Supplemental material: Maximizing influence in an unknown social network

Contents

1 Theoretical analysis of ARISEN

In this section, we present proofs of our guarantees for the performance of ARISEN.

1.1 Preliminaries

We study influence maximization using local information on a graph drawn from the stochastic block model (SBM). There is a fixed vertex set V, where $|V|= n$ is known to the algorithm. The vertices are partitioned into communities $C_1...C_L$ where each $C_i \subseteq V$. We assume that the communities are ordered as $|C_1|\geq |C_2|\geq .... \geq |C_L|$ The set of edges is sampled according to the following process:

1. Each edge (u, v) where u and v belong to the same community is independently present with probability p_w.
2. Each edge (u, v) where u and v belong to different communities is independently present with probability p_b.

Influence propagates according to the independent cascade model (ICM) where each edge has equal propagation probability q. This process can be viewed as follows. Each edge in the graph is independently kept with probability q and discarded with probability $1 - q$. Then, a node is influenced by a given seed set if it lies in the same connected component as a seed. The intuition for this view (which originated with Kempe et al. [5]) is that flipping all of the process's random coins in advance is equivalent to flipping them one at a time, as each node is influenced. The SBM and ICM can thus been seen as jointly inducing a graph where each within-community edge is present with probability $p_w q$ and each between-community edge is present with probability $p_b q$.

The algorithm has a budget of K nodes which it may select as seeds. We assume without loss of generality that $L \geq K$. If $L < K$, then all claims follow by analyzing the expected utility on $C_1...C_L$ instead of $C_1...C_K$. Let $f_E(S)$ give the expected number of nodes influenced in the independent cascade model by the set of nodes S when the set of realized edges are E. Let $OPT(E)$ give the greatest influence spread using any subset of K nodes when the realized edges are E. Our algorithm is denoted by $\mathcal{A}$; the set containing its selections given edge set E is denoted by $\mathcal{A}(E)$. Note that since $\mathcal{A}$ is randomized, $\mathcal{A}(E)$ is itself a random variable. We aim to prove that

$$\mathbb{E}[f_E(\mathcal{A}(E))] \geq \alpha\, \mathbb{E}[OPT(E)]$$

for some approximation ratio α. The expectations range over the randomness in the realization of E and the decisions of $\mathcal{A}$. Let $OPT = \mathbb{E}[OPT(E)]$.

We now state some facts about Erdős-Rényi random graphs which will be useful for our analysis. The following lemma can be found in any reference on random graphs (see, e.g. Janson, Luczak, and Rucinski [4]):

Lemma 1. *Let $\mathcal{G}(n, p)$ be the Erdős-Rényi graph on n vertices with connection probability p.*

- *If $np > \log n$, then with probability $1 - o(1)$, the graph is connected.*
- *If $1 < np < \log n$, then with probability $1 - o(1)$ the largest connected component has size $(1+o(1))\beta n$, where β is the unique solution to the equation $\beta - 1 + \exp(-\beta np) = 0$. All other components have size $O(\log n)$.*

- *If $np < 1$, then with probability $1 - o(1)$ the largest connected component has size $O(\log n)$.*

In our case, each community is internally an Erdős-Rényi random graph with size $|C_i|$ and connection probability p_w. The portion of each community with is internally connected under both the SBM and the ICM is the giant connected component of an Erdős-Rényi random graph with size $|C_i|$ and connection probability $p_w q$. With a slight abuse of notation, we use the function $\beta(x)$ to refer to the size of the giant connected component induced by the SBM/ICM in a community with size x. We impose the following:

Assumption 1. $p_w = O\left(\frac{\log n}{n}\right)$, and for all for all communities $|C_i|$, $p_w > \frac{\log|C_i|}{|C_i|}$. Intuitively, the subgraph formed by each community should be connected, but the graph is still relatively sparse. Our analysis can be extended to the dense case (e.g., $p_w = \Theta(1)$), but this is not the situation of interest for real world networks.

For the influence process, we require

Assumption 2. For all communities $|C_i|$, $p_w q|C_i| > 1$. This requires that the ICM and SBM jointly induce a giant connected component in each community, i.e., that an influence cascade can reach a linear portion of the community.

We also require that no community is too small. The technical condition we require is in terms of parameters ϵ and ρ which are introduced in the definition of our algorithm.

Assumption 3. For all communities C_i, $|C_i|(\epsilon^5\rho) = poly(n)$. As a corollary, this implies that all communities have size which scales polynomially (though possibly sublinearly) with n.

1.2 Summary

ARISEN and its motivation

The idea behind ARISEN is to improve on naive sampling by estimating the size of the community that each random sample lies in, and then choose the largest communities for seeding. Each community C_i is an Erdős-Rényi graph which has average degree $d_i = |C_i|p_w + (n - |C_i|)p_b$. An estimate of d_i, combined with knowledge of p_w and p_b, yields an estimate of $|C_i|$. d_i can be estimated by simulating a series of random walk through the community to obtain a sampled set of degrees.

Having estimated the size of the community that each sampled node lies in, a natural approach would be to choose the K samples with the largest estimated size as seed nodes. However, this idea fails because there is no way to tell (using only local information) whether two sampled nodes lie in the same community: they might lie in different communities which have very similar average degree. Hence, simply choosing the samples with the largest estimated size might just seed the same community K times, which gives an approximation ratio no better than $\frac{1}{K}$ in the worst case.

The idea that we use to overcome this issue is to independently choose each sample as a seed with probability *inversely* proportional to its size. Since large communities are sampled more often, this inverse weighting "evens out" the sampling bias towards large communities and ensures that, in expectation, each of the top K communities is seeded exactly once.

For reference, we recall in Algorithm 1 the algorithm that we will prove our guarantees for. This runs the ARISEN algorithm using only the InitializeWeights routine. As discussed later, RefineWeights can only improve ARISEN's influence (though it is not guaranteed to do so). Algorithm 1 divides the execution of ARISEN into four steps, which we will refer to during our analysis.

Algorithm 1 ARISEN (only InitializeWeights)

Require: R, T, B, K, n, p_w, p_b

Step 1:

for $i = 1...T$ **do**
  Sample v_i uniformly random from G.
  $H_i = R$ nodes on a random walk from v_i.
end for
for $i = 1...T$ **do**
  Form H_i' by discarding the first B nodes of H_i and keeping each remaining node v_j w.p. $\frac{1}{d(v_j)}$
  $\hat{d} = \frac{1}{R}\sum_{u \in H_i'} d(u)$
  $\hat{S}_i = \frac{\hat{d} - p_b n}{p_w - p_b}$
end for

Step 2:

$w_j = \frac{n}{\hat{S}_j T},\ j = 1...T.$

Step 3:

$\tau = \max\{\hat{S}_j | \sum_{\{i|\hat{S}_i \geq \hat{S}_j\}} w_i \geq K\}$
For any j with $\hat{S}_j < \tau$, set $w_j = 0$.

Step 4:

Sample $u_1...u_K \overset{iid}{\sim} \boldsymbol{w}$
return $u_1...u_K$

Proof overview: $p_b = 0$

We start out by proving our guarantee for $p_b = 0$, i.e., when the graph consists of disconnected communities. The idea is to prove that the number of random samples taken in Step 1 is sufficiently large that every community with size $|C_i| \geq \rho\epsilon n$ will be sampled $(1 \pm \epsilon)T\frac{|C_i|}{n}$ times, and that each time it is sampled, its estimated size $\hat{S}$ will be within a multiplicative ϵ of the true value $|C_i|$. This ensures that the top K communities will be assigned total weight close to 1 in Steps (2)-(3). Each one is hit with probability approximately $1 - (1 - \frac{1}{K})^K \geq 1 - 1/e$, and a random sample within the community lands in the giant connected component induced by the ICM with probability $\beta(|C_i|)$. Let $\beta_{max} = \beta(|C_1|)$ and $\beta_{min} = \beta((1-\epsilon)|C_K|)$. The above intuition motivates the approximation ratio $\frac{\beta_{min}^2}{\beta_{max}}(1 - 1/e)$, which the guarantee in the theorem converges to. The major challenge is to control the effects of sampling error. While standard concentration bounds suffice to show that the estimates taken in Step 1 are accurate to within relative error ϵ, the weights are truncated in Step 3. This has the potential to amplify small errors in sampling, so the bulk of the analysis is spent in ensuring that the total utility remains close to $\frac{\beta_{min}^2}{\beta_{max}}(1 - 1/e)$ after Step 3.

Proof overview: $p_b > 0$

There are two major modifications which need to be made to the proof when there are edges between communities. Here, we outline the obstacles and the steps that we take to resolve them.

First, Step 1 uses a random walk to collect samples from a given community. These samples are used to estimate the community's size in Step 1d. When there are edges between communities, the random walk may exit the community at some point, in which case the community's size will not be accurately estimated. However, we show (via connecting the behavior of the random walk to the conductance of the community in question) that each random walk stays within its starting community with probability at least $1 - o(1)$.

Second, our simple bound on OPT (Lemma 2) no longer holds when $p_b > 0$ since OPT may utilize between-community edges to influence more nodes than just the largest K communities. We provide a new bound on OPT which accounts for the presence of these edges. The intuition is that OPT can be bounded by the combined size of the largest K connected components in a subcritical Erdős-Rényi graph in which each community forms a node. However, formalizing this intuition requires a more intricate analysis.

1.3 Disconnected communities

We start out by assuming that $p_b = 0$, so that the graph consists of series of disconnected communities. Later, we generalize the result to account for edges between communities.

We will prove the following guarantee for the performance of this algorithm. The idea is that ρ controls the "resolution" at which the algorithm works: it competes with the optimal solution provided that OPT is at least ρnK. ϵ is a precision parameter which controls the degree of error caused by sampling.

Theorem 1. *Suppose that $\rho \leq \frac{\mu}{n}$. Then ARISEN can be implemented so that its approximation ratio is at least*

$$\frac{\beta_{min}^2}{\beta_{max}}(1 - e^{-(1-\epsilon)} - \epsilon - \frac{1}{K} - o(1))$$

using $O\left(\left(\frac{1}{\epsilon^5\rho}\right)\log^3\left(\frac{1}{\epsilon\rho}\right)\log^6 n\right)$ queries.

When n and K are large, and ϵ is small, this approximation guarantee converges to $\frac{\beta_{min}^2}{\beta_{max}}(1 - 1/e)$.
Proof.

We start out by stating a simple bound on OPT:

Lemma 2. *With probability $1 - o(1)$, $OPT \leq \sum_{i=1}^{K} \beta(|C_i|)|C_i|$.*

Proof. The set of nodes influenced by OPT is upper bounded by the total size of the K largest connected components when each within-community edge is sampled with probability qp_w. Via Lemma 1, with probability $1 - o(1)$, each community C_i has a giant connected component of size $\beta(|C_i|)|C_i|$, with all other components having size $O(\log|C_i|)$. Via Assumption 3, all communities have size scaling as $poly(n)$, so for any C_i, C_j, the giant connected component in C_i is larger than the second largest connected component in C_j. Hence, the K largest connected components correspond to the giant connected components of the K largest communities. □

All of our analysis will focus on communities which are sufficiently large, measured in terms of the parameters ρ and ϵ. We start out by noting that

Lemma 3. $\sum_{\{C_i|\ |C_i|\geq\epsilon\rho n\}}|C_i| \geq (1-\epsilon)\sum_{i=1}^{K}|C_i|$.

Proof. From the assumption that $\rho \leq \frac{\mu}{n}$, at most $K\epsilon\rho n \leq \epsilon\sum_{i=1}^{K}|C_i|$ of the nodes in the top K communities can lie in communities of size below $\epsilon\rho n$. □

Hence, the algorithm can compete with OPT by performing well on these large communities and ignoring those size less than $\epsilon\rho n$. Communities with size below this threshold are not guaranteed to have been sampled enough times, but even discarding them entirely will have little impact on the total utility. In what follows, we will assume for convenience that $|C_K| \geq \epsilon\rho n$ and then multiply the final guarantee by $(1-\epsilon)$.

Analysis of Step 1

Step 1 nests two levels of sampling: T nodes are sampled uniformly at random from the entire graph, and then R samples are taken from the community that each of these nodes lie in. At this point, we set $T = 6\left(\frac{1}{\epsilon^3\rho}\right)\log\frac{1}{\epsilon\rho}$. We will first show that, at the outer level, the number of times that each community is sampled is concentrated well. Then, we will show that the inner loop accurately estimates the size of each sampled community. The first claim follows from a straightforward application of the Chernoff bound. The second claim requires a more involved analysis of our rejection sampling procedure. The following two lemmas formalize these guarantees on the output of Step 1. Their proofs are given in Section 1.4.

Lemma 4. *Let X_j^i be the indicator variable for the event that sample j lands in community C_i. With probability at least $1-2\epsilon$,*

$$(1-\epsilon)T\frac{|C_i|}{n} \leq \sum_{j=1}^{T} X_i^j \leq (1+\epsilon)T\frac{|C_i|}{n}$$

holds for all i with $|C_i| \geq \epsilon\rho n$.

Lemma 5. *There are settings $R = O\left(\frac{1}{\epsilon^2}\log^2\left(\frac{T}{\epsilon}\right)\log^6 n\right)$ and $B = O(\log^{3/2} n)$ such that, given R random walk samples from a community C, the estimated size $\hat{d}$ satisfies $(1-2\epsilon)p_w|C| \leq \hat{d} \leq (1+2\epsilon)p_w|C|$ with probability at least $1-\frac{2\epsilon}{T}-o(1)$*

Corollary 1. *With probability at least $1-2\epsilon-o(1)$, $(1-2\epsilon)|C_i| \leq \hat{S}_i \leq (1+2\epsilon)|C_i|$ holds for every $i = 1...T$.*

Proof. We apply Lemma 5 to each of the T iterations. Since $(1-2\epsilon)p_w|C| \leq \hat{d} \leq (1+2\epsilon)p_w|C|$, we know that for $\hat{S} = \frac{1}{p_w}\hat{d}$ $(1-2\epsilon)|C| \leq \hat{d} \leq (1+2\epsilon)|C|$. To conclude, we take union bound over the failure probabilities at each iteration. □

We emphasize that Lemma 5 applies to *all* communities that are sampled, not just those which have size at least $\epsilon\rho n$. However, Lemma 4 only applies to communities with size at least $\epsilon\rho n$. That is, each sampled community's size estimate is accurate, but small communities may not be reliably sampled. Note that the total query cost is $R \cdot T$, which (given these settings for R and T), implies the bound in the theorem statement.

At this point in the analysis, we assume that we proceed with an $\epsilon' = \frac{\epsilon}{2}$, so that each $\hat{S}_i$ satisfies $(1-\epsilon)|C_i| \leq \hat{S}_i \leq (1+\epsilon)|C_i|$. This is purely for convenience; at the end of the proof we add up the total constant that ϵ should be divided by to obtain the guarantee in the theorem statement.

Analysis of Step 2

We now analyze the probability that each community in the top K (with size at least $\epsilon\rho n$) is seeded based on the above estimates. Consider any community $C_i \in \{C_1...C_K\}$ with $|C_i| \geq \epsilon\rho n$.

C_i is seeded if any of the sampled nodes from it are chosen in Step 4, and the probability of this event is determined by the total amount of weight which is allocated to nodes in C_i. In this step, we show that the total weight assigned to each of the top K communities is close to 1. Formally,

Lemma 6. *For any community C_i, let $w(C_i)$ be the total weight assigned to C_i. Suppose that C_i satisfies*

- $(1-\epsilon)T\frac{|C_i|}{n} \leq \sum_{j=1}^{T} X_i^j \leq (1+\epsilon)T\frac{|C_i|}{n}$
- $(1-\epsilon)|C_i| \leq \hat{S}_i \leq (1+\epsilon)|C_i|$ *each time C_i is sampled.*

Then, $1-2\epsilon \leq w(C_i) \leq 1+2\epsilon$.

Proof. We have

$$\begin{aligned} w(C_i) &= \sum_{j=1}^{T} \mathbf{1}\{j \in C_i\} w_j \\ &= \sum_{j=1}^{T} \mathbf{1}\{j \in C_i\} \frac{n}{\hat{S}_j T} \\ &\geq (1-\epsilon)T\frac{|C_i|}{n}\frac{n}{(1+\epsilon)|C_i|T} \\ &= \frac{1-\epsilon}{1+\epsilon} \\ &\geq 1-2\epsilon \end{aligned}$$

A similar argument shows that $w(C_i) \leq (1+2\epsilon)$ also holds. □

Corollary 2. *With probability at least $1-4\epsilon-o(1)$, $1-2\epsilon \leq w(C_i) \leq 1+2\epsilon$ holds for every community C_i sampled during Step 1 with $\hat{S}_i > 0$.*

Proof. Via Lemma 4 and Corollary 1 (and union bound), we can apply Lemma 6 to each community sampled in Step 1 with total probability at least $1-4\epsilon$. □

Analysis of Steps 3 and 4

Now, we need to analyze the impact of the truncation in Step 3. If the size of every community were perfectly estimated, then this step would set the weight of each community with size less than C_K to zero, leaving only $C_1 ... C_K$ to be seeded in Step 4. The following analysis controls the loss that can be suffered due to sampling errors.

For instance, it is possible that C_K could have "borderline" size arbitrarily close to $|C_{K+1}|$, in which case much of this weight may be truncated in favor of samples from C_{K+1}. For this to occur, a sample from C_{K+1} must have estimated size higher than a sample from C_K. But since the size of each sampled community is well-estimated, this implies that $|C_{K+1}|$ is actually very close to $|C_K|$, so not much is lost.

We now formalize this intuition. Recall that Step 3 calculates a threshold τ: the algorithm keeps all samples j where $\hat{S}_j \geq \tau$ and discards those with $\hat{S}_j < \tau$ by setting $w_j = 0$. Let $w(C_i)$ denote the total weight of community C_i before truncation and $w_T(C_i)$ denote its total weight after truncation. We define four sets of communities

- $Small = \{C_i|\ |C_i| < (1-\epsilon)|C_K|\}$. These are communities we would like to show never displace samples from communities in the other three sets.
- $A = \{C_i|\ \frac{1-\epsilon}{1+\epsilon}|C_K| \leq |C_i| < |C_K|\}$. These are communities with size less than $|C_K|$, but which we might not be able to detect and truncate due to sampling errors.
- $B = \{C_i|\ |C_K| \leq |C_i| \leq \frac{1+\epsilon}{1-\epsilon}|C_K|\}$. These are communities with size at least $|C_K|$, but which are small enough that they might be confused with communities in A.
- $Large = \{C_i|\ |C_i| > \frac{1+\epsilon}{1-\epsilon}|C_K|\}$. These are communities we would like to show are never truncated.

First, we show that communities in $Small$ and $Large$ behave well under truncation, in the sense that no samples from $Large$ are truncated and no samples from $Small$ displace samples from $B \cup Large$. In what follows we condition on the events in Corollaries 1 and 2.

Lemma 7. *w_T satisfies the following conditions:*

- $w_T(Small) \leq K - w_T(B) - w_T(Large)$
- $w_T(Large) = w(Large) \leq |Large| + 2\epsilon|Large|$

Proof. If a community C_i is in $Small$, then the size estimated for each of its samples satisfies

$$\hat{S} \leq (1+\epsilon)|C_i| < (1+\epsilon)\left(\frac{1-\epsilon}{1+\epsilon}\right)|C_K| \leq (1-\epsilon)|C_K|$$

Hence, samples from communities in $Small$ will always have estimated size less than every sample from $B \cup Large$, from which the first claim follows. The same logic shows that every sample from communities in $Large$ has estimated size higher than every sample from C_K. This implies (via $\sum_{i=1}^{K-1} w(C_i) \leq (K-1) + 2\epsilon(K-1)$) that each sample's estimated size lies above τ, proving the second claim. □

We recall here that choosing a random sample from a community C_i has probability $\beta(|C_i|)$ of hitting the giant connected component induced by the ICM, in which case it influences a fraction $\beta(|C_i|)$ of the nodes in the community. Hence, the total expected utility is

$$\sum_{C_i} \beta(|C_i|)^2 \Pr[C_i \text{ is seeded}]|C_i| \ \geq \sum_{C_i \in A \cup B \cup Large} \beta(|C_i|)^2 \Pr[C_i \text{ is seeded}]|C_i|$$

We refer to the value of the above summation restricted to a particular set of communities as the total utility obtained from that set. We now proceed to bound the total utility obtained from $A \cup B$, and then the total utility obtained from $Large$.

Lemma 8. *The total utility obtained from $A \cup B$ is at least*

$$\beta\left((1-2\epsilon)|C_K|\right)^2 \left(|B| - 4\epsilon K - 1\right)|C_K|\left(1 - e^{-(1-2\epsilon)}\right)$$

Proof. Via Lemma 7, at least $K - |Large| - 2\epsilon|Large| = |B| - 2\epsilon|Large|$ weight must be allocated to communities in A and B. Hence, the total expected utility obtained from these communities is

$$\begin{aligned}
\sum_{C_i \in A \cup B} \beta(|C_i|)^2 \Pr[C_i \text{ is seeded}]|C_i| &= \sum_{C_i \in A \cup B} \beta(|C_i|)^2 \left(1 - \left(1 - \frac{w(C_i)}{K}\right)^K\right)|C_i| \\
&\geq \sum_{C_i \in A \cup B} \beta(|C_i|)^2 \left(1 - e^{-w(C_i)}\right)|C_i| \\
&\geq \beta\left(\left(\frac{1-\epsilon}{1+\epsilon}\right)|C_K|\right)^2 \left(\frac{1-\epsilon}{1+\epsilon}\right)|C_K| \sum_{C_i \in A \cup B} 1 - e^{-w(C_i)} \\
&\geq \beta\left((1-2\epsilon)|C_K|\right)^2 (1-2\epsilon)\,|C_K| \sum_{C_i \in A \cup B} 1 - e^{-w(C_i)}. \qquad (1)
\end{aligned}$$

Given the above constraints on the total amount of weight allocated to each community, the value of Equation (1) is at least the value of the following optimization problem:

$$\begin{aligned}
&\min \beta\left((1-2\epsilon)|C_K|\right)^2 (1-2\epsilon)\,|C_K| \sum_{C_i \in A \cup B} 1 - e^{-w(C_i)} \\
&w(C_i) \leq 1 + 2\epsilon \quad \forall C_i \in A \cup B \\
&\sum_{C_i \in A \cup B} w(C_i) \geq |B| - 2\epsilon|Large|
\end{aligned}$$

Here the first constraint is due to Corollary 2, and the second is due to the argument at the start of this lemma. Let Q be the optimal value of the above optimization problem. The objective is the sum of identical concave functions in each variable $w(C_i)$. Hence, the minimum is achieved when as many of the $w(C_i)$ as possible are set to $1 + 2\epsilon$, with one community receiving the leftover weight. Specifically, $\left\lfloor \frac{|B| - 2\epsilon|Large|}{1+2\epsilon} \right\rfloor$ communities receive weight $1 + 2\epsilon$, and one community receives weight $(|B| - 2\epsilon|Large|) - (1 + 2\epsilon)\left\lfloor \frac{|B| - 2\epsilon|Large|}{1+2\epsilon} \right\rfloor$. We can lower bound Q by only considering the communities that received weight $1 + 2\epsilon$, which gives (via some straightforward algebra):

$$\begin{aligned}
Q &\geq \left\lfloor \frac{|B| - 2\epsilon|Large|}{1+2\epsilon} \right\rfloor \cdot \beta\left((1-2\epsilon)|C_K|\right)^2 (1-2\epsilon)\,|C_K|\left(1 - e^{-(1+2\epsilon)}\right) \\
&\geq \left(\frac{|B| - 2\epsilon|Large|}{1+2\epsilon} - 1\right) \cdot \beta\left((1-2\epsilon)|C_K|\right)^2 (1-2\epsilon)\,|C_K|\left(1 - e^{-(1+2\epsilon)}\right) \\
&\geq \left((1-2\epsilon)(|B| - 2\epsilon|Large|) - 1\right) \cdot \beta\left((1-2\epsilon)|C_K|\right)^2 (1-2\epsilon)\,|C_K|\left(1 - e^{-(1+2\epsilon)}\right) \\
&\geq \left(|B| - 2\epsilon|Large| - 2\epsilon|B| + 4\epsilon^2|Large| - 1\right) \cdot \beta\left((1-2\epsilon)|C_K|\right)^2 (1-2\epsilon)\,|C_K|\left(1 - e^{-(1+2\epsilon)}\right) \\
&\geq \left(|B| - 4\epsilon K - 1\right) \cdot \beta\left((1-2\epsilon)|C_K|\right)^2 |C_K|\left(1 - e^{-(1+2\epsilon)}\right) \\
&\geq \left(|B| - 4\epsilon K - 1\right) \cdot \beta\left((1-2\epsilon)|C_K|\right)^2 |C_K|\left(1 - e^{-(1-2\epsilon)}\right).
\end{aligned}$$

□

Lemma 9. *The total utility obtained from Large is at least* $\sum_{C_i \in Large} \beta(|C_i|)^2 |C_i|\left(1 - e^{-(1-2\epsilon)}\right)$.

Proof. Follows directly from Lemma 7, which implies that every $C_i \in Large$ satisfies $w(C_i) \geq 1-2\epsilon$. As a result, $\Pr[C_i \text{ is seeded}] \geq 1 - e^{-(1-2\epsilon)}$. □

After some more algebra, this leads to our final bound on the total utility. Recall that we defined $\beta_{min} = \beta((1-\epsilon)|C_K|)$

Lemma 10. *The total utility over all communities is at least*

$$\left(1 - 6\epsilon - \frac{1}{K}\right)\left(1 - e^{-(1-2\epsilon)}\right)\beta_{min}^2 \sum_{i=1}^{K}|C_i|$$

.

Proof.

$$\sum_{C_i \in A \cup B \cup Large} \beta(|C_i|)^2 \Pr[C_i \text{ is seeded}]|C_i|$$

$$\geq \sum_{C_i \in B} \beta((1-2\epsilon)|C_K|)^2|C_K|\left(1 - e^{-(1-2\epsilon)}\right) + \sum_{C_i \in Large} \beta(|C_i|)^2|C_i|\left(1 - e^{-(1-2\epsilon)}\right) - \beta((1-2\epsilon)|C_K|)^2 (4\epsilon K + 1)|C_K|\left(1 - e^{-(1-2\epsilon)}\right)$$

$$\geq (1-2\epsilon)\left(\sum_{C_i \in B} \beta((1-4\epsilon)|C_i|)^2|C_i|\left(1 - e^{-(1-2\epsilon)}\right) + \sum_{C_i \in Large} \beta(|C_i|)^2|C_i|\left(1 - e^{-(1-2\epsilon)}\right) - \beta((1-2\epsilon)|C_K|)^2 (4\epsilon K + 1)|C_K|\left(1 - e^{-(1-2\epsilon)}\right)\right)$$

$$\geq (1-2\epsilon)\left(\sum_{i=1}^{K} \beta((1-4\epsilon)|C_i|)^2|C_i|\left(1 - e^{-(1-2\epsilon)}\right) - (4\epsilon K + 1)\beta((1-4\epsilon)|C_K|)^2|C_K|\left(1 - e^{-(1-2\epsilon)}\right)\right)$$

$$\geq (1-2\epsilon)\left(\sum_{i=1}^{K} \beta((1-4\epsilon)|C_i|)^2|C_i|\left(1 - e^{-(1-2\epsilon)}\right) - (4\epsilon K + 1)\frac{1}{K}\sum_{i=1}^{K}\beta((1-4\epsilon)|C_i|)^2|C_i|\left(1 - e^{-(1-2\epsilon)}\right)\right)$$

$$= (1-2\epsilon)\left(1 - 4\epsilon - \frac{1}{K}\right)\left(1 - e^{-(1-2\epsilon)}\right)\sum_{i=1}^{K}\beta((1-4\epsilon)|C_i|)^2|C_i|$$

$$\geq \left(1 - 6\epsilon - \frac{1}{K}\right)\left(1 - e^{-(1-2\epsilon)}\right)\sum_{i=1}^{K}\beta((1-4\epsilon)|C_i|)^2|C_i|$$

$$\geq \left(1 - 6\epsilon - \frac{1}{K}\right)\left(1 - e^{-(1-2\epsilon)}\right)\beta((1-4\epsilon)|C_K|)^2\sum_{i=1}^{K}|C_i|$$

At this point, we can pretend that $\beta((1-4\epsilon)|C_K|) \geq \beta_{min}$ since we end up running the algorithm with a much smaller value of ϵ in any case. Details can be found at the conclusion of the proof. □

By Lemma 2, $OPT \leq \sum_{i=1}^{K}\beta(C_i)|C_i|$. Plugging in the bound from Lemma 10, we have (modulo the events we conditioned on earlier) the approximation ratio

$$\left(1 - 6\epsilon - \frac{1}{K}\right)\left(1 - e^{-(1-2\epsilon)}\right)\beta_{min}^2 \frac{\sum_{i=1}^{K}|C_i|}{\sum_{i=1}^{K}\beta(C_i)|C_i|}. \tag{2}$$

In the worst case, we can lower bound this as follows. Recall that we defined $\beta_{max} = \beta(|C_1|)$. We can define β_{min} using $(1-\epsilon)|C_K|$ because below we run the algorithm with a smaller value of ϵ in any case. We have

$$\begin{aligned}\frac{\beta_{min}^2 \sum_{i=1}^K |C_i|}{\sum_{i=1}^K \beta(C_i)|C_i|} &\geq \frac{\sum_{i=1}^K \beta_{min}^2 |C_i|}{\sum_{i=1}^K \beta_{max}|C_i|} \\ &= \frac{\beta_{min}^2}{\beta_{max}}.\end{aligned}$$

Now we are just a few details away from the final stated bound. First, the events in Corollary 2 hold with probability $1 - 4\epsilon - o(1)$. Second, the analysis used that the connected component induced by the ICM in each community has size β, which holds with probability $1-o(1)$. Via union bound, these events all occur with probability at least $1 - 4\epsilon - o(1)$. Second, we assume that all of the top K communities had size at least $\epsilon\rho n$. In fact, the "sufficiently large" communities that the algorithm obtains utility from have total size at least $(1-\epsilon)\sum_{i=1}^K |C_i|$ (via Lemma 3). Accounting for these facts, we have

$$\begin{aligned}\mathbb{E}[f_E(\mathcal{A}(E))] &\geq (1 - 4\epsilon - o(1))(1-\epsilon)\frac{\beta_{min}^2}{\beta_{max}}\left(1 - 6\epsilon - \frac{1}{K}\right)\left(1 - e^{-(1-2\epsilon)}\right) OPT \\ &\geq \frac{\beta_{min}^2}{\beta_{max}}\left(1 - 11\epsilon - \frac{1}{K} - o(1)\right)\left(1 - e^{-(1-2\epsilon)}\right) OPT.\end{aligned}$$

Finally, we note that we replaced ϵ by $\epsilon' = \frac{\epsilon}{2}$ earlier in the proof. Thus, running the algorithm with $\epsilon'' = \frac{\epsilon}{22}$ yields the stated bound.

1.4 Concentration lemmas

We now prove that the various estimates that the algorithm takes in Step 1 are sufficiently accurate. We make frequent use of the Chernoff bound:

Lemma 11 ([6])**.** *Let $X_1 ... X_N$ be independent binary random variables. Let $X = \sum_{i=1}^N X_i$ and $\mu = \mathbb{E}[X]$.*

- *For $0 < \delta < 1$, $\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\frac{\delta^2\mu}{3}}$.*
- *For $\delta > 1$, $\Pr[X \geq (1+\delta)\mu] \leq e^{-\frac{\delta\mu}{3}}$*

We now proceed to prove Lemmas 4 and 5.

Proof of Lemma 4. Note that $\mathbb{E}[\sum_{j=1}^T X_j^i] = T\frac{|C_i|}{n}$ Via the Chernoff bound, we have that

$$\begin{aligned}\Pr\left[\left|\sum_{j=1}^T X_j^i - T\frac{|C_i|}{n}\right| > \epsilon T\frac{|C_i|}{n}\right] &\leq 2\exp\left(-\frac{1}{3}\epsilon^2 T\frac{|C_i|}{n}\right) \\ &\leq 2\exp\left(-\frac{1}{3}\epsilon^2 T\epsilon\rho\right) \qquad (|C_i| \geq \epsilon\rho n) \\ &\leq 2\exp\left(-2\log\frac{1}{\epsilon\rho}\right) \\ &\leq 2(\rho\epsilon)^2.\end{aligned}$$

There are at most $\frac{1}{\epsilon\rho}$ communities of size at least $\epsilon\rho n$. By union bound, concentration holds for each of them with probability at least $1 - 2\epsilon\rho \geq 1 - 2\epsilon$. □

Proof of Lemma 5. We recall the lemma statement: There are settings $R = O\left(\frac{1}{\epsilon^2}\log^2\left(\frac{T}{\epsilon}\right)\log^6 n\right)$ and $B = O(\log^{3/2} n)$ such that, given R random walk samples from a community C, the estimated size $\hat{S}$ satisfies $(1-2\epsilon)|C| \leq \hat{S} \leq (1+2\epsilon)|C|$ with probability at least $1 - \frac{2\epsilon}{T} - o(1)$

Our estimator does the following:

1. Take R steps according to a random walk, discarding a burn in portion at the start while the walk mixes.
2. Keep each sample i with probability $\frac{1}{d_i}$
3. Estimate the average degree as the mean of the degrees of the samples that were kept.

It will be convenient to work with a graph that is guaranteed not to have any nodes of degree significantly higher than the expected degree. We show that with high probability, the maximum degree in the graph is $O(\log^2 n)$, and condition on this holding for all nodes in the rest of the proof.

Lemma 12. *Let d_{max} be the largest degree in the graph. With probability at least $1 - \frac{1}{n}$, $d_{max} = O(\log^2 n)$.*

Proof. Fix a node v. d_v is the sum of $|C_i|$ indicator random variables which give the presence or absence of each possible edge. Let Y_{vu} be the indicator random variable for the presence of the edge (v, u). $d_v = \sum_{u \in C_i} Y_{vu}$. Note that by Assumption 1 $\mathbb{E}[d_v] \geq \log|C_i| \geq 1$. Also by Assumption 1, $p_w = O\left(\frac{\log n}{n}\right)$, so $\mathbb{E}[d_v] = O(\log n)$ holds as well. Since the Y_{uv} are independent, we can use the Chernoff bound to argue

$$\begin{aligned}\Pr[d_v \geq (1 + 6\log n)\,\mathbb{E}[d_u]] &\leq \exp\left(-\frac{1}{3}6\log n\,\mathbb{E}[d_u]\right)\\ &\leq \exp\left(-2\log n\right)\\ &\leq \frac{1}{n^2}\end{aligned}$$

Taking union bound over n vertices, with probability $1 - \frac{1}{n}$, every vertex has degree at most a factor $1 + 6\log n \leq 7\log n$ greater than its expectation. Since $\mathbb{E}[d_v] = O(\log n)$ for all vertices, we conclude that $d_v = O(\log^2 n)$ holds for all vertices with the stated probability. □

We now introduce some notation dealing with Markov chains. Suppose a Markov chain has transition matrix $\boldsymbol{P}$. All Markov chains we consider will have a unique stationary distribution. Let this distribution be π. The total variational distance between probability distributions p and q is

$$d_{TV}(p, q) = \sup_x |p(x) - q(x)|.$$

The mixing time of a chain is the maximum time needed for the chain to converge to its stationary distribution:

$$t_{mix}(\epsilon) = \min\{t \mid \sup_x d_{TV}(1_x \boldsymbol{P}^t, \pi) \leq \epsilon\}$$
$$t_{mix} \coloneqq t_{mix}\left(\frac{1}{4}\right)$$

The spectral gap of the chain is $\gamma = 1 - \lambda_2$, where λ_2 is the second eigenvalue of $\boldsymbol{P}$. It is related to t_{mix} as follows:

Lemma 13 (Paulin [7] Proposition 3.3)**.** $\gamma \geq \frac{1}{1+t_{mix}/\log 2}$

The random walk carried out by ARISEN is a Markov chain whose state space is the vertices of the community C_i. C_i is an Erdős-Rényi graph with at most n nodes. We use the following bound on the spectral gap of this walk:

Lemma 14 (Hoffman et al. [3])**.** *For an Erdős-Rényi graph with average degree $d_{avg} > \frac{1}{2}\log n$, with probability $1 - o(1)$, $\gamma = \Omega\left(\frac{1}{\sqrt{d}}\right)$.*

Since we have by Assumption 1 that $\log|C_i| < d_{avg} = O(\log n)$, we conclude that $\gamma = \Omega\left(\frac{1}{\sqrt{\log n}}\right)$.

The stationary distribution of a random walk on an undirected graph places probability $\frac{d_v}{\sum_u d_u}$ probability on node v, i.e., the probability is proportional to each node's degree. This is a problem when we wish to estimate the average degree, because our samples will be biased towards high degree nodes. The solution we use is rejection sampling (step 2), which "unbiases" the samples by disproportionately rejecting high degree nodes.

To prove concentration for our estimator, we define a second Markov chain which builds in the rejection step. Let the transition matrix of this new chain be $\boldsymbol{P'}$. The new chain is defined by applying the following procedure to the random walk chain. Given that the new chain is at a node v, it proposes to take a step according to the normal random walk. Say that this step leads to node u. With probability $\frac{1}{d_u}$, this move is accepted (the next state of the chain is u). Otherwise, the new chain takes another random walk step from u and applies the same acceptance condition to the new node, continuing until some node is accepted. This new node is the next state of the chain. The distribution of a set of samples from this chain is, by construction, the same as the distribution of a set of samples produced by step 1 and accepted by step 2. This characterization lets us see the intuitive fact that the new chain's stationary distribution π' is uniform over the vertices: we would first sample a node with probability proportional to its degree, and then accept it with probability inversely proportional to its degree.

We will apply concentration bounds to the the new random walk. We use the following Bernstein-style concentration bound for the sum of a function of a Markov chain.

Lemma 15 (Paulin [7]) Theorem 3.3)**.** *Let $X_1 ... X_r$ be a stationary reversible Markov chain over state space Ω with spectral gap γ and stationary distribution π. Let $g \in L^2(\pi)$, with $|g(x) - \mathbb{E}_\pi[g]| \leq C$ for every $x \in \Omega$. Then we have*

$$\Pr_\pi\left[\left|\frac{1}{r}\sum_{i=1}^{r} g(X_i) - \mathop{\mathbb{E}}_{\pi}[g(x)]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{r\epsilon^2\gamma}{4\,Var_\pi(g) + 10\epsilon C}\right)$$

To account for the fact that the chain does not start at stationarity, we can use a burn in time of t_{mix} steps, which gives the following bound:

Lemma 16 (Paulin [7]) Proposition 3.10)**.** *Suppose that the chain starts from distribution q and we discard the first t_0 samples. Let $\boldsymbol{P}$ be the transition matrix. Then*

$$\Pr_q\left[\left|\frac{1}{r-t_0}\sum_{i=t_0+1}^{r} g(X_i) - \mathop{\mathbb{E}}_{\pi}[g(x)]\right| \geq \epsilon\right] \leq \Pr_\pi\left[\left|\frac{1}{r-t_0}\sum_{i=t_0+1}^{r} g(X_i) - \mathop{\mathbb{E}}_{\pi}[g(x)]\right| \geq \epsilon\right] + d_{TV}(q\boldsymbol{P}^{t_0}, \pi).$$

Let π be the stationary distribution of the original chain and π' be the stationary distribution of the new chain. Let t'_{mix} be the mixing time of the new chain. We formalize that the second chain does not take much longer to mix than the first:

Lemma 17. $t'_{mix}(\epsilon) \leq 4\log_2\left(\frac{n}{\epsilon}\right) t_{mix} = O\left(\log^{3/2}\frac{n}{\epsilon}\right)$.

Proof. Suppose that we have run the first chain for t steps, resulting in a distribution q over the vertices. We simulate the second chain from the first, giving a distribution q'. Suppose that $d_{TV}(q,\pi) \leq \delta$ for some δ. We would like to show that $d_{TV}(q',\pi') \leq \epsilon$. Let $E(C_i)$ give the set of edges within community C_i. For any node v, we have $|q(v) - \pi(v)| = |q(v) - \frac{d_v}{2|E(C_i)|}| \leq \delta$. Thus, $|\frac{q(v)}{d_v} - \frac{1}{2|E(C)|}| \leq \frac{\delta}{d_v} \leq \delta$. We know that $q'(v) \propto \frac{q(v)}{d_v}$ since the RHS is just the probability that a node is both proposed by the first chain and then accepted. Similarly, we know that $\frac{1}{2|E(C_i)|} \propto \pi'(v)$ since π' is uniform. From here, the proof is mostly just algebra. Let $Z_1 = \sum_v \frac{1}{2|E(C_i)|}$ and $Z_2 = \sum_v \frac{q(v)}{d_v}$ be normalization constants such that $q'(v) = \frac{q(v)}{Z_1 d_v}$ and $\pi'(v) = \frac{1}{Z_2 2|E(C_i)|}$. We know that $|Z_1 - Z_2| \leq \delta n$ via the triangle inequality combined with $|\frac{q(v)}{d_v} - \frac{1}{2|E(C)|}| \leq \delta\ \forall v$. Further, $\left|\frac{1}{Z_1} - \frac{1}{Z_2}\right| = \left|\frac{Z_1 - Z_2}{Z_1 Z_2}\right|$. We can bound $Z_1 = \sum_v \frac{q(v)}{d_v} \geq \frac{1}{n}\sum_v q(v) = \frac{1}{n}$ and $Z_2 = \sum_v \frac{1}{2|E(C_i)|} \geq \frac{|C_i|}{2|E(C_i)|} \geq \frac{1}{2n}$. Thus, $\left|\frac{1}{Z_1} - \frac{1}{Z_2}\right| \leq 2n^2|Z_1 - Z_2| \leq 2n^3\delta$. Using this, we show how to bound $\frac{q(v)}{Z_1 d_v} - \frac{1}{Z_2 2|E(C_i)|}$ (the argument for $\frac{1}{Z_2 2|E(C_i)|} - \frac{q(v)}{Z_1 d_v}$ is analogous). We have $\frac{q(v)}{Z_1 d_v} - \frac{1}{Z_2 2|E(C_i)|} \leq 2\delta^2 n^3 + \frac{q(v)}{Z_2 d_v} - \frac{1}{Z_1 2|E(C_i)|}$. A few lines of calculation yield that $\frac{q(v)}{Z_2 d_v} - \frac{1}{Z_1 2|E(C_i)|} \leq 2n^2(2\delta + n\delta^2)$. In the end, we get the final bound that $\left|\frac{q(v)}{Z_1 d_v} - \frac{1}{Z_2 2|E(C_i)|}\right| \leq 2\delta^2 n^3 + 2n^2(2\delta + n\delta^2) \leq 7\delta n^3$. Thus, it suffices to have $\delta \leq \frac{\epsilon}{7n^3}$. It is well known [6] that running any Markov chain for ct_{mix} steps results in a distribution with total variational distance at most 2^{-c} from the stationary distribution. Hence, we can take $4t_{mix}\log_2\frac{n}{\epsilon}$ steps and obtain $\delta \leq \frac{\epsilon^4}{n^4} < \frac{\epsilon}{7n^3}$. Thus, $t'_{mix}(\epsilon) \leq 4\log_2\left(\frac{n}{\epsilon}\right) t_{mix}$. □

Let γ' be the spectral gap of the transition matrix of the new chain. From Lemma 13, the above implies that $\gamma' = \Omega\left(\log^{-\frac{3}{2}} n\right)$.

After running the random walk for R steps, let R' be the number of steps taken in the new chain (the number of samples accepted by step 2). This is our effective sample size.

Lemma 18. *With probability at least $1-\epsilon$, $R' = \Omega\left(\frac{1}{\log\frac{1}{\epsilon}\log^2 n}R\right)$*

Proof. We know that each v satisfies $d_v \leq O(\log^2 n)$. Thus, the number of accepted samples stochastically dominates a Binomial random variable with R trials and success probability $\frac{1}{O(\log^2 n)}$. Let μ be the expected number of successes and X be the actual number of successes. Via the Chernoff bound, with probability at least $1-\epsilon$, we have that

$$X \geq \left(1 - \sqrt{\frac{2\log\frac{1}{\epsilon}}{\mu}}\right)\mu.$$

After some algebra, we can see that if we have $\mu = 2\log\frac{1}{\epsilon}R'$, then this implies $X \geq R'$ (provided $R' \geq 4$). Since $\mu = \Omega\left(\frac{R}{\log^2 n}\right)$, we obtain the statement in the lemma. □

Conditioning on having sufficiently large R' per Lemma 18, we can use Lemma 16 to obtain the following upper bound on the number of random walk steps needed to obtain the claimed failure probability:

Lemma 19. *Let d_{avg} be the average degree of C_i. In order to have*

$$\Pr\left[\left|\frac{1}{R' - t_{mix}}\sum_{i=t_{mix}+1}^{R'} d(v_i) - d_{avg}\right| \geq \epsilon\right] \leq 2\epsilon$$

it suffices to take $R = O\left(\frac{1}{\epsilon^2}\log^2\left(\frac{1}{\epsilon}\right)\log^6 n\right)$ random walk steps.

Proof. We will use Lemma 16 applied to the new chain with $g(v) = d_v$ since $\mathbb{E}_{\pi'}[d_v] = d_{avg}$. In order to do so, we need bounds on both the highest possible value of d_v and on $Var_{\pi'}(d_v)$. Lemma 12 supplies that $d_{max} = O(\log^2 n)$ is an upper bound on the maximum possible value (having conditioned on this holding for all nodes). For the variance, we note that d_v is a binomial random variable with variance $np(1-p) = O\left(n\frac{\log n}{n}\left(1 - \frac{\log n}{n}\right)\right) = O(\log n)$. Conditioning on its maximum value can only reduce its variance.

Via Lemma 14, the d_{TV} term in Lemma 16 is at most ϵ after the burn-in period. Hence, we can bound the failure probability for samples drawn from the stationary distribution using Lemma 15. We consider two cases

Case 1: $4Var_{\pi'}(d_v) \geq 10\epsilon d_{max}$. In this case, the failure probability is at most

$$\begin{aligned}
&2\exp\left(-\frac{R'\epsilon^2\gamma'}{8Var_{\pi}(d_v)}\right)\\
&\leq 2\exp\left(-\Omega\left(\frac{\epsilon^2\gamma' R}{Var_{\pi'}(d_v)\log\frac{1}{\epsilon}\log^2 n}\right)\right)\\
&\leq 2\exp\left(-\Omega\left(\frac{\epsilon^2 R}{\log\frac{1}{\epsilon}\log^{9/2} n}\right)\right)
\end{aligned}$$

So, there is a constant c_1 such that taking $R = c_1\frac{1}{\epsilon^2}\log^2\left(\frac{T}{\epsilon}\right)\log^5 n$ makes the failure probability at most $\frac{2\epsilon}{T}$.

Case 2: $10\epsilon d_{max} > 4Var_{\pi'}(d_v)$. In this case, the failure probability is at most

$$\begin{aligned}
&2\exp\left(-\frac{R'\epsilon^2\gamma'}{20\epsilon d_{max}}\right)\\
&\leq 2\exp\left(-\Omega\left(\frac{\epsilon\gamma' R}{d_{max}\log\frac{1}{\epsilon}\log^2 n}\right)\right)\\
&\leq 2\exp\left(-\Omega\left(\frac{\epsilon R}{\log\frac{1}{\epsilon}\log^{11/2} n}\right)\right)
\end{aligned}$$

So, there is a constant c_2 such that taking $R = c_2\frac{1}{\epsilon}\log^2\left(\frac{T}{\epsilon}\right)\log^6 n$ makes the failure probability at most $\frac{2\epsilon}{T}$. Between the two cases, we conclude that taking $R = O\left(\frac{1}{\epsilon^2}\log^2\left(\frac{T}{\epsilon}\right)\log^6 n\right)$ suffices as

claimed. Since the burn-in time is $O(\log^{3/2} n)$ per Lemma 14, the additional $O(\log^{3/2} n)$ steps at the start are absorbed into this figure. □

With this proof of this lemma completed, Lemma 5 immediately follows. □

This completes the proof of Theorem 1.

1.5 General case: $p_b > 0$

We now generalize the analysis of the above algorithm to handle edges between communities. We focus on the case where p_b is sufficiently small that the communities in the graph do not themselves form a giant connected component under the ICM. While it is clearly possible to prove guarantees for the case where p_b is above this threshold (since a linear portion of the network will be connected and could be hit just by random sampling), this is not the case we are interested in modeling from an applications perspective. To formalize the threshold for p_b, we require that every community has (in expectation) less than one live edge to other communities.

Assumption 4. $p_b q \cdot (n - |C_i|)|C_i| < 1 \ \forall C_i$.

Lastly, we assume

Assumption 5. $p_b < \frac{1}{n}$.

This implies that the between-community edges by themselves do not create a giant connected component in G (which is clearly what we expect in practice).

Let $\mu = \frac{1}{K}\sum_{i=1}^{K}|C_i|$ be the average size of the top K communities. We prove the following approximation guarantee:

Theorem 2. *Suppose that $\rho \leq \frac{\mu}{n}$ and choose $\epsilon < \frac{3}{8}$. Using the same number of samples as in Theorem 1, ARISEN influences at least*

$$\left(\frac{1 - c_{max}}{12 \log \frac{n}{\mu}}\right) \frac{\beta_{min}^2}{\beta_{max}} \left(1 - e^{-(1-\epsilon)} - \epsilon - \frac{1}{K} - o(1)\right) OPT$$

nodes in expectation.

Proof. We first deal with the accuracy of the size estimations in Step 1 and then provide a new bound on OPT. Subsequently, we wrap up the remaining details to obtain the stated guarantee.

Updated analysis of Step 1

We prove that the size estimates are correct with probability $1 - o(1)$ since there are sufficiently few between-community edges. Our analysis uses the connection between the conductance of a graph and the properties of random walks on it, so we start by introducing a few definitions. The *volume* of a set of nodes S, denoted by $\mu(S)$ is the sum of the degrees of the nodes in S:

$$\mu(S) = \sum_{i \in S} d_i$$

Let $E(S, S - V)$ denote the set of edges between nodes in S and those in $V - S$. They key ratio for our analysis is

$$\Phi(S) = \frac{|E(S, V-S)|}{\mu(S)}.$$

This is nearly the same as the normal definition of conductance, which has $\min(\mu(S), \mu(V-S))$ in the denominator. However, our analysis depends only on $\mu(S)$. The key lemma that we use relates $\Phi(S)$ to the properties of a random walk in S:

Lemma 20 (Spielman and Teng [8], Prop. 2.5)**.** *The probability that a random walk, started from a random node of S, stays entirely within S for t steps is at least $1 - \frac{1}{2}t\Phi(S)$.*

We remark that Spielman and Teng stated the lemma for the normal conductance (not our Φ), but their analysis trivially applies to Φ as we have defined it.

Lemma 20 will be used to control the probability that any of the nodes we sample in Step 1c lie outside of the starting community. Fix any C_i. We apply the Chernoff bound to the numerator and denominator of $\Phi(C_i)$ to show that it is close to $\frac{\log n}{n}$ with high probability over the draw of G from the SBM.

First, we show that with high probability, $|E(C_i, V-C_i)| \leq \frac{7\log n}{q}$. Let $Z = |E(C_i, V-C_i)|$ and note that Z is the sum of $(n-|C_i|)|C_i|$ indicator variables giving whether each possible between-community edge is present. From Assumption 4, we know that $\mathbb{E}[Z] < \frac{1}{q}$. Thus via the Chernoff bound we have

$$\begin{aligned}\Pr[Z > (1+6\log n)\,\mathbb{E}[Z]] &\leq \exp\left(-\frac{1}{3}\left(\frac{6\log n}{q}\right)\right) \\ &\leq \frac{1}{n^2}\end{aligned}$$

There are at most n total communities, so taking union bound over all of them gives total failure probability at most $\frac{1}{n}$. Conditioned on concentration holding, we have $Z \leq (6\log n + 1)\,\mathbb{E}[Z] \leq \frac{7\log n}{q}$.

Next, we examine $\mu(C_i)$. We have $\mathbb{E}[\mu(C_i)] = p_w|C_i|^2$. By assumption, $p_w|C_i|^2 \geq |C_i|$. Via Chernoff bound,

$$\begin{aligned}\Pr\left[\mu(C_i) \leq \left(1 - \sqrt{\frac{6\log n}{|C_i|}}\right)\mathbb{E}[\mu(C_i)]\right] &\leq \exp\left(-\frac{1}{3}\left(\frac{6\log n}{|C_i|}\right)|C_i|\right) \\ &\leq \frac{1}{n^2}.\end{aligned}$$

Again via union bound, the total failure probability over all communities is at most $\frac{1}{n}$.

Conditioning on the bounds on both the numerator and denominator holding, we have

$$\Phi(C_i) \leq \frac{7\log n}{\left(1 - \sqrt{\frac{6\log n}{|C_i|}}\right) q p_w |C_i|^2}$$

Since $qp_w|C_i| \geq 1$ by Assumption 2, this implies

$$\Phi(C_i) \leq \frac{7\log n}{\left(1-\sqrt{\frac{6\log n}{|C_i|}}\right)|C_i|} = \frac{7\log n}{|C_i|-\sqrt{6|C_i|\log n}}$$

Now we can apply Lemma 20 to bound the probability that the random walk leaves C_i. In any single iteration, we take R random walk steps, which leave C_i with probability at most $\frac{1}{2}R\Phi(C_i)$. There are T iterations in total, so via union bound the total probability that any random walk leaves its starting community is at most $\frac{1}{2}\Phi(C_{min})RT$ where C_{min} is the smallest community. This yields

$$\begin{aligned}
\frac{1}{2}\Phi(C_{min})RT &= O\left(\left(\frac{1}{\epsilon^5\rho}\right)\log^3\frac{1}{\epsilon\rho}\log^6 n\right)\frac{\frac{7}{2}\log n}{|C_i|-\sqrt{6|C_i|\log n}} \\
&= O\left(\frac{\left(\frac{1}{\epsilon^5\rho}\right)\log^3\frac{1}{\epsilon\rho}\log^6 n}{|C_{min}|}\right) && \text{(since } |C_{min}| = poly(n)\text{)} \\
&= o(1) && \left(\text{by Assumption 3, } \frac{1}{|C_{min}|(\epsilon^5\rho)} = \frac{1}{poly(n)}\right)
\end{aligned}$$

We conclude that the total probability of any random walk leaving its starting community is at most $o(1)$. Conditioning on this extra event, Corollary 1 for the $p_b = 0$ case still holds, which is the only guarantee needed on the output of Step 1.

Bounding OPT

We prove the following guarantee on the relative sizes of $\sum_{i=1}^K |C_i|$ and OPT in the $p_b > 0$ setting:

Lemma 21. *Let $\mu = \frac{1}{K}\sum_{i=1}^K |C_i|$ denote the average size of the top K communities. Then we have*

$$\sum_{i=1}^K |C_i| \geq \left(\frac{1-c_{max}}{12\log\frac{n}{\mu}}\right) OPT$$

Proof. Let $X_1 ... X_K$ be the sizes of the K largest connected components induced by the SBM and ICM. We have $OPT \leq \mathbb{E}\left[\sum_{i=1}^K X_i\right]$. Each X_i contains the giant connected component in one or more communities. Let C_i^* be the (random) community which is the largest community whose giant component is contained in X_i. Let $\boldsymbol{C^*}$ be a vector which collects $|C_1^*|...|C_K^*|$. Clearly, we have $\sum_{i=1}^K |C_i| \geq \mathbb{E}\left[\sum_{i=1}^K |C_i^*|\right]$. We will now bound the amount by which $\mathbb{E}\left[\sum_{i=1}^K X_i\right]$ can exceed $\mathbb{E}\left[\sum_{i=1}^K |C_i^*|\right]$, which in turn lets us bound $\sum_{i=1}^K |C_i|$ in terms of OPT.

The crucial step is to bound a single X_i relative to $|C_i^*|$. We show

Lemma 22. $\mathbb{E}[X_i|\boldsymbol{C^*}] \leq \left(\frac{12}{1-c_{max}}\right)\log\left(\frac{n}{|C_i^*|}\right)|C_i^*|$

Proof. We analyze a branching process, similar to that used to analyze the subcritical Erdős-Rényi graph. This process starts at a single node, and then reveals the status of all potential edges to the remaining nodes. Each edge that exists creates a new child and the process then explores the

edges of each child. The size of the connected component is the total number of nodes explored by the branching process.

Our analysis will collapse the giant connected component of each community into a single node in a higher-level branching process. This allows us to bound the total number of nodes of G that can be absorbed into a connected component. The major challenge for us to analyze the branching process is that the communities need not have equal sizes, so we cannot apply the analysis of the Erdős-Rényi graph exactly. We prove that the number of nodes reached in the true branching process is stochastically dominated by one in which every community in the graph has size $|C_i^*|$.

Conditioning on $\boldsymbol{C^*}$ (as in the lemma statement) complicates the branching process because if a given community is reached, then it has a chance to reach a community with size above $|C_i^*|$, or to reach a community in one of the other components. Hence, conditioning on $\boldsymbol{C^*}$ reduces the probability that the branching process will reach any of the other communities in the graph. However, the true process is stochastically dominated by a branching process on the subgraph induced by the communities with size at most $|C_i^*|$; call this graph G_A. Essentially, in this process we ignore that conditioning on $|C_i^*|$ can indirectly limit the number of nodes reached, and that the other components could "compete" with X_i for nodes. To formalize this reasoning, we define two branching processes:

BP-cond: This is the "true" branching process. Pick a node in C_i^* to start from. From the starting node, reveal the status (live or not) of all edges from this node's community to other communities. These revelations follow a distribution which conditions on (1) not reaching a community with size greater than C_i^* or (2) reaching a community which belongs to the other $K-1$ components. Note that the BP-cond's corresponding to each of the K largest components could have a complicated joint distribution but we do not need to fully describe it (as will be seen below).

BP-A: Pick a node in the largest community of G_A. Follow the branching process from that node using only edges between nodes in G_A (but ignoring the two conditions for BP-cond)

Let Z_{cond} (resp Z_A) be a random variable giving the total number of nodes in communities reached by BP-cond (resp. BP-A). We have

Claim 1. *Z_{cond} is stochastically dominated by Z_A*

Proof. Let Y_e $\forall e \in V \times V$ be an indicator variable for the event that edge e is live (i.e., it is present in both the SBM and ICM). $\boldsymbol{Y}$ is a vector which collects all of the Y_e. Let $h(\boldsymbol{Y})$ denote the total number of nodes reached by the branching process when the status of the edges are specified by $\boldsymbol{Y}$. Note that h is monotone nondecreasing in $\boldsymbol{Y}$. The distribution of Z_{cond} or Z_A can be simulated by drawing $\boldsymbol{Y}$ from the distribution induced by the corresponding branching process and then returning $h(\boldsymbol{Y})$. Consider any subset $E' \subseteq \{e \in V \times V\}$. We couple Z_{cond} and Z_A by having them share Y_e for all $e \notin E'$. We have

$$\Pr_{\boldsymbol{Y} \sim \text{BP-cond} | \{Y_e | e \notin E'\}} \left[Y_e = 1 \ \forall e \in E' \right] \leq \Pr_{\boldsymbol{Y} \sim \text{BP-A} | \{Y_e | e \notin E'\}} \left[Y_e = 1 \ \forall e \in E' \right]. \tag{3}$$

To see this, note that under BP-cond, the probability that $Y_e = 1 \ \forall e \in E'$ is either

- The probability of this event under the Y_e's original marginal distribution (drawn from the SBM and ICM) if setting them equal to 1 would not violate either condition for BP-cond.
- 0 if setting them to 1 would create a violation

However, BP-A always follows the first case, which assigns at least as high a probability to the event $Y_e = 1 \ \forall e \in E'$. The claim then follows from Equation 3 combined with the monotonicity of h. □

This claim allows us to analyze BP-A in place of BP-cond. However, BP-A is still difficult to deal with because the sizes of the communities may be different. So, we introduce a process which simulates a graph where all communities have size $|C_i^*|$.

BP-B: Let G_B be a graph divided into communities of size $|C_i^*|$, with $\frac{n}{|C_i^*|}$ communities in total. Follow the same process as in in BP-A (starting from the same node), except on G_B instead of G_A. Z_B gives the total number of nodes reached.

Claim 2. *Z_A is stochastically dominated by Z_B.*

Proof. The idea is to interpolate between BP-A and BP-B by considering a series of local moves in which we split one of the communities in G_B into two smaller communities. Consider a series of graphs $G_B = G_1....G_W = G_A$ with the following property: G_{i+1} is equal to G_i except that a single community C_i of G_i is split into two communities C_i^1 and C_i^2. With each G_i, we can associate a branching process BP-i and corresponding Z_i. We will show that for any i, Z_i stochastically dominates Z_{i+1}. Since for any G_A there exists a sequence of local moves that can produce it from G_B, this will show that Z_B stochastically dominates Z_A.

To prove that Z_i stochastically dominates Z_{i+1}, we couple BP-i and BP-$(i+1)$ by sharing the status (live or not) of every edge in the graph between them. If BP-$(i+1)$ reaches either C_i^1 or C_i^2, then BP-i reaches $C_i = C_i^1 \cup C_i^2$. Hence, every community that is visited by BP-$(i+1)$ is also visited by BP-i. This establishes that Z_i stochastically dominates Z_{i+1}, as desired. □

BP-B is a nonuniform branching process in which the distribution of the number of children at each step depends on the total number of communities which remain to be explored. Note that G_B has $\frac{n}{|C_i^*|}$ communities in total. Suppose that BP-B has explored k communities so far. Define q_{eff} to be the "effective" probability of a live edge between two communities:

$$q_{\text{eff}} = 1 - (1 - p_b q)^{|C_i^*|^2}$$

By definition, we have $q_{\text{eff}} \frac{n}{|C_i^*|} \leq c_{max} < 1$. The number of children spawned by the kth community is distributed as $Bin(\frac{n}{|C_i^*|} - k, q_{\text{eff}})$. Since this nonuniform process is difficult to analyze, we note that it is stochastically dominated by a final branching process:

BP-uniform: A Galtson-Watson branching process with offspring distribution $Bin(\frac{n}{|C_i^*|}, q_{\text{eff}})$.

X_i represents the ith largest connected component in G, which we have established is stochastically dominated by the corresponding component generated by BP-B. For simplicity, we upper bound the ith largest component by the single largest component. In G_B there are at most $\frac{n}{|C_i^*|}$ components. The maximum of $\frac{n}{|C_i^*|}$ draws from BP-B is stochastically dominated by the maximum of $\frac{n}{|C_i^*|}$ draws of $Z_{uniform}$.

Claim 3. *Draw $Z_1...Z_N$ iid as $Z_{uniform}$. Then*

$$\mathbb{E}\left[\max Z_i\right] \leq 12 \left(\frac{1}{1 - \frac{n}{|C_i^*|} q_{eff}} \right) \log N$$

Proof. For any j, let ξ_j be iid from $Bin(\frac{n}{|C_i^*|}, q_{\text{eff}})$. Draief and Massoulie [2] (Lemma 1.9) give the following tail bound for Z_i:

$$\Pr[Z_i \geq K] \leq \Pr\left[\sum_{j=1}^{K} \xi_j \geq K \right]$$

$\sum_{j=1}^{K} \xi_j$ is distributed as $Bin(K\frac{n}{|C_i^*|}, q_{\text{eff}})$, so via Chernoff bound we have

$$\Pr\left[\sum_{j=1}^{K} \xi_j \geq K\right] \leq \exp\left(-\frac{1}{3}K\left(1 - \frac{n}{|C_i^*|}q_{\text{eff}}\right)\right)$$

So, we see that Z_i is stochastically dominated by an exponential random variable with mean $\lambda = \frac{1}{3}\left(1 - \frac{n}{|C_i^*|}q_{\text{eff}}\right)$. Dasarathy [1] (Eq. 7) show that the expected maximum of N exponential variables is upper bounded by $\frac{2\log N}{\lambda\left(1-\frac{1}{N}\right)}$. Noting that $1 - \frac{1}{N} \geq \frac{1}{2}$ and $\lambda \geq \frac{1}{3}(1 - c_{max})$, the claim follows. □

By substituting $N = \frac{n}{|C_i^*|}$ into Claim 3 and multiplying by $|C_i^*|$ (the size of each community in G_A), we conclude the proof of the lemma. We remark here that the reason we have a factor $1 - c_{max}$ and not $(1 - c_{max})^2$ is that we have bounded the *expectation* of the maximum of the N variables, not given a bound that holds with high probability.

□

With the key lemma in hand, we are now ready to proceed to the proof of our bound on OPT. Let $OPT(\boldsymbol{C^*})$ be a random variable which gives the expected optimal value conditioned on $\boldsymbol{C^*}$.

$$\begin{aligned}
OPT &= \mathop{\mathbb{E}}_{\boldsymbol{C^*}}\left[\mathbb{E}\left[OPT(\boldsymbol{C^*})\middle|\boldsymbol{C^*}\right]\right] \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{C^*}}\left[\sum_{i=1}^{K} X_i\middle|\boldsymbol{C^*}\right] \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{C^*}}\left[\sum_{i=1}^{K} \frac{12}{1-c_{max}}\log\left(\frac{n}{|C_i^*|}\right)|C_i^*|\right] \text{ (Lemma 22)} \\
&\leq \sum_{i=1}^{K} \frac{12}{1-c_{max}}\log\left(\frac{n}{|C_i|}\right)|C_i| \qquad (|C_i| \geq |C_i^*|).
\end{aligned}$$

Given the guarantee that $\sum_{i=1}^{K} \frac{12}{1-c_{max}}\log\left(\frac{n}{|C_i|}\right)|C_i| \geq OPT$, we now analyze how small $\sum_{i=1}^{K}|C_i|$ can be. We are interested in the value of the optimization problem

$$\begin{aligned}
&\min_{|C_1|\ldots|C_K|} \sum_{i=1}^{K}|C_i| \\
&\text{s.t. } \sum_{i=1}^{K} \frac{12}{1-c_{max}}\log\left(\frac{n}{|C_i|}\right)|C_i| \geq OPT
\end{aligned}$$

This can be reformulated as the convex program

$$\begin{aligned}
&\min_{|C_1|\ldots|C_K|} \sum_{i=1}^{K}|C_i| \\
&\text{s.t. } -\sum_{i=1}^{K} \frac{12}{1-c_{max}}\log\left(\frac{n}{|C_i|}\right)|C_i| \leq -OPT
\end{aligned}$$

We structurally characterize the optimal solution as follows. Let v^* denote the optimal value of the above convex program. Note that Slater's condition holds, and so we have strong duality. Consider the Lagrange dual function

$$\mathcal{L}(\lambda) = \inf_{|C_1|...|C_K|} \sum_{i=1}^{K} |C_i| + \lambda \left(OPT - \sum_{i=1}^{K} \frac{12}{1 - c_{max}} \log\left(\frac{n}{|C_i|}\right) |C_i| \right)$$

where the dual problem is

$$\max_{\lambda \geq 0} \mathcal{L}(\lambda).$$

Let λ^* be the optimal value of the Lagrange multiplier. We write

$$\begin{aligned} v^* &= \mathcal{L}(\lambda^*) \\ &= \inf_{|C_1|...|C_K|} \sum_{i=1}^{K} |C_i| + \lambda^* \left(OPT - \sum_{i=1}^{K} \frac{12}{1 - c_{max}} \log\left(\frac{n}{|C_i|}\right) |C_i| \right) \end{aligned} \tag{4}$$

Examining Equation 4, let $a_1...a_K$ be values of $|C_1|...|C_K|$ which achieve v^*. We must have that $a_1...a_K$ maximize $\sum_{i=1}^{K} \log\left(\frac{n}{a_i}\right) a_i$ subject to $\sum_{i=1}^{K} a_i = v^*$ (otherwise a smaller value could have been achieved). Since $\frac{12}{1-c_{max}} \log\left(\frac{n}{a_i}\right) a_i$ is concave, Jensen's inequality gives

$$\sum_{i=1}^{K} \log\left(\frac{n}{a_i}\right) a_i \leq K \log\left(\frac{n}{\frac{1}{K}\sum_{i=1}^{K} a_i}\right) \left(\frac{1}{K} \sum_{i=1}^{K} a_i\right).$$

That is, $\sum_{i=1}^{K} \log\left(\frac{n}{a_i}\right) a_i$ is maximized when $a_1 = a_2 = ... = a_K = \frac{1}{K}\sum_{i=1}^{K} a_i$. Thus, the optimal value v^* can be obtained when we restrict the space of feasible $|C_1|...|C_K|$ to points where all are equal. Let $\mu = \frac{1}{K}\sum_{i=1}^{K} |C_i|$ denote the average size of the top K communities. We rephrase the original optimization problem as

$$\begin{aligned} &\min_{\mu} \mu K \\ &\text{s.t. } \mu K \left(\frac{12}{1 - c_{max}} \log\left(\frac{n}{\mu}\right) \right) \geq OPT \end{aligned}$$

The constraint in this problem gives a lower bound on the possible size of μK. Thus we have

$$\sum_{i=1}^{K} |C_i| = \mu K \geq \left(\frac{1 - c_{max}}{12 \log \frac{n}{\mu}} \right) OPT$$

which concludes the proof of the lemma. □

In order to conclude the proof of the theorem, we wrap up a few details which allow us to use the machinery of the $p_b = 0$ case. First, we show that (conditioned on the random walk steps all staying in their starting community), the estimated degrees are still accurate. Formally, Lemma 5 still holds because the process is identical to the $p_b = 0$ case if the walk never leads its starting community. Our assumption that $p_b < \frac{1}{n}$ ensures that between-community edges do not impact the requirement that $d_{max} = O(\log^2 n)$ (Lemma 12), since a simple Chernoff bound guarantees that with high probability, no edge has more than $O(\log n)$ edges to nodes in other communities.

What has changed is that our estimate of the average degree accounts for between community edges: $\hat{S} = \frac{\hat{d} - np_b}{p_w - p_b}$. We now show that the equivalent of Corollary 1 holds when $p_b > 0$.

Lemma 23. *When $p_b > 0$, with probability at least $1 - 2\epsilon - o(1)$, the estimated size $\hat{S}$ for each sample from a community of size $|C|$ satisfies $(1-4\epsilon)|C| \leq \hat{S} \leq (1+4\epsilon)|C|$.*

Proof. Let $\hat{d}$ be the estimated average degree. Conditioning on concentration holding via Lemma 5, we have that

$$(1-2\epsilon)(p_w|C| + (n-|C|)p_b) \leq \hat{d} \leq (1+2\epsilon)(p_w|C| + (n-|C|)p_b)$$

We estimate the size as $\hat{S} = \frac{\hat{d} - np_b}{p_w - p_b}$. We now show the upper bound on $\hat{S}$; the argument for the lower bound is exactly the same.

$$\begin{aligned}
\hat{S} &\leq \frac{(1+2\epsilon)(p_w|C| + (n-|C|)p_b) - np_b}{p_w - p_b} \\
&\leq \frac{(1+2\epsilon)(p_w|C| + (n-|C|)p_b - np_b) + 2\epsilon np_b}{p_w - p_b} \\
&= (1+2\epsilon)|C| + \frac{2\epsilon np_b}{p_w - p_b}.
\end{aligned}$$

So, we just need to bound the size of $\frac{2\epsilon np_b}{p_w - p_b}$ relative to $|C|$. We know by Assumption 5 that $2\epsilon np_b < 2\epsilon$. Further, using Assumption 1,

$$\begin{aligned}
p_w - p_b &> \frac{\log|C|}{|C|} - \frac{1}{n} \\
&\geq \frac{\log|C|}{|C|} - \frac{1}{|C|} \\
&\geq \frac{1}{|C|}
\end{aligned}$$

from which we conclude that $\frac{2\epsilon np_b}{p_w - p_b} \leq 2\epsilon|C|$. Thus, $\hat{S} \leq (1+4\epsilon)|C|$. □

Thus, the samples in the $p_b > 0$ case satisfy the conditions of Corollary 1 from the $p_b = 0$ case. Lastly, via the bound on $\sum_{i=1}^{K}|C_i|$ in Lemma 21, we can apply Lemma 10 with $\sum_{i=1}^{K}|C_i| = \left(\frac{1-c_{max}}{12\log\frac{n}{\mu}}\right)OPT$ to obtain a bound on the total influence attained by the algorithm. Lemma 10 only counts influence that spreads between communities, not using between-community edges. Clearly, the algorithm's utility can only increase if we counted influence along these edges. The bound from Lemma 10 then directly implies that the algorithm influences at least

$$\left(\frac{1-c_{max}}{12\log\frac{n}{\mu}}\right)\frac{\beta_{min}^2}{\beta_{max}}\left(1 - e^{-(1-\epsilon)} - \epsilon - \frac{1}{K} - o(1)\right)OPT$$

nodes in expectation, which concludes the proof.

□

2 Estimating the surrogate objective g

In this section, we explain more detail our procedure for estimating the surrogate objective (g). Recall that we defined $g(X) = \sum_{i=1}^{L} f(X, C_i)$, i.e, the influence spread of X considering only within-community edges. We would like a way of estimated $\mathbb{E}[g(X)]$ using only local information. Note that the influence spread within each C_i depends only on the nodes in $X \cap C_i$, which we write as X_{C_i} for short. So, $\mathbb{E}[g(X)]$ can be rewritten as $\mathbb{E}\left[\sum_{i=1}^{L} f(X_{C_i}, C_i)\right]$. If we knew X_{C_i}, then we could calculate $\mathbb{E}\left[f(X_{C_i}, C_i)\right]$ by simulating draws from the SBM for the unobserved portions of C_i conditioned on the presence of the subgraphs that the algorithm visited. Thus, the main challenge is that we do not know what community each node belongs to.

We start out by rewriting the influence bound in terms of the marginal contribution made by each v_i. Let $\chi(v)$ give the community of vertex v. We can write the bound as

$$g(X) = \sum_{i=1}^{T} \mathop{\mathbb{E}}_{X \sim \boldsymbol{w}} \left[f(X_{\chi(v_i)} \cap \{v_1 ... v_i\}, \chi(v_i) - f(X_{\chi(v_i)} \cap \{v_1 ... v_{i-1}\}, \chi(v_i)) \right]$$

where $X \sim \boldsymbol{w}$ denotes a seed set X with each element independently sampled with probability proportional to $\boldsymbol{w}$. Taken at face value, this does not seem like an improvement because we still do not know $X_{\chi(v_i)}$ for each term. However, since we have an estimate for the size of $\chi(v_i)$, we know (approximately) how many other times $\chi(v_i)$ will have been sampled as well (approximately) the weight that each of these samples will have received. For each node, we can simulate a set $sim(v_i)$ which contains v_i plus a sample from the distribution of the other nodes that ARISEN sampled from $\chi(v_i)$ in its random walks. The only issue is that we do not know where each node of $sim(v_i)$ lies in the order $\{v_1 ... v_T\}$, i.e., whether it takes "precedence" over v_i when we compute the marginal contributions. The final ingredient we need to overcome this obstacle is to realize that there is nothing special about the ordering $\{v_1 ... v_T\}$; we can equivalently rearrange the nodes in any order. In fact, we take the expectation over a uniformly random permutation π of the ordering: we first draw π and then sum in the order $v_{\pi(1)} .... v_{\pi(T)}$. Via linearity of expectation, we can take a different permutation for each term $i = 1....T$, where the permutation in term i need only a establish an ordering among the elements of $sim(v_i)$. For any set X, let $[X]_{\pi}^{i}$ represent the first i elements of X in the permutation π. Then we can write the influence bound as

$$g(X) = \sum_{i=1}^{T} \mathop{\mathbb{E}}_{\pi, sim(v_i), X} \left[f([X \cap sim(v_i)]_{\pi}^{i}, \chi(v_i)) - f([X \cap sim(v_i)]_{\pi}^{i-1}, \chi(v_i)) \right].$$

In this final form, we can calculate each term by averaging over simulations of $sim(v_i)$, an ordering π on $sim(v_i)$, and set of seed nodes from $sim(v_i)$ that are chosen (given the simulated weights). As discussed earlier, we can the compute f by averaging over simulations of the draw of C_i, and simulating the ICM on each simulated community. Complete pseudocode for EstVal is given in Algorithm 2. The proof that EstVal accurately estimates g follows immediately from the construction given above.

Algorithm 2 EstVal

for $i = 1...len(\boldsymbol{w})$ **do**
  for $j = 1...M$ **do**
    Simulate G_i^j from $\mathcal{G}(p_w, \hat{S}_i)$ conditioned on H_i appearing.
    for $k = 1...P$ **do**
      π = a uniformly random permutation on $V(G_i^j)$
      $N \sim$ Binom $(T, \frac{\hat{S}_i}{n})$
      Draw $u_1...u_N$ uniformly random from $V(G_i^j) \setminus V(H_i)$
      for $\ell = 1...N$ **do**
        w_ℓ^{samp} = weight $\boldsymbol{w}$ assigns to a node with value $f(u_\ell, G_i^j)$
      end for
      X = a random subset of $u_1...u_N$ when $K-1$ nodes are chosen from all samples, the total weight is $||\boldsymbol{w}||_1$, and $u_1...u_N$ have corresponding weights from w^{samp}
      for $u \in X$ **do**
        if $\pi(s_i) > \pi(u)$, remove u from X
      end for
      $val+ = \frac{1}{MP}\left(1 - \left(1 - \frac{w_i}{||\boldsymbol{w}||_1}\right)^K\right)\left[f(\{s_i\} \cup X, G_i^j) - f(X, G_i^j)\right]$
    end for
  end for
end for
return val

3 Additional experimental results

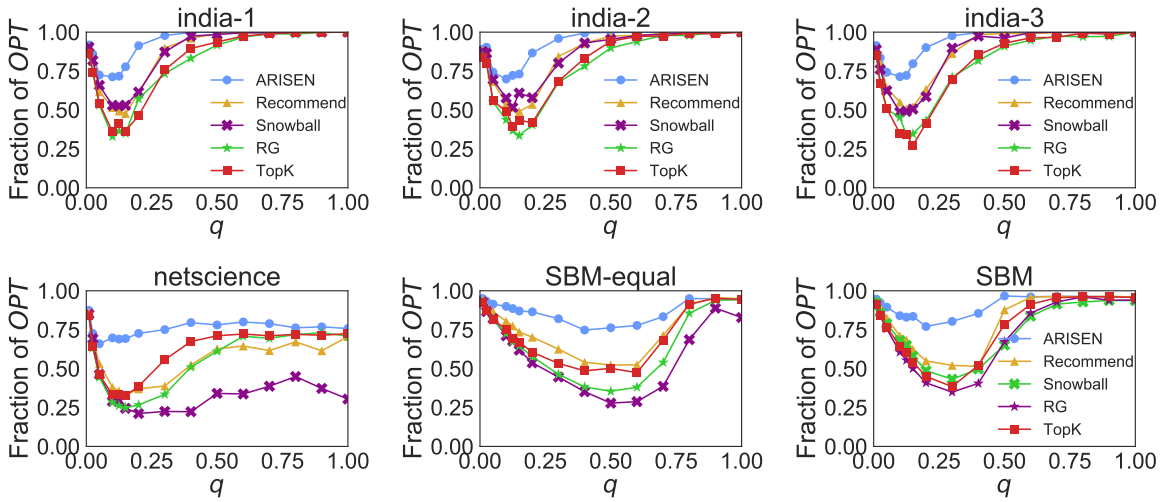
3.1 Parameter settings

In all runs we set $B = 0$ (no burn-in). The values for R and T can be found in the table below.

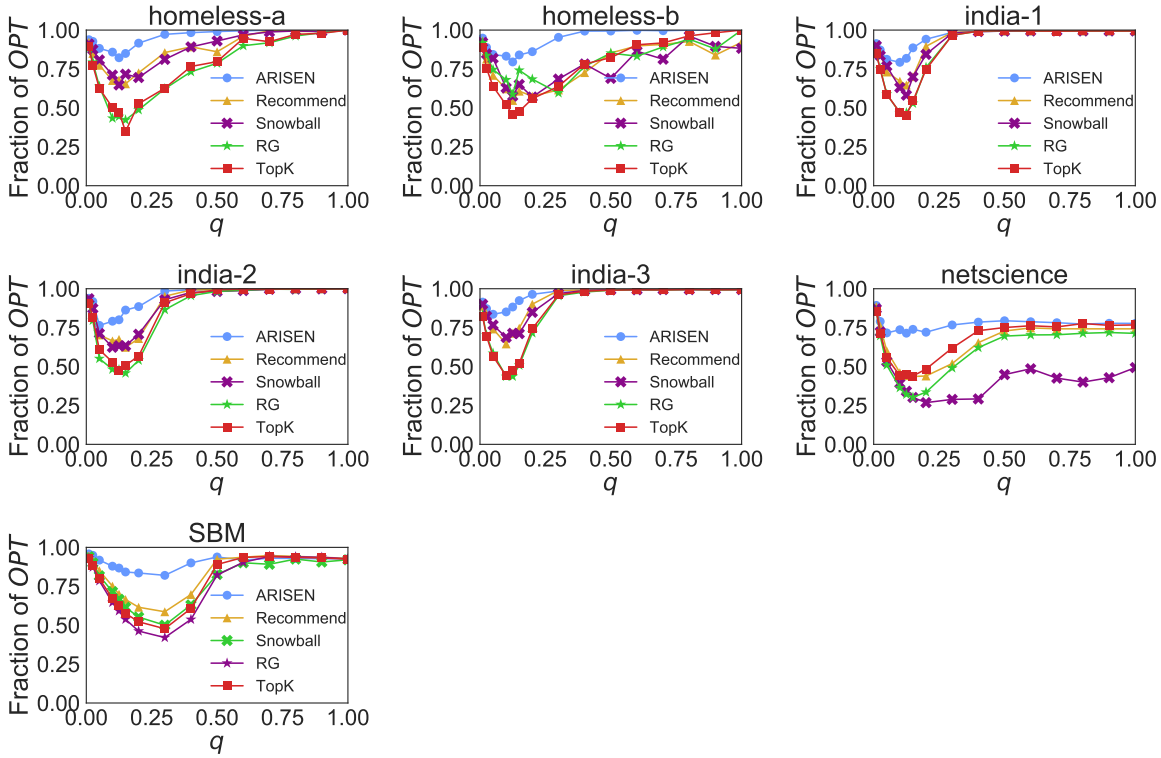
Network	K	T	R
homeless-a	$0.01 \cdot n$	5	10
homeless-a	$0.015 \cdot n$	5	10
homeless-a	$0.02 \cdot n$	5	10
homeless-b	$0.01 \cdot n$	7	12
homeless-b	$0.015 \cdot n$	7	12
homeless-b	$0.02 \cdot n$	7	12
india-1	$0.005 \cdot n$	10	15
india-1	$0.01 \cdot n$	10	15
india-1	$0.015 \cdot n$	10	15
india-1	$0.02 \cdot n$	10	15
india-2	$0.005 \cdot n$	7	12
india-2	$0.01 \cdot n$	10	12
india-2	$0.015 \cdot n$	10	12
india-2	$0.02 \cdot n$	10	12
india-2	$0.005 \cdot n$	6	25
india-2	$0.01 \cdot n$	12	25
india-2	$0.015 \cdot n$	18	25
india-2	$0.02 \cdot n$	25	25
netscience	$0.005 \cdot n$	40	25
netscience	$0.01 \cdot n$	40	25
netscience	$0.015 \cdot n$	40	25
netscience	$0.02 \cdot n$	40	25
SBM	$0.005 \cdot n$	6	25
SBM	$0.01 \cdot n$	12	25
SBM	$0.015 \cdot n$	18	25
SBM	$0.02 \cdot n$	25	25

3.2 Influence spread

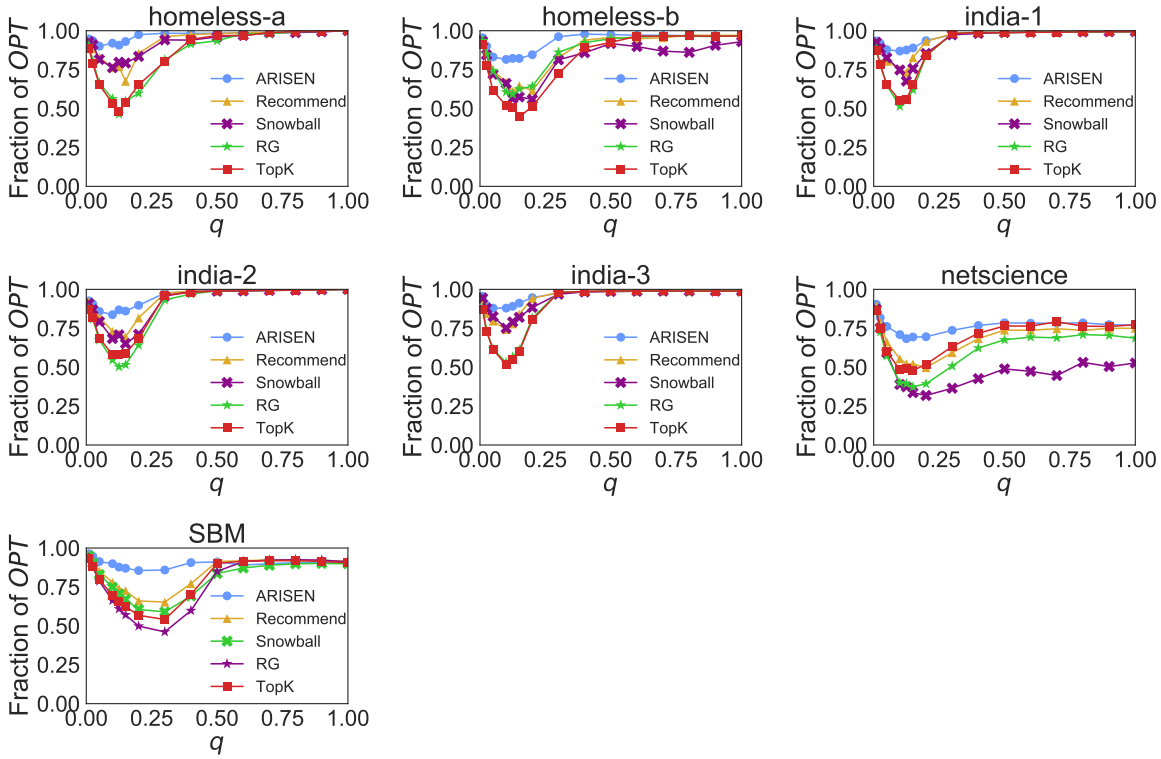
$K = 0.005 \cdot n$



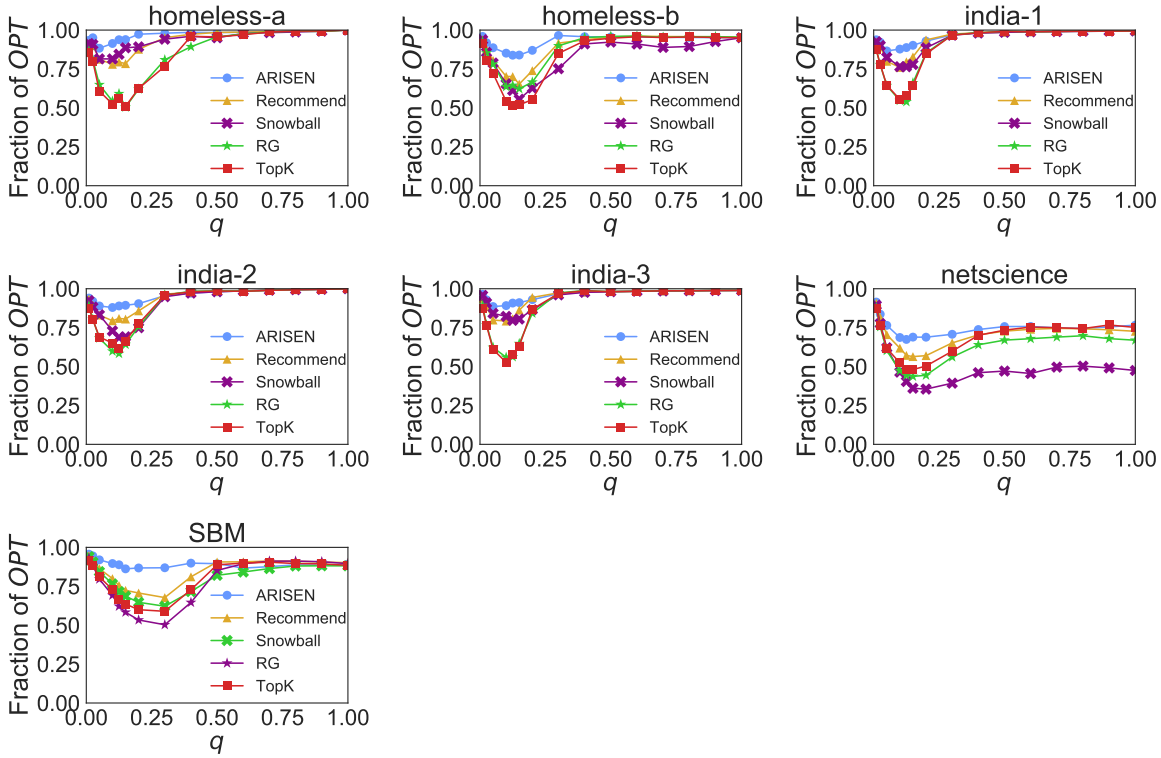
$K = 0.01 \cdot n$



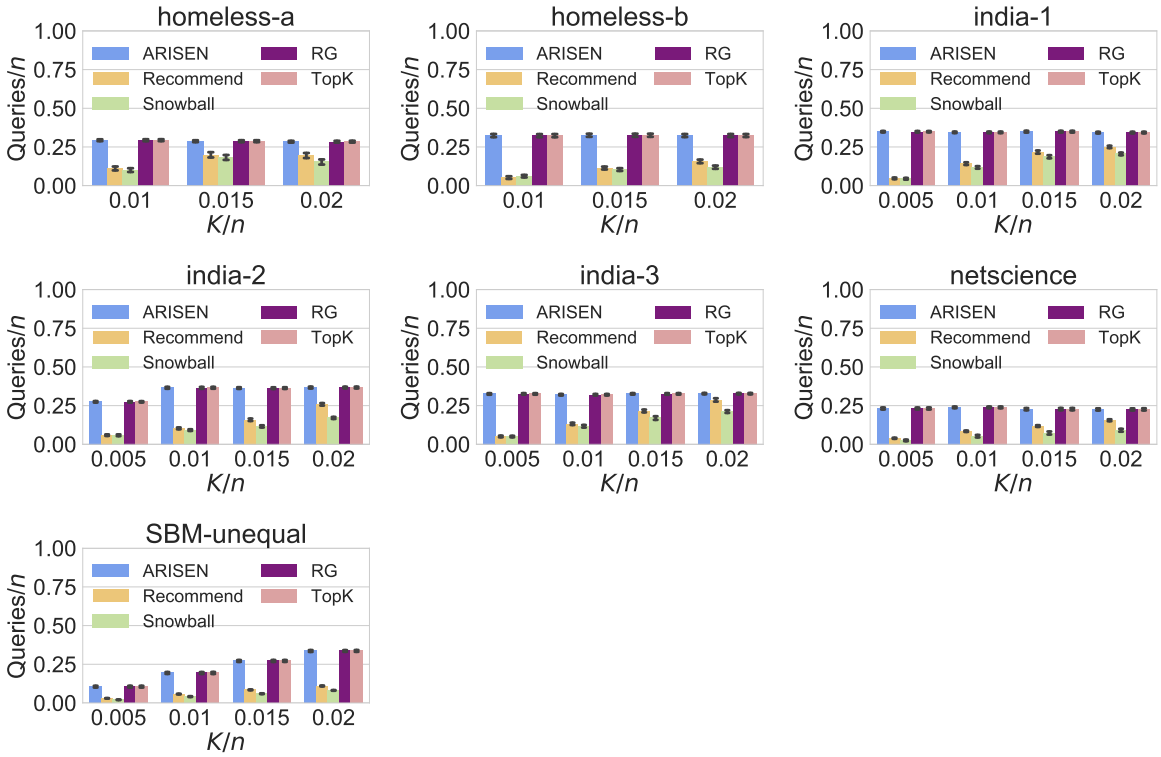
$K = 0.015 \cdot n$



$K = 0.02 \cdot n$



3.3 Query cost



References

[1] Gautam Dasarathy. A simple probability trick for bounding the expected maximum of n random variables. 2011.

[2] Moez Draief and Laurent Massouli. *Epidemics and rumours in complex networks*. Cambridge University Press, 2010.

[3] Christopher Hoffman, Matthew Kahle, and Elliot Paquette. Spectral gaps of random graphs and applications to random topology. *arXiv preprint arXiv:1201.0425*, 2012.

[4] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random graphs*, volume 45. John Wiley & Sons, 2011.

[5] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146. ACM, 2003.

[6] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Chapman & Hall/CRC, 2010.

[7] Daniel Paulin et al. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

[8] Daniel A Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.