A BATCH SIZE ANALYSIS

In this section, we prove Lemma 4.1 and Corollary 4.2. We first present the following simplifying assumption that states that for any two vertices v_1, v_2 the intersection of their neighborhoods has weak correlation.

Assumption A.1 *Let $S \subset V$ be a random set drawn such that each $v \in V$ is picked to S with probability $p > \log n/d_{\min}$. There exist constants $c_1, c_2 \in \mathbb{R}$ such that*

$$\frac{1}{c_1}\mathop{\mathbf{E}}_{S}[|N(v_1)\cap S|]\mathop{\mathbf{E}}_{S}[|N(v_2)\cap S|] \le \mathop{\mathbf{E}}_{S}[|N(v_1)\cap S|\cdot|N(v_2)\cap S|] \le c_2\cdot\mathop{\mathbf{E}}_{S}[|N(v_1)\cap S|]\mathop{\mathbf{E}}_{S}[|N(v_2)\cap S|].$$

We start by stating a straightforward lemma from probability theory, whose proof is a direct application of Chernoff bounds.

Lemma A.1 *Let U be a finite set, and consider a random set $\mathrm{S} \subseteq U$ drawn such that each element in U is picked to S with probability p independently. Then, with probability at least $1-o(1/\mathrm{poly}(n))$*

$$|\mathrm{S}| \in \left[p|U| - \sqrt{p|U|\log n}, p|U| + \sqrt{p|U|\log n}\right]$$

For a fixed vertex u, we let $\tilde{x}_u = Wx_u$, so that our estimator

$$\xi = \frac{1}{|\mathrm{S}_1|}\sum_{v\in\mathrm{S}_1}\frac{\mathbf{1}_{N(v)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_2\cap N(v)|}\sum_{u\in N(v)\cap\mathrm{S}_2}\frac{\tilde{A}_{v,u}\cdot\tilde{x}_u}{\alpha_v}$$

We start with the proof of Proposition 4.1. Note that we considered the more general case where S_1 is picked according to probability $p = m_1/n$ and S_2 is picked according to probability $q = m_2/n$.

Proof of Proposition 4.1: We start with computing the mean of our estimator.

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|}\sum_{v\in\mathrm{S}_1}\chi_v\right] &= \mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|}\sum_{v\in\mathrm{S}_1}\frac{\mathbf{1}_{N(v)\cap\mathrm{S}_2\neq\emptyset}}{\alpha_v\cdot|N(v)\cap\mathrm{S}_2|}\sum_{u\in N(v)\cap\mathrm{S}_2}\tilde{A}_{v,u}\tilde{x}_u\right]\\
&= \mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|}\sum_{v,u\in V}\mathbf{1}_{v\in\mathrm{S}_1}\cdot\mathbf{1}_{N(v)\cap\mathrm{S}_2\neq\emptyset}\cdot\mathbf{1}_{u\in\mathrm{S}_2\cap N(v)}\cdot\frac{\tilde{A}_{v,u}\tilde{x}_u}{\alpha_v\cdot|\mathrm{S}_2\cap N(v)|}\right]\\
&= \sum_{(v,u)\in E}\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v\in\mathrm{S}_1}}{|\mathrm{S}_1|}\frac{\mathbf{1}_{N(v)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{u\in\mathrm{S}_2}}{\alpha_v\cdot|N(v)\cap\mathrm{S}_2|}\right]\cdot\tilde{A}_{v,u}\tilde{x}_u,
\end{aligned}$$

where for an event Z we let $\mathbf{1}_{\mathrm{Z}}$ denote the indicator random variable for Z.

Note that by the fact that each vertex is picked to S_1 (respectively S_2) independently w.p p we can apply Lemma A.1 and conclude that with very high probability $|\mathrm{S}_1| = pn \pm \sqrt{pn\log n} = \Theta(pn)$

By the independence of S_1 and S_2, and an application of Jensen's inequality, we can establish the following bound:

$$\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v\in\mathrm{S}_1}}{|\mathrm{S}_1|}\frac{\mathbf{1}_{u\in\mathrm{S}_2}}{\alpha_v|N(v)\cap\mathrm{S}_2|}\right] = \mathop{\mathbf{E}}_{\mathrm{S}_1}\left[\frac{\mathbf{1}_{v\in\mathrm{S}_1}}{|\mathrm{S}_1|}\right]\mathop{\mathbf{E}}_{\mathrm{S}_2}\left[\frac{\mathbf{1}_{N(v)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{u\in\mathrm{S}_2}}{\alpha_v|N(v)\cap\mathrm{S}_2|}\right] \ge \Omega\left(\frac{\alpha_v}{n\alpha_v|N(v)|}\right) = \Omega\left(\frac{1}{n|N(v)|}\right),$$

where we used the fact that, $\mathbf{Pr}[\mathbf{1}_{N(v)\cap\mathrm{S}_2\neq\emptyset}] = \alpha_v$, leading to mean of $\Omega\left(\frac{1}{n}\sum_{(v,u)\in E}\frac{\tilde{A}_{v,u}\tilde{x}_u}{|N(v)|}\right)$.

Next, we compute the second moment of our estimator.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|^2}\sum_{v_1,v_2\in\mathrm{S}_1}\chi_{v_1}\chi_{v_2}\right] \\
&\quad=\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|^2}\sum_{(v_1,u_1),(v_2,u_2)\in E}\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\cdot\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{\alpha_{v_1}|N(v_1)\cap\mathrm{S}_2|\alpha_{v_2}|N(v_2)\cap\mathrm{S}_2|}\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_2,u_2}\tilde{x}_{u_2}\right] \\
&\quad=\sum_{(v_1,u_1),(v_2,u_2)\in E}\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_1|^2\cdot\alpha_{v_1}\alpha_{v_2}|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right]\cdot\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_2,u_2}\tilde{x}_{u_2}.
\end{aligned}$$

Similarly to before, we inspect the above expectation. In here, there are four cases corresponding to the following sets.

1. $\mathcal{C}_1=\{(v_1,u_1),(v_2,u_2)\in E\mid v_1\neq v_2,u_1\neq u_2\}$: in a similar way to before, by Lemma A.1, we obtain

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_1|^2\cdot\alpha_{v_1}\alpha_{v_2}|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right] \\
&\quad=\frac{p^2}{\alpha_{v_1}\alpha_{v_2}(pn)^2}\mathop{\mathbf{E}}_{\mathrm{S}_2}\left[\frac{\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}}{|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right] \\
&\quad=\frac{p^2q^2}{(pn)^2}\mathop{\mathbf{E}}_{\mathrm{S}_2}\left[\frac{1}{|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right],
\end{aligned}$$

   where we used the fact that for v_i, $\mathbf{Pr}[\mathbf{1}_{N(v_i)\cap\mathrm{S}_2\neq\emptyset}]=\alpha_{v_i}$.
   In order to analyze the above expectation, we use our assumption that neighborhoods are uncorrelated to get

$$\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_1|^2\cdot\alpha_{v_1}\alpha_{v_2}|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right]\lesssim\frac{1}{n^2|N(v_1)|\cdot|N(v_2)|}.$$

2. $\mathcal{C}_2=\{(v_1,u_1),(v_2,u_2)\in E\mid v_1\neq v_2,u_1=u_2\}$: similarly to the previous case,

$$\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_1|^2\cdot\alpha_{v_1}\alpha_{v_2}|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right]\lesssim\frac{1}{qn^2|N(v_1)||N(v_2)|}.$$

3. $\mathcal{C}_3=\{(v_1,u_1),(v_2,u_2)\in E\mid v_1=v_2,u_1\neq u_2\}$. This case requires extra care, since in this case, the neighborhoods of v_1 and v_2 are correlated (actually the same).

$$\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_1|^2\cdot\alpha_{v_1}\alpha_{v_2}|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right]\simeq\frac{pq^2}{\alpha_v(pn)^2}\mathbf{E}\left[\frac{1}{|N(v)\cap\mathrm{S}_2|^2}\right].$$

   By Lemma A.1, with very high probability

$$|N(v)\cap\mathrm{S}_2|\in[q|N(v)|-\sqrt{q|N(v)|\log n},q|N(v)|+\sqrt{q|N(v)|\log n}],$$

   and by our constraint that $q=\Omega(\log n/d_{\min})=\Omega(\log n/|N(v)|)$, we have that with high probability

$$\mathop{\mathbf{E}}_{\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_2\cap N(v)|^2}\right]\simeq\frac{1}{(q|N(v)|\pm\sqrt{q|N(v)|\log n})^2}\lesssim O\left(\frac{1}{q^2|N(v)|^2}\right).$$

   so that

$$\frac{pq^2}{\alpha_v(pn)^2}\mathbf{E}\left[\frac{1}{|N(v)\cap\mathrm{S}_2|^2}\right]\lesssim\frac{1}{pn^2\alpha_v|N(v)|^2}$$

4. $\mathcal{C}_4=\{(v_1,u_1),(v_2,u_2)\in E\mid v_1=v_2,u_1=u_2\}$. Similarly to the previous case,

$$\mathop{\mathbf{E}}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{\mathbf{1}_{v_1,v_2\in\mathrm{S}_1}\cdot\mathbf{1}_{u_1,u_2\in\mathrm{S}_2}\mathbf{1}_{N(v_1)\cap\mathrm{S}_2\neq\emptyset}\mathbf{1}_{N(v_2)\cap\mathrm{S}_2\neq\emptyset}}{|\mathrm{S}_1|^2\cdot\alpha_{v_1}\alpha_{v_2}|N(v_1)\cap\mathrm{S}_2||N(v_2)\cap\mathrm{S}_2|}\right]=\frac{pq}{\alpha_v(pn)^2}\mathbf{E}\left[\frac{1}{|N(v)\cap\mathrm{S}|^2}\right]\lesssim\frac{1}{n^2\alpha_vpq|N(v)|^2}.$$

Combining the above and subtracting the expectation squared yields,

$$\mathrm{Var}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|}\sum_{v\in\mathrm{S}_1}\chi_v\right] \lesssim \frac{1}{n^2}\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\frac{\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2}{q|N(v_1)||N(v_2)|}$$
$$+\frac{1}{n^2}\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\frac{\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}}{p\alpha_v|N(v)|^2}+\frac{1}{n^2}\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_4}\frac{\tilde{A}_{v,u}^2x_{u_1}^2}{pq\alpha_v|N(v)|^2}-\left(\frac{1}{n}\sum_{(v,u)\in E}\frac{\tilde{A}_{v,u}\tilde{x}_u}{|N(v)|}\right)^2$$

After rearranging we get

$$\frac{1}{n^2}\Bigg(\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\left(\frac{1}{q|N(v_1)||N(v_2)|}-\frac{1}{|N(v_1)||N(v_2)|}\right)\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2$$
$$+\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\left(\frac{1}{p\alpha_v|N(v)|^2}-\frac{1}{|N(v)|^2}\right)\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}+\sum_{(v,u)\in E}\left(\frac{1}{pq\alpha_v|N(v)|^2}-\frac{1}{|N(v)|^2}\right)\tilde{A}_{v,u}^2\tilde{x}_u^2\Bigg).$$

If we let $p=m_1/n$ and $q=m_2/n$ so that $\mathbf{E}_{\mathrm{S}_1}[|\mathrm{S}_1|]=m_1$ and $\mathbf{E}_{\mathrm{S}_2}[|\mathrm{S}_2|]=m_2$, we get that

$$\mathrm{Var}_{\mathrm{S}_1,\mathrm{S}_2}\left[\frac{1}{|\mathrm{S}_1|}\sum_{v\in\mathrm{S}_1}\chi_v\right]$$
$$\lesssim\frac{1}{n^2}\Bigg(\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\frac{\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2}{q|N(v_1)||N(v_2)|}+\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\frac{\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}}{p\alpha_v|N(v)|^2}+\sum_{(v,u)\in E}\frac{\tilde{A}_{v,u}^2\tilde{x}_u^2}{pq\alpha_v|N(v)|^2}\Bigg)$$
$$\lesssim\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\left(\frac{1}{nm_2}\right)\frac{\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2}{|N(v_1)||N(v_2)|}+\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\left(\frac{1}{m_1\cdot n}\right)\frac{\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}}{\alpha_v|N(v)|^2}$$
$$+\sum_{(v,u)\in E}\left(\frac{1}{m_1\cdot m_2}\right)\frac{\tilde{A}_{v,u}^2\tilde{x}_u^2}{\alpha_v|N(v)|^2}.$$

Note that the variance decreases as m_1, m_2 increase. Therefore, if we assume for simplicity that $m_1=m_2=m$ the variance is bounded by

$$\mathrm{Var}_{\mathrm{S}_1,\mathrm{S}_2}[\boldsymbol{\xi}]\lesssim\frac{1}{nm}\Bigg(\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\frac{\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2}{|N(v_1)||N(v_2)|}+\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\frac{\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}}{\alpha_v|N(v)|^2}\Bigg)$$
$$+\sum_{(v,u)\in E}\left(\frac{1}{m^2}\right)\frac{\tilde{A}_{v,u}^2\tilde{x}_u^2}{\alpha_v|N(v)|^2}. \tag{10}$$

Let's consider the pseudo precision of our estimator. Note that the cost generating the estimator ξ is approximately $2m\cdot\bar{d}$, where $\bar{d}$ is the average degree of the graph. This is since we have roughly $2m$ vertices to generate and for each vertex sampled we need to aggregate its neighbors information (which is approximately $\bar{d}$).

$$\rho(\xi)=(\mathrm{Var}(\xi)\cdot\mathrm{Cost}(\xi))^{-1}\gtrsim\Bigg(\frac{2\bar{d}}{n}\Bigg(\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\frac{\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2}{|N(v_1)||N(v_2)|}+\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\frac{\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}}{\alpha_v|N(v)|^2}\Bigg)$$
$$+\sum_{(v,u)\in E}\left(\frac{2\bar{d}}{m}\right)\frac{\tilde{A}_{v,u}^2\tilde{x}_u^2}{\alpha_v|N(v)|^2}\Bigg)^{-1}. \tag{11}$$

Let

$$\phi = \frac{2\bar{d}}{n}\left(\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_2}\frac{\tilde{A}_{v_1,u_1}\tilde{A}_{v_2,u_1}\tilde{x}_{u_1}^2}{|N(v_1)||N(v_2)|}+\sum_{(v_1,u_1),(v_2,u_2)\in\mathcal{C}_3}\frac{\tilde{A}_{v_1,u_1}\tilde{x}_{u_1}\tilde{A}_{v_1,u_2}\tilde{x}_{u_2}}{\alpha_v|N(v)|^2}\right),$$

and note that as m increase, the efficiency converges to ϕ.

If we define δ as

$$\delta(m) = \frac{2\bar{d}\sum_{(v,u)\in E}\frac{\tilde{A}_{v,u}^2\tilde{x}_u^2}{\alpha_v|N(v)|^2}}{m\phi}, \tag{12}$$

we get that $\rho(\xi) \geq \frac{1}{\phi(1+\delta)}$, as claimed. ■

Now we show that if we assume that the graph is d-regular, we can get a clean relation between the efficiency of our estimator and the size of the batch.

Proof of Corollary 4.2: Fix any $\delta > 0$. By Equation (12),

$$m = \frac{2\bar{d}\sum_{(v,u)\in E}\frac{\tilde{A}_{v,u}^2\tilde{x}_u^2}{\alpha_v|N(v)|^2}}{\delta\phi}$$

so that the pseudo precision is at least

$$\rho(\xi) \geq \left(\phi(1+\delta(m))\right)^{-1}.$$

By assuming that the graph is d-regular, and using the fact that for all $v \in V$, $1 - e^{-qd} \leq \alpha_v \leq qd$ and $qd > 1$,

$$m = O\left(\frac{1}{\delta}\cdot\frac{n\cdot\left(\frac{1}{1-e^{-qd}}\frac{nd}{d^2}\right)}{\frac{1}{d^2}n\cdot\binom{d}{2}+\frac{1}{dq}\frac{1}{d^2}n\binom{d}{2}}\right) = O\left(\frac{1}{\delta}\cdot\frac{n/d\cdot\frac{1}{1-e^{-qd}}}{\frac{1}{2}+\frac{1}{2dq}}\right) = O\left(\frac{n}{\delta d}\right).$$

■

Remark on graph-wise sampling: Note that the case where $\mathrm{S}_1 = \mathrm{S}_2$ is equivalent to subgraph sampling. The same analysis can be adapted to subgraph sampling, and yields the same proposed minibatch size.