Appendix

Appendix A: Missing Proofs

Lemma 1. *With probability at least $1-\frac{\beta}{k}$, we have $R_t \cap (C^* \setminus C_t) \neq \varnothing$ at t-th iteration.*

Proof. It is clearly that $\Pr\left[R_t \cap (C^* \setminus C_t) = \varnothing\right] = 0$ if $|R_t| = |V_t|$. Otherwise, we have

$$\begin{aligned} \Pr\left[R_t \cap (C^* \setminus C_t) = \varnothing\right] v &= \left(1 - \frac{|C^* \setminus C_t|}{|V \setminus C_t|}\right)^{|R_t|} \\ \leqslant e^{(-\frac{k-t+1}{|V_t|-t+1}\frac{|V_t|-t+1}{k-t+1}\ln\frac{k}{\beta})} &= \frac{\beta}{k} \end{aligned}$$

□

Theorem 4. *For k-median problem, the Algorithm 2 with probability $(1-2\beta)(1-2\exp(\frac{-|U|\eta^2}{2}))$, returns a solution C such that $F(C) \geqslant (1-\frac{1}{e})F(C^*) - \frac{4k^2}{\epsilon}\ln\frac{k|V|}{\beta} - (2-\frac{1}{e})\eta$ by evaluating F at most $O(|V|\ln k \ln\frac{k}{\beta})$ times. Moreover, this algorithm preserves ϵ-differential privacy.*

Proof. Similar to the proof of the Theorem 3, Algorithm 2 is ϵ-differentially private and we have

$$F_U(C^*) - F_U(C_t) \leqslant (1-\frac{1}{k})\left[F_U(C^*) - F_U(C_{t-1})\right] + \alpha \quad (7)$$

at each iteration with probability $1-\frac{\beta}{k}$ when $|R_t| = |V_t|$. However, R_t is equally likely to contain each element of $C^* \setminus C_t$, we have $\Pr\left[R_t \cap (C^* \setminus C_t) \neq \varnothing\right] = 1 - \frac{\beta}{k}$. Hence the Equation 7 would be hold with probability $(1-\frac{\beta}{k})^2$. Let C_U^* denote the optimal solution for the funtion F_U. After k iterations, Algorithm 2 will return $C = C_k$ with quality at least $F_U(C) \geqslant (1-\frac{1}{e})F_U(C_U^*) - \frac{4k^2}{\epsilon}\ln\frac{k|V_t|}{\beta}$ with probability at least $p = 1-2\beta$. Next we state the derivation of p. By a union bound over $t \in [k]$, we have

$$\begin{aligned} p &= 1 - k(1-(1-\frac{\beta}{k})^2) = 1 - k(\frac{2\beta}{k} - \frac{\beta^2}{k^2}) \\ &= 1 - 2\beta + \frac{\beta^2}{k} \geqslant 1 - 2\beta. \end{aligned}$$

According to the Proposition 1, with probability $(1-2\beta)(1-2\exp(\frac{-|U|\eta^2}{2}))$, we have

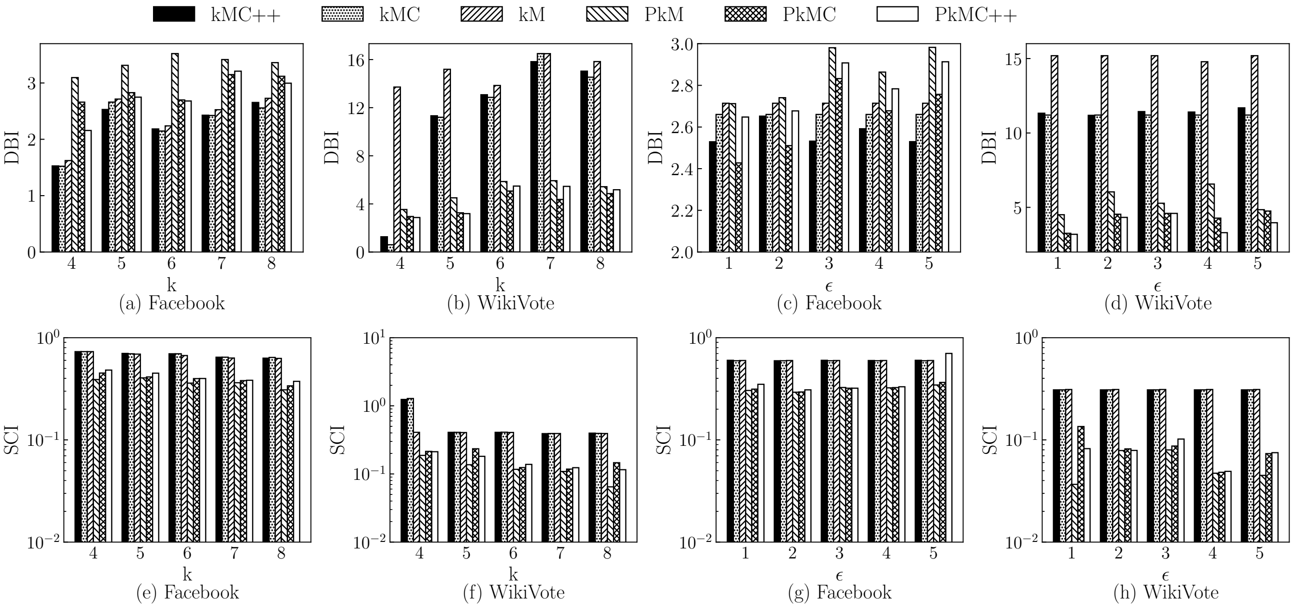
$$\begin{aligned} F(C) \geqslant F_U(C) - \eta &\geqslant (1-\frac{1}{e})F_U(C_U^*) - \frac{4k^2}{\epsilon}\ln\frac{k|V_t|}{\beta} - \eta \\ &\geqslant (1-\frac{1}{e})F_U(C^*) - \frac{4k^2}{\epsilon}\ln\frac{k|V_t|}{\beta} - \eta \\ &\geqslant (1-\frac{1}{e})(F(C^*) - \eta) - \frac{4k^2}{\epsilon}\ln\frac{k|V_t|}{\beta} - \eta \\ &= (1-\frac{1}{e})F(C^*) - \frac{4k^2}{\epsilon}\ln\frac{k|V_t|}{\beta} - (2-\frac{1}{e})\eta \end{aligned}$$

where the first and fourth inequalities used Proposition 1, the second inequality used Equation 7, the third inequality is because C_U^* is the optimal solution for the function F_U.

Finally, we focus on analyzing the number of evaluations of function F, it is at most

$$\begin{aligned} \sum_{t=1}^{k}\frac{|V_t|-t+1}{k-t+1}\ln\frac{k}{\beta} &\leqslant \sum_{t=1}^{k}\frac{|V|-t+1}{k-t+1}\ln\frac{k}{\beta} \\ &= \sum_{t=1}^{k}\frac{|V|-k+t}{t}\ln\frac{k}{\beta} = O(|V|\ln k\ln\frac{k}{\beta}) \end{aligned}$$

The theorem follows. □



(a) Facebook (b) WikiVote (c) Facebook (d) WikiVote

(e) Facebook (f) WikiVote (g) Facebook (h) WikiVote

Fig. 5. Comparison on DBI and SCI vs. ϵ and k (for k-median problem).

Appendix B: Additional Experimental Results

Evaluation Indicators. In this section, for k-median problem, we choose two cluster validation indices Davis Bouldin index (DBI) and Silhouette coefficient index (SCI) to compare the performance of the algorithms. Let d_i denote the average distance of all vertices in the i-th cluster to the cluster center c_i. Let $a(v)$ denote the average distance between vertex v and all other vertices within the same cluster. Let $b(v)$ denote the smallest average distance of vertex v to all vertices in any other cluster. Thus, the index expression is as follows.

- DBI: $DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\frac{d_i+d_j}{d_{c_i,c_j}}$, the lower value indicates the better clustering.
- SCI: $SCI = \frac{1}{|V|}\sum_{v\in V}\frac{b(v)-a(v)}{\max(a(v),b(v))}$. the higher value indicates the better clustering.

Result for k-median problem. Fig. 5 (a), (b), (e), (f) show the changes of DBI and SCI when varying the number k. Figure 5 (c), (d), (g), (h) show the changes of DBI and SCI when varying privacy budget ϵ. Notice that some value of SCI for WikiVote dataset is negative, we add 0.4(for (e) and (f)) or 0.3(for (g) and (h)) to each value in order to facilitate observation. Regarding DBI, kMC(PkMC) and kMC++(PkMC++) are almost lower than kM(PkM) in general, which means that our algorithms can achieve better clustering utility. Similarly, our algorithm obtains better results on the evaluation of SCI. The higher SCI is, the better clustering is. Therefore, by considering the equivalent problem of the original k-median problem, we practically prove that our method achieves better results than the original problem, and simultaneously our solution has an approximate ratio guarantee.