6 SUPPLEMENTARY MATERIAL

6.1 Proof of Theorem 3.5

We need Lemma 1 of [1] which characterizes a bound for the query complexity of the random sampling phase. We first define a *witness set* of the cut set C as the node set that contains at least one node for each $V_i, i \in [T]$.

Lemma 6.1 (Lemma 1 in [1])**.** Consider a β-balancedness graph $G = (V, E)$. For all $\alpha > 0$, a subset W chosen uniformly at random is a witness of the cut-set C with probability at least $1-\alpha$ as long as

$$|W| \geq \frac{\log\left(\frac{1}{\beta\alpha}\right)}{\log\left(1/(1-\beta)\right)}$$

Moreover, we will need the following lemma. Basically it ensures that once HS^2-point discovers a cut hyperedge from a cut component, then HS^2-point will discover all remaining cut hyperedges in this cut component and the shortest pathes that include these hyperedges are at most with length κ.

Lemma 6.2. Suppose a hypergraph G with a cut set C is κ-clustered. Moreover, suppose C_{rs} is a cut component. If $e \in C_{rs}$ is discovered, which means a pair of nodes $u \in \Omega_r(e), v \in \Omega_s(e)$ are labeled, then at least one remaining cut hyperedge in C_{rs} lies in a path of length at most κ from a pair of nodes with labels r and s respectively.

Proof. By definition 3.4, we know that the hyperedges in C_{rs} will form a strongly connected component in H_κ. This means for any $e \in C_{rs}$, there is at least one $e' \in C_{rs}$ such that the arc ee' exists in H_κ. Recall there exists arc ee' in H_κ if and only if $\Delta(e, e') \leq \kappa$. By definition 3.3 this means for any node pair $u \in \Omega_r(e), v \in \Omega_s(e)$, the length of the shortest path including e' but excluding e will be less then κ. Note that in the definition of $\Delta(e, e')$, we use the supremum taking over the node set $\Omega_r(e), \Omega_s(e)$. This is because it ensures that no matter which node pair $u, v \in e$ we have, $\Delta(e, e')$ can always upper bound the length of the shortest path including e' but excluding e with endpoints u, v. In contrast, we use the infimum taking over the node set $\Omega_r(e'), \Omega_s(e')$. This is because it only needs to search for the shortest path. Hence once we find a cut hyperedge e in C_{rs}, we are guarantee to find at least one cut hyperedge $e' \in C_{rs}$ through a path of length $l_S \leq \kappa$ after we remove e. □

Now let us prove Theorem 3.5. The proof uses a similar outline as the proof given in [1]. However, we need to take care of the hypergraph structures that are described by the definition 3.1 to 3.4. We will also derive a tighter bound for the number of runs R, which finally yields a lower query complexity than that in [1].

We can divide the query complexity in two parts which are associated with the random sampling phase and the aggressive search phase respectively. The goal of the random sampling phase is to find a witness set. By applying Lemma 6.1, we can bound from above the number of random queries needed.

For the aggressive search phase, let $l_S(G, L, f)$ be the length of the shortest path among all paths connecting nodes with different labels after we collects the labels of nodes in L. After each step of the aggressive search, the $l_S(G, L, f)$ will roughly get halved. Thus it will take no more than $\lceil \log_2 l_S(G, L, f) \rceil + 1$ steps to find a cut hyperedge. In order to bound the required number of active queries, let us split up the aggressive search phase into "runs", where each run ends when a new boundary node has been discovered. Let R be the number of runs, and it's obvious that $R \leq |\partial C|$ since we will at most discovere all the boundary nodes, which is $|\partial C|$. Moreover, we also have $R \leq |C|$. This is because we will discover a new boundary node if and only if we discover at least a cut hyperedge. Hence together we have $R \leq \min(|C|, |\partial C|)$. The observation of $R \leq |C|$ is missed by [1]. However, this part is extremely important for the hypergraph setting according to the later discussion in Remark 6.1.

For each $i \in [R]$, let $G^{(i)}$ and $L^{(i)}$ be the graph and label set up to run i. Then the total number of active queries can be upper bounded by $\sum_{i=1}^{R}(\lceil \log_2(l_S(G^{(i)}, L^{(i)}, f)) \rceil + 1)$. By observation, in each run it is trivial $l_S(G^{(i)}, L^{(i)}, f) \leq n$. From Lemma 6.2, once we discover a cut hyperedge from C_{rs}, we're able to find at least one undiscovered cut hyperedge from C_{rs} through a path of length at most κ according to Lemma 6.2. After running $|C_{rs}|-1$ times, we may fully discover C_{rs}. In all these $|C_{rs}|-1$ runs in R, $l_S \leq \kappa$. In all, the runs that we first discover each cut components are long runs, whose l_S can be upper bounded naively by n, and the number of long runs is not greater than m. Once we discover the first cut hyperedge in C_{rs}, the rest $|C_{rs}|-1$ runs are short runs whose l_S can be upper bounded by κ. Therefore, we have

$$\begin{aligned}
&\sum_{i=1}^{R}(\lceil \log_2(l_S(G^{(i)}, L^{(i)}, f)) \rceil + 1) \\
&\leq (R + m\lceil \log_2 n \rceil + (R-m)\lceil \log_2 \kappa \rceil) \\
&\leq m(\lceil \log_2 n \rceil - \lceil \log_2 \kappa \rceil) + \min(|C|, |\partial C|)(\lceil \log_2 \kappa \rceil + 1)
\end{aligned}$$

Hence we complete the proof.

Remark 6.1. Note that in [1] they only use the bound $R \leq |\partial C|$ and miss the bound $R \leq |C|$. As they focused on standard graphs, $|C|$ can be lower bounded

by $\frac{|\partial C|}{2}$. Therefore, in standard graphs, the bound $R \leq |\partial C|$ will at most loose by a constant factor of 2. However, in hypergraphs, it's possible that $|C|$ is substantially smaller than $|\partial C|$, when the sizes of hyperedges are large. So $R \leq |C|$ is crucial for the tight analysis in the hypergraph scenario.

6.2 Proof of Proposition 3.6

6.2.1 checking the equal parameters

We check the parameters one by one.

We start from proving that if C is κ-clustered, then $C^{(ce)}$ is also κ-clustered. We note that performing CE does not change the length of the shortest path of arbitrary node pair $v_1, v_2 \in V$. This is because CE will replace a hyperedge by a clique, which makes all nodes in the hyperedge become fully connected. Hence the $C^{(ce)}$ is still κ-clustered.

Now, we prove that if G has m non-empty cut components, then $G^{(ce)}$ will also have m non-empty cut components. We note that for any non-empty cut component $C_{i,j}$ in G, there is at least one hyperedge $e \in C_{i,j}$. By definition, we know that $e \cap V_i \neq \emptyset$ and $e \cap V_j \neq \emptyset$. So after CE, in the clique corresponding to this hyperedge e, there must be at least one edge such that one of its endpoint is from V_i and the other one is from V_j, which makes $C_{i,j}$ still non-empty in $G^{(ce)}$. On the other hand, for arbitrary i, j, the cut component $C_{i,j}$ is empty in G if and only if there is no hyperedge between V_i, V_j. Hence $C_{i,j}$ will still be empty in $G^{(ce)}$. Together we show that if there are m non-empty cut components in G, there are exactly m non-empty cut components in $G^{(ce)}$.

It is easy to see $G^{(ce)}$ keep β-balanced as f does not change in CE.

Now, we prove that $|\partial C| = |\partial C^{(ce)}|$. For any $e \in C$, let's denote $e = \{v_1, ..., v_d\}$. By definition we know that $v_1, ..., v_d \in \partial C$. Suppose $e \in C$ and the nodes $v_1, ..., v_d$ can be partitioned into t non-empty set $S_1, ..., S_t$ according to their labels. Without loss of generality, let $v_1 \in S_1$. Then after CE of e we know that the edges $(v_1, v), v \in S_j, j \in \{2, 3, ..., t\}$ will be in the set $C^{(ce)}$. By definition of $C^{(ce)}$, we know that all $v \in S_j, j \in \{2, 3, ..., t\}$ will be in the cut set $\partial C^{(ce)}$. We can repeat the same argument for all nodes in S_1 and know that $S_1 \subset \partial C^{(ce)}$. In the end, we can show that $\forall v \in e$, $v \in \partial C^{(ce)}$. By definition we also have $\forall v \in e$, $v \in \partial C$. Therefore, we claim that $\partial C = \partial C^{(ce)}$ which furthermore $|\partial C| = |\partial C^{(ce)}|$.

6.2.2 proof for the inequality

Now, we prove that $\min(|C|, |\partial C|) \leq \min(|C^{(ce)}|, |\partial C^{(ce)}|)$. As above, we have proved $|\partial C| = |\partial C^{(ce)}|$. The case when $|\partial C^{(ce)}| \leq |C^{(ce)}|$ is an easy case. So, we only need to prove for the case when $|\partial C^{(ce)}| > |C^{(ce)}|$. We claim that if $|\partial C^{(ce)}| > |C^{(ce)}|$, then $|C| \leq |C^{(ce)}|$, which is proved as follows.

Let us first introduce an auxiliary graph G' that can be useful in the proof. $G' = (\partial C^{(ce)}, C^{(ce)})$ is a subgraph of $G^{(ce)}$ with the node set $\partial C^{(ce)}$ and the edge set $C^{(ce)}$.

In the following, we show that when $|\partial C^{(ce)}| > |C^{(ce)}|$, then it's impossible for G' to have any cliques of size greater or equal to 3. Note that by the definition of $C^{(ce)}$ and $\partial C^{(ce)}$, the auxiliary graph G' is connected. Moreover, as for the condition $|\partial C^{(ce)}| > |C^{(ce)}|$, we know that the average degree of G' is strictly less then 2. This is because

$$2 > \frac{2|C^{(ce)}|}{|\partial C^{(ce)}|} = \frac{\sum_{v \in \partial C^{(ce)}} d_v}{|\partial C^{(ce)}|}$$

where d_v is the degree of node v in G'. Hence it's impossible to have any cliques of sizes that are greater than or equal to 3 in G'.

By using the above observation and the definition of clique expansion, we know that when $|\partial C^{(ce)}| > |C^{(ce)}|$, all hyperedges in C are actually edges. Equivalently, we have $C = C^{(ce)}$, which implies $|C| = |C^{(ce)}| < |\partial C^{(ce)}|$. This concludes the proof.

By the end of this subsection, we would like to show that it's possible to have $\min(|C|, |\partial C|) < \min(|C^{(ce)}|, |\partial C^{(ce)}|)$ for some hypergraphs. Let C contain only one hyperedge e such that $|e| = 4$. Then it's obvious to see that $1 = |C| < |C^{(ce)}| = 6$ and $|\partial C^{(ce)}| = |\partial C| = 4$. Hence in this special example we have $\min(|C|, |\partial C|) < \min(|C^{(ce)}|, |\partial C^{(ce)}|)$.

6.3 Proof of Theorem 4.2

Before we start our proof, we need to prepare preliminary results. The first one is Theorem 3 in [10] that characterizes the theoretical performance of Algorithm 2 in [10].

Theorem 6.3 (Theorem 3 in [10])**.** Given a set of M points which can be partition into k clusters. The Algorithm 2 in [10] will return all clusters of size at least $\frac{64k \log M}{(1-2p)^4}$ with probability at least $1 - \frac{2}{M}$. The corresponding query complexity is $O(\frac{Mk^2 \log M}{(1-2p)^4})$.

Basically we use this theorem to analyze Phase 1 of Algorithm 4. The next one is a lemma that characterizes a lower bound of the KL divergence of two Bernoulli distributions.

Lemma 6.4 ([33])**.** Let us denote $D(x||y)$ be the KL divergence of two Bernoulli distributions with parameters $x, y \in [0,1]$ respectively. We have

$$D(x||y) \geq \frac{(y-x)^2}{2\min\{x,\, y\}} \tag{7}$$

Remark 6.2. Note that the bound is tighter than directly using Pinsker's inequality [34] when $y \leq 1/8$.

Now we start to prove Theorem 4.2. First we will show that Phase 1 of Algorithm 4 will return the correct partition $S_1, ..., S_k$ with high probability. From Theorem 6.3 we know that we have to ensure our sampled M points contain all underlying true clusters with size at least $O(\frac{Mk^2 \log M}{(1-2p)^4})$. Since we sample these M points uniformly at random, thus $(S_1, ..., S_k)$ is the multivariate hypergeometric random vector with parameters $(n, np_1, ..., np_k, M)$ and $\forall i,\ p_i = \frac{|\{v \in V | f(v)=i\}|}{n}$. It's well known ([35],[36]) that when $M \leq n/2$, the tail bound for the multivariate hypergeometric distribution is

$$\begin{aligned}
&\mathbb{P}(S_i \leq M(p_i - \frac{p_i}{2})) \leq \exp(-MD(\frac{p_i}{2}||p_i)) \\
&\leq \exp(\frac{-Mp_i}{8}) \\
&\Rightarrow \mathbb{P}(S_i \leq \frac{M\beta}{2}) \leq \exp(\frac{-M\beta}{8}),
\end{aligned} \tag{8}$$

where we use Lemma 6.4 for the second inequality. For the case $M \geq n/2$, we could apply trick of symmetry and have ([35],[36],[37])

$$\begin{aligned}
&\mathbb{P}(S_i \leq M(p_i - \frac{p_i}{2})) \\
&\leq \exp(-(n-M)D(p_i + \frac{p_i M}{2(n-M)}||p_i)) \\
&\leq \exp(-(n-M)\frac{(\frac{p_i M}{2(n-M)})^2}{p_i(2 + \frac{M}{n-M})}) \\
&= \exp(-\frac{p_i M^2}{4(2n-M)}) \\
&\leq \exp(-\frac{Mp_i}{12}),
\end{aligned}$$

where the second inequality is via Lemma 6.4 and the last inequality uses the assumption $M \geq n/2$. Hence, for all $M \leq n$, we have

$$\mathbb{P}(S_i \leq \frac{M\beta}{2}) \leq \exp(\frac{-M\beta}{12}) \tag{9}$$

Since we need (9) holds for all i, we apply the union bound over all k events which gives

$$\mathbb{P}(\bigcap_{i=1}^{k}\{S_i \geq \frac{M\beta}{2}\}) \geq 1 - k\exp(\frac{-M\beta}{12}) \tag{10}$$

Now, we need M is large enough such that $\frac{M\beta}{2}$ meets the requirement of Theorem 6.3. Moreover, we also need M to be large enough such that this event holds with probability at least $1 - \frac{\delta}{4}$. For the first requirement, we have

$$\frac{M\beta}{2} \geq \frac{64k \log M}{(2p-1)^4} \Rightarrow \frac{M}{\log M} \geq \frac{128k}{\beta(2p-1)^4}$$

This is exactly our first requirement on M in (4). For the high probability requirement, we have

$$k\exp(\frac{-M\beta}{12}) \leq \frac{\delta}{4} \Rightarrow M \geq \frac{12}{\beta}\log\frac{4k}{\delta}$$

This is exactly the second requirement on M in (4). Moreover, we also need Algorithm 2 of [10] successfully recover all the true clusters with probability at least $1 - \frac{\delta}{4}$, and thus we have

$$\frac{2}{M} \leq \frac{\delta}{4} \Rightarrow M \geq \frac{8}{\delta}$$

This is exactly the third requirement on M in (4).
Now assume that Algorithm 2 of [10] indeed returns all true clusters. We will analyze Phase 2. Start from assuming all S_i's are correctly clustered. Then for any new node v, from the algorithm we designed we will query for comparing v with all the M nodes that have been clustered. Before we continue, let us introduce some error events which is useful for the following analysis. Let $Er^{(i)}$ be the event that a node with label i is incorrectly clustered by the normalized majority voting. Let $Er_j^{(i)} = \{\frac{M_j}{|S_j|} > \frac{M_i}{|S_i|}\}$, for $j \neq i$, where M_j is the number of nodes in S_j that respond positively to the pairwise comparisons with node v. Note that we have $M_j \sim Bin(p, |S_j|)$ for $j \neq i$ and $M_i \sim Bin(1-p, |S_i|)$. All these M_l's are mutually independent.
We start from analyzing the normalized majority voting for the unlabeled node v. Then we have

$$\begin{aligned}
&\mathcal{O}_p(v,u) \sim Ber(1-p)\ \forall u \in S_i; \\
&\mathcal{O}_p(v,u) \sim Ber(p)\ \forall u \notin S_i
\end{aligned}$$

where we recall that $\mathcal{O}_p(x,y)$ is the query answer for the point pair (x,y) from the noisy oracle $\mathcal{O}_p$. So the error probability $\mathbb{P}(Er^{(i)})$ that we misclassify the point v can be upper bounded by

$$\mathbb{P}(Er^{(i)}) \leq (k-1)\max_{j \neq i}\mathbb{P}(Er_j)$$

where we used the union bound. Moreover, we can upper bound $\mathbb{P}(Er_j^{(i)})$ as following (recall that $p < 1/2$)

$$\begin{aligned}
\mathbb{P}(Er_j^{(i)}) &= \mathbb{P}(\frac{M_j}{|S_j|} > \frac{M_i}{|S_i|}) \\
&\leq \mathbb{P}(\frac{M_j}{|S_j|} \geq \frac{1}{2}) + \mathbb{P}(\frac{1}{2} > \frac{M_j}{|S_j|})
\end{aligned}$$

Let's denote $\lambda = \frac{1}{2} - p > 0$. So we have $\frac{1}{2} = \lambda + p = \bar{p} - \lambda$ where $\bar{p} = 1 - p$. Hence by Chernoff's bound the first term can be upper bounded by

$$\mathbb{P}(\frac{M_j}{|S_j|} \geq \frac{1}{2}) \leq \exp(-|S_j| \cdot D(p + \lambda || p))$$

and similarly the second term can be upper bounded by

$$\mathbb{P}(\frac{1}{2} > \frac{M_i}{|S_i|}) \leq \exp(-|S_i| \cdot D(\bar{p} - \lambda || \bar{p}))$$

Hence we have

$$\begin{aligned}\mathbb{P}(Er^{(i)}) \leq (k-1)[&\max_{j \neq i} \exp(-|S_j| \cdot D(p + \lambda || p)) \\ &+ \exp(-|S_i| \cdot D(\bar{p} - \lambda || \bar{p}))]\end{aligned}$$

Recall that from (10), we have $\min_{i \in [k]} |S_i| \geq \frac{M\beta}{2}$ with probability at least $1 - \frac{\delta}{4}$. Moreover, we observe that $D(0.5||p) = \min\{D(p + \lambda||p), D(\bar{p} - \lambda||\bar{p})\}$ by the symmetry of KL-divergence for Bernoulli distribution. Thus, the error probability for any new point can be upper bounded as

$$\mathbb{P}(Er) \leq \max_i \mathbb{P}(Er^{(i)}) \leq 2(k-1) \exp(\frac{-M\beta D(0.5||p)}{2})$$

Note that from Theorem 3.5 we will need to query $\mathcal{Q}^*(\frac{\delta}{4})$ nodes in the aggressive search Phase if we want the exact result holds for probability at least $1 - \frac{\delta}{4}$ in noiseless case. Hence by using the union bound, the error probability for exact recovery of these $\mathcal{Q}^*(\frac{\delta}{4})$ points is upper bounded by

$$2\mathcal{Q}^*(\frac{\delta}{4})(k-1) \exp(\frac{-M\beta D(0.5||p)}{2})$$

Requiring this to be smaller than $\frac{\delta}{4}$, then we have

$$M \geq \frac{2}{\beta D(0.5||p)} \log(\frac{8(k-1)\mathcal{Q}^*(\frac{\delta}{4})}{\delta})$$

This is exactly the forth requirement on M in (4). Further, via the union bound, the overall algorithm will succeed with probability at least $1 - \delta$. Note that if we have exact recovery on these $\mathcal{Q}^*(\frac{\delta}{4})$ nodes, then we can indeed find the cut set C by Theorem 3.5, which concludes the proof.