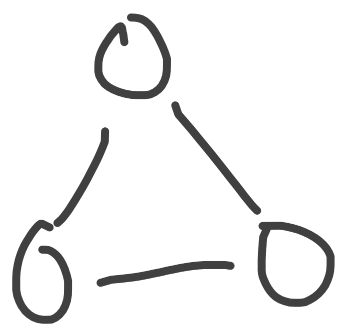
March 29, 2018

Reax papers due tonight

Counting triangles (in graphs)

1. exact
2. approximate

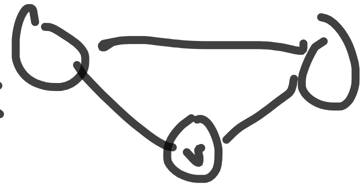
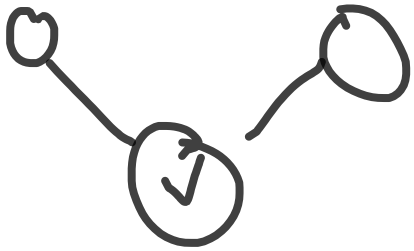
Why triangles?



Simplest non-trivial small subgraph pattern (motifs)

Social Network Analysis

clustering coefficient

$C_v =$ #  / # 

$= \dfrac{T_v}{\binom{d_v}{2}}$ ← triangles containing v

fraction of pairs of friends that are friends themselves

How do we count triangles (fast)?

Model: Undirected graph G

$O(1)$ time to check if $(i,j) \in G$

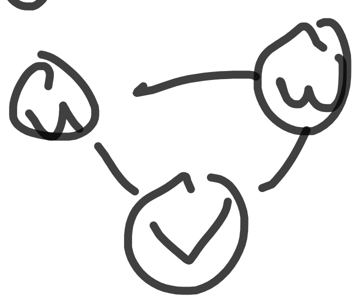
$O(d_v)$ time to get all of the neighbors of node v

---

first approach (brute force)

look at all triples of nodes

for $u, v, w \in V$

check if  is in the graph

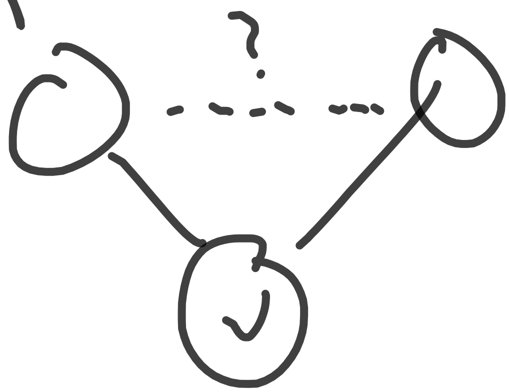
$\Theta(n^3)$ $n = |V|$

$m = |E|$

on <u>every</u> graph

Second approach (neighbor pairs)

Idea:

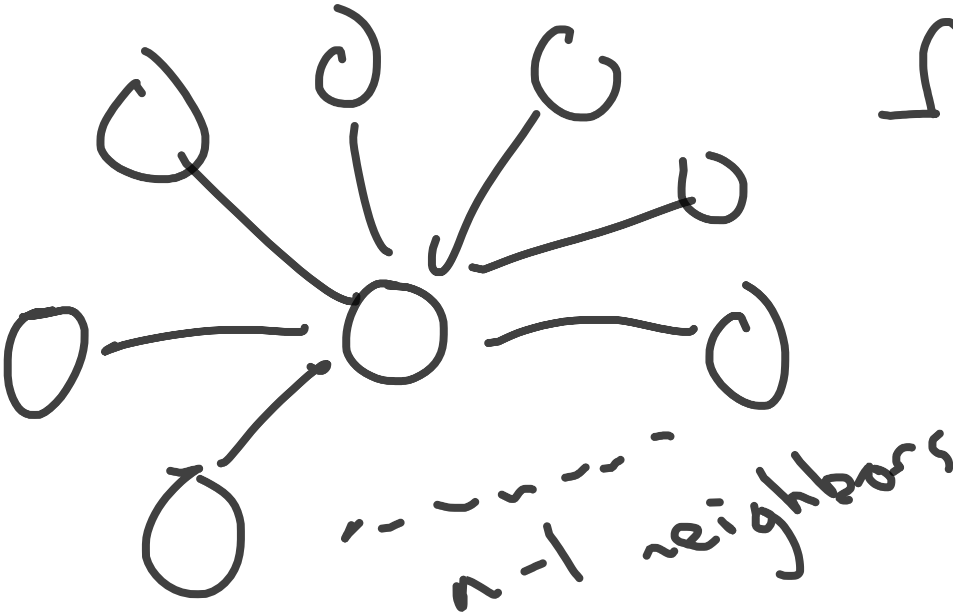


can access neighbor list efficiently

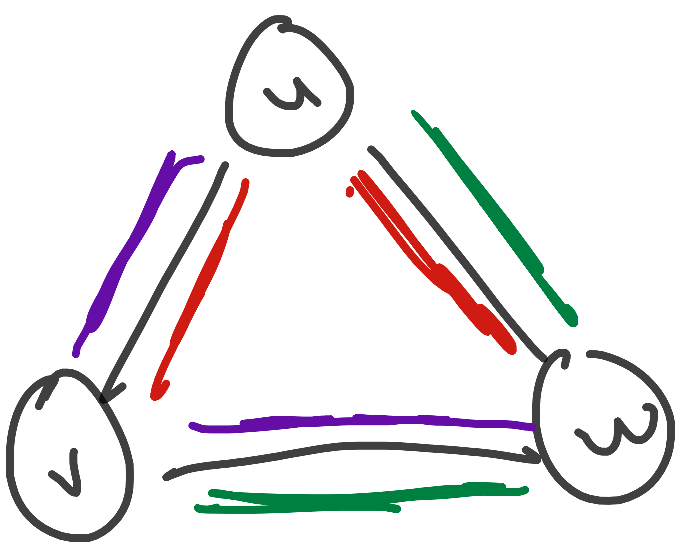
for $v \in V$
  for neighbors u, w of v
    if u — w
      increment counter

$$\Theta\left(\sum_v d_v^2\right)$$

heavy-tailed degree distributions
$\Rightarrow$ some high degree nodes in G



$\Omega(n^2)$

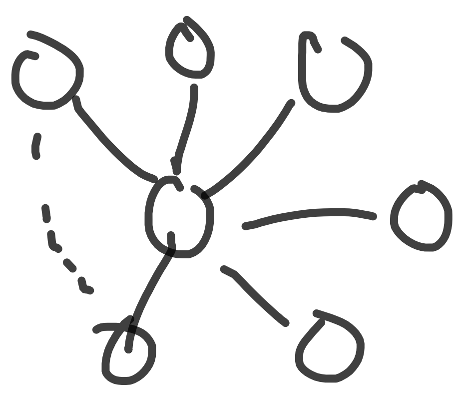


only count once in a "smart" way

Third approach (low degree centers)

1. sort the nodes by degree
   $\Rightarrow$ ordering σ
   $\Rightarrow \sigma(u) < \sigma(v) \Rightarrow d_u \leq d_v$

2. for each $v \in V$
   for each neighbor pair
   u, w with $\sigma(v) \leq \sigma(u), \sigma(w)$
   check if
   u — w exists



$\Theta(m)$ time

Claim: $\Theta(m^{3/2})$ worst-case for every graph

complete graph: $m = n^2 \Rightarrow n^3$ ✓

Proof:

split vertices into big or small

v big if $d_v > \sqrt{m}$ B

v small if $d_v \leq \sqrt{m}$ S

how much work do we do when processing small nodes?

$d_v^2 \leq m$

Work:

$O\left(\sum_{v \in S} d_v^2\right)$

$\sum_{v \in S} d_v \leq 2m$

Worst-case:

$$d_v = \sqrt{m} \quad v \in S$$

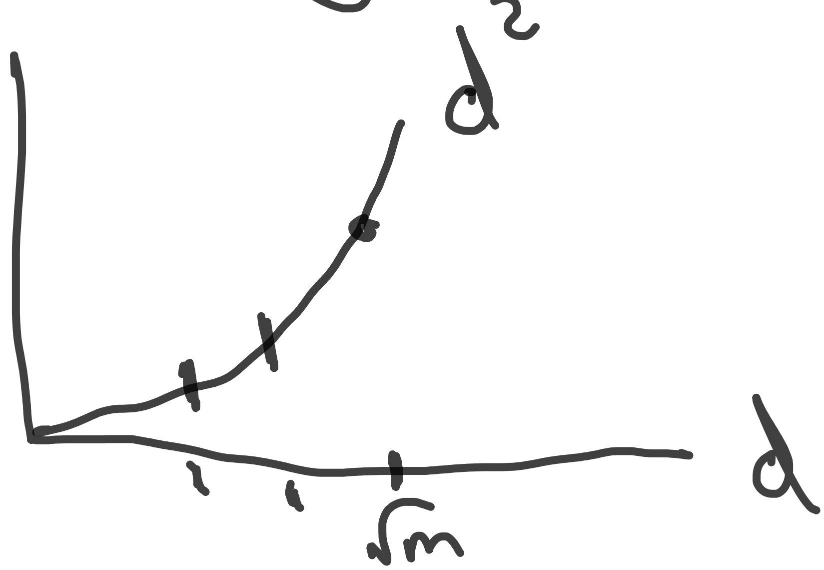
At most $2\sqrt{m}$ of these

$$\sum_{v \in S} d_v = \sum_{v \in S} \sqrt{m}$$

if more $2\sqrt{m}$ of these

$$> \sqrt{m} \cdot 2\sqrt{m} = 2m \;\divideontimes$$

could have more nodes with smaller degree



$$\sum_{v \in S} d_v^2 = \sum_{v \in S} m$$

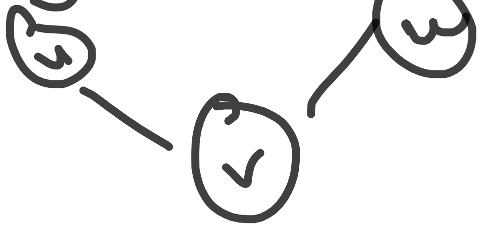
$$\le O(\sqrt{m} \cdot m)$$

$$= O(m^{3/2})$$

What about big nodes?

At most $2\sqrt{m}$ big nodes (same reasoning as before)

When the algorithm processes a big node as a center



$v \in B$

neighbor pairs that are

processed are also big

Bound big node work by

$O\left(\binom{|B|}{3}\right)$

$|B| \leq 2\sqrt{m}$

$\rightarrow O(m^{3/2})$ $\square$

Only really needed a big/small split

Degree ordering works well in practice

Can prove additional guarantees

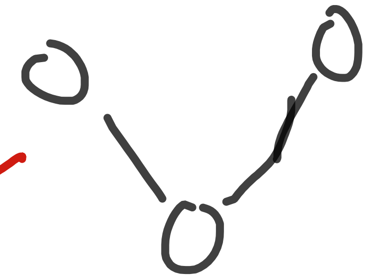
if you assume something about the degree dist (power law)

see Latapy

Approximate counting

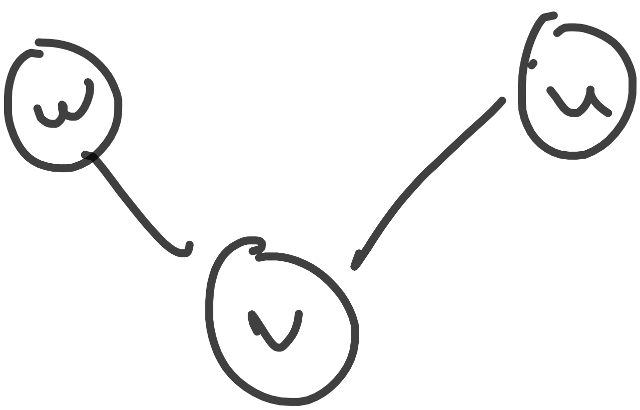
Global clustering coeff.

$$C = \frac{3T}{\sum_v \binom{d_v}{2}}$$



(Seshadhri et al. 2013)

"Wedge sampling"



wedge centered at node v

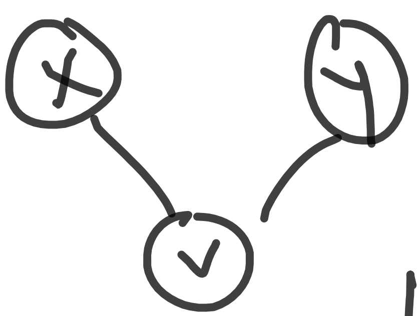
$\binom{d_v}{2}$ such wedges

$$p_v = \frac{\binom{d_v}{2}}{\sum_w \binom{d_w}{2}}$$

(fraction of wedges centered at v)

UniformWedgeSampl()

1. choose $v \sim \{p_w\}_w$
2. choose two neighbors of v unif. at random

Pr(select 

)

$$= p_v \cdot \frac{1}{\binom{d_v}{2}} = \frac{\cancel{\binom{d_v}{2}}}{\sum_w \binom{d_w}{2}} \frac{1}{\cancel{\binom{d_v}{2}}}$$

Approx GCCE (k)

1. Sample k wedges unif. at random
2. Output fraction of the wedges that induce a triangle

Why does this work well?

1. Unbiased estimator

$$X_i = \begin{cases} 1 & i\text{th wedge forms } \Delta \\ 0 & \text{o/w} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1)$$

$= \Pr(\text{UAR wedge forms a } \Delta)$

$$= \frac{3T}{\#\text{wedges}} = C$$

② Concentration of the estimator

Hoeffding's inequality

$X_1, \dots, X_k$ indep random vars

with $0 \leq X_j \leq 1$

$$\bar{X} = \frac{1}{k} \sum_{j=1}^{k} X_j$$

then for any $\varepsilon \in (0,1)$

$\Pr(|\bar{X} - \mathbb{E}(\bar{X})| \geq \varepsilon)$

$\leq 2\exp(-2k\varepsilon^2)$

in our case, indicator RVs

$X_i = \begin{cases} 1 & \text{ith wedge form } \Delta \\ 0 & \text{o/w} \end{cases}$

k samples, output

$\bar{X} = \frac{1}{k} \sum X_k$

= fraction of sampled wedges forming a triangle

$\mathbb{E}(\bar{X}) = C$ (global clust. coeff)

Choose # of samples

$$k = \frac{\log(2/\delta)}{2\varepsilon^2}$$

then $|\bar{X} - C| \leq \varepsilon$

with prob. at least $1-\delta$

Can also use to get estimate on # of triangles T

$$\left|\bar{X} \cdot \frac{\sum \binom{d_w}{2}}{3} - T\right| < \varepsilon \frac{\sum \binom{d_w}{2}}{3}$$

Other things to notice

① Can use uniform wedge sampler as uniform Δ sampler

(just reject)

2 Can also get avg. ccf

$$\bar{C} = \frac{1}{n} \sum_{w} \frac{T_w}{\binom{d_w}{2}}$$

$T_w = \#\Delta$ containing w

(i) Pick k nodes unif at random

(ii) choose uniform wedge for each node

(iii) output fraction of closed wedges