Heterogeneous Graphlets

Ryan A. Rossi[1], Nesreen K. Ahmed[2], Aldo Carranza[1], David Arbour[1], Anup Rao[1], Sungchul Kim[1] and Eunyee Koh[1]

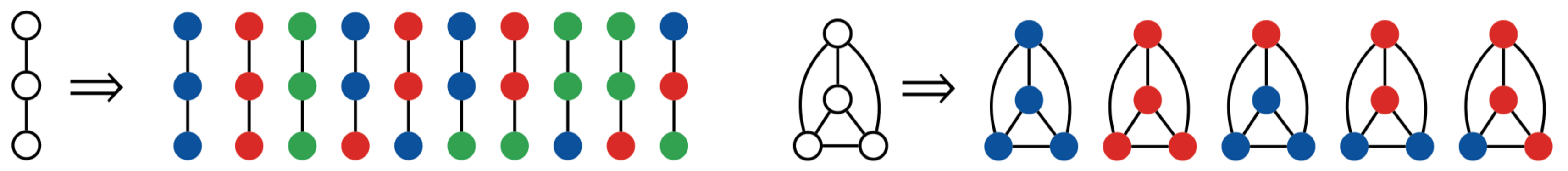
[1]Adobe Research, San Jose, CA
[2]Intel Labs, Santa Clara, CA

Overview

We generalize the notion of graphlets (network motifs) to heterogeneous networks by introducing the notion of a small induced typed subgraph called typed graphlet. 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 best known untyped algorithm. Unlike existing methods that take hours on small networks, the proposed approach takes only seconds on large networks with millions of edges. This gives rise to new opportunities and applications for typed graphlets on large real-world networks.

Heterogeneous Graphlets

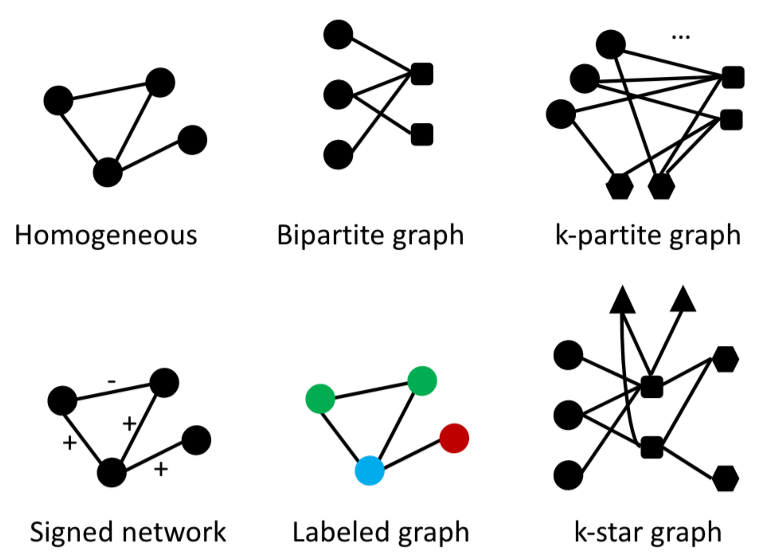
A **heterogeneous graphlet** of a graph G = (V, E, ϕ, ξ) is an induced heterogeneous subgraph H = (V', E', ϕ', ξ') of G such that (1) (V', E') is a graphlet of (V,E), (2) $\phi' = \phi|_{V'}$, that is, ϕ' is the restriction of ϕ to V' and (3) $\xi' = \xi|_{E'}$, that is, ξ' is the restriction of ξ to E'



(a) Typed 3-paths with $L = 3$ types (b) Typed 4-cliques with $L = 2$ types

Heterogeneous Graphlet Instance. An *instance* of a *heterogeneous graphlet* H = (V', E', ϕ', ξ') of graph G is a heterogeneous graphlet F = (V'', E'', ϕ'', ξ'') of G such that

1. (V'', E'') is isomorphic to (V', E')
2. $\mathcal{T}_{V''} = \mathcal{T}_{V'}$ and $\mathcal{T}_{E''} = \mathcal{T}_{E'}$ i.e., node & edge type multisets are correspondingly equal



Heterogeneous graphlets are useful for a variety of different classes of graphs

Graph Type	$\lvert\mathcal{T}_V\rvert$	$\lvert\mathcal{T}_E\rvert$
Homogeneous	1	1
Bipartite	2	1
K-partite	k	$k-1$
Signed	1	2
Labeled	k	ℓ
Star	k	$k-1$

For a single K node graphlet, the number of heterogeneous graphlets with L types/colors is:

$$\left(\!\!\binom{L}{K}\!\!\right) = \binom{L+K-1}{K}$$

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

Algorithm 1 Heterogeneous Graphlets

Input: a graph G
Output: nonzero typed graphlet counts $\mathcal{X}_{ij}$ for each edge $(i, j) \in E$

```
parallel for each (i, j) ∈ E do
    T_ij^t = Γ_i^t ∩ Γ_j^t,   for t = 1, ..., L          ▹ typed triangles
    S_i^t = Γ_i^t \ T_ij^t,   for t = 1, ..., L          ▹ typed 3-paths centered at i
    S_j^t = Γ_j^t \ T_ij^t,   for t = 1, ..., L          ▹ typed 3-paths centered at j
    S_ij^t = S_i^t ∪ S_j^t,   for t = 1, ..., L          ▹ typed 3-paths
    Store nonzero counts of the 3-node typed graphlets derived above
    Let T_ij = ∪_t T_ij^t,  S_i = ∪_t S_i^t,  and S_j = ∪_t S_j^t
    Given S_i and S_j, derive typed path-based motifs via Algorithm 2
    Given T_ij, derive typed triangle-based motifs via Algorithm 3
    for t, t' ∈ {1, ..., L} such that t ≤ t' do
        Derive remaining typed graphlet orbits in constant time via
        Eq. 13-16 and update counts x and set of motifs M_ij
    for c ∈ M_ij do X_ij = X_ij ∪ {(c, x_c)}   ▹ nonzero typed motif counts
end parallel
```

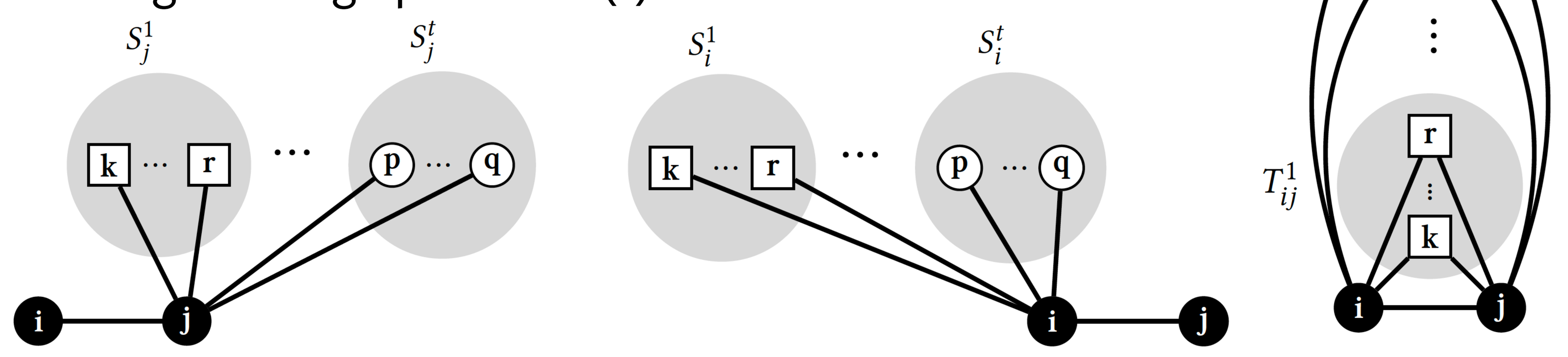
3-node typed graphlets (lines 2–6)

4-node typed graphlets (lines 8–11)

Worst-case for a single edge: $\mathcal{O}(\Delta(|S_i| + |S_j| + |T_{ij}|))$

Combinatorial Relationships

Intuition: use typed lower-order sets along with a few graphlet counts to derive the remaining higher-order heterogeneous graphlets in o(1) constant time



Example: Given an edge with types ϕ_i and ϕ_j, select type t and t', then directly compute counts using equations involving k-1 node typed graphlet counts:

Clique-based Graphlets (use k-1 node typed cliques)

Typed **chordal-cycle center** orbit count:

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

4-cliques with type vector **t**

Path-based Graphlets (use k-1 node typed paths)

Typed **4-path center** orbit count:

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

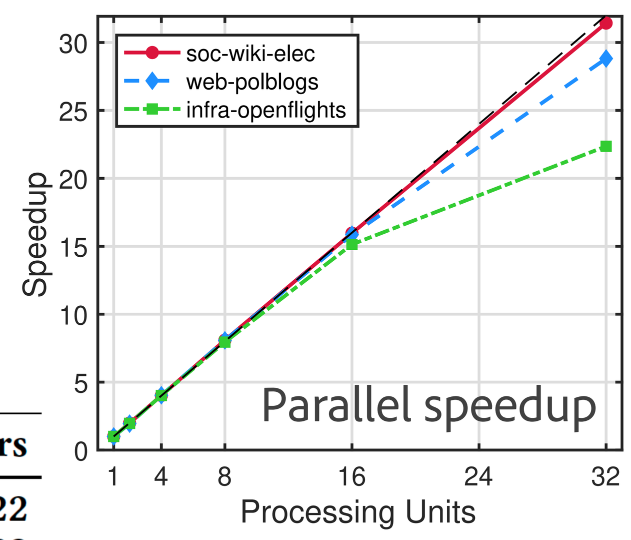
4-cycles with type vector **t**

Evaluation & Results

Approach uses between 42x and 776x less space

	citeseer	cora	movielens	web-spam
GC	30.1MB	50.4MB	ETL	ETL
ESU	13.4MB	46.2MB	ETL	ETL
G-Tries	161.9MB	448.6MB	ETL	ETL
Ours	**316KB**	**578KB**	**22.5MB**	**128.9MB**

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



	$\lvert E\rvert$	Δ	$\lvert\mathcal{T}_V\rvert$	$\lvert\mathcal{T}_E\rvert$	SECONDS GC	ESU	G-Tries	Ours
citeseer	4.5k	99	6	21	46.27	5937.75	144.08	**0.022**
cora	5.3k	168	7	28	467.20	10051.07	351.40	**0.032**
fb-relationship	44.9k	106	6	20	1374.60	54,837.69	3789.17	**0.701**
web-polblogs	16.7k	351	2	1	28,986.70	26,577.10	1,563.04	**1.055**
ca-DBLP	11.3k	69	3	3	149.20	1,188.11	18.90	**0.100**
inf-openflights	15.7k	242	2	2	9262.20	18,839.36	458.01	**0.578**
soc-wiki-elec	100.8k	1.1k	2	2	ETL	ETL	26,468.85	**5.316**
webkb	459	122	5	14	85.82	7,158.10	187.22	**0.006**
terrorRel	8.6k	36	2	3	192.6	3130.7	241.1	**0.039**
pol-retweet	48.1k	786	2	3	ETL	ETL	ETL	**0.296**
web-spam	465k	3.9k	3	6	ETL	ETL	ETL	**210.97**
movielens	170.4k	3.6k	3	3	ETL	ETL	ETL	**5.23**
citeulike	1.4M	11.2k	3	2	ETL	ETL	ETL	**126.53**
yahoo-msg	739.8k	9.4k	2	2	ETL	ETL	ETL	**35.22**
dbpedia	921.7k	24.8k	4	3	ETL	ETL	ETL	**56.02**
digg	477.3k	219	2	2	ETL	ETL	ETL	**5.592**
bibsonomy	1.2M	211	3	3	ETL	ETL	ETL	**3.631**
epinions	2.6M	775	2	2	ETL	ETL	ETL	**85.27**
flickr	6.8M	216	2	2	ETL	ETL	ETL	**120.79**
orkut	37.4M	166	2	2	ETL	ETL	ETL	**1241.01**

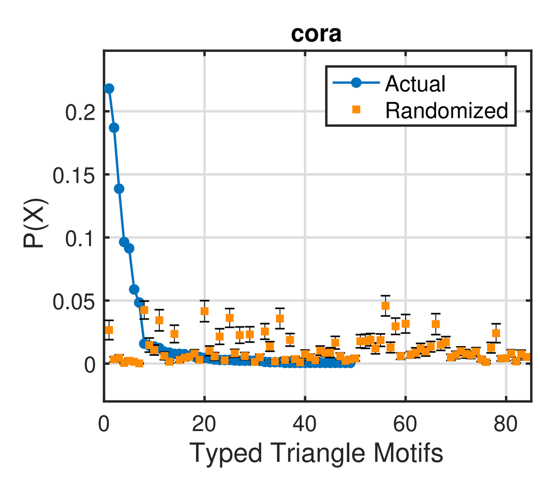
* ETL = Exceeded Time Limit (24 hours / 86,400 seconds)

Orders of magnitude faster and more space-efficient

Political retweets (type=political view)

p = [0.608 0.003 0.001 0.388]

99.65% of the 24,815 triangles form between users of the same political learning



- The 7 triangles with homo. types are also the triangles with largest frequency, accounting for 83.86% of all triangles in G.
- Only 49 out of the 84 possible heterogeneous triangles actually occur in G

Main Findings & Contributions

1. Generalize the notion of graphlet to heterogeneous/attributed/labeled graphs
2. Described a computational framework for computing them
3. Proposed algorithm with worst-case time complexity that matches the best known algorithm for untyped graphlets
4. Demonstrated the effectiveness of heterogeneous graphlets for exploratory analysis
   - Node classification, link prediction, visitor stitching (our recent work)