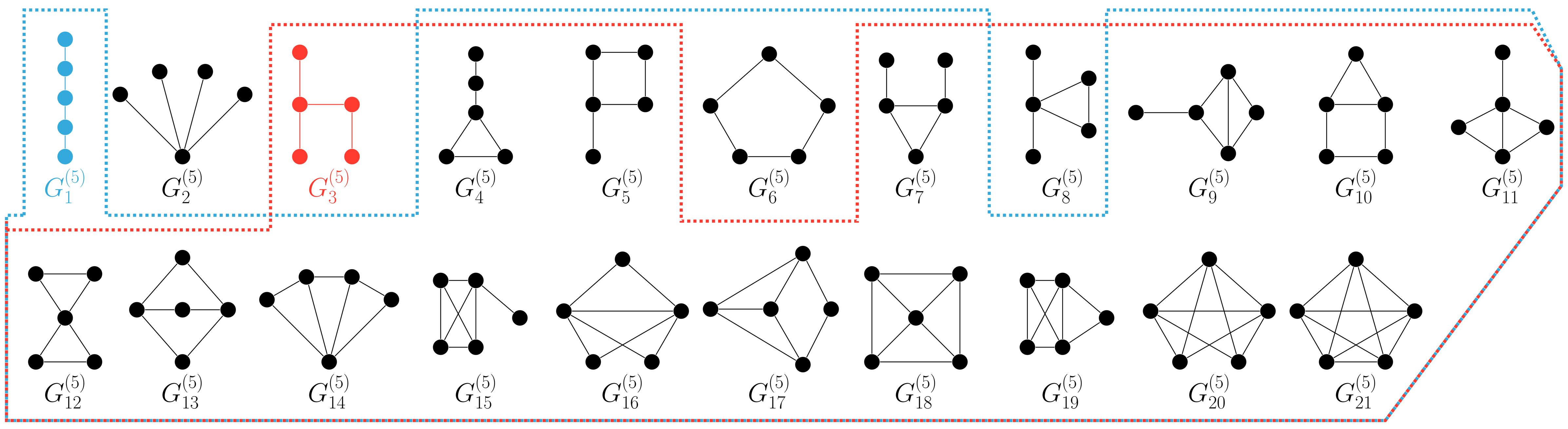
MOSS-5: A Fast Method of Approximating Counts of 5-Node Graphlets in Large Graphs

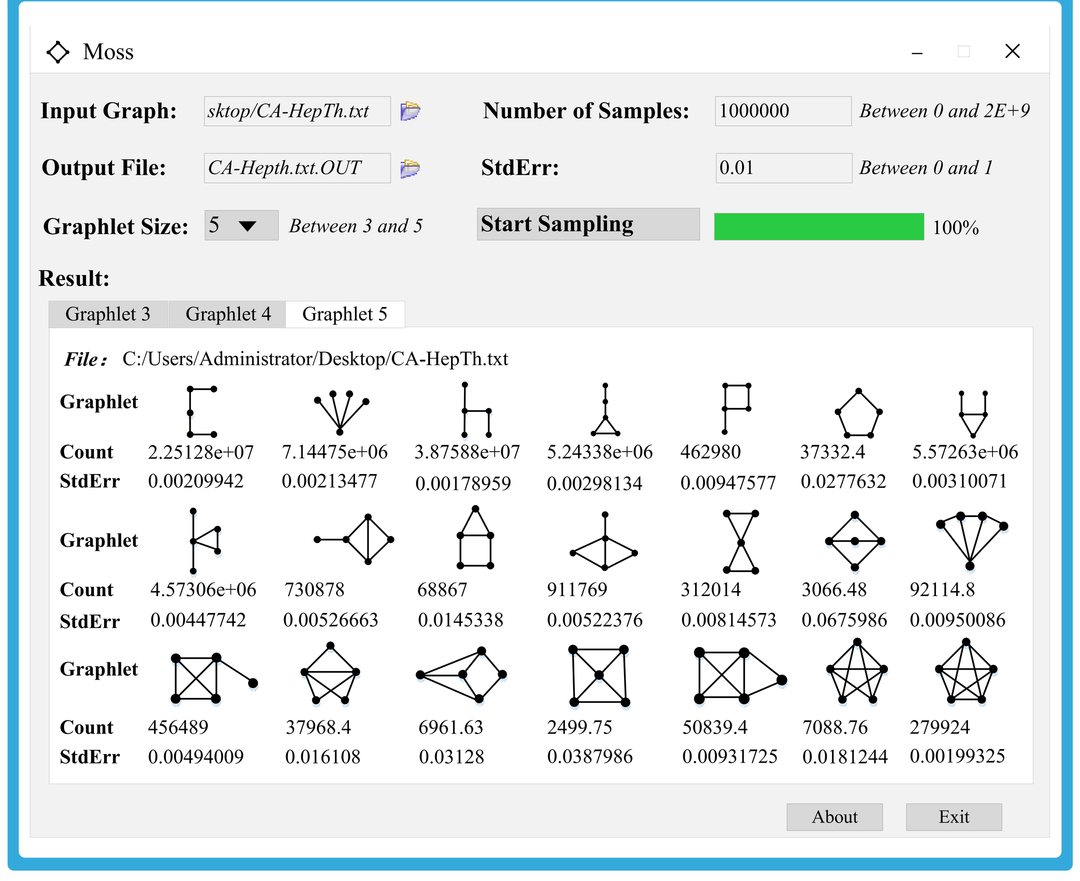
Pinghui Wang Junzhou Zhao Xiangliang Zhang Zhenguo Li Jiefeng Cheng
John C.S. Lui Don Towsley Jing Tao Xiaohong Guan

5-Node Graphlets in Undirected Graphs



- Connected subgraph patterns are useful for a variety of graph mining and learning tasks, e.g., graph similarity computation, clustering, malware detection, protein/compound function prediction, etc.

Our Developed Tool

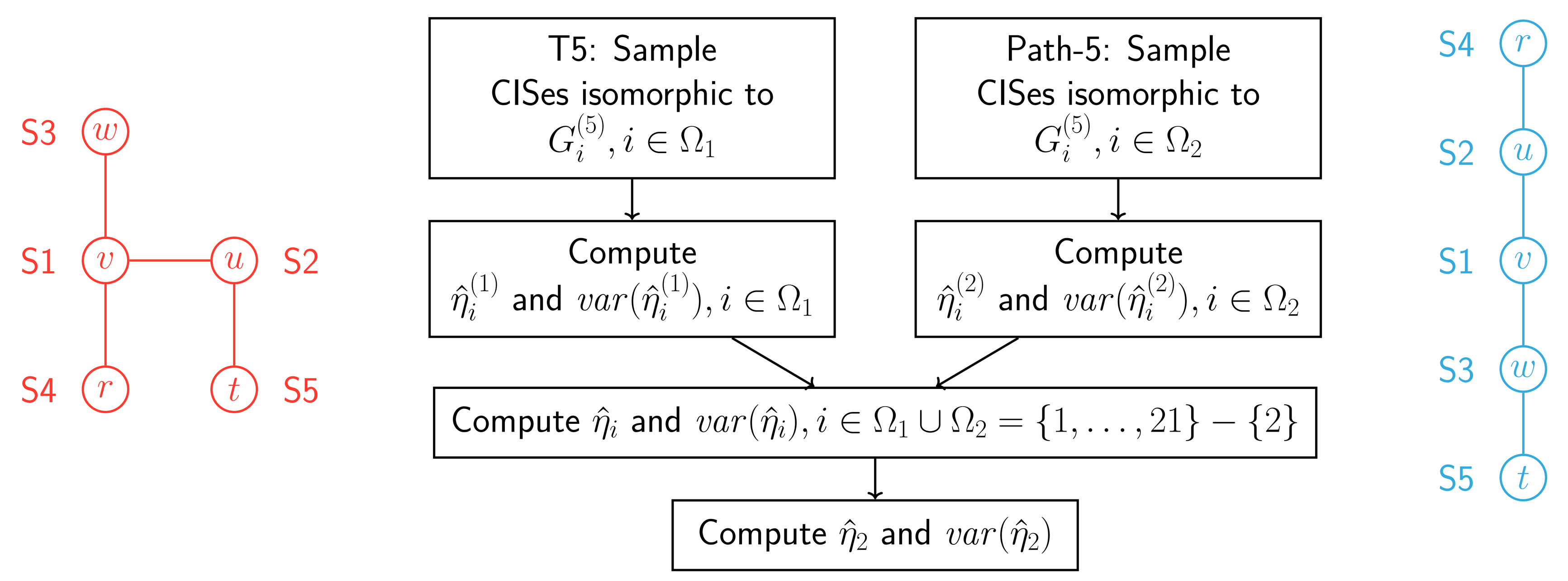


Problem Formulation

- Undirected graph $G = (V, E)$ where V and E are sets of nodes and edges
- Graphlets are connected induced subgraphs (CISes) induced by G
- 5-node graphlets: $G_1^{(5)}, \ldots, G_{21}^{(5)}$
- Let $C_i^{(5)}$ be the set of 5-node CISes in G isomorphic to graphlet $G_i^{(5)}$
- **Goal**: estimate graphlet count of $G_i^{(5)}$: $\eta_i \triangleq |C_i^{(5)}|$, for $i = 1, \ldots, 21$
- **Challenge**: combinatorial explosion.
- E.g., Epinions graph have merely 10^5 nodes and 10^6 edges, but contains more than 10^{13} 5-node CISes. Thus, brute-force enumeration approach is not practical.

Overview of Our Approach: MOSS-5

- We design **sampling methods** to sample 5-node CISes efficiently, and estimate their counts with desired accuracy.
- **Observation 1**: 5-node CISes contain at least one subgraph isomorphic to graphlet $G_3^{(5)}$ except CISes in $C_1^{(5)} \cup C_2^{(5)} \cup C_6^{(5)}$.
- **Observation 2**: 5-node CISes contain at least one subgraph isomorphic to graphlet $G_1^{(5)}$ except CISes in $C_2^{(5)} \cup C_3^{(5)} \cup C_8^{(5)}$.
- Let $\Omega_1 \triangleq \{1, \ldots, 21\} - \{1, 2, 6\}$, and $\Omega_2 \triangleq \{1, \ldots, 21\} - \{2, 3, 8\}$.



T-5

- Assign each node v with a weight $\Gamma_v^{(1)} \triangleq (d_v - 1)(d_v - 2)\sum_{x \in N_v}(d_x - 1)$. Let $\Gamma^{(1)} \triangleq \sum_v \Gamma_v^{(1)}$ and $\rho_v^{(1)} \triangleq \Gamma_v^{(1)}/\Gamma^{(1)}$. T-5 consists of following 6 steps:
- S1: sample a node v from V according to dist. $\rho^{(1)} \triangleq \{\rho_v^{(1)} \colon v \in V\}$;
- S2: sample a node u from N_v according to dist. $\sigma^{(v)} \triangleq \{\sigma_u^{(v)} \colon u \in N_v\}$ where $\sigma_u^{(v)} \triangleq (d_u - 1)/\sum_{x \in N_v}(d_x - 1)$;
- S3: sample a node w from $N_v - \{u\}$ at random;
- S4: sample a node r from $N_v - \{u, w\}$ at random;
- S5: sample a node t from $N_u - \{v\}$ at random;
- S6: return the CIS s consisting of nodes v, u, w, r, and t.

Theorem: *T-5 samples a CIS $s \in C_i^{(5)}$ with probability $p_i^{(1)} = 2\phi_i^{(1)}/\Gamma^{(1)}$.*

- Run T-5 K_1 times, obtain CISes $S^{(1)} \triangleq \{s_k^{(1)} \colon k = 1, \ldots, K_1\}$.
- Let $G^{(5)}(s)$ be the graphlet ID of $s \in S^{(1)}$, and $m_i^{(1)} \triangleq \sum_{s \in S^{(1)}} \mathbf{1}(G^{(5)}(s))$
- Obviously $\mathbb{E}[m_i^{(1)}] = K_1 p_i^{(1)} \eta_i$
- Therefore, $\hat{\eta}_i^{(1)} \triangleq m_i^{(1)}/(K_1 p_i^{(1)})$ is an unbiased estimator of η_i for $i \in \Omega_1$.

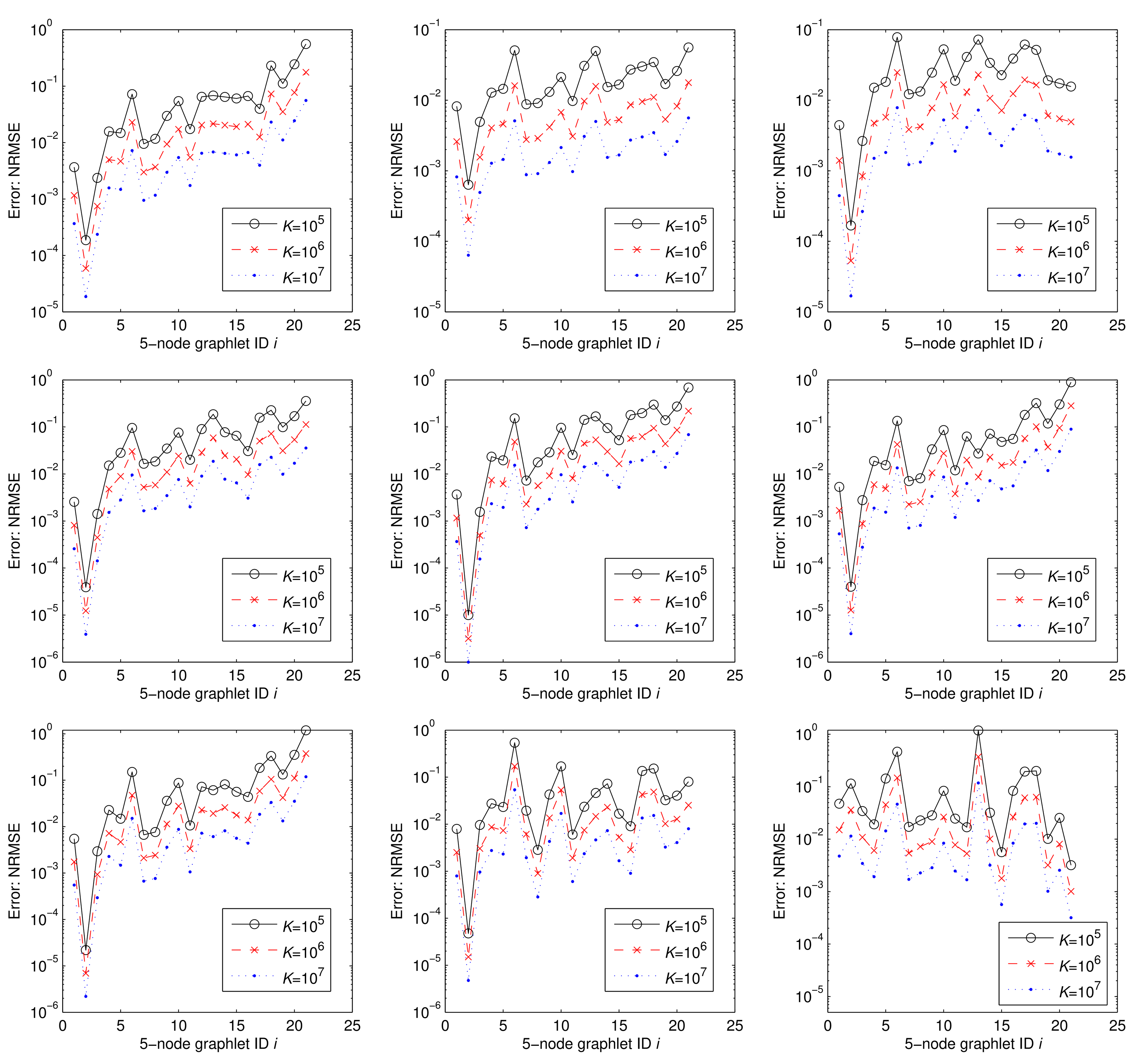
Path-5

- Assign each node v with a weight $\Gamma_v^{(2)} \triangleq (\sum_{x \in N_v}(d_x - 1))^2 - \sum_{x \in N_v}(d_x - 1)^2$. Let $\Gamma^{(2)} \triangleq \sum_v \Gamma_v^{(2)}$ and $\rho_v^{(2)} \triangleq \Gamma_v^{(2)}/\Gamma^{(2)}$. Path-5 consists of following 6 steps:
- S1: sample a node v from V according to dist. $\rho^{(2)} \triangleq \{\rho_v^{(2)} \colon v \in V\}$;
- S2: sample a node u from N_v according to dist. $\tau^{(v)} \triangleq \{\tau_u^{(v)} \colon u \in N_v\}$ where $\tau_u^{(v)} \triangleq (d_u - 1)(\sum_{y \in N_v - \{u\}}(d_y - 1))/\Gamma_v^{(2)}$;
- S3: sample a node w from $N_v - \{u\}$ according to dist. $\mu^{(v,u)} \triangleq \{\mu_w^{(v,u)} \colon w \in N_v - \{u\}\}$ where $\mu_w^{(v,u)} \triangleq (d_w - 1)/\sum_{y \in N_v - \{u\}}(d_y - 1)$
- S4: sample a node r from $N_u - \{v\}$ at random;
- S5: sample a node t from $N_w - \{v\}$ at random;
- S6: return the CIS s consisting of nodes v, u, w, r, and t.

Theorem: *T-5 samples a CIS $s \in C_i^{(5)}$ with probability $p_i^{(2)} = 2\phi_i^{(2)}/\Gamma^{(2)}$.*

- Similar to T-5, $\hat{\eta}_i^{(2)} \triangleq m_i^{(2)}/(K_2 p_i^{(2)})$ is an unbiased estimator of η_i for $i \in \Omega_2$
- The number of all 5-node subgraphs in G isomorphic to $G_2^{(5)}$ is $\Lambda_4 \triangleq \sum_v \binom{d_v}{4}$
- Let $\phi_i^{(3)}$ denote the number of subgraphs in s that are isomorphic to $G_2^{(5)}$, and $\Omega_3 \triangleq \{i \colon \phi_i^{(3)} > 0\}$. Therefore, $\sum_{i \in \Omega_3} \phi_i^{(3)} \eta_i = \Lambda_4$.
- Because $\phi_2^{(3)} = 1$, we can estimate η_2 as $\hat{\eta}_2 \triangleq \Lambda_4 - \sum_{i \in \Omega_3 - \{2\}} \phi_i^{(3)} \hat{\eta}_i$.

Evaluation



Graphs: Orkut, Flickr, LiveJournal, Pokec, Wiki-talk, Xiami, YouTube, Web-Google, and HepPh