Computing the Clustering Coefficient

Definition

$$\mathrm{CC}(v) = \frac{|\{u, w\ u \in N(v) \text{ and } w \in N(v) \text{ and } (u,w) \in E\}|}{\deg(v) \times (\deg(v) - 1)} = \frac{2 * t_v}{\deg(v) \times (\deg(v) - 1)}$$

Every triangle is counted twice. t_v is the number of triangles which include node v. An algorithm to approximate this quantity in a streaming model is

- Initialize count = 0
- Set $S = \emptyset$
- Set $p = m/M$
- $deg(v) = 0$ for all $v \in V$:
- $dt_v^S = 0$ for all $v \in V$:
  - For each $(u, v) \in E_{stream}$:
    - Set $deg(v) = deg(v) + 1$
    - Set $deg(u) = deg(u) + 1$
    - $N_(u, v)^S = N_(u)^S \cap N_(v)^S$:
    - For each $w \in N_(u, v)^S$:
      - Set $t_v^S = t_v^S + 1$
      - Set $t_u^S = t_u^S + 1$
      - Set $t_w^S = t_w^S + 1$
  - if $SampleProb(p)$:
    - $S = S \cup (u, v)$

return $t_v^S \frac{2}{deg(v)(deg(v)-1)}$ for all $v \in V$

Proposition:

$$E(t_v^S \frac{2}{deg(v)(deg(v) - 1)}) = \mathrm{CC}(v)$$

Proof:
Let

$$X_i = \begin{cases} 1 & \text{if i-th triangle is counted} \\ 0 & \text{else} \end{cases}$$

Consider:

$$E(\sum_{i=0}^{t_v^G} X_i) = \sum_{i=0}^{t_v^G} Pr(X_i = 1) = \sum_{i=0}^{t_v^G} p^2 = t_v^G p^2,$$

where t_v^G is the number of triangles in the Graph G which involve v. Thus

$$E(\frac{1}{p^2}\sum_{i=0}^{t_v^G} X_i) = \frac{1}{p^2}\sum_{i=0}^{t_v^G} Pr(X_i = 1) = \frac{1}{p^2}\sum_{i=0}^{t_v^G} p^2 = t_v^G,$$

and therefore

$$E(\frac{1}{p^2}\sum_{i=0}^{t_v^G} X_i \frac{2}{deg(v)(deg(v)-1))}) = \frac{2 * t_v^G}{\deg(v)(\deg(v)-1)} = \mathrm{CC}(v).$$