Fast and Accurate Estimation of Typed Graphlets

Ryan A. Rossi
Adobe Research

Anup Rao
Adobe Research

Tung Mai
Adobe Research

Nesreen K. Ahmed
Intel Labs

ABSTRACT

Typed graphlets are small typed (labeled, colored) induced subgraphs and were recently shown to be the fundamental building blocks of rich complex heterogeneous networks. In many applications, speed is more important than accuracy, and it is sufficient to trade-off a tiny amount of accuracy for a significantly faster method. In this work, we propose fast and accurate estimators for typed graphlets. The typed graphlet estimation techniques naturally support general heterogeneous graphs with any arbitrary number of types, which include bipartite, k-partite, k-star, labeled graphs, and attributed networks as special cases. The experiments demonstrate the effectiveness of the typed graphlet estimation techniques.

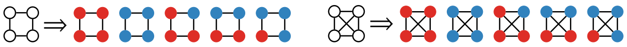
1 ESTIMATION FRAMEWORK

Graphlets have recently been generalized to labeled and heterogeneous networks [3]. Since then they have found many important applications including clustering [1], exploratory analysis [3], and link prediction [1]. However, depending on the application constraints, speed may be more important than accuracy. To address this problem, we propose two general classes of estimation methods for typed graphlets. First we introduce a class of methods that sample edges and the graphlets centered at those edges (Sec. 1.1). Then we propose a class of estimation methods based on sampling typed paths (Sec. 1.2).

Problem and preliminaries. 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. We denote the number of occurrences of the i-th typed *induced* subgraph with types $\mathbf{t}$ as $C_{i,\mathbf{t}}$. Further, let $N_{i,\mathbf{t}}$ denote the count of the i-th typed *non-induced* subgraph with types $\mathbf{t}$. We use the following linear relationship between induced and non-induced typed subgraph counts:

$$\underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 & 2 & 4 \\ 0 & 1 & 2 & 4 & 6 & 12 \\ 0 & 0 & 1 & 0 & 4 & 12 \\ 0 & 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} C_{1,\mathbf{t}} \\ C_{2,\mathbf{t}} \\ C_{3,\mathbf{t}} \\ C_{4,\mathbf{t}} \\ C_{5,\mathbf{t}} \\ C_{6,\mathbf{t}} \end{pmatrix} = \begin{pmatrix} N_{1,\mathbf{t}} \\ N_{2,\mathbf{t}} \\ N_{3,\mathbf{t}} \\ N_{4,\mathbf{t}} \\ N_{5,\mathbf{t}} \\ N_{6,\mathbf{t}} \end{pmatrix} \tag{1}$$

where A_{ij} is the # of distinct copies of the i-th typed subgraph in the j-th subgraph.



(a) Typed 4-cycles with 2 types

(b) Typed 4-cliques with 2 types

Figure 1: Examples of *typed graphlets*

1.1 Typed Edge Sampling & Estimation

Since there are no existing methods for estimating typed graphlets, we begin by introducing a simple class of estimation methods for typed graphlets based on edge sampling. Let $J \subseteq E$ be a subset of edges sampled via a uniform or weighted distribution $\mathbb{F}$. Further, let $\mathbf{X}_H = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots \end{bmatrix} \in \mathbb{R}^{|J| \times |\mathcal{H}|}$ denote the typed graphlet counts that occur at each sampled edge in J for a specific induced subgraph H (*e.g.*, 4-clique, 4-cycle) and $\mathcal{H}$ is the set of typed graphlets of H. Given $\mathbf{X}_H$ we derive unbiased estimates for the typed graphlet counts as

$$\widehat{\mathbf{x}}_H = \left(\frac{|J|}{|E|} \right)^{-1} \frac{\mathbf{e}^T \mathbf{X}_H}{|E(H)|} \tag{2}$$

where $|E(H)|$ is the number of edges in the graphlet H and $\mathbf{e} \in \mathbb{R}^{|J|}$ is a vector of 1's. These methods essentially sample an edge via $\mathbb{F}$ and all the neighbors of that edge. Afterwards, we compute all typed graphlets that occur between them, and use these counts to obtain unbiased estimates of the overall global counts of all typed graphlets in G. For comparison with the other class of estimation methods proposed in Section 1.2 that samples typed graphlets directly (as opposed to edges), we sample edges until the number of typed graphlets found from the edges and their neighbors match the number of graphlets sampled by the other methods.

1.2 Typed Path Sampling & Estimation

Following the idea of path sampling from [2], we describe a general class of estimation methods based on sampling typed paths. Let Γ_i^t denote the set of nodes adjacent to i of type t and $d_i^t = |\Gamma_i^t|$.

Definition 1 (Typed Wedges). *Given an edge $(i,j) \in E$ with types ϕ_i and ϕ_j, the (i,j)-entry of the typed wedge matrix with types t and t' is:*

$$\Lambda_e^{tt'} = \Lambda_{ij}^{tt'} = \begin{cases} \left(d_i^t - 1\right)\left(d_j^{t'} - 1\right) & \textit{if } t = \phi_j \wedge t' = \phi_i \\ \left(d_i^t - 1\right) d_j^{t'} & \textit{if } t = \phi_j \wedge t' \neq \phi_i \\ d_i^t \left(d_j^{t'} - 1\right) & \textit{if } t \neq \phi_j \wedge t' = \phi_i \\ d_i^t d_j^{t'} & \textit{otherwise} \end{cases} \tag{3}$$

Note that $(d_i^t - 1)$ excludes the neighbor j if $\phi_j = t$ (same type).

Claim 1. *Let $W = \sum_{(i,j) \in E} (d_i - 1)(d_j - 1)$ and $W^{tt'} = \sum_{(i,j) \in E} \Lambda_{ij}^{tt'}$, then $W = \sum_t \sum_{t'} W^{tt'}$*

This is straightforward to see and implies the total wedges W in G is equal to the sum of typed wedges $W^{tt'}$ for all $t, t' \in \{1, \ldots, L\}$.

Claim 2. *Fix any typed 4-path with types (t, t_i, t_j, t'), the probability that Algorithm 1 outputs this typed 4-path is exactly $1/W^{tt'}$.*

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382683

Algorithm 1 Typed Path Sample

```
Output: the four sampled nodes $(i', i, j, j')$ with types $(t_1, t, t', t_2)$ that form a typed 4-path with edges $\{(i', i), (i, j), (j, j')\}$
Compute $\Lambda_e^{tt'}$ (Eq. 3) for all typed edges and set $p_e^{tt'} = \Lambda_e^{tt'} / W^{tt'}$
Select $e = (i, j)$ of type $t_e = (t, t')$ with probability $p_e^{tt'}$
Select $i' \in \Gamma_i^{t_1}$ with type $t_1$ uniformly at random s.t. $i' \neq j$ if $\phi_j = t_1$.
Select $j' \in \Gamma_j^{t_2}$ with type $t_2$ uniformly at random s.t. $j' \neq i$ if $\phi_i = t_2$.
```

Algorithm 2 Estimation via Typed Paths

```
Input: graph $G$, # samples $k$
Output: estimated counts for all typed 4-node graphlets
Obtain $k$ samples (sets of vertices) by running Alg. 1 $k$ times where $S_j$ denotes the $j$-th set of vertices.
parallel for $j = 1, \dots, k$ do
    Determine subgraph induced by $S_j$ (and type vector $\mathbf{t}$)
    If this is the $i$-th graphlet with types $\mathbf{t}$, increment $F_{i,\mathbf{t}}^{tt'}$
    Increment $k^{tt'}$ where $t, t'$ are the other two node types
for $i \in [2, 6]$ and type vector $\mathbf{t}$ do
    for all $t, t' \in \{1, \dots, L\}$ do Set $\widehat{C}_{i,\mathbf{t}} = \widehat{C}_{i,\mathbf{t}} + (F_{i,\mathbf{t}}^{tt'} / k^{tt'}) \cdot \frac{W^{tt'}}{A_{2,i}}$
    Set $\widehat{C}_{i,\mathbf{t}} = \widehat{C}_{i,\mathbf{t}}/2$
Set $\widehat{C}_{1,\mathbf{t}} = N_{1,\mathbf{t}} - \widehat{C}_{3,\mathbf{t}} - 2\widehat{C}_{5,\mathbf{t}} - 4\widehat{C}_{6,\mathbf{t}}$, $\forall \mathbf{t}$ s.t. $N_{1,\mathbf{t}}$ is computed via Eq. 4
```

Proof. There are 4 cases. If t is the type of node j ($t = \phi_j$) and t' is the type of node i ($t' = \phi_i$), then the typed 4-path is selected with probability $(d_i^t - 1)(d_j^{t'} - 1)/W^{tt'} \cdot 1/(d_i^t - 1) \cdot 1/(d_j^{t'} - 1) = \frac{1}{W^{tt'}}$. If t is the type of node j ($t = \phi_j$), but t' is not the type of node i ($t' \neq \phi_i$), then the typed 4-path is selected with probability $(d_i^t - 1) d_j^{t'} / W^{tt'} \cdot 1/(d_i^t - 1) \cdot 1/d_j^{t'} = 1/W^{tt'}$. If t is not the type of node j ($t \neq \phi_j$) and t' is the type of node i ($t' = \phi_i$), then the typed 4-path is selected with probability $d_i^t (d_j^{t'} - 1)/W^{tt'} \cdot 1/d_i^t \cdot 1/(d_j^{t'} - 1) = 1/W^{tt'}$. If t is not the type of node j ($t \neq \phi_j$) and t' is *not* the type of node i ($t' \neq \phi_i$), then the typed 4-path is selected with probability $d_i^t d_j^{t'} / W^{tt'} \cdot 1/d_i^t \cdot 1/d_j^{t'} = 1/W^{tt'}$. ■

The approach is summarized in Alg. 2. In particular, a set of 4 nodes representing a non-induced typed path are sampled via Alg. 1. Next, we obtain the typed graphlet induced by this set of nodes and repeat the above k times to estimate the actual global counts as shown in Alg. 2. To obtain the count of typed 4-stars with type vector $\mathbf{t}$, we first derive

$$N_{1,\mathbf{t}} = \sum_{i \in V^t} d_i^{t_1} \cdot (d_i^{t_2} - 1) \cdot (d_i^{t_3} - 2)/6 \tag{4}$$

where $\mathbf{t} = \begin{bmatrix} t_1 & t_2 & t_3 & t \end{bmatrix}$. Note Eq. 4 is for when types are the same. Due to space, we omit the other cases as they are just as straightforward. Then we use this quantity to derive the typed 4-node star count in $o(1)$ time via $\widehat{C}_{1,\mathbf{t}} = N_{1,\mathbf{t}} - \widehat{C}_{3,\mathbf{t}} - 2\widehat{C}_{5,\mathbf{t}} - 4\widehat{C}_{6,\mathbf{t}}$.

2 EXPERIMENTS

Since this work is the first to propose estimators for typed graphlets, we compare the typed path sampling (TPS) method (Sec. 1.2) to the simpler typed edge sampling (TES) method (Sec. 1.1).

Accuracy. For comparison, we report the mean relative error in Table 1 defined as $\frac{1}{|\mathcal{H}|} \sum_{H \in \mathcal{H}} |\widehat{C}_H - C_H| / C_H$ where $\mathcal{H}$ is the set of typed graphlets for an induced subgraph (*e.g.*, 4-clique), C_H is the exact count of H and $\widehat{C}_H$ is the estimated count. Overall, TPS significantly outperforms TES across the different typed graphlets as shown in Table 1. Due to space, typed 4-star results were omitted.

Table 1: Mean relative error of typed graphlet estimates. We set $k = 50000$ and perform 100 runs.

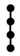
Data	Methods					
fb-political	TES	0.012	0.033	0.036	0.036	0.070
	TPS	**0.002**	**0.021**	**0.010**	**0.024**	**0.034**
yahoo-msg	TES	0.774	0.867	1.233	2.784	0.430
	TPS	**0.001**	**0.046**	**0.003**	**0.270**	**0.025**
web-polblogs	TES	0.010	0.083	0.011	0.100	0.164
	TPS	**0.002**	**0.006**	**0.005**	**0.007**	**0.008**
soc-wiki-elec	TES	0.684	1.160	0.788	1.353	1.552
	TPS	**0.002**	**0.003**	**0.003**	**0.005**	**0.008**
soc-digg	TES	0.460	0.412	0.890	0.713	1.474
	TPS	$<10^{-3}$	**0.004**	**0.003**	**0.006**	**0.011**

Table 2: Typed graphlet estimates and relative error using $k = 50000$ (typed 4-cliques).

			$\widehat{C}$		$\frac{\lvert C - \widehat{C} \rvert}{C}$		Std	
graph	types	C	TES	TPS	TES	TPS	TES	TPS
fb-political	1111	8.14K	7.35K	8.37K	0.0973	0.0288	5K	507
	2111	7.64K	8.13K	7.86K	0.0634	0.0285	3.4K	545
	2211	6.27K	6.85K	6.50K	0.0924	0.0371	2.6K	448
	2221	6.12K	6.55K	6.29K	0.0701	0.0281	2.3K	455
	2222	4.46K	4.59K	4.68K	0.0274	0.0483	2.4K	462

Variance of estimates. In Table 2, we observe the standard deviation/variance of TPS to be about an order of magnitude smaller than TES. Results for other graphs have been omitted due to space.

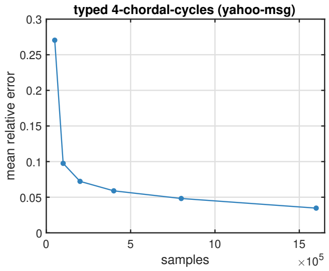


Figure 2: Convergence of estimates. Increasing the sample size k decreases error of TPS.

Convergence of estimates. To show convergence, we vary the number of samples $k \in \{50K, 100K, 200K, 400K, 800K, 1.6M\}$ and report the mean relative error for typed 4-chordal-cycles, since these are the most difficult to estimate for yahoo-msg (Table 1). In Fig. 2, the error decreases towards zero as the sample size increases.

Speedup. TPS and TES take less than a second for all graphs. This is hundreds of times faster than the exact algorithm.

REFERENCES

[1] Aldo G. Carranza, Ryan A. Rossi, Anup Rao, and Eunyee Koh. 2018. Higher-order Spectral Clustering for Heterogeneous Graphs. In *arXiv:1810.02959*. 15.

[2] Madhav Jha, C Seshadhri, and Ali Pinar. 2015. Path sampling: A fast and provable method for estimating 4-vertex subgraph counts. In *WWW*. 495–505.

[3] Ryan A. Rossi, Nesreen K. Ahmed, Aldo Carranza, David Arbour, Anup Rao, Sungchul Kim, and Eunyee Koh. 2019. Heterogeneous Network Motifs. In *arXiv:1901.10026*. 18.