.*

**PROOF.** From Theorem 3 we have that  $O(\Delta(|S_i| + |S_j| + |T_{ij}|))$ , which is exactly the time complexity of the best known untyped graphlet counting algorithm (PGD [Ahmed et al. 2015, 2016]). ■

## 7.2 Space Complexity

Since our approach generalizes to graphs with an arbitrary number of types  $L$ , the specific set of typed graphlets is unknown. As demonstrated in Table 2, it is impractical to store the counts of all possible  $k$ -node typed graphlets for any graph of reasonable size as typically done in traditional methods for untyped graphlets [Ahmed et al. 2015; Marcus and Shavitt 2012]. The space complexity required to store the counts of all possible typed graphlets is at least:

$$O(MT_{\max}) \quad (48)$$

where  $M = |E|$  is the number of edges in  $G$  and  $T_{\max}$  is the number of different possible typed graphlets with  $L$  types. Thus,  $MT_{\max}$  is the total space to store  $M$  vectors of length  $T_{\max}$ , *i.e.*, one  $T_{\max}$ -dimensional vector per edge. To understand the space requirements and how it is impractical for any moderately sized graph, suppose we have a graph with  $M = 10,000,000$  edges and  $L = 7$  types. Counting all 3- and 4-node typed graphlet orbits for every edge would require 90.72 GB of space to store the large  $MT_{\max}$  matrix (assuming 4 bytes per count/entry). This is obviously impractical for any graph of even moderate size. In contrast, Algorithm 1 is orders of magnitude more space-efficient.

**LEMMA 7.** *The space complexity of typed graphlets (Algorithm 1) is  $O(M\bar{T})$ .*

**PROOF.** For an edge  $(i, j) \in E$ , it takes  $|X_{ij}|$  space to store the counts of the nonzero typed graphlets. Let  $\bar{T} = \frac{1}{M} \sum_{(ij) \in E} |X_{ij}|$  denote the average number of typed graphlets with nonzero counts per edge. Therefore, the total space required to store the nonzero typed graphlet counts for all  $M = |E|$  edges is only  $O(M\bar{T})$ . The space of all other data structures used in Algorithm 1 is small in comparison, *e.g.*,  $\Psi$  takes at most  $O(|V|)$  space, whereas  $T_{ij}$ ,  $S_i$ , and  $S_j$  take  $O(\Delta)$  space in the worst-case (by Property 1) and can be reused for every edge. In addition, the size of  $\mathbf{x}$  is independent of the graph size  $(|V| + |E|)$  and can also be reused. ■

From Lemma 7, it is straightforward to see that

$$O(M\bar{T}) \ll O(MT_{\max}) \quad (49)$$

The space required by the proposed approach (Algorithm 1) is nearly-optimal and orders of magnitude lower than methods used for colored graphlets such as GC [Gu et al. 2018], which by definition solve a much simpler problem since there are strictly fewer colored graphlet counts to store. This is also shown empirically in Table 8.

Table 5. Runtime results for counting typed graphlets (ours) compared to state-of-the-art methods for colored graphlets (which is a different but simpler problem). Since these methods are unable to handle large or even medium-sized graphs as shown below, we include a number of very small graphs (e.g., cora, citeseer, webkb) for comparison; and count all  $\{2, 3, 4\}$ -node typed graphlets (ours) and colored graphlets. Note  $\Delta = \max$  node degree;  $|\mathcal{T}_V| = \text{number of node types}$ ;  $|\mathcal{T}_E| = \text{number of edge types}$ .

	$ V $	$ E $	$\Delta$	$ \mathcal{T}_V $	$ \mathcal{T}_E $	SECONDS			Typed Graphlets
						GC	ESU	G-Tries	
citeseer	3.3k	4.5k	99	6	21	46.27	5937.75	144.08	<b>0.022</b>
cora	2.7k	5.3k	168	7	28	467.20	10051.07	351.40	<b>0.032</b>
fb-relationship	7.3k	44.9k	106	6	20	1374.60	54,837.69	3789.17	<b>0.701</b>
web-polblogs	1.2k	16.7k	351	2	1	28,986.70	26,577.10	1,563.04	<b>1.055</b>
ca-DBLP	2.9k	11.3k	69	3	3	149.20	1,188.11	18.90	<b>0.100</b>
inf-openflights	2.9k	15.7k	242	2	2	9262.20	18,839.36	458.01	<b>0.578</b>
soc-wiki-elec	7.1k	100.8k	1.1k	2	2	ETL	ETL	26,468.85	<b>5.316</b>
webkb	262	459	122	5	14	85.82	7,158.10	187.22	<b>0.006</b>
terrorRel	881	8.6k	36	2	3	192.6	3130.7	241.1	<b>0.039</b>
pol-retweet	18.5k	48.1k	786	2	3	ETL	ETL	ETL	<b>0.296</b>
web-spam	9.1k	465k	3.9k	3	6	ETL	ETL	ETL	<b>210.97</b>
movielens	28.1k	170.4k	3.6k	3	3	ETL	ETL	ETL	<b>5.23</b>
citeulike	907.8k	1.4M	11.2k	3	2	ETL	ETL	ETL	<b>126.53</b>
yahoo-msg	100.1k	739.8k	9.4k	2	2	ETL	ETL	ETL	<b>35.22</b>
dbpedia	495.9k	921.7k	24.8k	4	3	ETL	ETL	ETL	<b>56.02</b>
digg	217.3k	477.3k	219	2	2	ETL	ETL	ETL	<b>5.592</b>
bibsonomy	638.8k	1.2M	211	3	3	ETL	ETL	ETL	<b>3.631</b>
epinions	658.1k	2.6M	775	2	2	ETL	ETL	ETL	<b>85.27</b>
flickr	2.3M	6.8M	216	2	2	ETL	ETL	ETL	<b>120.79</b>
orkut	6M	37.4M	166	2	2	ETL	ETL	ETL	<b>1241.01</b>
ER (10K,0.001)	10k	50.1k	26	5	15	183.32	5,399.14	241.27	<b>0.48</b>
CL (1.8)	9.2k	44.2k	218	5	15	31,668	45,399.14	5,241.27	<b>1.46</b>
KPGM (log 12,14)	3.3k	43.2k	1.3k	5	15	ETL	ETL	63,843.86	<b>8.94</b>
SW (10K,6,0.3)	10k	30k	12	5	15	21.48	5,062.67	206.92	<b>0.24</b>

\* ETL = Exceeded Time Limit (24 hours / 86,400 seconds)

## 8 EXPERIMENTS

The experiments are designed to investigate the runtime performance (Section 8.1), space-efficiency (Section 8.2), parallelization (Section 8.3), and scalability (Section 8.4-8.5) of the proposed approach. Results for position-aware typed graphlets are provided in Section 8.6. We also demonstrate the

Table 6. Runtime speedup results. Note “ $\infty$ ” indicates the baseline method (GC, ESU, or G-Tries) did not terminate within 24 hours and thus the precise speedup is unknown.  $|\mathcal{T}_V|$  = number of node types;  $|\mathcal{T}_E|$  = number of edge types.









	$ V $	$ E $	$ \mathcal{T}_V $	$ \mathcal{T}_E $	SPEEDUP (TYPED GRAPHLETS VS.)		
					GC	ESU	G-Tries
citeseer	3.3k	4.5k	6	21	2103x	269897x	6549x
cora	2.7k	5.3k	7	28	14600x	314095x	10981x
fb-relationship	7.3k	44.9k	6	20	1960x	78227x	5405x
web-polblogs	1.2k	16.7k	2	1	27475x	25191x	1481x
ca-DBLP	2.9k	11.3k	3	3	1492x	11881x	189x
inf-openflights	2.9k	15.7k	2	2	16024x	32594x	792x
soc-wiki-elec	7.1k	100.8k	2	2	$\infty$	$\infty$	45793x
webkb	262	459	5	14	14303x	1193016x	31203x
terrorRel	881	8.6k	2	3	4938x	80274x	6182x
pol-retweet	18.5k	48.1k	2	3	$\infty$	$\infty$	$\infty$
web-spam	9.1k	465k	3	6	$\infty$	$\infty$	$\infty$
movielens	28.1k	170.4k	3	3	$\infty$	$\infty$	$\infty$
citeulike	907.8k	1.4M	3	2	$\infty$	$\infty$	$\infty$
yahoo-msg	100.1k	739.8k	2	2	$\infty$	$\infty$	$\infty$
dbpedia	495.9k	921.7k	4	3	$\infty$	$\infty$	$\infty$
digg	217.3k	477.3k	2	2	$\infty$	$\infty$	$\infty$
bibsonomy	638.8k	1.2M	3	3	$\infty$	$\infty$	$\infty$
epinions	658.1k	2.6M	2	2	$\infty$	$\infty$	$\infty$
flickr	2.3M	6.8M	2	2	$\infty$	$\infty$	$\infty$
orkut	6M	37.4M	2	2	$\infty$	$\infty$	$\infty$
ER (10K,0.001)	10k	50.1k	5	15	381x	11248x	502x
CL (1.8)	9.2k	44.2k	5	15	21690x	31095x	3589x
KPGM (log 12,14)	3.3k	43.2k	5	15	$\infty$	$\infty$	7141x
SW (10K,6,0.3)	10k	30k	5	15	89x	21094x	862x

utility of *typed graphlets* for two important use cases: (i) exploratory analysis/mining (Section 8.7) and for (ii) improving a real-world predictive modeling application (Section 8.8). To demonstrate the *effectiveness* of the approach, we use a variety of heterogeneous and attributed network data from different application domains. All data can be accessed at NetworkRepository [Rossi and Ahmed 2015].

## 8.1 Runtime Comparison

Since this is the first work to propose and investigate typed graphlets, there are no existing methods for direct comparison. Nevertheless, we compare the proposed framework to a few recent methods that focus on counting colored graphlets, which is a fundamentally simpler problem. We first demonstrate how fast the proposed framework is for deriving typed graphlets by comparing the runtime (in seconds) of our approach against three methods for colored graphlets (a similar but simpler problem), namely, ESU (using fanmod) [Wernicke and Rasche 2006], G-Tries [Ribeiro and Silva 2014], and GC [Gu et al. 2018]. Note that these methods do not solve the typed graphlet problem (formally defined in Section 3). See Figure 2 for an intuitive example of the key differences

Table 7. Comparing the number of unique *typed graphlets* that occur for each induced subgraph (e.g., there are 40 typed triangles with different type structures in citeseer).

Network data	$ E $	$ \mathcal{T}_V $	$ \mathcal{T}_E $								
citeseer	4.5k	6	21	56	40	124	119	66	98	56	19
cora	5.3k	7	28	82	49	202	190	76	157	73	19
fb-relationship	44.9k	6	20	50	47	112	109	85	106	89	77
web-polblogs	16.7k	2	1	4	4	5	5	5	5	5	5
ca-DBLP	11.3k	3	3	10	10	15	15	15	15	15	15
inf-openflights	15.7k	2	2	4	4	5	5	5	5	5	5
soc-wiki-elec	100.8k	2	2	4	4	5	5	5	5	5	5
webkb	459	5	14	31	21	59	59	23	51	32	8
terrorRel	8.6k	2	3	4	4	5	0	4	5	5	5
pol-retweet	48.1k	2	3	4	4	5	5	5	5	5	4
web-spam	465k	3	6	10	10	15	15	15	15	15	15
movielens	170.4k	3	3	7	1	6	9	6	3	3	0
citeulike	1.4M	3	2	5	0	3	6	3	0	0	0
yahoo-msg	739.8k	2	2	3	2	3	4	3	3	3	2
dbpedia	921.7k	4	3	8	0	6	10	5	0	0	0
digg	477.3k	2	2	4	3	4	5	4	4	4	2
bibsonomy	1.2M	3	3	7	1	6	9	6	3	3	0
epinions	2.6M	2	2	3	2	3	4	3	3	3	2
flickr	6.8M	2	2	3	2	3	4	3	3	3	2
orkut	37.4M	2	2	4	3	4	4	3	4	3	2
ER (10K,0.001)	50.1k	5	15	35	30	70	70	69	66	1	0
CL (1.8)	44.2k	5	15	35	35	70	70	70	70	70	68
KPGM (log 12,14)	43.2k	5	15	35	35	70	70	70	70	70	70
SW (10K,6,0.3)	30k	5	15	35	35	70	70	70	70	70	69

between colored graphlets and our proposed formalization called typed graphlets. Since these methods are inherently serial (and difficult to parallelize), we use a serial version of the proposed approach for comparison. We also note that the three methods count colored graphlets for every node whereas the proposed approach derives typed graphlets for every edge. See Section 2 for other key differences. Nevertheless, these methods are used for comparison since they are the closest to our own work and solve conceptually simpler problems than the one described in this paper.<sup>4</sup>

For comparison, we use a wide variety of real-world graphs from different domains. In Table 5, we report the time (in seconds) required by each method. We also report the speedup obtained from our approach over the other methods in Table 6. To be able to compare with the existing methods, we included a variety of very small graphs for which the existing methods could solve in a reasonable amount of time. Note ETL indicates that a method did not terminate within 24 hours. Strikingly, the existing methods are unable to handle medium to large graphs with hundreds of thousands or more nodes and edges as shown in Table 5. Even small graphs can take hours to finish using existing methods (Table 5). For instance, the small citeseer graph with only 3.3k nodes and 4.5k edges takes 46.27 seconds using the best existing method whereas ours finishes in a tiny fraction of a second, notably,  $2/100$  seconds. This is about 2,100 times faster than the next best method. Similarly, on the small cora graph with 2.7K nodes and 5.3K edges, GC takes 467

<sup>4</sup>As an aside, we do not compare to methods for counting *untyped graphlets* since these obviously solve a fundamentally different problem and thus are outside the scope of this work. Furthermore, we also do not focus on graphs with edge types or counting typed graphlets with 5-nodes (or larger) as these are outside the scope of this paper and left for future research.

seconds whereas G-Tries takes 351 seconds. However, our approach finishes counting all typed graphlets with  $\{2, 3, 4\}$ -nodes in only 0.03 seconds. This is 10,000 times faster than the next best method. Unlike existing methods, our approach is shown to be significantly faster and able to handle large-scale graphs. The significant speedups obtained by our approach are largely due to the combinatorial relationships between the typed graphlets that we introduce in this work and leverage for deriving many of the typed graphlets in  $o(1)$  time. On flickr, our approach takes about 2 minutes to count the occurrences of all typed graphlets for all 6.8 million edges. Across all graphs, the proposed method achieves significant speedups over the existing methods as shown in Table 6. These results demonstrate the effectiveness of our approach for *counting typed graphlets* in large real-world networks. As such, the proposed approach brings new opportunities to leverage typed graphlets for real-world applications on much larger networks.

The results in Table 5-6 demonstrate the effectiveness of the proposed approach on a wide variety of heterogeneous and attributed network data from different application domains. As an aside, the first 11 graphs in Table 5-6 (*i.e.*, citeseer, cora, fb-relationship, web-polblogs, ca-DBLP, inf-openflights, soci-wiki-elec, webkb, terrorRel, pol-retweet, and web-spam) are all attributed networks as the “type” corresponds to different attributes (*e.g.*, political views, paper/research topic, gender, protein function, among others). This is in contrast to the other 9 real-world networks in Table 5-6 (*i.e.*, movielens, citeulike, yahoo-msg, dbpedia, digg, bibsonomy, epinions, flickr, orkut) where the types correspond to node or edge types such as users, movies, papers, ratings, among others.

Typed graphlet statistics are shown in Table 7. This includes the number of typed graphlets with nonzero counts for each induced subgraph. For instance, in cora, there are 49 typed triangle graphlets with nonzero counts out of the 84 possible typed triangle graphlets that could actually occur with  $L = 7$  types. From these results, we make an important observation. In real-world graphs we observe that certain typed graphlets do not occur at all in the graph. We define such typed graphlets that do not occur in  $G$  as *forbidden typed graphlets* as their appearance in the future would indicate something strong. For instance, perhaps an anomaly or malicious activity. Other interesting insights and applications of typed graphlets are discussed and explored further in Section 8.7 and Section 8.8.

We also generated synthetic graphs from 4 different graph models including: Erdős-Rényi (ER) [Erdős and Rényi 1960], Chung-Lu (CL) [Chung and Lu 2002], Kronecker Product Graph Model (KPGM) [Leskovec et al. 2010], and Watts-Strogatz Small-World (SW) graph model [Watts and Strogatz 1998]. Since these models generate graphs without types, we assign them uniformly at random such that  $\frac{N}{L}$  nodes are assigned to every type. Unless otherwise mentioned, we set  $L = 5$ . Results are provided at the bottom of Table 5. Just as before, we observe significant speedups across all graphs and methods as shown in Table 6. Other experiments using synthetic graphs are discussed in Section 8.5.

## 8.2 Space Efficiency Comparison

We theoretically showed the space complexity of our approach in Section 7.2. In this section, we empirically investigate the space-efficiency of our approach compared to ESU (using fanmod) [Wernicke and Rasche 2006], G-Tries [Ribeiro and Silva 2014], and GC [Gu et al. 2018]. Table 8 reports the space used by each method for a variety of real-world graphs. Strikingly, the proposed approach uses between 42x and 776x less space than existing methods as shown in Table 8. These results indicate that our approach is space-efficient and practical for large networks. As an aside, typed graphlet statistics are also shown in Table 7.

Table 8. Comparing the *space* used by the proposed typed graphlet approach. Since there are no existing methods for typed graphlet, we compare against colored graphlet methods that solve a simpler problem.

	citeseer	cora	movielens	web-spam
<b>GC</b>	30.1MB	50.4MB	ETL	ETL
<b>ESU</b>	13.4MB	46.2MB	ETL	ETL
<b>G-Tries</b>	161.9MB	448.6MB	ETL	ETL
<b>Typed graphlets</b>	<b>316KB</b>	<b>578KB</b>	<b>22.5MB</b>	<b>128.9MB</b>
<b>Position-aware typed graphlets</b>	<b>417KB</b>	<b>806KB</b>	<b>32.2MB</b>	<b>192.1MB</b>

\* ETL = Exceeded Time Limit (24 hours / 86,400 seconds)

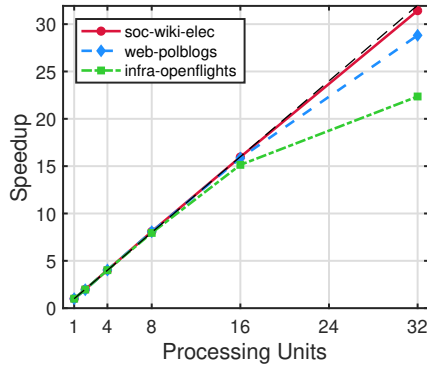


Fig. 8. Parallel speedup of the proposed approach. Notably, the approach exhibits nearly linear parallel scaling as the number of cores increases.

### 8.3 Parallel Speedup

This section evaluates the parallel scaling of the proposed approach. As an aside, this work describes the first parallel approach for typed graphlet counting. In these experiments, we used a two processor, Intel Xeon E5-2686 v4 system with 256 GB of memory. None of the experiments came close to using all the memory. Parallel speedup is simply  $S_p = \frac{T_1}{T_p}$  where  $T_1$  is the execution time of the sequential algorithm, and  $T_p$  is the execution time of the parallel algorithm with  $p$  processing units (cores). In Figure 8, we observe nearly linear speedup as we increase the number of cores. These results indicate the effectiveness of the parallel algorithm for counting typed graphlets in general heterogeneous graphs.

### 8.4 Scalability

To evaluate the scalability of the proposed framework as the size of the graph grows (*i.e.*, number of nodes and edges increase), we generate Erdős-Rényi graphs of increasing size (from 100 to 1 million nodes) such that each graph has an average degree of 10. In Figure 9, we observe that our approach scales linearly as the number of nodes and edges grow large. As an aside, our approach takes less than 2 minutes to derive all typed  $\{2, 3, 4\}$ -node graphlets for a large graph with 1 million nodes and 10 million edges. Note that existing methods are not shown in Figure 9 since they are unable to handle medium to large-sized graphs as shown previously in Table 5.

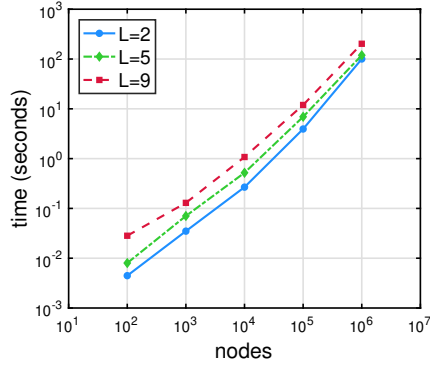


Fig. 9. Scalability of the proposed framework as the size of the graph increases. Erdős-Rényi graphs with an average degree of 10 are used.

### 8.5 Synthetic Graph Experiments

In these experiments, we generate synthetic graphs. For each graph, we vary the number of node types  $L$  from 2 to 9, and measure the runtime performance as the number of node types increases as well as the impact in terms of space as  $L$  increases. Given  $L \in \{2, \dots, 9\}$  node types, we assign types to nodes uniformly at random such that  $\frac{N}{L}$  nodes are assigned to every type.

Table 9. Comparing the number of unique typed graphlets that occur for each induced subgraph as we vary the number of types  $L$  using a KPGM graph with 3.3k nodes, 43.2k edges, average degree = 26, and max degree = 1.3K. Speedup is shown in parenthesis.

$L$									time (sec.)	
									serial	parallel - 4 cores
2	4	4	5	5	5	5	5	5	8.11	2.08 (3.89x)
5	35	35	70	70	70	70	70	70	8.94	2.26 (3.95x)
9	165	165	495	495	495	495	495	495	10.37	2.62 (3.95x)

**8.5.1 Impact on Performance.** We first investigate the runtime performance of our approach as the number of types  $L$  increases from 2 to 9. We use both a serial and parallel implementation of our method for comparison. Results are shown in Table 9. Notably, the parallel speedup of the parallel algorithm is constant regardless of  $L$ . Therefore, it is not impacted by the increase in  $L$ . Furthermore, the runtime of both the serial and parallel algorithm increases slightly as  $L$  increases. Notice the additional work depends on the number of unique typed graphlets (the sum of columns 2-9 in Table 9) and not directly on  $L$  itself. The total amount of unique typed graphlets substantially increases as  $L$  increases from 2 to 9 as shown in Table 9. This is primarily due to the random assignment of types to nodes. However, in sparse real-world graphs the total unique typed graphlets is typically much smaller as shown in Table 7.

To further understand how the structure of the graph impacts runtime, we generate an ER and Chung-Lu (CL) graph with 100K nodes, 1M edges, and average degree 10. We vary the number of types  $L$  and assign types to nodes uniformly at random as discussed previously. Notice that both the ER and CL graph are generated such that they each have 100K nodes, 1 million edges, with average degree 10. However, both graphs are structurally very different. For instance, the degrees among the nodes in the ER graph are more uniform whereas the degree of the nodes in CL are



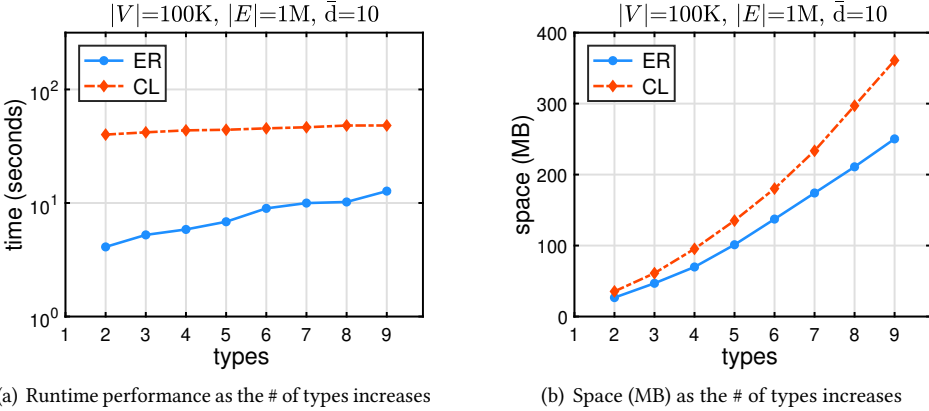


Fig. 10. Comparing runtime performance *and* space as the number of types increases. **Left:** In the case of CL graphs with a skewed degree distribution, the runtime is nearly constant as the number of types increases. **Right:** Both ER and CL have similar space requirements. Since node types are randomly assigned to nodes, this represents a type of worst-case since every edge in  $G$  is likely to have significantly more distinct typed graphlets compared to sparse real-world graphs (as shown in Table 7 and Section 8.7). Note  $\bar{d} = \frac{1}{n} \sum d_i$ .

skewed such that a few nodes have very large degree while the others have relatively small degree. We observe in Figure 10(a) that for CL graphs with a skewed degree distribution, the runtime of the approach as the number of types increases is essentially constant. This result is important as most real-world graphs also have a skewed degree distribution (social networks, web graphs, information networks, etc.) [Faloutsos et al. 1999; Girvan and Newman 2002]. However, even in the case where the degrees are more uniform across the nodes, our approach still performs well as shown in Figure 10(a).

**8.5.2 Impact on Space.** Figure 10(b) shows the memory (space) required by our approach as the number of types increases from  $L \in \{2, \dots, 9\}$ . Both ER and CL graphs are shown to have similar space requirements. This is likely due to the random assignment of types to nodes. This assignment represents a type of worst case since every edge is likely to have significantly more distinct typed graphlets compared to sparse real-world graphs. This difference can be seen in Table 7.









### 8.6 Position-Aware Typed Graphlet Results

In these experiments, we investigate position-aware typed graphlets introduced formally in Section 5. The difference between typed graphlets (Def. 5) and *position-aware typed graphlets* (Def. 8) is shown in Figure 7 using an intuitive example. In Table 10, we show the number of unique position-aware typed graphlets that occur for each induced subgraph using both synthetic and real-world graphs from a wide range of domains. Note the first eleven graphs in Table 10 are attributed graphs whereas the next nine graphs are heterogeneous networks.

Runtime results for position-aware typed graphlets is shown in Table 11. We compare the runtime of position-aware typed graphlets to typed graphlets. Notably, in all cases, the runtime of position-aware typed graphlets and typed graphlets is very close as shown in Table 11. Recall from Section 5 that this is what is expected since the algorithmic difference is trivial. To better understand the differences in runtime, we show the runtime in seconds of typed graphlets (x-axis) vs. position-aware typed graphlets (y-axis) for all the graphs in Table 11. Note the diagonal line in Figure 11 represents the expected runtime assuming that typed graphlets and position-aware

typed graphlets have the same runtime performance. In Figure 11, most of the graphs lie on the line, while a few have minor deviations in either direction. Position-aware typed graphlets are typically faster to compute for most of the real-world graphs as shown in Table 11. However, in terms of the four synthetic graphs in Table 11 (last four graphs), we find that typed graphlets is faster for three of the four with the KPGM graph being the exception. We also note that compared to the

Table 10. Comparing the number of unique *position-aware typed graphlets* that occur for each induced subgraph.

Network data	$ E $	$ \mathcal{T}_V $	$ \mathcal{T}_E $								
citeseer	4.5k	6	21	212	114	1090	885	278	663	243	49
cora	5.3k	7	28	299	132	1839	1626	289	1129	334	52
fb-relationship	44.9k	6	20	199	180	1148	1105	761	1113	895	703
web-polblogs	16.7k	2	1	8	8	16	16	16	16	16	16
ca-DBLP	11.3k	3	3	27	27	81	81	81	81	81	81
infra-openflights	15.7k	2	2	8	8	16	16	16	16	16	16
soc-wiki-elec	100.8k	2	2	8	8	16	16	16	16	16	16
webkb	459	5	14	97	63	444	396	98	382	209	44
terrorRel	8.6k	2	3	8	8	16	0	10	16	16	16
pol-retweet	48.1k	2	3	8	8	16	16	16	16	16	8
web-spam	465k	3	6	27	27	81	81	81	81	81	81
movielens	170.4k	3	3	18	6	42	42	18	30	18	0
citeulike	1.4M	3	1	12	0	16	20	8	0	0	0
yahoo-msg	739.8k	2	2	4	3	7	8	5	7	6	3
dbpedia	921.7k	4	3	20	0	36	36	14	0	0	0
digg	477.3k	2	2	6	4	10	12	7	8	5	3
bibsonomy	1.2M	3	3	18	6	42	42	18	30	18	0
epinions	2.6M	2	2	6	4	10	12	7	10	8	4
flickr	6.8M	2	2	6	4	10	12	7	10	8	4
orkut	37.4M	2	2	8	6	15	15	7	14	10	5
ER (10K,0.001)	50.1k	5	15	125	117	625	625	623	623	3	0
CL (1.8)	44.2k	5	15	125	125	625	625	625	625	625	625
KPGM (log 12,14)	43.2k	5	15	125	125	625	625	625	625	625	625
SW (10K,6,0.3)	30k	5	15	125	125	625	625	625	625	625	623

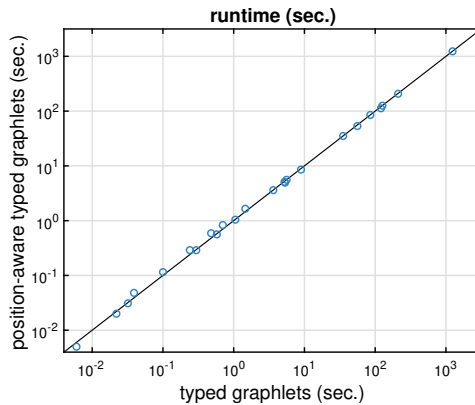


Fig. 11. Runtime (in seconds) for typed graphlets and position-aware typed graphlets. See text for discussion.

Table 11. Runtime results (in seconds) for counting position-aware typed graphlets (Def. 8) compared to typed graphlets (Def. 5). Best result is bold. Note  $\Delta$  = max node degree;  $|\mathcal{T}_V|$  = number of node types.

	$ V $	$ E $	$\Delta$	$ \mathcal{T}_V $	Typed graphlets	Position-aware (Def. 8)
citeseer	3.3k	4.5k	99	6	0.022	<b>0.020</b>
cora	2.7k	5.3k	168	7	0.032	<b>0.031</b>
fb-relationship	7.3k	44.9k	106	6	<b>0.701</b>	0.832
web-polblogs	1.2k	16.7k	351	2	1.055	<b>1.042</b>
ca-DBLP	2.9k	11.3k	69	3	<b>0.100</b>	0.115
inf-openflights	2.9k	15.7k	242	2	0.578	<b>0.562</b>
soc-wiki-elec	7.1k	100.8k	1.1k	2	5.316	<b>4.939</b>
webkb	262	459	122	5	0.006	<b>0.005</b>
terrorRel	881	8.6k	36	2	<b>0.039</b>	0.048
pol-retweet	18.5k	48.1k	786	2	0.296	<b>0.289</b>
web-spam	9.1k	465k	3.9k	3	210.97	<b>207.36</b>
movielens	28.1k	170.4k	3.6k	3	5.23	<b>5.12</b>
citeulike	907.8k	1.4M	11.2k	3	126.53	<b>125.66</b>
yahoo-msg	100.1k	739.8k	9.4k	2	35.22	<b>35.08</b>
dbpedia	495.9k	921.7k	24.8k	4	56.02	<b>53.36</b>
digg	217.3k	477.3k	219	2	5.592	<b>5.578</b>
bibsonomy	638.8k	1.2M	211	3	3.631	<b>3.607</b>
epinions	658.1k	2.6M	775	2	85.27	<b>85.05</b>
flickr	2.3M	6.8M	216	2	120.79	<b>112.45</b>
orkut	6M	37.4M	166	2	1241.01	<b>1236.21</b>
ER (10K,0.001)	10k	50.1k	26	5	<b>0.48</b>	0.59
CL (1.8)	9.2k	44.2k	218	5	<b>1.46</b>	1.65
KPGM (log 12,14)	3.3k	43.2k	1.3k	5	8.94	<b>8.56</b>
SW (10K,6,0.3)	10k	30k	12	5	<b>0.24</b>	0.29

other three methods for counting the simpler notion of colored graphlets, both typed graphlets and position-aware typed graphlets remain significantly faster.

In Table 8, we also report the space used by position-aware typed graphlets. Note that by definition position-aware typed graphlets use at least as much space as typed graphlets, and typically use more space than typed graphlets as shown in Table 8. Despite that typed graphlets and position-aware typed graphlets are more complex than colored graphlets (and theoretically should use less space), the proposed framework for both uses significantly less space than these other methods.

## 8.7 Exploratory Analysis

This section demonstrates the use of heterogeneous graphlets for mining and exploratory analysis.

**8.7.1 Political retweets.** The political retweets data consists of 18,470 Twitter users classified into 2 types that encode the users political leanings (*i.e.*, left, right). The graph has 61,157 links representing retweets. There are 24,815 triangles in the political retweet network. Triangles in this graph indicate that users retweeted by an individual also retweet each other (*i.e.*, triangle = three users that have all mutually retweeted each other). Triangles may represent users with similar interests. However, triangles alone do not reveal any additional information about the users. Another interesting question is as follows: are users with a particular political leaning more likely to form retweet triangles with users of the same political leaning or vice-versa? Unfortunately,

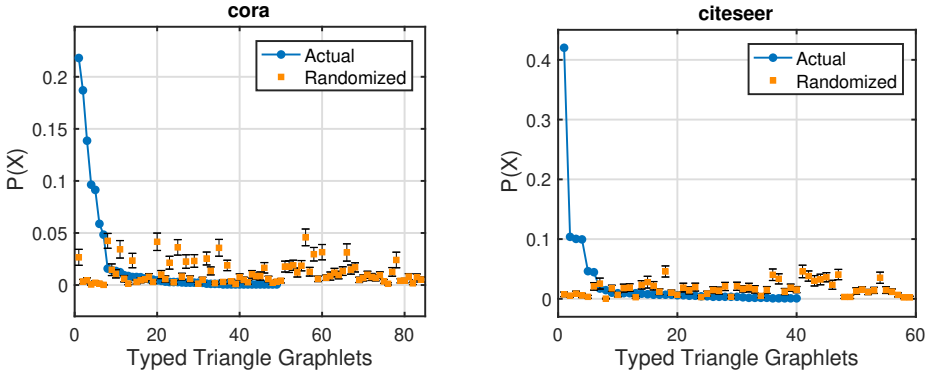


Fig. 12. Comparing the actual typed triangle distribution to the randomized typed triangle distribution. We compute 100 random permutations of the node types and run the approach on each permutation then average the resulting counts to obtain the mean randomized typed triangle distribution. There are three key findings. First, we observe a significant difference between the actual and randomized typed triangle distributions. Second, many of the typed triangles that occur when the types are randomized, do not occur in the actual typed triangle distribution. Third, we find the typed triangle distribution to be skewed as a few typed triangles occur very frequently while the vast majority have very few occurrences.

untyped triangles alone cannot be used to answer such questions. To answer such questions, typed graphlets are used by encoding the political leanings of a Twitter user as the type.<sup>5</sup> Interestingly, the 24,815 (untyped) triangles are distributed as follows:

$$p = \left[ \begin{matrix} \triangleleft & \triangle & \triangle & \triangle \\ 0.608 & 0.003 & 0.001 & 0.388 \end{matrix} \right]$$

Notably, we observe that 60.86% and 38.79% of the 24,815 triangles are formed among users with the same political leanings. This implies that three users with the same political leanings are more likely to retweet each other than with users of different political leanings. These results indicate the presence of homophily [McPherson et al. 2001] as users tend to retweet similar others. Furthermore, these homogeneous typed triangles ( $\triangleleft$ ,  $\triangle$ ) account for 99.65% of the 24,815 triangles. Intuitively, this implies that the network consists of two tightly-knit communities of users of the same political leanings. The two communities are sparsely connected. Typed triangles obviously contain significantly more information than untyped triangles. This includes not only information about the local properties but also about the global structure of the network as shown above. Obviously, untyped graphlets are unable to provide such insights as they do not encode the types, attribute values, or class labels associated with a graphlet. They only reveal the structural information independent of any important external information associated with the node.

We also investigated typed 4-clique graphlets. Strikingly, only 4 of the 5 typed 4-clique graphlets that arise from 2 types actually occur in the graph. In particular, the typed 4-clique graphlet with 2 right users and 2 left users does not even appear in the graph. This typed graphlet might indicate collusion between individuals from different political parties or some other extremely rare anomalous activity. The other typed 4-cliques that are extremely rare are the typed 4-clique graphlet with 3 right (left) users and a single left (right) user.

8.7.2 *Cora citation network.* The Cora citation network consists of 2708 scientific publications classified into one of seven types (class labels) that indicate the paper topic. The citation network

<sup>5</sup>Typed graphlets can be used with any attribute.

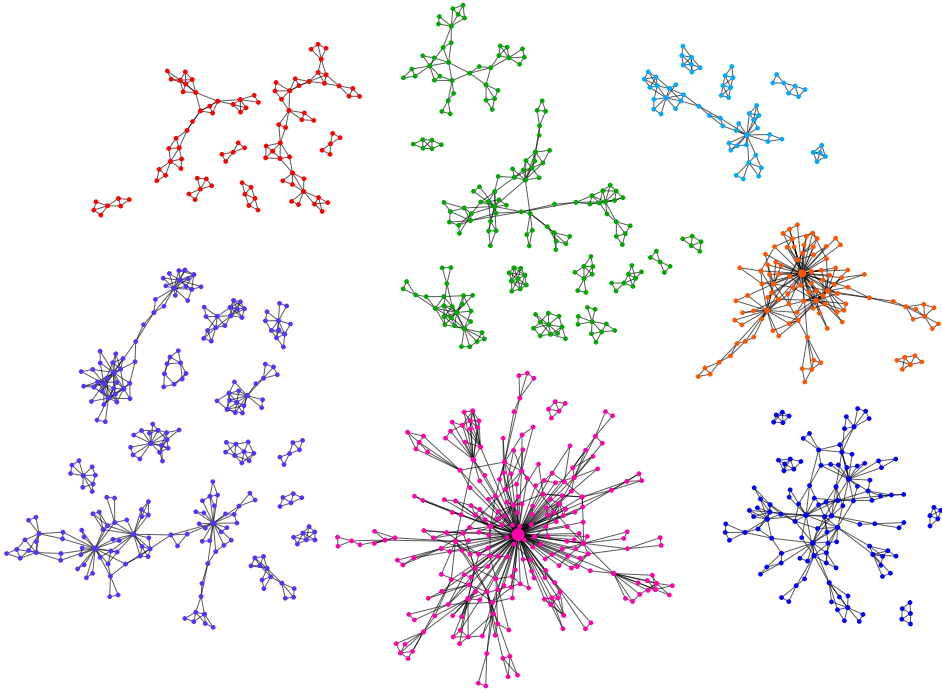


Fig. 13. A common prediction task in the cora citation network is to predict the research area (type) of a paper (node). We visualize the network resulting from the edges with nonzero typed triangle counts and find something striking. Typed triangles shatter the graph into many different components that are tightly connected and overwhelmingly homogeneous with respect to research area (type/label) of the nodes. This can be used to filter noisy links from the graph to improve classification performance. Node color encodes the research area of the papers.

consists of 5429 links. Using the proposed heterogeneous graphlets, we find 129 typed 3-node graphlets among the 168 possible typed 3-node graphlets that could occur. Notably, we observe the most frequent typed triangle graphlets are of a single type. Indeed, the first 7 typed triangle graphlets with largest frequency in Figure 12 are of a single type and account for 83.86% of all typed triangle graphlets. This finding indicates *strong homophily* among nodes with similar types (Figure 13). Unlike untyped graphlets, typed graphlets simultaneously capture the labeling and structural properties that lie at the heart of homophily [La Fond and Neville 2010; McPherson et al. 2001]. Therefore, typed graphlets provide a principled foundation for studying homophily in social networks. In Figure 12, we observe a large gap that clearly separates the 7 single-typed triangle graphlets from the other typed triangle graphlets with heterogeneous types. Furthermore, only 49 out of the 84 possible typed triangle graphlets (Table 2) actually occur in  $G$ .

In Figure 14, we also investigate a variety of typed 4-node graphlet distributions (from most dense to least dense). Strikingly, only 19 of the 210 possible typed 4-clique graphlets actually occur in  $G$  when node types are randomly shuffled. In the case of 4-node typed cycles, we observe 66 of the actual 210 possible typed 4-cycle graphlets appear when node types are randomly shuffled.

**8.7.3 Citeseer citation network.** The citeseer citation network consists of papers with citation links between them. Each paper is associated with one of six types representing the paper topic (e.g., ML). The 5 most frequently occurring typed triangle graphlets are those with a single type

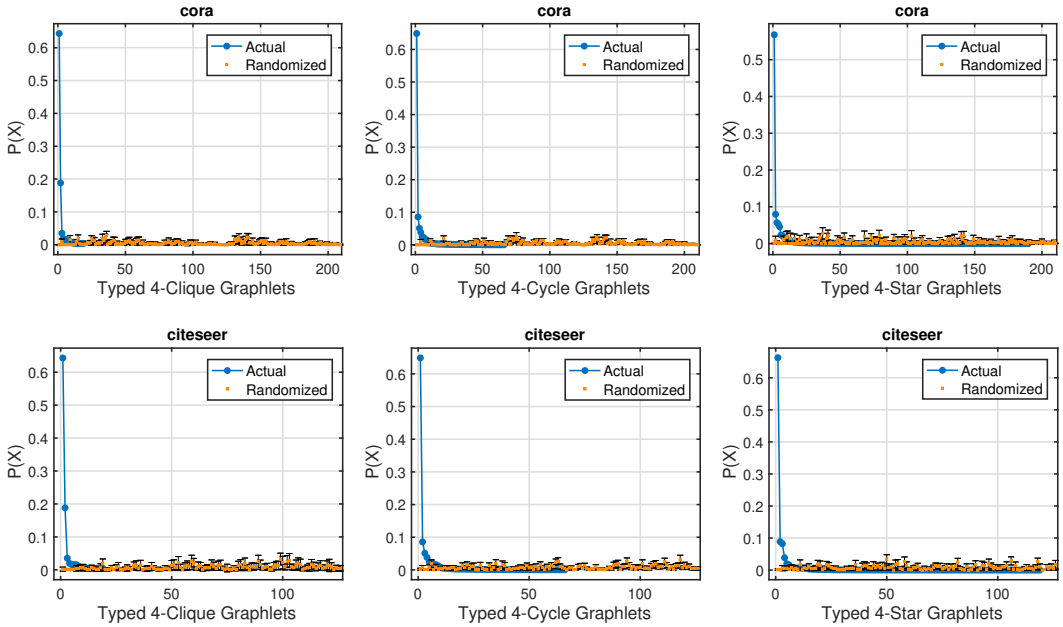


Fig. 14. Comparing the actual *typed* 4-clique, 4-cycle and 4-star graphlet distributions to the randomized *typed* distributions. We compute 100 random permutations of the node types and run the approach on each permutation then average the resulting counts to obtain the mean randomized *typed* graphlet distribution. There are three key findings. First, we observe a significant difference between the actual and randomized *typed* graphlet distributions. Second, many of the *typed* graphlets that occur when the types are randomized, do not occur in the actual *typed* graphlet distribution. Third, we find the *typed* graphlet distribution to be extremely skewed as a few *typed* graphlets occur very frequently while the vast majority have very few occurrences (and many *typed* graphlets are even forbidden, in the sense that they do not occur at all in the graph).

(Figure 12). Overall, these *typed* triangle graphlets account for 77.02% of all *typed* triangle graphlets that occur in  $G$ . This finding indicates *strong homophily* among nodes. Furthermore, since there are six types corresponding to paper topics, there are 56 potential *typed* triangle graphlets. However, among the 56 possible *typed* triangle graphlets of six types, only 40 actually appear in  $G$  as shown in Figure 12. The others are forbidden *typed* triangle graphlets. In addition, Figure 14 compares the actual *typed* 4-clique, 4-cycle, and 4-star distributions to the randomized distributions. Notably, while 126 different *typed* 4-cliques appear when node types are randomly shuffled, only 19 distinct *typed* 4-cliques (with different type configurations) actually appear in  $G$ . This finding is consistent with the cora citation network discussed in Section 8.7.2.

## 8.8 Use Case: Link Prediction

This section quantitatively demonstrates the effectiveness of *typed* graphlets for link prediction. Given a partially observed graph  $G$ , the link prediction task is to predict the missing edges. This general problem has applications in recommendation systems, *e.g.*, recommending movies to users (movielens) or suggesting potential friends (yahoo), among other important applications.

**8.8.1 Higher-Order Typed Graphlet Embedding.** Algorithm 5 summarizes the method for deriving *higher-order typed graphlet node embeddings*. In particular, given a *typed*-graphlet  $H$  of interest,

**Algorithm 5** Higher-Order *Typed Graphlet* Node Embeddings**Input:** a graph  $G$ , typed graphlet  $H$ , embedding dimension  $D$ **Output:** Higher-order node embedding matrix  $\mathbf{Z} \in \mathbb{R}^{N \times D}$  for  $H$ 

- 1  $(\mathbf{W}_{G^H})_{ij} \leftarrow \# \text{ instances of } H \text{ containing } i \text{ and } j, \forall (i, j) \in E$
- 2  $\mathbf{D}_{G^H} \leftarrow \text{typed-graphlet degree matrix } (\mathbf{D}_{G^H})_{ii} = \sum_j (\mathbf{W}_{G^H})_{ij}$
- 3  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_D \leftarrow \text{eigenvectors of } D \text{ smallest eigenvalues of } \mathbf{L}_{G^H} = \mathbf{I} - \mathbf{D}_{G^H}^{-1/2} \mathbf{W}_{G^H} \mathbf{D}_{G^H}^{-1/2}$
- 4  $Z_{ij} \leftarrow U_{ij} / \sqrt{\sum_{j=1}^D U_{ij}^2}$
- 5 **return**  $\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \dots \ \mathbf{z}_n]^T \in \mathbb{R}^{N \times D}$

Table 12. Link prediction edge types and semantics. The edge type predicted by the models is bold.

Graph	$ \mathcal{T}_V $	$ \mathcal{T}_E $	Heterogeneous Edge Types
movielens	3	3	<b>user-by-movie</b> , user-by-tag, tag-by-movie
dbpedia	4	3	<b>person-by-work</b> (produced work), person-has-occupation, work-by-genre (work-associated-genre)
yahoo-msg	2	2	<b>user-by-user</b> (communicated with), user-by-location (communication location)

Algorithm 5 outputs a matrix  $\mathbf{Z}$  of node embeddings. For graphs with many connected components, Algorithm 5 is called for each connected component of the typed graphlet graph  $G^H$  and the resulting embeddings are stored in the appropriate locations in the overall embedding matrix  $\mathbf{Z}$ .

**8.8.2 Experimental Setup.** We evaluate the higher-order typed graphlet node embedding approach that explicitly leverages typed graphlets (Algorithm 5) against the following methods: DeepWalk (DW) [Perozzi et al. 2014], LINE [Tang et al. 2015], GraRep [Cao et al. 2015], spectral embedding (untyped edge graphlet) [Long et al. 2006], and spectral embedding using untyped-graphlets. All methods output ( $D=128$ )-dimensional node embeddings. For DeepWalk (DW) [Perozzi et al. 2014], we perform 10 random walks per node of length 80 as mentioned in [Grover and Leskovec 2016]. For LINE [Tang et al. 2015], we use 2nd-order proximity and perform 60 million samples. For GraRep (GR) [Cao et al. 2015], we use ( $k=2$ )-steps. In contrast, the spectral embedding methods do not have any hyperparameters besides  $D$  which is fixed for all methods.

**8.8.3 Results.** We generate a labeled dataset of positive and negative edges. Positive edge examples are obtained by removing 50% of edges uniformly at random, whereas negative examples are generated by randomly sampling an equal number of node pairs  $(i, j) \notin E$ . For each method, we learn node embeddings using the remaining graph. Given embedding vectors  $\mathbf{z}_i$  and  $\mathbf{z}_j$  for node  $i$  and  $j$ , we derive a  $D$ -dimensional edge embedding vector  $\mathbf{z}_{ij} = \sigma(\mathbf{z}_i, \mathbf{z}_j)$  where  $\sigma$  is defined as one of the following *edge embedding functions*:

$$\sigma \in \left\{ \frac{\mathbf{z}_i + \mathbf{z}_j}{2}, \mathbf{z}_i \odot \mathbf{z}_j, |\mathbf{z}_i - \mathbf{z}_j|, (\mathbf{z}_i - \mathbf{z}_j)^{\odot 2}, \max(\mathbf{z}_i, \mathbf{z}_j), \mathbf{z}_i + \mathbf{z}_j \right\}$$

Note  $\mathbf{z}_i \odot \mathbf{z}_j$  is the element-wise product,  $\mathbf{z}^{\circ 2}$  is the Hadamard power, and  $\max(\mathbf{z}_i, \mathbf{z}_j)$  is the element-wise max. Using the edge embeddings, we then learn a logistic regression model to predict if an edge in the test set exists in  $E$  or not. Experiments are repeated for 10 random seed initializations and the average performance is reported. All methods are evaluated against four different evaluation metrics including  $F_1$ , Precision, Recall, and AUC.

Table 12 summarizes the heterogeneous network data and the type/label of the edge predicted by the models. The results are provided in Table 13. We report the best result among the different edge embedding functions and untyped/typed graphlets. In Table 13, the typed graphlet approach is shown to outperform all other methods across *all* four evaluation metrics. In all cases, the approach that leverages typed graphlets outperforms the other methods (Table 13) with an overall mean gain (improvement) in  $F_1$  of 18.7% (and up to 48.4% improvement) across all graph data. In terms of AUC, the typed graphlet approach achieves a mean gain of 14.4% (and up to 45.7% improvement) over all methods. Furthermore, we posit that an approach similar to the one proposed in [Rossi et al. 2018] could be used to achieve even better predictive performance by leveraging multiple typed graphlets simultaneously.

## 9 CONCLUSION

In this work, we introduced the notion of typed graphlet that generalizes the notion of graphlet to heterogeneous networks. We proposed a fast, parallel, and space-efficient framework for counting typed graphlets. The proposed typed graphlet algorithms count only a few typed graphlets and derives the others in  $o(1)$  constant time using new non-trivial combinatorial relationships that involve counts of lower-order typed graphlets. Thus, the proposed approach avoids explicit enumeration of any nodes involved in those typed graphlets. For every edge, we count a few typed graphlets and obtain the exact counts of the remaining ones in  $o(1)$  constant time. Theoretically, the worst-case time complexity of the proposed approach is shown to match the best untyped graphlet algorithm. Since this is the first investigation into typed graphlets, there are no existing methods for comparison. However, we compared our approach to colored graphlet counting methods that solve a strictly simpler problem. Empirically, our approach is shown to outperform the state-of-the-art

Table 13. Link prediction results. These results demonstrate the effectiveness of typed graphlets for prediction.

		DeepWalk	LINE	GraRep	Spectral	Untyped Graphlets	Typed Graphlets
movielens	<b>F<sub>1</sub></b>	0.8544	0.8638	0.8550	0.8774	0.8728	<b>0.9409</b>
	<b>Prec.</b>	0.9136	0.8785	0.9235	0.9409	0.9454	<b>0.9747</b>
	<b>Recall</b>	0.7844	0.8444	0.7760	0.8066	0.7930	<b>0.9055</b>
	<b>AUC</b>	0.9406	0.9313	0.9310	0.9515	0.9564	<b>0.9900</b>
dbpedia	<b>F<sub>1</sub></b>	0.8414	0.7242	0.7136	0.8366	0.8768	<b>0.9640</b>
	<b>Prec.</b>	0.8215	0.7754	0.7060	0.7703	0.8209	<b>0.9555</b>
	<b>Recall</b>	0.8726	0.6375	0.7323	0.9669	0.9665	<b>0.9733</b>
	<b>AUC</b>	0.8852	0.8122	0.7375	0.9222	0.9414	<b>0.9894</b>
yahoo	<b>F<sub>1</sub></b>	0.6927	0.6269	0.6949	0.9140	0.8410	<b>0.9303</b>
	<b>Prec.</b>	0.7391	0.6360	0.7263	0.9346	0.8226	<b>0.9432</b>
	<b>Recall</b>	0.5956	0.5933	0.6300	0.8904	0.8699	<b>0.9158</b>
	<b>AUC</b>	0.7715	0.6745	0.7551	0.9709	0.9272	<b>0.9827</b>



in terms of runtime, space-efficiency, and scalability as it is able to handle large networks. While these methods take hours on small graphs with thousands of edges, our typed graphlet counting approach takes only seconds on networks with millions of edges. Finally, the proposed approach is able to handle *large* general heterogeneous networks while lending itself to an efficient and highly scalable parallel implementation. The proposed approach gives rise to new opportunities and applications for typed graphlets. Future work should use the ideas introduced in this paper to extend and derive equations for typed graphlets of 5 nodes and larger. This is similar to how recent work [Dave et al. 2017; Pinar et al. 2017] extended the ideas introduced by Ahmed et al. [2015, 2016] to 5-node untyped graphlets.

## REFERENCES

- Evrin Acar, Tamara G Kolda, and Daniel M Dunlavy. 2011. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv:1105.3422* (2011).
- Nesreen K. Ahmed, Nick Duffield, Theodore L. Willke, and Ryan A. Rossi. 2017a. On Sampling from Massive Graph Streams. In *VLDB*. 1430–1441.
- Nesreen K. Ahmed, Jennifer Neville, Ryan A. Rossi, and Nick Duffield. 2015. Efficient Graphlet Counting for Large Networks. In *ICDM*. 10.
- Nesreen K. Ahmed, Jennifer Neville, Ryan A. Rossi, Nick Duffield, and Theodore L. Willke. 2016. Graphlet Decomposition: Framework, Algorithms, and Applications. *KAIS* (2016), 1–32.
- Nesreen K. Ahmed, Ryan A. Rossi, Theodore L. Willke, and Rong Zhou. 2017b. A Higher-order Latent Space Network Model. In *Proceedings of the AAAI PAIR (Plan, Activity, and Intent Recognition) Workshop*. 1–7.
- Nesreen K. Ahmed, Ryan A. Rossi, Rong Zhou, John Boaz Lee, Xiangnan Kong, Theodore L. Willke, and Hoda Eldardiry. 2018. Learning Role-based Graph Embeddings. In *StarAI IJCAI*.
- Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *DMKD* 29, 3 (2015), 626–688.
- Arindam Banerjee, Sugato Basu, and Srujana Merugu. 2007. Multi-way clustering on relation graphs. In *SDM*. SIAM, 145–156.
- Danielle Smith Bassett and ED Bullmore. 2006. Small-world brain networks. *The neuroscientist* 12, 6 (2006), 512–523.
- Austin R Benson, David F Gleich, and Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* 353, 6295 (2016), 163–166.
- Ed Bullmore and Olaf Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10, 3 (2009), 186–198.
- Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. GraRep: Learning graph representations with global structural information. In *CIKM*. ACM, 891–900.
- Aldo G. Carranza, Ryan A. Rossi, Anup Rao, and Eunye Koh. 2018. Higher-order Spectral Clustering for Heterogeneous Graphs. In *arXiv:1810.02959*. 15.
- Lianhua Chi and Xingquan Zhu. 2017. Hashing techniques: A survey and taxonomy. *ACM Computing Surveys (CSUR)* 50, 1 (2017), 11.
- F. Chung and L. Lu. 2002. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics* 6, 2 (2002), 125–145.
- Joseph Crawford and Tijana Milenković. 2015. GREAT: GRaphlet Edge-based network AlignmenT. In *BIBM*. 220–227.
- Vachik S Dave, Nesreen K Ahmed, and Mohammad Al Hasan. 2017. E-CLoG: counting edge-centric local graphlets. In *2017 IEEE International Conference on Big Data (Big Data)*. 586–595.
- N. Eagle and A. Pentland. 2006. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10, 4 (2006), 255–268.
- Paul Erdős and A Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci* 5 (1960), 17–61.
- M. Faloutsos, P. Faloutsos, and C. Faloutsos. 1999. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. ACM, 262.
- M Girvan and M E J Newman. 2002. Community structure in social and biological networks. *PNAS* 99, 12 (2002), 7821–7826.
- Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*. 855–864.
- Shawn Gu, John Johnson, Fazle E Faisal, and Tijana Milenković. 2018. From homogeneous to heterogeneous network alignment via colored graphlets. *Scientific reports* 8, 1 (2018), 12524.
- Wayne Hayes, Kai Sun, and Nataša Pržulj. 2013. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics* 29, 4 (2013), 483–491.

- Yuriy Hulovatyy, Huili Chen, and T Milenković. 2015. Exploring the structure and function of temporal networks with dynamic graphlets. *Bioinformatics* 31, 12 (2015), i171–i180.
- Chia-Chen Hung, Hope Chan, and Eric Hsiao-Kuang Wu. 2008. Mobility pattern aware routing for heterogeneous vehicular networks. In *WCNC*. 2200–2205.
- Xiangnan Kong, Jiawei Zhang, and Philip S Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*. 179–188.
- Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. 2011. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment* 2011, 11 (2011), P11005.
- Mehmet Koyutürk, Yohan Kim, Umut Topkara, Shankar Subramaniam, Wojciech Szpankowski, and Ananth Grama. 2006. Pairwise alignment of protein interaction networks. *JCB* 13, 2 (2006), 182–199.
- Timothy La Fond and Jennifer Neville. 2010. Randomization Tests for Distinguishing Social Influence and Homophily Effects. In *WWW*. 601–610.
- Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. 2010. Kronecker graphs: An approach to modeling networks. *JMLR* 11, Feb (2010), 985–1042.
- Ryan N Lichtenwalter and Nitesh V Chawla. 2012. Vertex Collocation Profiles: Subgraph Counting for Link Analysis and Prediction. In *WWW*. 1019–1028.
- Bo Long, Zhongfei (mark Zhang, Xiaoyun Wu, and Philip S. Yu. 2006. Spectral clustering for multi-type relational data. In *ICML*.
- Dror Marcus and Yuval Shavitt. 2012. RAGE—a rapid graphlet enumerator for large networks. *Computer Networks* 56, 2 (2012), 810–819.
- Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Ann. Rev. of Soc.* 27, 1 (2001).
- Tijana Milenković and Nataša Pržulj. 2008. Uncovering Biological Network Function via Graphlet Degree Signatures. *Cancer Informatics* 6 (2008), 257.
- R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. 2002. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 5594 (2002), 824–827.
- C.C. Noble and D.J. Cook. 2003. Graph-based anomaly detection. In *SIGKDD*. 631–636.
- G. Palla, I. Derényi, I. Farkas, and T. Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 7043 (2005), 814–818.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*. 701–710.
- Ali Pinar, C. Seshadhri, and Vaidyanathan Vishal. 2017. ESCAPE: Efficiently Counting All 5-Vertex Subgraphs. In *WWW*. 1431–1440.
- Nataša Pržulj. 2007. Biological network comparison using graphlet degree distribution. *Bioinfo.* 23, 2 (2007), e177–e183.
- N Pržulj, Derek G Corneil, and Igor Jurisica. 2004. Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 18 (2004), 3508–3515.
- F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. 2004. Defining and identifying communities in networks. *PNAS* 101, 9 (2004), 2658–2663.
- Pedro Ribeiro and Fernando Silva. 2014. Discovering colored network motifs. In *Complex Networks V*. Springer, 107–118.
- Ryan A. Rossi and Nesreen K. Ahmed. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *AAAI*. 4292–4293. <http://networkrepository.com>
- Ryan A. Rossi, Nesreen K. Ahmed, Eunye Koh, Sungchul Kim, Anup Rao, and Yasin Abbasi-Yadkori. 2018. HONE: Higher-Order Network Embeddings. *arXiv:1801.09303* (2018).
- Ryan A. Rossi, Sonia Fahmy, and Nilothpal Talukder. 2013. A Multi-Level Approach for Evaluating Internet Topology Generators. In *Proceedings of IFIP Networking*. 1–9.
- Ryan A. Rossi and Rong Zhou. 2016. Parallel Collective Factorization for Modeling Large Heterogeneous Networks. In *SNAM*. 30.
- Ryan A. Rossi and Rong Zhou. 2018. GraphZIP: A Clique-based Sparse Graph Compression Method. *Journal of Big Data* 5, 1 (2018), 14.
- Nino Shervashidze, Tobias Petri, Kurt Mehlhorn, Karsten M Borgwardt, and Svn Vishwanathan. 2009. Efficient graphlet kernels for large graph comparison. In *AISTATS*.
- Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. 2011. Weisfeiler-lehman graph kernels. *JMLR* 12, Sep (2011), 2539–2561.
- Ryan W Solava, Ryan P Michaels, and T Milenković. 2012. Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics* 28, 18 (2012), i480–i486.
- Richard P Stanley. 1986. *What Is Enumerative Combinatorics?* Springer.

- Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB* 4, 11 (2011), 992–1003.
- Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *WWW*. 1067–1077.
- S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. 2010. Graph kernels. *JMLR* 11 (2010), 1201–1242.
- Jian-Wei Wang and Li-Li Rong. 2009. Cascade-based attack vulnerability on the US power grid. *Safety Science* 47, 10 (2009), 1332–1336.
- D.J. Watts and S.H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature* 393, 6684 (1998), 440–442.
- Sebastian Wernicke and Florian Rasche. 2006. FANMOD: a tool for fast network motif detection. *Bioinformatics* 22, 9 (2006), 1152–1153.
- Zhijun Yin, Rui Li, Qiaozhu Mei, and Jiawei Han. 2009. Exploring social tagging graph for web object classification. In *SIGKDD*. ACM, 957–966.
- 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*. 283–292.
- L. Zhang, R. Hong, Y. Gao, R. Ji, Q. Dai, and X. Li. 2016. Image Categorization by Learning a Propagated Graphlet Path. *TNNLS* 27, 3 (2016), 674–685.
- Luming Zhang, Mingli Song, Zicheng Liu, Xiao Liu, Jiajun Bu, and Chun Chen. 2013. Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. In *CVPR*.