2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Almost Exact Recovery in Gossip Opinion Dynamics over Stochastic Block Models

Yu Xing and Karl H. Johansson

Abstract— **We study community detection based on state observations from gossip opinion dynamics over stochastic block models (SBM). It is assumed that a network is generated from a two-community SBM where each agent has a community label and each edge exists with probability depending on its endpoints' labels. A gossip process then evolves over the sampled network. We propose two algorithms to detect the communities out of a single trajectory of the process. It is shown that, when the influence of stubborn agents is small and the link probability within communities is large, an algorithm based on clustering transient agent states can achieve almost exact recovery of the communities. That is, the algorithm can recover all but a vanishing part of community labels with high probability. In contrast, when the influence of stubborn agents is large, another algorithm based on clustering time average of agent states can achieve almost exact recovery. Numerical experiments are given for illustration of the two algorithms and the theoretical results of the paper.**

I. Introduction

Networks exist ubiquitously in various fields such as computer science, biology, and sociology. It is common that nodes in a network connect densely within subgroups but sparsely in general. Such subgroups are referred to as communities [1]. Community detection is one of the central questions in network science and studies how to find communities of a network. Often only state dynamics evolving over the network are observable, rather than the network itself. Hence, a growing number of studies have been investigating community detection based on state observations [2], [3], [4], [5], [6], [7]. Lacking network information makes community detection difficult, and it is still not clear how to detect communities based on a single trajectory of networked dynamics, which is considered in this paper.

A. Related Work

Community detection has been studied for two decades in multiple domains including physics and computer science [1], [8]. Traditional methods apply agglomerative or divisive clustering to pairs of nodes with given weights [9]. A popular concept for communities, called modularity, is introduced in [10]. The modularity measures the quality of a given graph partition from a random partition. The Louvain method [11] is a renowned fast community detection algorithm based on modularity. A statistical approach modeling communities is to introduce generative network models that have planted community structures. A canonical model is the stochastic block model (SBM), in which each node has a pre-assigned community label and each edge exists with independent probability depending on its endpoints' labels. Spectral clustering and belief propagation methods are commonly used for the detection problem [12]. Another detection approach is to execute dynamical processes over the network, for example, the Infomap algorithm [13]. The authors in [14] introduces a bounded-confidence model, in which agents eventually form clusters coinciding with communities of the network.

In the aforementioned approaches, the network is assumed to be known. However, in practice it is likely the dynamic state data are available, instead of the network itself. As a result, there is a growing interest in community detection for networked systems based on state observations. Maximum likelihood methods applied to cascade data are introduced in [2], [7]. The paper [2] also proposes a two-step procedure: first the underlying network is recovered and then agents are grouped from the network estimates. Blind community detection [4], [5], [6] uses sample covariance matrices of agent states to recover the community structure. Estimating covariance matrices requires capturing a single snapshot from each of multiple trajectories. The paper [3] investigates learning the network topology and the community structure at the same time, for epidemics and an Ising model. The papers [15], [16] consider a gossip model over a weighted graph with two communities, where agents within the same community have the same interaction probability, different from the interaction probability between communities. It is shown that the community structure can be recovered by clustering the state time average of the process. There is a need to investigate how detection algorithms based on a single trajectory can be applied to general graphs.

In this paper we study the detection of communities from the gossip model with stubborn agents. The problem is related to recently increasing research of estimating network structure from social dynamics [17]. Network information is useful in applications, but directly collecting such data is hard because networks are topic specific [18] and perturbed by noise [19]. Detecting communities for a coarse characterization of a network is a better choice than estimating all edges of that network, which is computationally expensive. The gossip update rule describes the stochastic nature of personal encounters and is a key building block of more complex opinion models [20]. In social network modeling,

This work was supported by the Knut and Alice Wallenberg Foundation (Wallenberg Scholar Grant), the Swedish Research Council (Distinguished Professor Grant 2017-01078), the Swedish Foundation for Strategic Research (CLAS Grant RIT17-0046).

The authors are with Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, and also with Digital Futures, Stockholm, Sweden. Email: {yuxing2,kallej}@kth.se

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383465

media, influential bloggers, and opinion leaders can be seen as stubborn agents, whose existence may have a great effect on the dynamics and result in opinion fluctuation [21].

B. Contributions

This paper studies community detection based on a single trajectory of gossip opinion dynamics over a two-community SBM. Two detection algorithms are proposed. The first algorithm (Algorithm 1) is based on applying the k-means algorithm to agent states in a given transient time interval. It shown in Theorem 1 that, if the influence of stubborn agents is small and the link probability within communities is large, the algorithm can recover all but a vanishing proportion of community labels of the SBM with high probability (i.e., almost exact recovery). The time interval depends on the relative magnitude of link probability within and between communities. The second algorithm (Algorithm 2) deals with the case where the influence of stubborn agents is large. The algorithm computes the time average of agent states and returns community estimates by clustering the time average after a given time step. Theorem 2 states that this algorithm can also achieve almost exact recovery.

The results generalize the earlier works [15], [16], in which the graph with two communities is deterministic and fixed. The difficulty lies in theoretically characterizing the relation between agent states and the community structure of the SBM. We verify almost exact recovery by using new concentration results developed in [22], [23].

The results show that a single trajectory of the gossip model can be enough for achieving almost exact recovery. Also, the proposed algorithm based on transient states indicates that excitation from stubborn agents may not be necessary to guarantee recovery. The analysis framework provides insight into design and analysis of community detection algorithm based on state observations. Given a process evolving over a structured network, we can first study how the community structure influences the dynamics, and then exploit the obtained properties to design detection algorithms.

C. Outline

Section II introduces the SBM and the gossip model. Section III defines almost exact recovery of the SBM and formulates the problem. Section IV provides two algorithms and their performance analysis. Numerical experiments are presented in Section V. Section VI concludes the paper.

Notation. Let $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$, and $\mathbb{N}$ be the n-dimensional Euclidean space, the set of $n\times m$ real matrices, and the set of nonnegative integers, respectively. Denote $\mathbb{N}_+ = \mathbb{N}\setminus\{0\}$. For $x\in\mathbb{R}$, $\log x$ is the natural logarithm of x. Denote the n-dimensional all-one vector and the unit vector with i-th entry being one by $\mathbf{1}_n$ and $e_i^{(n)}$, respectively. The superscript (n) is omitted if the context is clear. I_n is the $n\times n$ identity matrix, and $\mathbf{1}_{m,n}$ ($\mathbf{0}_{m,n}$) is the $m\times n$ all-one (all-zero) matrix. Denote the Euclidean norm of a vector and the spectral norm of a matrix by $\|\cdot\|$. For $x\in\mathbb{R}^n$, x_i is its i-th entry, and for $A\in\mathbb{R}^{n\times n}$, a_{ij} or $[A]_{ij}$ is its (i,j)-th entry. The cardinality of a set $\mathcal{S}$ is $|\mathcal{S}|$. The function $\mathbb{I}_{[\text{property}]}$ is the indicator function, which is one if the property in the bracket holds, and is zero otherwise. Denote the probability of an event A by $\mathbb{P}\{A\}$ and the expectation of a random vector X by $\mathbb{E}\{X\}$. For real numbers $a(n)$ and $b(n)>0$, $n\in\mathbb{N}$, denote $a(n)=O(b(n))$ if $|a(n)|\le Cb(n)$ for all $n\in\mathbb{N}$ and some $C>0$, $a(n)=o(b(n))$ if $|a(n)|/b(n)\to 0$. If further $a(n)>0$, $n\in\mathbb{N}$, denote $a(n)=\omega(b(n))$ if $b(n)=o(a(n))$, $a(n)=\Omega(b(n))$ if $b(n)=O(a(n))$, and $a(n)=\Theta(b(n))$ if both $a(n)=O(b(n))$ and $a(n)=\Omega(b(n))$. We will use subscripts to emphasize the dependence on n, for example, $a(n)=o_n(b(n))$. Denote $x\vee y:=\max\{x,y\}$ and $x\wedge y:=\min\{x,y\}$, $x,y\in\mathbb{R}$. An undirected graph $\mathcal{G}=(\mathcal{V},\mathcal{E},A)$ has the agent set $\mathcal{V}$, the edge set $\mathcal{E}$, and the adjacency matrix $A=[a_{ij}]$ with $a_{ij}=1$ ($a_{ij}=0$) if $\{i,j\}\in\mathcal{E}$ ($\{i,j\}\notin\mathcal{E}$). The degree of $i\in\mathcal{V}$ is $d_i=\sum_{j\in\mathcal{V}}a_{ij}$.

II. Preliminaries

A. Two-Community SBM

In this subsection, we define a two-community SBM. The SBM characterizes the community structure of real networks. For a graph $\mathcal{G}$, we assume that its agent set $\mathcal{V}$ can be represented by the union of two disjoint sets $\mathcal{V}_{\text{r}1}$ and $\mathcal{V}_{\text{r}2}$. Let the vector $\mathcal{C}\in\{1,2\}^n$ be such that $\mathcal{C}_i=1$ if $i\in\mathcal{V}_{\text{r}1}$ and $\mathcal{C}_i=2$ if $i\in\mathcal{V}_{\text{r}2}$. In other words, agents in $\mathcal{V}_{\text{r}1}$ (in $\mathcal{V}_{\text{r}2}$) have the label 1 (label 2). We call $\mathcal{V}_{\text{r}1}$ and $\mathcal{V}_{\text{r}2}$ the communities of the graph and $\mathcal{C}$ the community structure.

Definition 1 (SBM): Let $n\in\mathbb{N}_+$ be an even number, $\boldsymbol{l}=[l_{\text{s}}\ l_{\text{d}}]^{\text{T}}=[l_{\text{s}}(n)\ l_{\text{d}}(n)]^{\text{T}}\in(0,1)^2$. The SBM$(n,\boldsymbol{l})$ is a random graph. The SBM assigns agents $1,\dots,n/2$ (agents $1+n/2$, $\dots$, n) with label 1 (label 2). Then it generates an undirected graph $\mathcal{G}=(\mathcal{V},\mathcal{E},A)$, without self-loops, by independently adding $\{i,j\}$ with $i\neq j$ to $\mathcal{E}$ with probability $l_{\text{s}}\mathbb{I}_{[\mathcal{C}_i=\mathcal{C}_j]}+l_{\text{d}}\mathbb{I}_{[\mathcal{C}_i\neq\mathcal{C}_j]}$. ■

From this definition, we know that a graph generated from an SBM has communities $\mathcal{V}_{\text{r}1}=\{1,\dots,n/2\}$, $\mathcal{V}_{\text{r}2}=\{1+n/2,\dots,n\}$, and community structure $\mathcal{C}=[\mathbf{1}_{n/2}^{\text{T}}\ 2\mathbf{1}_{n/2}^{\text{T}}]^{\text{T}}$.

Remark 1: The size of communities can be a random variable instead of a fixed number as assumed in the definition (see Remark 3 of [12]). ■

B. Gossip Model with Stubborn Agents

This subsection introduces the gossip model with stubborn agents. The model captures random personal encounters. The gossip model is a random process over an undirected graph $\mathcal{G}=(\mathcal{V},\mathcal{E},A)$. The agent set $\mathcal{V}=\mathcal{V}_{\text{r}}\cup\mathcal{V}_{\text{s}}$ (disjoint) has both regular and stubborn agents. For convenience, denote $\mathcal{V}_{\text{r}}=\{1,\dots,n_{\text{r}}\}$ and $\mathcal{V}_{\text{s}}=\{1+n_{\text{r}},\dots,n_{\text{s}}+n_{\text{r}}\}$, so $|\mathcal{V}|=n=n_{\text{r}}+n_{\text{s}}$. Here n_{r} (n_{s}) is the number of regular (stubborn) agents, and n is the network size. Regular agents have opinion vector $X(t)\in\mathbb{R}^{n_{\text{r}}}$ at time $t\in\mathbb{N}$, and $X_i(t)$ is the opinion of the agent i. Stubborn agents have opinion vector $z^{(\text{s})}\in\mathbb{R}^{n_{\text{s}}}$ with $z_j^{(\text{s})}$ being the opinion of the stubborn agent $j+n_{\text{r}}$. An edge $\{i,j\}$ is chosen at each time t independent of previous updates, with an interaction probability $w_{ij}=w_{ji}=a_{ij}/\alpha$, where $\alpha=\sum_{i=1}^n\sum_{j=i+1}^n a_{ij}=|\mathcal{E}|$ and $W=[w_{ij}]\in\mathbb{R}^{n\times n}$ is the interaction probability matrix. The two corresponding

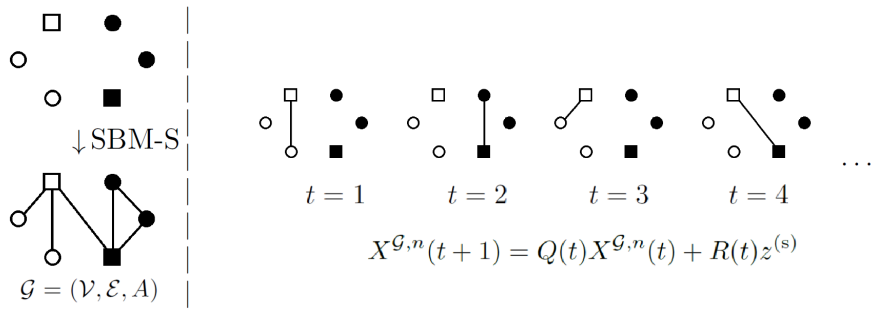


Fig. 1. Illustration of the gossip model over an SBM-S. On the left side of the figure, a network is generated from the SBM-S and then fixed. Circles and squares represent regular and stubborn agents, respectively. The black and white are two communities. On the right, the gossip model evolves over the generated network. An edge is selected at a time.

agents then update according to the following rule: If i and j are regular, then $X_i(t+1) = X_j(t+1) = (X_i(t)+X_j(t))/2$. If one of them is stubborn, for example j, then j does not update, but $X_i(t+1) = (X_i(t) + z_j^{(\mathrm{s})})/2$. All other agents keep their states at t. The compact form of the update is

$$X(t+1) = Q(t)X(t) + R(t)z^{(\mathrm{s})}, \tag{1}$$

with $\{[Q(t)\ \ R(t)]\}$ a sequence of i.i.d. random matrices. With probability w_{ij} if $i, j \in \mathcal{V}_\mathrm{r}$ then $[Q(t),\ R(t)] = [I_{n_\mathrm{r}} - \frac{1}{2}(e_i^{(n_\mathrm{r})} - e_j^{(n_\mathrm{r})})(e_i^{(n_\mathrm{r})} - e_j^{(n_\mathrm{r})})^\mathrm{T},\ \mathbf{0}_{n_\mathrm{r},n_\mathrm{s}}]$, and if $i \in \mathcal{V}_\mathrm{r}, j \in \mathcal{V}_\mathrm{s}$ then $[I_{n_\mathrm{r}} - \frac{1}{2}e_i^{(n_\mathrm{r})}(e_i^{(n_\mathrm{r})})^\mathrm{T},\ \frac{1}{2}e_i^{(n_\mathrm{r})}(e_j^{(n_\mathrm{s})})^\mathrm{T}]$. Here we replace $e_{j-n_\mathrm{r}}^{(n_\mathrm{s})}$ with $e_j^{(n_\mathrm{s})}$ for $j \in \mathcal{V}_\mathrm{s}$ for convenience.

C. SBM with Stubborn Agents

To define a gossip model over an SBM, in this section we introduce an SBM with stubborn agents based on Definition 1.

Definition 2 (SBM with stubborn agents, SBM-S): Let the number of regular agents $n_\mathrm{r} \in \mathbb{N}_+$ be an even number, the number of stubborn agents $n_\mathrm{s} \in \mathbb{N}_+$, the network size $n = n_\mathrm{r} + n_\mathrm{s}$, the link probability between regular agents $\boldsymbol{l} = [l_\mathrm{s}\ l_\mathrm{d}]^\mathrm{T} = [l_\mathrm{s}(n)\ l_\mathrm{d}(n)]^\mathrm{T} \in (0,1)^2$, and the link probability matrix between regular and stubborn agents $L^{(\mathrm{s})} = [l_{ij}^{(\mathrm{s})}] = [l_{ij}^{(\mathrm{s})}(n)] \in [0,1)^{n_\mathrm{r}\times n_\mathrm{s}}$. The SBM-S$(n_\mathrm{r}, n_\mathrm{s}, \boldsymbol{l}, L^{(\mathrm{s})})$ is a random graph which first generates an undirected graph on the n_r regular agents from SBM$(n_\mathrm{r}, \boldsymbol{l})$ and then add each edge $\{i,j\}$ to the graph with probability $l_{ij-n_\mathrm{r}}^{(\mathrm{s})}$ for $i \in \mathcal{V}_\mathrm{r} = \{1,\ldots,n_\mathrm{r}\}$ and $j \in \mathcal{V}_\mathrm{s} = \{1+n_\mathrm{r},\ldots,n\}$. ■

The SBM-S includes stubborn agents in a network. The probability $l_{ij}^{(\mathrm{s})}$ measures the possibility of the regular agent i connected to the stubborn agent $j+n_\mathrm{r}$. Let $r_0 := n_\mathrm{r}/n \in (0,1)$ be the proportion of the regular, and $s_0 := n_\mathrm{s}/n \in (0,1)$ that of the stubborn. Hereafter by a gossip model we mean a gossip model that evolves over a sample graph generated by an SBM-S (see Fig. 1 for an illustration).

III. Problem Formulation

We study how to detect the community structure of an SBM-S out of state observations. To measure the performance of a detection algorithm, denote the accuracy of an estimate $\hat{\mathcal{C}}$ of the community structure $\mathcal{C}$ by

$$\mathrm{Acc}(\mathcal{C},\hat{\mathcal{C}}) := \frac{1}{n}\max\left\{\sum_{i=1}^n \mathbb{I}_{[\mathcal{C}_i=\hat{\mathcal{C}}_i]}, \sum_{i=1}^n \mathbb{I}_{[\mathcal{C}_i=3-\hat{\mathcal{C}}_i]}\right\}. \tag{2}$$

The first summation in (2) represents the number of identical entires between $\mathcal{C}$ and $\hat{\mathcal{C}}$. Note that $\mathcal{C}_i \in \{1,2\}$, so $3-\hat{\mathcal{C}}_i$ swaps the community label of i, and the second term in (2) represents the number of identical entires between $\mathcal{C}$ and the swapped estimated labels. That is, the accuracy is defined up to a permutation of labels. Now define almost exact recovery of an algorithm detecting communities in SBMs as follows [12].

Definition 3: For an SBM with n agents and a community structure $\mathcal{C}$, suppose that a detection algorithm outputs an estimation of the community structure $\hat{\mathcal{C}}$. We say that the algorithm achieves almost exact recovery, if

$$\mathbb{P}\{\mathrm{Acc}(\mathcal{C},\hat{\mathcal{C}}) = 1 - o_n(1)\} = 1 - o_n(1). \quad ■$$

Remark 2: Almost exact recovery means that the algorithm can detect most community labels (up to a permutation) except for a vanishing part with probability approaching one, as the network size increases. ■

The problem considered in this paper is as follows:

Problem: For an SBM-S and the gossip model over this SBM-S, given a trajectory of the model, $\{X(0),\ldots,X(t),\ldots\}$, propose algorithms that use the trajectory data to achieve almost exact recovery of the regular agents' community structure. ■

In the next section, we will show that almost exact recovery can be achieved from clustering transient agent states (Theorem 1). If the influence of stubborn agents is large enough, then almost exact recovery can be achieved from clustering state time averages of the model (Theorem 2).

IV. Detection Algorithms and Performance

In this section we propose two detection algorithms using a single trajectory of the gossip model, and show that the algorithm using transient states (Algorithm 1) achieves almost exact recovery (Theorem 1), and so does the algorithm using state time average (Algorithm 2 and Theorem 2).

First, we introduce the following assumptions.

Assumption 1: Suppose that the following hold.

(i) There exists $l^{(\mathrm{s})} \geq 0$ such that $\sum_{1\leq j\leq n_\mathrm{s}} l_{ij}^{(\mathrm{s})} = l^{(\mathrm{s})} =$

Algorithm 1 (Recovery Based on Transient States)

Input: Trajectory $\{X(t), t \in \mathbb{N}\}$.
Output: Community estimate $\hat{\mathcal{C}}$.

1: Select a vector $X(t)$ with $t \in (\Theta(n \log n), o(nl_{\rm s}/l_{\rm d}))$.
2: Obtain $\hat{\mathcal{C}}$ by applying k-means to $X(t)$.

$l^{(\rm s)}(n)$ for all $i \in \mathcal{V}_{\rm r}$.
(ii) The vectors $X(0)$ and $z^{(\rm s)}$ are deterministic, and satisfy that $|X_i(0)| \leq c_x$ and $|z_j^{(\rm s)}| \leq c_x$, for all $1 \leq i \leq n_{\rm r}$ and $1 \leq j \leq n_{\rm s}$, and some constant $c_x > 0$.
(iii) The proportion of regular agents r_0 is a constant and $r_0 n$ is an even number. ■

Remark 3: In (i), $l^{(\rm s)}$ represents the total influence of stubborn agents on one regular agent. We impose the first assumption for simplicity. It is possible to analyze the general problem by using matrix perturbation theory and the upper and lower bounds of $\sum_{1 \leq j \leq n_{\rm s}} l_{ij}^{(\rm s)}$. The assumption (ii) implies that the process is bounded. Finally, note that r_0 in (iii) is fixed but can be any number in $(0, 1)$. The case where $r_0 n$ is an odd number can be studied similarly to the case of even numbers. ■

We also need the following assumption for the link probabilities of the SBM-S.

Assumption 2: It holds that $l_{\rm s} = \omega((\log n)/n)$ and $l^{(\rm s)} = \omega(\log n)$. ■

Remark 4: The assumption show that we consider the logarithm regima of SBMs, where an SBM generates connected graphs [12]. The condition $l^{(\rm s)} = \omega(\log n)$ is required because the influence of stubborn agents needs to be large enough to ensure concentration. This condition may be removed, which is left to future work. ■

We now propose the first detection algorithm (Algorithm 1), based on the intuition that regular agents within communities may have similar states when the process has not evolved too long, if the influence of stubborn agents is small. Although simple, proving the recovery result needs analysis of concentration of $X(t)$, which is not trivial (see [22] for the details). In application, $l_{\rm s}$ and $l_{\rm d}$ are unknown, so we can use $X(t)$ with $t = \Theta(n \log n)$ for clustering. Improvement of the algorithm will be studied in the future. The recovery performance of Algorithm 1 is given in the following theorem.

Theorem 1: Suppose that Assumptions 1 and 2 hold. If $l_{\rm s}/(\log n) = \omega(l_{\rm d})$, $(l_{\rm s} \wedge l_{\rm d})n/(\log n) = \omega(l^{(\rm s)})$, $l_{\rm d} = \omega([l_{\rm s}(\log n)/n]^{1/2})$ and $\sum_{i \in \mathcal{V}_{\rm r1}} X_i(0) \neq \sum_{j \in \mathcal{V}_{\rm r2}} X_j(0)$, then Algorithm 1 achieves almost exact recovery. ■

Proof Sketch: The main idea of the proof is to compare the gossip dynamics over an SBM-S with a gossip dynamics over an averaged graph. The averaged graph $\bar{\mathcal{G}} = (\mathcal{V}, \bar{\mathcal{E}}, \mathbb{E}\{A\})$ is obtained by averaging all possible graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ generated from the SBM-S, where $\mathbb{E}\{A\}$ is the weighted adjacency matrix and $[\mathbb{E}\{A\}]_{ij} = l_{\rm s}\mathbb{I}_{[\mathcal{C}_i = \mathcal{C}_j]} + l_{\rm d}\mathbb{I}_{[\mathcal{C}_i \neq \mathcal{C}_j]}$ for $i \neq j \in \mathcal{V}_{\rm r}$. A gossip dynamics taking place over this averaged graph has an interaction probability matrix $\mathcal{W} = \mathbb{E}\{A\}/\mathbb{E}\{\alpha\}$.

As shown in Theorem 1 of [22], agent states of the gossip dynamics over the averaged graph are close to opinion averages within their corresponding communities, under the assumptions of the current theorem. By using Bernstein concentration inequalities [24], [25], we can show that the variance of agent states $X(t)$ of the original gossip dynamics is close to that over the averaged graph. Hence agent states of the original gossip dynamics are also close to the corresponding community averages. Note that the k-means problem on the real line can be solved optimally [26]. By a counting argument, it is possible to show that the optimal partition of $X(t)$ is the same as the community structure, except for a vanishing set of agents. The conclusion then follows. ■

Algorithm 2 (Recovery Based on State Time Average)

Input: Trajectory $\{X(t), t \in \mathbb{N}\}$, time step for clustering $T \in \mathbb{N}_+$.
Output: Community estimate $\hat{\mathcal{C}}$.

1: Set $S(0) = X(0)$.
2: **for** $t = 1, \ldots, T$ **do**
3: Compute

$$S(t) = \frac{t}{t+1} S(t-1) + \frac{1}{t+1} X(t).$$

4: **end for**
5: Obtain $\hat{\mathcal{C}}$ by applying k-means to $S(T)$.

Remark 5: The theorem indicates that, if the influence of stubborn agents is small and the link probability within communities is large, then it is possible to recover most labels based on agent states at a time step smaller than $nl_{\rm s}/l_{\rm d}$. The possible time interval depends on the relative magnitude of link probability within and between communities. More careful analysis can remove the term $\log n$ in the condition. The condition $(l_{\rm s} \wedge l_{\rm d})n/(\log n) = \omega(l^{(\rm s)})$ implies that link probability between regular agents grows faster with n than the influence of stubborn agents. This represents the maximum allowable effect of stubborn agents on regular agents for guaranteeing almost exact recovery. ■

When the influence of stubborn agents is large, the process can reach its steady states quickly [22], so it is difficult to find an interval suggested by Algorithm 1. For this case, we compute the state time average and cluster this vector to obtain an estimate of the community structure (Algorithm 2). In the algorithm, the state time average $S(t) = (\sum_{j=0}^{t-1} X(i))/t$ is computed recursively. In practice, the link probabilities are unknown, so it may be hard to decide how to set T in Algorithm 2. A possible way is to cluster $S(t)$ for every $t \in \mathbb{N}$ and to terminate the process when the change of community estimates is below a given threshold.

To guarantee exact recovery of Algorithm 2 and simplify analysis, we need the following technical assumption.

Assumption 3: For the SBM-S and the stubborn agent states $z^{(\rm s)}$, it holds that $L^{(\rm s)} z^{(\rm s)} = [\zeta_1 \mathbf{1}_{n_{\rm r}/2}^{\rm T} \ \zeta_2 \mathbf{1}_{n_{\rm r}/2}^{\rm T}]^{\rm T}$ with $\zeta_1, \zeta_2 \in \mathbb{R}$ and $\zeta_1 \neq \zeta_2$. ■

Remark 6: The assumption means that the weighted average of stubborn-agent states for each regular agent is the same within communities, and the averages are different between the two communities. We introduce this condition

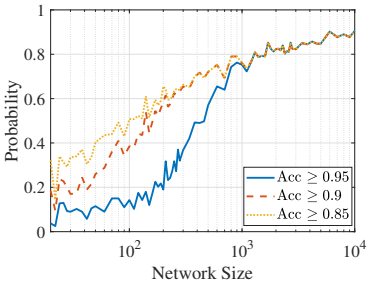


Fig. 2. Accuracy of Algorithm 1.

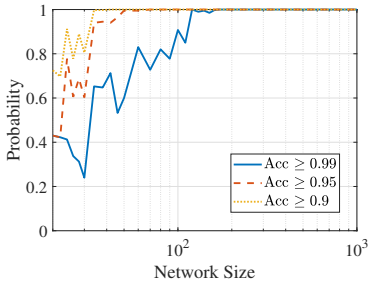


Fig. 3. Accuracy of Algorithm 2.

to ensure that the expected steady agent states between communities are different. Thus the community structure can be recovered based on these different values. The condition can be replaced by a bound controlling the difference between $[L^{(s)}z^{(s)}]_i$ and $[L^{(s)}z^{(s)}]_j$ for $i \in \mathcal{V}_{r1}$ and $j \in \mathcal{V}_{r2}$. ■

The almost exact recovery by Algorithm 2 is stated in the following theorem.

Theorem 2: Suppose that Assumptions 1–3 hold. If $l_s = \omega(l_d)$ and $l^{(s)} \geq l_s n$, then Algorithm 2 achieves almost exact recovery if $T = \Omega(n^3 l^{(s)})$. ■

Proof Sketch: Similar to the proof of Theorem 1, here we compare the expected final opinions $\mathbf{x}^{\mathcal{G},n} := (I - \mathbb{E}\{Q(t)\})^{-1}\mathbb{E}\{R(t)\}z^{(s)}$ with the expected final opinions of the gossip dynamics over the averaged graph. Theorem 4.3 of [23] guarantees that with high probability the difference between these two vectors can be bounded by a vanishing error. Moreover, Theorem 4.11 and Remark 4.12 of [23] yield that the state time average $S(t)$ is close to $\mathbf{x}^{\mathcal{G},n}$ for $t = \Omega(n^3 l^{(s)})$. Hence applying k-means to $S(t)$ yields the desired partition with high probability. ■

Remark 7: The theorem indicates that almost exact recovery can be achieved if the stubborn agent influence and the link probability between communities are large enough. The bound for the clustering time $\Theta(n^3 l^{(s)})$ can be relaxed, as observed in the numerical experiments in Section V. ■

V. Numerical Simulation

In this section we present numerical experiments to validate the main results. We test the proposed algorithms for gossip dynamics over an SBM-S and dynamics over a real network.

To generate an SBM-S with n agents, set the proportion of regular agents to be $r_0 = 0.9$, so the proportion of stubborn agents is $s_0 = 0.1$. Recall the link probability within communities l_s and the link probability between communities l_d. For the link probability between regular and stubborn agents, let

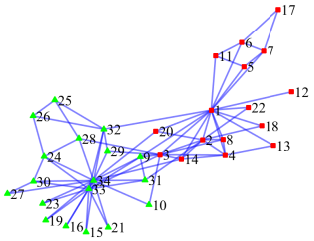
$$L^{(s)} = \begin{bmatrix} l_1^{(s)}\mathbf{1}_{n_r/2,n_s/2} & \mathbf{0}_{n_r/2,n_s/2} \\ \mathbf{0}_{n_r/2,n_s/2} & l_1^{(s)}\mathbf{1}_{n_r/2,n_s/2} \end{bmatrix}.$$

Intuitively this matrix shows that the stubborn agents also form two communities, and each stubborn community influences only one regular community, with link probability $l_1^{(s)}$. In this way we define an SBM-S$(n_r, n_s, \boldsymbol{l}, L^{(s)})$, where $\boldsymbol{l} = [l_s \ l_d]^T$. For the gossip model, we set the states of the first half of stubborn agents to be 1 and the other half to be -1.

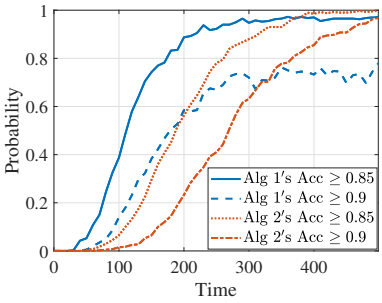
To show the almost exact recovery of Algorithm 1, we set $l_s = (\log n)^{2.5}/n$, $l_d = (\log n)/n$, and $l_1^{(s)} = \log n/n$. Generate the initial agent states in $\mathcal{V}_{r1}$ independently from the uniform distribution on $(-1, 0)$, and those in $\mathcal{V}_{r2}$ independently and uniformly from $(0, 1)$. Then run the gossip dynamics over the SBM-S for different n from 10 to 10^4. For each n, 20 graph samples of the SBM-S are generated and for each graph sample 20 trajectories are collected. Algorithm 1 is applied to $X(t)$ with $t = [n \log n]$ for each trajectory, where $[\cdot]$ is the rounding function. Recall the accuracy of an algorithm is the proportion of correctly detected community labels up to a permutation, as given in (2). Fig. 2 shows that the probability of the algorithm achieving a given accuracy increases with the number of agents. In addition, the accuracy of the algorithm also increases with the network size, when considering a given probability. These observations indicate that the algorithm achieves almost exact recovery.

For the almost exact recovery of Algorithm 2, we set set $l_s = (\log n)^2/n$, $l_d = (\log n)/n$, and $l_1^{(s)} = (\log n)^{2.5}/n$. Generate the initial states of all regular agents independently from the uniform distribution on $(-1, 1)$. Run the gossip model over the SBM-S for different n from 10 to 10^3. As previously, 20 graph samples are generated for each n and 20 trajectories are collected for each graph sample. Algorithm 2 is applied to each trajectory with the final step $T = [n(\log n)^{2.5}]$. Almost exact recovery of the algorithm is shown in Fig. 3. The result also indicates that Algorithm 2 performs better than Algorithm 1, and Algorithm 2 can recover all labels even for relatively small n. This may be because stubborn agents produce more excitation. Also note that T is much smaller than the condition given in Theorem 2, suggesting the possiblity of improving the results.

To test the performance of the proposed algorithms, we also conduct a numerical experiment over Zachary's karate club network, shown in Fig. 4(a). This network with two communities has been widely used as a benchmark in community detection. In the network, an edge represents interactions between agents. We assume that gossip dynamics take place over the network and we can observe only agent opinions but not interactions or the network. Agents 1 and 34 are leaders of the two communities, so they are set to be stubborn agents holding opinions 1 and -1, respectively. Other agents are regular ones with initial opinions



(a) The community structure of Zachary's karate club network. Red squares and green triangles show two communities.



(b) Accuracy of Algorithms 1 and 2 for community detection in gossip dynamics over Zachary's karate club network.

Fig. 4. Numerical experiment over Zachary's karate club network.

uniformed generated from $(-1, 1)$. During evolution, an edge in Fig. 4(a) is uniformly randomly selected at each time. We run the model for 400 times and apply Algorithms 1 and 2 to every time step of each trajectory. Fig. 4(b) shows that Algorithm 1 achieves high accuracy with high probability at the initial phase of the process, but Algorithm 2 performs better when the process reaches steady state. The experiments indicate the applicability of both algorithms.

VI. Conclusion

We studied community detection based on state observations from gossip opinion dynamics over a two-community SBM. Two algorithms were proposed and both of them use only a single trajectory of the process. When the influence of stubborn agents is small and the link probability within communities is large, it was shown that the algorithm based on clustering transient agent states can achieve almost exact recovery. In contrast, when the influence of stubborn agents is large, the algorithm based on clustering state time average can achieve almost exact recovery. Future work includes investigating the general SBM case and generalizing the algorithm to other networked dynamics.

References

[1] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[2] L. Prokhorenkova, A. Tikhonov, and N. Litvak, "When less is more: Systematic analysis of cascade-based community detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 16, no. 4, pp. 1–22, 2022.

[3] T. P. Peixoto, "Network reconstruction and community detection from dynamics," *Physical Review Letters*, vol. 123, no. 12, p. 128301, 2019.

[4] H.-T. Wai, S. Segarra, A. E. Ozdaglar, A. Scaglione, and A. Jadbabaie, "Blind community detection from low-rank excitations of a graph filter," *IEEE Transactions on Signal Processing*, vol. 68, pp. 436–451, 2019.

[5] M. T. Schaub, S. Segarra, and J. N. Tsitsiklis, "Blind identification of stochastic block models from dynamical observations," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 2, pp. 335–367, 2020.

[6] T. M. Roddenberry, M. T. Schaub, H.-T. Wai, and S. Segarra, "Exact blind community detection from signals on multiple graphs," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5016–5030, 2020.

[7] M. Ramezani, A. Khodadadi, and H. R. Rabiee, "Community detection using diffusion information," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 2, pp. 1–22, 2018.

[8] S. Fortunato and M. E. Newman, "20 years of network community detection," *Nature Physics*, pp. 1–3, 2022.

[9] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[10] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[12] E. Abbe, "Community detection and stochastic block models: Recent developments," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[13] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[14] I.-C. Morarescu and A. Girard, "Opinion dynamics with decaying confidence: Application to community detection in graphs," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1862–1873, 2010.

[15] Y. Xing, X. He, H. Fang, and K. H. Johansson, "Community detection for gossip dynamics with stubborn agents," in *IEEE Conference on Decision and Control*, pp. 4915–4920, 2020.

[16] Y. Xing, X. He, H. Fang, and K. H. Johansson, "Community structure recovery and interaction probability estimation for gossip opinion dynamics," *Automatica*, vol. 154, p. 111105, 2023.

[17] C. Ravazzi, F. Dabbene, C. Lagoa, and A. V. Proskurnikov, "Learning hidden influences in large-scale dynamical social networks: A data-driven sparsity-based approach, in memory of Roberto Tempo," *IEEE Control Systems Magazine*, vol. 41, no. 5, pp. 61–103, 2021.

[18] S. K. Cowan and D. Baldassarri, "'It could turn ugly': Selective disclosure of attitudes in political discussion networks," *Social Networks*, vol. 52, pp. 1–17, 2018.

[19] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1, pp. 211–222, 2012.

[20] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. Part I," *Annual Reviews in Control*, vol. 43, pp. 65–79, 2017.

[21] D. Acemoğlu, G. Como, F. Fagnani, and A. Ozdaglar, "Opinion fluctuations and disagreement in social networks," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 1–27, 2013.

[22] Y. Xing and K. H. Johansson, "Transient behavior of gossip opinion dynamics with community structure," *arXiv preprint arXiv:2205.14784*, 2022.

[23] Y. Xing and K. H. Johansson, "Concentration in gossip opinion dynamics over random graphs," *arXiv preprint arXiv:2301.05352*, 2023.

[24] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[25] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[26] A. Blum, J. Hopcroft, and R. Kannan, *Foundations of Data Science*. Cambridge University Press, 2020.