MOSS-5: A Fast Method of Approximating Counts of 5-Node Graphlets in Large Graphs (Extended abstract)

Pinghui Wang[1], Junzhou Zhao[2], Xiangliang Zhang[2], Zhenguo Li[3],
Jiefeng Cheng[4], John C.S. Lui[5], Don Towsley[6], Jing Tao[1], and Xiaohong Guan[1]
[1]MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, China
[2]King Abdullah University of Science and Technology, Thuwal, SA
[3]Huawei Noah's Ark Lab, Hong Kong
[4]Tencent Cloud Security Lab, Shenzhen, China
[5]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[6]Department of Computer Science, University of Massachusetts Amherst, MA, USA
Email: {phwang, jtao, jzzhao, xhguan}@mail.xjtu.edu.cn, xiangliang.zhang@kaust.edu.sa,
li.zhenguo@huawei.com, geoffcheng@tencent.com, cslui@cse.cuhk.edu.hk, towsley@cs.umass.edu

***Abstract*—Despite recent efforts in counting 3-node and 4-node graphlets, little attention has been paid to characterizing 5-node graphlets. In this paper, we develop a computationally efficient sampling method to estimate 5-node graphlet counts. We not only provide a fast sampling method and unbiased estimators of graphlet counts, but also derive simple yet exact formulas for the variances of the estimators which are of great value in practice—the variances can be used to bound the estimates' errors and determine the smallest necessary sampling budget for a desired accuracy. We conduct experiments on a variety of real-world datasets, and the results show that our method is several orders of magnitude faster than the state-of-the-art methods with the same accuracy.**

I. Introduction

For complex networks such as online social networks, computer networks, and biological networks, designing tools for estimating the counts (or frequencies) of 3-, 4-, and 5-node connected subgraph patterns (i.e., graphlets) is fundamental for detecting evolution and anomaly patterns in a large graph and computing graph similarities for graph classification, which have been widely used for a variety of graph mining and learning tasks. Despite recent progress in counting triangles and 4-node graphlets, little attention has been given to developing fast tools for characterizing and counting 5-node graphlets. Formally, let $G = (V, E)$ be an undirected graph, where V and E are the node set and edge set respectively. All undirected graphs' 5-node graphlets $G_1^{(5)}, \ldots, G_{21}^{(5)}$ studied in this paper are shown in Fig. 1. Denote by $C^{(5)}$ the set of 5-node CISes in G, and $C_i^{(5)}$ the set of 5-node CISes in G isomorphic to graphlet $G_i^{(5)}$. The graphlet count of $G_i^{(5)}$ is defined as $\eta_i = |C_i^{(5)}|$, $1 \leq i \leq 21$. Recently, Pinar et al. [1] propose a fast method *ESCAPE* for counting 5-node graphlets by utilizing the relationships between 3-, 4-, and 5-node graphlets counts. However, ESCAPE is not scalable to large graphs, which requires more than 10 hours to handle graphs with millions of nodes and edges. To address this challenge, we propose a novel sampling method MOSS-5 to fast estimate the counts of 5-node graphlets.

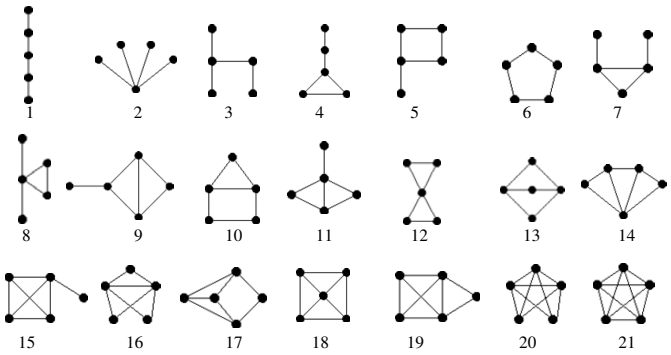

Figure 1. 5-node undirected graphlets $G_i^{(5)}$, $i = 1, \ldots, 21$.

II. Our Method

We observe that 1) except CISes in $C_1^{(5)} \cup C_2^{(5)} \cup C_6^{(5)}$, 5-node CISes include at least one subgraph isomorphic to graphlet $G_3^{(5)}$; 2) except CISes in $C_2^{(5)} \cup C_3^{(5)} \cup C_8^{(5)}$, 5-node CISes include at least one subgraph isomorphic to graphlet $G_1^{(5)}$. Let $\Omega_1 = \{1, \ldots, 21\} - \{1, 2, 6\}$ and $\Omega_2 = \{1, \ldots, 21\} - \{2, 3, 8\}$. Inspired by the above two observations, we develop a method MOSS-5 consisting of two sub-methods: T-5 and Path-5, where T-5 is customized to fast sample 5-node CISes isomorphic to $G_i^{(5)}$, $i \in \Omega_1$, and Path-5 is customized to fast sample 5-node CISes isomorphic to $G_j^{(5)}$, $j \in \Omega_2$. For any $i \in \Omega_1$, we provide an unbiased estimate $\hat{\eta}_i^{(1)}$ of η_i based on sampled CISes of T-5. For any $j \in \Omega_2$, similarly, we provide an unbiased estimate $\hat{\eta}_j^{(2)}$ of η_j based on sampled CISes of Path-5. Based on $\hat{\eta}_i^{(1)}$ and $\hat{\eta}_j^{(2)}$, we propose a more accurate estimator $\hat{\eta}_k$ of η_k, $k \in \Omega_1 \cup \Omega_2 = \{1, \ldots, 21\} - \{2\}$ and provide an unbiased estimator $\hat{\eta}_2$ of η_2.

The T-5 Sampling Method. Denote $\Gamma_v^{(1)} = (d_v - 1)(d_v - 2)\sum_{x\in N_v}(d_x - 1)$, $v \in V$. We assign a weight $\Gamma_v^{(1)}$ to each node $v \in V$. Define $\Gamma^{(1)} = \sum_{v\in V}\Gamma_v^{(1)}$ and $\rho_v^{(1)} = \frac{\Gamma_v^{(1)}}{\Gamma^{(1)}}$. Let N_v denote the set of neighbors of v in graph G. To sample a 5-node CIS, T-5 uses six steps: **Step 1**) sample a node v from V according to the distribution $\rho^{(1)} = \{\rho_v^{(1)} : v \in V\}$; **Step 2**) Sample a node u from N_v according to the distribution $\sigma^{(v)} = \{\sigma_u^{(v)} : u \in N_v\}$, where $\sigma_u^{(v)}$ is defined as $\sigma_u^{(v)} = \frac{d_u - 1}{\sum_{x\in N_v}(d_x-1)}$, $u \in N_v$; **Step 3**) sample a node w from $N_v - \{u\}$ at random; **Step 4**) sample a node r from $N_v - \{u, w\}$ at random; **Step 5**) sample a node t from $N_u - \{v\}$ at random; **Step 6**) return the CIS s that includes nodes v, u, w, r, and t. We run the above procedure K_1 times to obtain K_1 CISes $s_1^{(1)}, \ldots, s_{K_1}^{(1)}$. For a CIS s isomorphic to graphlet $G_i^{(5)}$, $1 \le i \le 21$, denote $\phi_i^{(1)}$ as the number of subgraphs in s that are isomorphic to graphlet $G_3^{(5)}$. Using the sampling procedure once (i.e., $K_1 = 1$), T-5 returns a CIS $s \in C_i^{(5)}$ sampled with probability $p_i^{(1)} = \frac{2\phi_i^{(1)}}{\Gamma^{(1)}}$, $1 \le i \le 21$. We let $G^{(5)}(s)$ be the 5-node graphlet ID of s when s is a 5-node CIS, and -1 otherwise. We define $m_i^{(1)} = \sum_{k=1}^{K_1}\mathbf{1}(G^{(5)}(s_k^{(1)}) = i)$. For $i \in \Omega_1$, $p_i^{(1)}$ is larger than zero and thus we estimate η_i as $\hat{\eta}_i^{(1)} = \frac{m_i^{(1)}}{K_1 p_i^{(1)}}$. For $i \in \Omega_1$, $\hat{\eta}_i^{(1)}$ is an unbiased estimator of η_i, i.e., $\mathbb{E}(\hat{\eta}_i^{(1)}) = \eta_i^{(1)}$, and the variance of $\hat{\eta}_i^{(1)}$ is $\mathrm{Var}(\hat{\eta}_i^{(1)}) = \frac{\eta_i}{K_1}\left(\frac{1}{p_i^{(1)}} - \eta_i\right)$.

The Path-5 Sampling Method. Let $\Gamma_v^{(2)} = \left(\sum_{x\in N_v}(d_x - 1)\right)^2 - \sum_{x\in N_v}(d_x - 1)^2$, $v \in V$. We assign a weight $\Gamma_v^{(2)}$ to each node $v \in V$. Define $\Gamma^{(2)} = \sum_{v\in V}\Gamma_v^{(2)}$ and $\rho_v^{(2)} = \frac{\Gamma_v^{(2)}}{\Gamma^{(2)}}$. To sample a 5-node CIS, Path-5 mainly consists of six steps: **Step 1**) sample a node v from V according to the distribution $\rho^{(2)} = \{\rho_v^{(2)} : v \in V\}$; **Step 2**) sample a node u from N_v according to the distribution $\tau^{(v)} = \{\tau_u^{(v)} : u \in N_v\}$, where we define $\tau_u^{(v)} = \frac{(d_u-1)(\sum_{y\in N_v-\{u\}}(d_y-1))}{\Gamma_v^{(2)}}$, $u \in N_v$; **Step 3**) sample a node w from $N_v - \{u\}$ according to the distribution $\mu^{(v,u)} = \left\{\mu_w^{(v,u)} : w \in N_v - \{u\}\right\}$, where we define $\mu_w^{(v,u)} = \frac{d_w - 1}{\sum_{y\in N_v-\{u\}}(d_y-1)}$, $w \in N_v - \{u\}$; **Step 4**) sample a node r from $N_u - \{v\}$ at random; **Step 5**) sample a node t from $N_w - \{v\}$ at random; **Step 6**) return the CIS s that includes nodes v, u, w, r, and t. We run the above procedure K_2 times to obtain K_2 CISes $s_1^{(2)}, \ldots, s_{K_2}^{(2)}$. For a CIS s isomorphic to graphlet $G_i^{(5)}$, $1 \le i \le 21$, let $\phi_i^{(2)}$ denote the number of subgraphs in s that are isomorphic to $G_1^{(5)}$. Using the sampling procedure once (i.e., $K_2 = 1$), Path-5 samples a CIS $s \in C_i^{(5)}$ with probability $p_i^{(2)} = \frac{2\phi_i^{(2)}}{\Gamma^{(2)}}$, $1 \le i \le 21$. Denote $m_i^{(2)} = \sum_{k=1}^{K_2}\mathbf{1}(G^{(5)}(s_k^{(2)}) = i)$. For $i \in \Omega_2$, $p_i^{(2)}$ is larger than zero and we then estimate η_i as $\hat{\eta}_i^{(2)} = \frac{m_i^{(2)}}{K_2 p_i^{(2)}}$. For $i \in \Omega_2$, $\hat{\eta}_i^{(2)}$ is an unbiased estimator of η_i and its variance is $\mathrm{Var}(\hat{\eta}_i^{(2)}) = \frac{\eta_i}{K_2}\left(\frac{1}{p_i^{(2)}} - \eta_i\right)$.

Hybrid Estimator of 5-Node Graphlet Counts. We estimate η_i as $\hat{\eta}_i^{(1)}$ and $\hat{\eta}_i^{(2)}$ for $i \in \Omega_1 - \Omega_2$ and $i \in \Omega_2 - \Omega_1$ respectively. When $i \in \Omega_1 \cap \Omega_2$, we estimate η_i based on its two unbiased estimates $\hat{\eta}_i^{(1)}$ and $\hat{\eta}_i^{(2)}$. Formally, let

$$\lambda_i^{(1)} = \frac{\mathrm{Var}(\hat{\eta}_i^{(2)})}{\mathrm{Var}(\hat{\eta}_i^{(1)}) + \mathrm{Var}(\hat{\eta}_i^{(2)})}, \lambda_i^{(2)} = \frac{\mathrm{Var}(\hat{\eta}_i^{(1)})}{\mathrm{Var}(\hat{\eta}_i^{(1)}) + \mathrm{Var}(\hat{\eta}_i^{(2)})}.$$

For $i \in \Omega_1 \cup \Omega_2 = \{1, 3, 4, 5, \ldots, 21\}$, we estimate η_i as

$$\hat{\eta}_i = \begin{cases} \lambda_i^{(1)}\hat{\eta}_i^{(1)} + \lambda_i^{(2)}\hat{\eta}_i^{(2)}, & i \in \Omega_1 \cap \Omega_2, \\ \hat{\eta}_i^{(1)}, & i \in \Omega_1 - \Omega_2, \\ \hat{\eta}_i^{(2)}, & i \in \Omega_2 - \Omega_1. \end{cases}$$

For a CIS s isomorphic to graphlet $G_i^{(5)}$, $1 \le i \le 21$, let $\phi_i^{(3)}$ denote the number of subgraphs in s that are isomorphic to graphlet $G_2^{(5)}$. Let $\Lambda_4 = \sum_{v\in V}\binom{d_v}{4}$. Then, the number of all 5-node subgraphs (not necessarily induced) in G isomorphic to graphlet $G_2^{(5)}$ is Λ_4. Let $\Omega_3 = \{j : \phi_j^{(3)} > 0\}$ and $\Omega_3^* = \Omega_3 - \{2\}$. We observe that $\sum_{i\in\Omega_3}\phi_i^{(3)}\eta_i = \Lambda_4$. Since $\phi_2^{(3)} = 1$, we estimate η_2 as

$$\hat{\eta}_2 = \Lambda_4 - \sum_{i\in\Omega_3^*}\phi_i^{(3)}\hat{\eta}_i.$$

Experimental results. Table I shows the expected smallest computational time of MOSS-5 required to obtain all estimates $\hat{\eta}_1, \ldots, \hat{\eta}_{21}$ with NRMSE smaller than 0.1. To compute $\eta_1, \ldots, \eta_{21}$, the state-of-the-art exact computing method ESCAPE requires 52 hours, 32 hours, and 23 hours for graphs Flickr, com-Orkut, and LiveJournal respectively. We can see that the computational time of ESCAPE does not strictly increase with the graph size. For example, graph ca-HepPh is more than ten times smaller than graphs YouTube and Web-Google. To compute $\eta_1, \ldots, \eta_{21}$, however, ESCAPE requires much more time for ca-HepPh than for YouTube and Web-Google. From Table I, we see that our method MOSS-5 is 2 to 18,945 times faster than ESCAPE when providing accurate estimates with NRMSE smaller than 0.1. When $\max_{i=1,\ldots,21}\mathrm{NRMSE}(\hat{\eta}_i) = 0.1$. the average NRMSE varies from 0.01 to 0.04 for all graphs studied in this paper.

Table I
MOSS-5 IN COMPARISON WITH EXACT COUNTING METHOD ESCAPE.

Graph	ESCAPE	MOSS-5, $\max_{i=1,\ldots,21}\mathrm{NRMSE}(\hat{\eta}_i) = 0.1$	
	(time)	time	$\frac{1}{21}\sum_{i=1}^{21}\mathrm{NRMSE}(\hat{\eta}_i)$
Flickr	189,450 s	10 s	0.039
com-Orkut	116,029 s	103 s	0.015
LiveJournal	82,445 s	31 s	0.037
Pokec	3,696 s	31 s	0.024
Wiki-Talk	1,877 s	47 s	0.018
Xiami	518 s	82 s	0.013
Web-Google	112 s	25 s	0.013
YouTube	193 s	96 s	0.011
ca-HepPh	589 s	64 s	0.011

REFERENCES

[1] A. Pinar, C. Seshadhr, and V. Visha, "Escape: Efficiently counting all 5-vertex subgraphs," in *WWW*, 2017.