Sketch-based community detection in evolving networks

Andre Beckus
Department of Electrical and Computer Engineering,
University of Central Florida, Orlando, FL 32816 USA.

George K. Atia
Department of Electrical and Computer Engineering and Department of Computer Science,
University of Central Florida, Orlando, FL 32816 USA.
(Dated: December 6, 2022)

We consider an approach for community detection in time-varying networks. At its core, this approach maintains a small sketch graph to capture the essential community structure found in each snapshot of the full network. We demonstrate how the sketch can be used to explicitly identify six key community events which typically occur during network evolution: growth, shrinkage, merging, splitting, birth and death. Based on these detection techniques, we formulate a community detection algorithm which can process a network concurrently exhibiting all processes. One advantage afforded by the sketch-based algorithm is the efficient handling of large networks. Whereas detecting events in the full graph may be computationally expensive, the small size of the sketch allows changes to be quickly assessed. A second advantage occurs in networks containing clusters of disproportionate size. The sketch is constructed such that there is equal representation of each cluster, thus reducing the possibility that the small clusters are lost in the estimate. We present a new standardized benchmark based on the stochastic block model which models the addition and deletion of nodes, as well as the birth and death of communities. When coupled with existing benchmarks, this new benchmark provides a comprehensive suite of tests encompassing all six community events. We provide analysis and a set of numerical results demonstrating the advantages of our approach both in run time and in the handling of small clusters.

I. INTRODUCTION

The detection of community structure in networks has garnered a great deal of attention, leading to a vast array of algorithms. Much of the focus has been on static networks, where the goal is to identify groups of nodes within which connections are dense and between which connections are relatively sparse. However, it is often the case that networks evolve with time. For example, edges in social media networks appear and disappear to reflect ever-changing friendships, and gene expression networks continuously evolve in response to external stimuli [1, 2]. In this dynamic setting, new sequential algorithms are needed to track the community structure underlying each temporal snapshot of the network. Here, we propose a sketch-based approach.

Sketching involves the construction of a small synopsis of a full dataset [3]. Notably, this technique has been used in static community detection [4, 5], where a sketch sub-graph is generated by sampling nodes from the full network. The sketch is clustered using an existing community detection algorithm, and the community membership of the nodes in the full network are inferred based on the estimated communities in the sketch. Here, we propose the use of an evolving sketch to detect and handle the six canonical community events observed in dynamic networks [6]: growth, shrinkage, merging, splitting, birth and death. This dynamic approach addresses two pervasive issues in community detection.

One important concern in community detection is the ability to process large graphs. Many static methods become infeasibly slow when processing a large network, thus motivating a search for efficient algorithms [7]. The extra time dimension inherent to the dynamic setting only makes this search for efficiency more pressing. However, time-evolving networks also offer a distinct advantage not found in the static domain. Specifically, evolving networks often possess temporal smoothness in which the community structure changes gradually [8]. In this case, previous snapshots offer prior information which can aid in the clustering of subsequent snapshots. We present a method which relies on a small sketch to convey information regarding previous snapshots. By using a small sketch, the algorithm can detect the main community events without requiring the full graph to be examined, thus reducing the required computational complexity. If the sketch size and number of clusters are fixed, the complexity of our algorithm scales linearly in network size.

Another typical issue found in community detection is the detection of small clusters [9]. If a community shrinks too small, it may become lost, i.e., the community may be absorbed into a larger community in the estimated partition. We show that once a community is captured in the sketch, it can be tracked even if the community becomes very small.

We use dynamic benchmarks as a means for evaluating the proposed algorithm with respect to the canonical network events. The first four events are included in the benchmarks of [10], which are based on the well-known Stochastic Block Model (SBM) [11]. Here, we propose a new dynamic SBM benchmark which captures the last two events of birth and death. An important feature of this proposed benchmark is that the size of the network...
varies with time, a characteristic not found in the existing benchmarks. In addition to modeling the birth event, this benchmark incrementally adds new nodes to the network which join existing communities, a feature also not seen in [10].

This paper is organized as follows. In Sec. II, we summarize existing community detection algorithms for evolving networks. Sec. III describes the network model, and Sec. IV summarizes SBM benchmarks which capture key evolutionary processes. In Sec. V, we describe the sketch-based approach, and formulate techniques by which sketches can detect events and track evolutionary processes. Sec. VI presents the proposed algorithm based on these tracking techniques. We analyze the algorithm in Sec. VII and present numerical results in Sec. VIII. Appendix A describes the static clustering algorithm. Appendix B provides details on the main algorithm itself. Appendix C derives the results found in the analysis of Sec. VII. Appendix D provides additional details regarding the algorithms we compare against in the numerical results.

II. RELATED WORK: COMMUNITY DETECTION IN EVOLVING NETWORKS

A number of algorithms have been proposed for community detection in evolving networks (see [8, 12] for comprehensive surveys). One straightforward approach entails the independent clustering of each snapshot using a static clustering algorithm. The communities in the current snapshot are matched to the previous communities such that there is continuity in the community identities. This category of algorithm contains a number of variants beginning with the classic work of [13].

More recently, many algorithms take a more sophisticated “dependent” approach, in which previous snapshots are accounted for in the clustering of the current snapshot. These algorithms have the potential to outperform independent community detection algorithms, since they incorporate previous knowledge directly in the clustering step.

One approach commonly seen in this category is the representation of each snapshot using a compact graph. In [14], a small weighted graph is constructed after clustering a given snapshot, with each community represented by a single “supernode”. The weights of the edges between supernodes indicate the cumulative number of edges between the corresponding communities. These supernodes are then incorporated into the next snapshot’s graph, thus carrying forward information from the previous estimates. A similar idea can be seen in dynamic methods built around the static Louvain algorithm [15], for example as seen in [16]. The extension of the Louvain algorithm to time-varying networks follows naturally from its reliance on supernodes. Our approach also uses a small representative graph, however using an altogether different idea of sketching, as described in Sec. V.

The model used in this paper is based on the SBM [11]. Several recent algorithms have been developed based on dynamic SBM-based models. The dynamic models of [17, 18] specify that nodes move between a fixed set of communities according to a stationary transition probability matrix. In addition to allowing the movement of nodes between communities, the models of [19, 20] also allow the edge probabilities of the communities to vary. Nonetheless, these works focus on the case where individual nodes only change community membership, i.e., the communities undergo the grow and shrink processes. Although [21] is able to track communities which are also merging and splitting, it still does not allow varying numbers of nodes across the snapshots. The algorithm of [22] allows nodes to join or leave the graph, but requires that all snapshots be known when invoking the algorithm. We emphasize that our proposed algorithm is online in nature, i.e., it performs community detection iteratively on one snapshot at a time, while carrying forward the clustering results from previous snapshots.

III. TEMPORAL NETWORK MODEL

At time $t$, the network snapshot is represented by graph $G(t) = (V(t), E(t))$, where $V(t)$ is the set of nodes in existence at time $t$, and $E(t)$ is the set of edges between these nodes. Let $\hat{C}(t) = \{ \hat{C}_u(t) \mid u \in \{1, \ldots, \hat{q}(t)\} \}$ be the partition at time $t$, with $\hat{C}_u(t)$ denoting the set of nodes in community $u$, and $\hat{q}(t)$ the number of communities.

In each snapshot, an edge exists between nodes within a community with probability $p_{in}$. Nodes in community $u$ are connected to nodes in a different community $u'$ with probability $p_{u,u'}(t)$. An evolutionary process may vary the intercommunity edge density $p_{u,u'}(t)$ so long as the resulting graph adheres to the SBM. We discuss such a process in Sec. IV.B. A pair of communities $u, u'$ are considered to be merged if

$$p_{in} - p_{u,u'}(t) < \sqrt{\frac{2(p_{in} + p_{u,u'}(t))}{|\hat{C}_u(t)| + |\hat{C}_u'(t)|}}.$$  \(1\)

When communities are of equal size, this condition corresponds to the asymptotic weak detectability limit (see [10] for a discussion of this bound in the context of merging and splitting communities). For simplicity, here, we average the community sizes when they are of unequal size.

The following events may occur at time $t$.

- **Node movement between communities** A set of nodes $V_{u \rightarrow u'}(t)$ belonging to community $u$ may move to community $u'$. The edges connected to these nodes are regenerated according to the SBM based on the new community memberships.
• New nodes and community birth A set of nodes $V^+(t)$ may join the graph. A subset $V_{\text{birth}}(t) \subseteq V^+(t)$ of these nodes join new communities. The remaining nodes $V^+(t) \setminus V_{\text{birth}}(t)$ join existing communities. The edges of nodes in $V^+(t)$ are generated according to the SBM.

• Removed nodes and community death A set of nodes $V^-(t)$ may be removed from the graph. Death occurs when all nodes in a particular community are removed.

• Merge and split of communities The merge event occurs for communities $u, u'$ when $p_{u,u'}(t)$ increases such that $[1]$ becomes true. Likewise, the split event occurs when $p_{u,u'}(t)$ decreases such that $[1]$ becomes false.

Note that many of the model variables are functions of time $t$. Where there is no ambiguity, we omit this time parameter to simplify the exposition.

IV. EVOLUTIONARY PROCESSES: BENCHMARKS

For the purpose of illustrating and analyzing the proposed algorithm, we consider here specific examples of evolutionary processes. These are realized by four benchmark networks, i.e., parameterized sequences of snapshots with known community partitions for validating and comparing community detection algorithms. The grow-shrink and merge-split benchmarks are defined in [10], whereas we present the birth-death process here for the first time.

Each benchmark consists of an evolving network containing $2n$ total nodes. The underlying process is driven by a periodic triangular waveform

$$x(t) = \begin{cases} 2t^*, & 0 \leq t^* < 1/2, \\ 2 - 2t^*, & 1/2 \leq t^* < 1, \end{cases}$$

where

$$t^* \equiv (t/\tau + \phi) \mod 1,$$

$\tau$ is the period of the waveform, and $\phi$ controls the phase of the waveform. We will assume that $\phi = 0$ unless otherwise specified.

A. Grow-shrink benchmark

The grow-shrink benchmark moves nodes between a pair of communities denoted $A$ and $B$, thus growing and shrinking the communities. At each time step the first community contains

$$n_A = n - nf(2x(t + \tau/4) - 1)$$

nodes, whereas the second community contains $n_B = 2n - n_A$ nodes. Nodes lost from the first community are transferred to the second community, and vice-versa. The parameter $f \in [0, 1]$ controls the variation in community sizes. For $t \in \{0, \tau/2, \tau\}$ the sizes of the communities are equal. At time $t = \tau/4$, a fraction $f$ of nodes in community $A$ will have moved to community $B$, whereas at time $t = 3\tau/4$ the opposite holds.

B. Merge-split benchmark

The merge-split benchmark has two communities denoted $A$ and $B$, each of size $n$, with intracommunity edge density $p_m$. Initially, the intercommunity edge density is $p_{A,B}(0) = p_{\text{out}}$. New edges are gradually added between the two communities until they are completely merged at time $t = \tau/2$ with $p_{A,B}(\tau/2) = p_m$. Then, the process reverses and the new edges are removed until the communities are completely split again at time $t = \tau$.

The intercommunity edges are placed in the following way. The number of intercommunity edges $m_{\text{um}}$ in the unmerged state are drawn according to a binomial distribution with parameters $n^2$ and $p_{\text{out}}$. The number of edges $m_m$ in the merged state is similarly drawn, except using probability $p_m$. The number of edges at time $t$ is then determined by

$$m^*(t) = \lfloor 1 - x(t) \rfloor m_{\text{um}} + x(t)m_m,$$

where the edges are placed uniformly at random. In this way, the edge density between the two communities is $p_{A,B}(t) = m^*(t)/n^2$. The communities are considered merged at the detectability limit $[1]$.

C. Birth-death benchmark

We now propose a new benchmark which realizes the birth and death of communities, as well as the addition and removal of nodes from the network. A schematic diagram of the birth-death benchmark is shown in Fig. 1(a). The benchmark contains two communities which pass into and out of existence. The size of the first community is

$$n_A = \begin{cases} 0, & x(t + \tau/4) \geq 1 - \gamma/2, \\ n \lfloor 1 - x(t + \tau/4) \rfloor, & \text{otherwise,} \end{cases}$$

where $n_A = 0$ designates a non-existent community, and parameter $\gamma \in [0, 1]$ controls the minimum size of the community. The community starts at time $t = 0$ with $n/2$ nodes. Nodes are removed from the network, until the community shrinks to size $\gamma n/2$ at time $t = \tau(1 - \gamma)/4$. At this point, the community dies and all of its remaining nodes are deleted from the network. At time $t = \tau(1 + \gamma)/4$, a new set of $\gamma n/2$ nodes is added to the network and used to re-create the community. New nodes
are gradually created and added to the community until it reaches size \( n \). At this point, nodes are again removed from the community until it contains \( n/2 \) nodes, and the process repeats.

The second community is of size

\[
n_2 = \begin{cases} 
0, & x(t + \tau/4) < \gamma/2 \\
ax(t + \tau/4), & \text{otherwise}.
\end{cases}
\]

This community undergoes essentially the same process as the first community except with a phase shift of \( \tau/2 \).

### D. Mixed benchmark

To model concurrent processes capturing all of the events, we present a mixed benchmark which is created by "stacking" the grow-shrink, merge-split, and birth-death benchmarks. A schematic of this mixed benchmark is shown in Fig. 1(b). The benchmark has a maximum of \( 6n \) nodes. The first \( 4n \) nodes contain the grow-shrink and merge-split benchmarks as previously described, whereas the last \( 2n \) nodes participate in the birth-death process (the actual number of nodes varies with time due to addition and deletion of nodes in the birth-death benchmark). We show an example of this mixed benchmark in Fig. 2(a).

![Fig. 1](image1.png)

**FIG. 1.** (a) Schematic representation of the birth-death benchmark, showing the two communities labeled \( A \) and \( B \). (b) Schematic representation of the mixed benchmark, which stacks the grow-shrink, merge-split, and birth-death benchmarks.

![Fig. 2](image2.png)

**FIG. 2.** (a) Planted partitions for full graphs of the mixed benchmark network. Each vertical slice indicates the planted partition at time \( t \). The model parameters are \( n = 200, \hat{q} = 6, f = 0.9, \gamma = 0.2 \). (b) Sketches produced with \( n' = 40 \). Each vertical slice indicates the planted partitions in sketch \( S(t) \). White regions indicate that the corresponding node does not exist at time \( t \).

### V. SKETCH-BASED TRACKING OF EVOLUTIONARY PROCESSES

Our algorithm relies on a small representative sketch of the full network. The sketch captures important information which can be used to detect network events and track the processes by which the network evolves. Meanwhile, the smaller size of the sketch allows these checks to be performed quickly without requiring a complete assessment of the entire network. The estimates of the communities in snapshot \( t \) are \( C(t) = \{ C_u(t) \mid u \in \{1, \ldots, q(t)\} \} \), where \( q(t) \) is the estimated number of communities at time \( t \). We first describe the sketch, and then describe how this sketch can be used to detect specific events.

The sketch consists of a set of nodes sampled from the full network. At each time step, this set is updated such that it contains an equal number of nodes \( n' \) from each community. A schematic of the sketch is shown in Fig. 1(b). The benchmark has a maximum of \( 6n \) nodes. The first \( 4n \) nodes contain the grow-shrink and merge-split benchmarks as previously described, whereas the last \( 2n \) nodes participate in the birth-death process (the actual number of nodes varies with time due to addition and deletion of nodes in the birth-death benchmark). We show an example of this mixed benchmark in Fig. 2(a).
build the sketches based on estimates of the true communities. We will present an actual sketch produced by the proposed algorithm in Sec. VIII D.

A. Inferring community membership of nodes

We show in this section how the sketch may be used to infer community membership of any node $i \in V(t)$ in network snapshot $G(t)$. To this end, we calculate

$$s_{i,v}(t) = \left[ \frac{|\{(i,j) \in E(t) | j \in C'_v(t-1)\}|}{|C'_v(t-1)|} \right]$$

(8)

to evaluate the connectivity of node $i$ to each sketch community $v$. Let $u$ be the true community assignment of node $i$. Since

$$\mathbb{E}[s_{i,v}(t)] = \begin{cases} p_{in}, & v = u, \\ p_{u,u}(t), & v \neq u, \end{cases}$$

(9)

it follows that $s_{i,v}(t)$ provides a point estimate of the probability that there is an edge between node $i$ and any node $j \in C'_v(t-1)$. Node $i$ can then be assigned to the community $u'$ with which connectivity is greatest, i.e., where

$$u' = \arg \max_{1 \leq v \leq q(t)} s_{i,v}(t).$$

(10)

The proposed algorithm uses (10) to assign communities to new nodes joining the network, as well as to identify nodes which have changed community membership.

We finish this section by noting that the variance in $s_{i,v}(t)$ is

$$\text{Var}(s_{i,v}(t)) = \begin{cases} \frac{p_{in}(1-p_{in})}{|C'_v(t-1)|}, & v = u, \\ \frac{p_{u,u}(t)(1-p_{u,u}(t))}{|C'_v(t-1)|}, & v \neq u. \end{cases}$$

(11)

The variance grows as the sketch communities shrink, thus motivating the use of equal-sized communities in the sketch.

B. Detecting the split event

Suppose that community $u$ is undergoing a split into two separate communities $u$ and $u'$. To detect the emerging clusters we can use the spectrum of the non-backtracking matrix as described in [23]. Let $G_u$ be the sub-graph of $G(t)$ induced by the latest estimate $C_u(t-1)$, and $A$ be the adjacency matrix of $G_u$. Given diagonal matrix $D$ containing the degrees of nodes in $A$, and identity matrix $I$, define

$$B' = \begin{pmatrix} 0 & D - I \\ -I & A \end{pmatrix}.$$ 

(12)

Suppose the emerging communities are each of size $n$, and define $\lambda_1, \lambda_2$ as the largest and second largest eigenvalues of $B'$, respectively. If

$$n(p_{in} - p_{u,u'}(t))^2 > p_{in} + p_{u,u'}(t),$$

(13)

then in the limit as $n \to \infty$ with $np_{in}$ and $np_{u,u'}(t)$ constant, $\lambda_1 \to n(p_{in} + p_{u,u'}(t))$ and $\lambda_2 \to n(p_{in} - p_{u,u'}(t))$ such that

$$\lambda_2 > \sqrt[5]{\lambda_1}.$$ 

(14)

Although condition (14) is only valid in the limit of infinite sized graphs, it can still serve as a reliable split indicator for a given sequence of network realizations. We show an example of this in Fig. 3. The planted partitions are shown in Fig. 3(a), and the dashed blue line in Fig. 3(b) shows the corresponding gap $\lambda_2 - \sqrt[5]{\lambda_1}$ for each time step. The value of this gap increases as the process moves in either direction away from the fully merged state at $t=50$, and towards the fully split states at $t \in \{0, \tau\}$. Decision threshold (14) is shown as a horizontal dashed line. As can be seen, the split is detected fairly close to the full graph detectability limit.

Rather than calculating the eigenvalues for the full network (at great computational cost), we propose to instead detect the split using the sketch. We apply the same procedure as described above, but instead substitute $C_u(t-1)$ with $C'_u(t-1)$. The estimate based on the sketch is shown in Fig. 3(b) as a solid orange line. Note that the time of detection in the sketch diverges from that in the full graph as the community sizes in the sketch decrease. We analyze this dependence on sketch size in Sec. VII.

C. Detecting the merge event

Suppose that communities $u, u'$ are merging. To detect the merge event, we exploit the fact that the two communities are already known at time $t-1$. This means that we can estimate $p_{in}$ and $p_{u,u'}(t-1)$ and use these estimates to directly check condition (1) to detect a merge event. The sketch allows us to quickly calculate point estimates of the edge probabilities using the expressions

$$\hat{p}_{in} = \frac{2 \sum_{v=1}^{q(t)} \left| \{(i,j) \in E(t) \mid i,j \in C'_v(t-1)\} \right|}{\sum_{v=1}^{q(t)} |C'_v(t-1)| \cdot |C'_v(t-1)| - 1},$$

(15)

$$\hat{p}_{u,u'}(t) = \left| \frac{\{(i,j) \in E(t) \mid i \in C'_u(t-1), j \in C'_v(t-1)\}}{|C'_u(t-1)| |C'_v(t-1)|} \right|. $$

(16)

Figure 3(c) shows the actual (dashed blue line) and estimated (solid orange line) values of $p_{in}$ for the example in Fig. 3(a). The actual (dashed purple line) and estimated (solid green line) values of $p_{u,u'}(t)$ are also shown in the sample plot. In both cases, the estimates track the actual values well.
FIG. 3. (a) Planted partitions of the merge-split benchmark. For reference, the detectability limits that formally define the splits in the benchmark are shown as vertical dashed and dotted lines. Network parameters are $q = 2$, $n = 200$, $p_{in} = 0.5$, $p_{out} = 0.05$. (b) Actual and estimated values of $\lambda_2 - \sqrt{\lambda_1}$ at each time step. When the gap is greater than the decision threshold (horizontal dashed red line), the community is considered split. The sketch is constructed using $n = 50$ nodes sampled uniformly at random from each of the two communities at each time step. (c) Actual and estimated values of $\hat{p}_{in}$ and $p_{in, u'}$ at each time step.

D. Detecting the birth event

Consider a node $i \in V^+(t)$, which is joining the network. If the node joins an existing community $u$, then $\mathbb{E}[s_{i,u}(t)] = p_{in}$, and we can detect this occurrence by checking if $s_{i,u}(t) \geq \hat{p}_{in} - 3\bar{\sigma}$, where $\hat{p}_{in}$ is the estimate from (15), and $\bar{\sigma} = \sqrt{\hat{p}_{in}(1 - \hat{p}_{in})/n'}$ is an estimate of the standard deviation of $s_{i,u}(t)$. On the other hand, if $i \in V_{\text{birth}}(t)$, then the expectation $\mathbb{E}[s_{i,v}(t)]$ will equal the intercommunity edge density between the new community and any existing community $v$. This suggests that we can identify the set of nodes that are joining newborn communities using the expression

$$V_{\text{birth}}(t) = \{i \in V^+(t) \mid s_{i,v}(t) < \hat{p}_{in} - 3\bar{\sigma}, \forall v \in \{1, \ldots, q(t)\}\}.$$  

VI. PROPOSED ALGORITHM

We first discuss preliminaries. The proposed algorithm invokes a function $\text{Static-Cluster}(G)$, which performs clustering of a static graph $G$ to produce community estimates $C = \{C_1, \ldots, C_q\}$. We implement this function using spectral techniques based on the non-backtracking matrix [23], along with enhancements to provide more robust estimation of the number of communities (details of the function are given in Appendix A). The computational cost of this function is dominated by the eigendecomposition, which is cubic in the size of graph $G$. We now summarize the main steps of the proposed algorithm. We provide an assessment of the computational complexity for each step, and comment on the overall complexity at the end. A detailed algorithm listing is provided in Appendix B.

Main-Algorithm

Input: Initial sketch size $N'$. Sketch community size $n'$.
Graph snapshots $G(t), t = 0, 1, \ldots$

1. **Cluster initial snapshot.** Build sketch $G'$ by sampling $N'$ nodes from $G(0)$ uniformly at random. Invoke $\text{Static-Cluster}(G')$ to obtain community estimates $C'$ for the sketch. Use (10) to infer the community memberships $C(0)$ of all nodes in $G(0)$ based on community estimates $C'$.

**Complexity:** By executing $\text{Static-Cluster}$ solely on the sketch, we reduce the running time of this expensive step to only $O(N'^2 + N' |\mathcal{V}(0)|)$. The first term corresponds to clustering of the sketch, and the second term corresponds to inference on the full graph.

2. **For each snapshot $G(t), t = 1, 2, \ldots$ do**

3. **Update sketch.** Update the sketch to include $n'$ nodes sampled uniformly at random from each community.

4. **Birth detection.** Identify newborn communities by calculating $V_{\text{birth}}$ as in (17). Since there may be more than one community born at the same time, we cluster the graph induced by $V_{\text{birth}}$ using $\text{Static-Cluster}$. To keep running time low, we use the same sketch-based approach as in Step 1.

**Complexity:** The clustering of the nodes in $V_{\text{birth}}$ incurs the dominant cost. We use a sketch consisting of $n'$ nodes from $V_{\text{birth}}$, and so the clustering will take time $O(q(t)3n'^3 + q(t)n' |\mathcal{V}(t)|)$

5. **Infer community membership of new and moved nodes.** Use the estimator (10) to infer community membership of each node $i \in V(t) \setminus V_{\text{birth}}$. Note that this set includes existing nodes, which may have changed community
m \text{embership, as well as new nodes } V^+(t) \text{ which are joining existing communities. }

\text{Complexity:} \text{ Calculation of the similarity metric } s_{i,u}(t) \text{ for a single community } u \text{ and single node } i \text{ takes time } O(n^2). \text{ Therefore, this step is } O(q(t)n^2|V(t)|) \text{ in total.}

(6) \text{Split detection.} \text{ For each community } u, \text{ build graph } G' \text{ induced by } C_u(t). \text{ From this induced graph, build } B' \text{ as defined in (12). Calculate the eigenvalues } \lambda_1, \lambda_2 \text{ of } B'. \text{ If } \lambda_2 > \sqrt{\lambda_1}, \text{ then a split event is declared. In this case, invoke Static-Cluster}(G') \text{ to identify the emerging communities in the sketch, and then use (10) to identify nodes in the full graph belonging to these emerging communities.}

\text{Complexity:} \text{ In the worst case, for each sketch community we must perform an eigendecomposition, estimate the partitions, and infer community membership in the full graph. Thus, this step is } O(q(t)n'(n^2 + |V(t)|)) \text{ in total.}

(7) \text{Merge detection.} \text{ For each pair of communities } u, u', \text{ consider the communities merged if}

\begin{align}
\tilde{p}_{in} - \tilde{p}_{u,u'} &< d \sqrt{\frac{2(\tilde{p}_{in} + \tilde{p}_{u,u'})}{|C_u| + |C_u'|}}, \tag{18}
\end{align}

where we use estimates of the intracomunity edge density \( \tilde{p}_{in} \), and the intercommunity edge density \( \tilde{p}_{u,u'} \). Condition (18) is similar to (1), except with an additional scaling parameter \( d \) in the right hand side. If \( d = 1 \) then a shrinking density gap \( \tilde{p}_{in} - \tilde{p}_{u,u'} \) causes erratic behavior during node inference, resulting in nodes incorrectly being moved between the pair of merging communities. This in turn corrupts the estimates \( \tilde{p}_{in}, \tilde{p}_{u,u'} \). We set \( d = 2 \) to trigger the merge earlier and avoid this issue.

\text{Complexity:} \text{ Constructing the estimates takes } O(n^2) \text{ time, whereas checking the merge condition for all pairs takes } O(q(t)^2) \text{ time.}

(8) \text{Build estimate } C(t) \text{ using results of Steps (4)-(7)}

\text{Output:} \text{Partitions } C(t), t = 0, 1, \ldots

Suppose that \( \overline{q} \) and \( \overline{N} \) are the maximum number of communities and nodes, respectively, in any given snapshot. We furthermore assume that \( N' \) is at most \( \overline{p}n' \). Then, the computational complexity for estimating a single partition \( C(t) \) at time \( t \geq 0 \) is \( O(\overline{p}n'(\overline{q}^2n^2 + \overline{N})) \). For the first iteration, this is the time required for executing step (4), whereas for each subsequent iteration, this is the total time required to execute steps (4)-(7). Contrast this with clustering the full snapshot graph, which is \( O\left(\overline{N}^3\right) \) for each iteration. If \( \overline{q} \ll \overline{N} \) and we use a small sketch, this results in an order-wise improvement in complexity.

\section{Analysis}

In this section, we provide performance guarantees for the proposed algorithm, as well as guidelines for setting sketch size. To simplify analysis we take the sketch to be balanced at all time steps, i.e., \( |C_u(t)| = n' \) for each community \( u \) and time \( t \). Furthermore, we suppose that the graph at the previous time step has been correctly clustered, i.e., \( \mathcal{C}(t-1) = \hat{\mathcal{C}}(t-1) \). Unless otherwise specified, it is assumed that \( p_{u,u'}(t) = p_{out} \) for any two communities \( u \) and \( u' \). The average degree of such a snapshot with \( \overline{q} \) communities is \( c = n'(p_{in} + p_{out}) \). The following approximation is made to provide clearer results.

\textbf{Assumption 1.} Each of \( s_{i,u}(t), \tilde{p}_{in}, \) and \( \tilde{p}_{u,u'} \) is well approximated by a normal random variable having the same mean and variance.

This assumption follows from the fact that the listed variables are driven by binomial random variables. The underlying distributions of these random variables will generally have enough symmetry to be well approximated by normal distributions [21]. More details are provided in Appendix C.

We now provide definitions used in this section. Denote by \( \Phi^{-1}(\cdot) \) the inverse cumulative distribution function of the standard normal distribution. Specifically, given a standard normal random variable \( Z \) and probability \( \alpha \), we have \( \Pr \left(Z \leq \Phi^{-1}(\alpha)\right) = \alpha \). Consider a graph with \( N \) nodes and \( \hat{q} \) equal-sized communities \( \hat{C} = \{ C_v \mid v \in \{1, \ldots , \hat{q}\} \} \). The agreement with an estimated community partition \( C = \{ C_v \mid v \in \{1, \ldots , \hat{q}\} \} \) is defined as [25]

\begin{align}
A(\hat{C}, C) &= \frac{1}{N} \max_{\pi} \sum_{v=1}^{\hat{q}} \left| C_v \cap C_{\pi(v)} \right|, \tag{19}
\end{align}

where \( \pi \) ranges over the permutations on \( \hat{q} \) elements (this permutation is necessary since the community indices may be ordered arbitrarily). Exact recovery is solved by an algorithm if it produces community estimates such that \( \Pr \left( A(\hat{C}, C) = 1 \right) \rightarrow 1 \) as \( N \rightarrow \infty \). In this section, we use 20 trials for each experiment. Detailed derivations for the results in this section are deferred to Appendix C.

\subsection{Estimating communities in the initial sketch}

This section provides guidelines for choosing initial sketch size. For simplicity, we consider the symmetric case in which every community has \( n \) nodes. Suppose that an initial sketch has been constructed by sampling \( N' \) nodes from \( G(0) \). If \( a = \frac{p_{in} N'}{\ln(N')} \) and \( b = \frac{p_{out} N'}{\ln(N')} \) are held constant, then exact recovery of the planted partition is efficiently solvable in the initial sketch provided \( \sqrt{\overline{a}} - \sqrt{\overline{b}} \sqrt{\overline{q}} > \sqrt{\overline{q}}(0) \) [23]. Although this bound is only exact in the limit, it can still be used to estimate values of \( N' \).
for which agreement will remain high. Specifically, for fixed \(N', p_{in}, p_{out}\), this bound becomes
\[
\frac{N'}{\ln(N')} > \hat{q}(0)\left(\sqrt{p_{in}} - \sqrt{p_{out}}\right)^{-2}. \tag{20}
\]
Either a small density gap \(p_{in} - p_{out}\) or a large number of communities \(\hat{q}(0)\) can make the initial estimate unreliable. These issues can be mitigated by increasing the sketch size.

We demonstrate the efficacy of (20) in Fig. 4. We produce a sketch from a graph with two communities of size \(n = 2500\), and plot agreement between the estimated and planted communities. The blue line indicates the boundary of (20). Indeed, the agreement remains high (exceeding 0.998%) whenever this condition holds.

We note that if the initial snapshot is imbalanced, i.e., with communities of different size, the SPIN (Sampling Inversely proportional to Node degree) sampling method may be used in place of uniform random sampling. This method can improve the success rate by sampling more uniformly across communities.

### B. Birth detection

Suppose that one or more new communities are born at time \(t\), and take a node \(i \in V_{birth}\) which belongs to one of these communities. Let \(\sigma_{\hat{p}_{in}} = \sqrt{\frac{2p_{in}(1-p_{out})}{\hat{q}(t)n'/n}}\) be the standard deviation of estimator \(\hat{p}_{in}\). Then, the probability that node \(i\) is correctly identified as belonging to \(V_{birth}\) is at least \(\alpha\) if
\[
n' \geq \left(\frac{\Phi^{-1}(1-\beta)\sqrt{p_{out}(1-p_{out})} + 3\sqrt{p_{in}(1-p_{in})}}{p_{in} - p_{out}}\right)^2. \tag{21}
\]

where \(\beta = (1-\alpha) / (2(\hat{q}(t) - 1))\) and
\[
p_{in} = p_{in} \pm \Phi^{-1}\left(1 - \frac{\beta}{2}\right)\sigma_{\hat{p}_{in}}. \tag{22a}
\]
Note that the sufficient number of samples in (21) is independent of community size in the full graph. This allows for the detection of new communities even when they are of a very small size. This advantage is illustrated further in the numerical results of Sec. VIII B.

Fig. 5 shows results in which a new community with 500 nodes joins a graph containing two existing communities of size \(n = 2500\). The plots indicate the fraction of nodes in \(V_{birth}\) which are correctly identified as belonging to the newborn community. The red line shows the boundary of condition (21) with \(\alpha = 0.99\), and shows excellent agreement with the numerical results.

As the density gap \(p_{in} - p_{out}\) shrinks, a larger sketch will be required to reliably detect which nodes belong to newborn communities. In fact, as \(n' \to \infty\), we have \(\sigma_{\hat{p}_{in}} \to 0\) such that \(p_{in} \to \hat{p}_{in}\), and the right side of (21) converges to
\[
\frac{\left(\Phi^{-1}(1-\beta)\sqrt{p_{out}(1-p_{out})} + 3\sqrt{p_{in}(1-p_{in})}\right)^2}{p_{in} - p_{out}}. \tag{23}
\]
In this regime, the denominator depends solely on the square of the density gap.

### C. Inferring community membership

We next consider the required sketch size to successfully infer community membership of individual nodes using (10). Define \(m_{u \to u'} = |V_{\hat{u} \to \hat{u'}}(t) \cap S(t-1)|\), i.e., the number of nodes in the sketch that are moving from \(u\) to \(u'\). The analysis here will use the following simplification.
Assumption 2. In place of random variable \( m'_{u\rightarrow u'} \), we use its expected value \( \mathbb{E}[m'_{u\rightarrow u'}] = n' |V_{u\rightarrow u'}| / |C_u(t-1)| \).

We denote the minimum community size in a given snapshot by \( n_{\text{min}} = \min_{1 \leq v \leq \hat{q}(t)} |C_v(t)| \).

Suppose that at most \( m' \) nodes move between any two pairs of communities, i.e., \( |V_{u\rightarrow w}(t)| \leq m' \) for \( 1 \leq u, w \leq \hat{q}(t) \). Then, \( \alpha(t) \) correctly identifies the community of a given node \( i \not\in \hat{S}(t) \) with probability \( \geq \alpha \) provided that

\[
n' \geq \bar{n}_1 \left[ \Phi^{-1} \left( 1 - \frac{1 - \alpha}{\bar{m}} \right) \right]^2 \tag{24}
\]

where

\[
\bar{m} = \left( p_{\text{in}} - p_{\text{out}} \right) \left( 1 - \frac{m' \hat{q}(t)}{n_{\text{min}}} \right), \tag{25}
\]

\[
\bar{n}_1 = p_{\text{in}}(1 - p_{\text{in}}) + p_{\text{out}}(1 - p_{\text{out}}) \tag{26}
\]

\[
+ \bar{m} n_{\text{min}} (1 - p_{\text{in}}) - p_{\text{out}}(1 - p_{\text{out}}), \tag{27}
\]

Variable \( \bar{m} \) serves as a lower bound on the expected value of \( s_{u,v}(t) - s_{u,v}(t) \) for \( v \neq u \), whereas \( \bar{n}_1 \) serves as an upper bound on the standard deviation. Both a small density gap \( \bar{m} \) and a large number of moving nodes \( \bar{m} \) can make inference less reliable. In these situations, an increased sketch size will be required to keep the probability of misclassification low.

Fig. 6 shows the inference success rate of \( \alpha(t) \) for a network containing two communities of size \( n = 2500 \). In Fig. 6(a), \( \bar{m} = 0 \) (no nodes move between communities), whereas Fig. 6(b) varies \( \bar{m} \). The red lines indicate the boundary of \( \bar{n}_1 \) with \( \alpha = 0.99 \), and show excellent agreement with the numerical results. As \( \bar{m} \rightarrow 0 \), this boundary converges to

\[
\left[ \Phi^{-1} \left( 1 - \frac{1 - \alpha}{\bar{m}} \right) \right]^2 \left( \frac{p_{\text{in}}(1 - p_{\text{in}}) + p_{\text{out}}(1 - p_{\text{out}})}{(p_{\text{in}} - p_{\text{out}})^2} \right), \tag{28}
\]

which is independent of community size in the full graph. This advantage will be illustrated further in the numerical results of Sec. VIII B. In this case, the primary driver of performance becomes the density gap \( p_{\text{in}} - p_{\text{out}} \).

D. Split detection

Consider a network with a single community undergoing a split into two equal-sized communities \( u \) and \( u' \). An important consideration is the smallest value of \( p_{u,u'} \) at which the communities will be considered split according to (14). Using a similar argument as for the initial sketch, in practice we may use the exact recovery limit to approximate this lower bound. Likewise, we can use the asymptotic detectability threshold as an approximate upper bound. Following this line of reasoning, it is likely that the split will be detected for some \( p_{u,u'} \) bounded according to

\[
p_{\text{in}} + \frac{1}{2n'} \sqrt{\frac{2p_{\text{in}}}{n'}} + \frac{1}{4n'^2} > p_{u,u'} \tag{29}
\]

Increased sketch size will tend to allow earlier detection of the split, i.e., for smaller values of \( p_{\text{in}} - p_{u,u'} \).

Fig. 7(a) shows a plot of the estimated number of communities from (14) for a sketch with two communities containing \( n' \) nodes each (the detected number of communities is 2 if the condition holds, and 1 otherwise). Along the horizontal axis, we vary \( p_{u,u'} \) within \([0, p_{\text{in}}]\), where \( p_{\text{in}} = 0.5 \). The blue and green lines show the lower and upper bounds in (29), respectively. The split is indeed detected for a value of \( p_{u,u'} \) within these bounds.
E. Merge detection

Finally, suppose that two equal-sized communities \( u, u' \) are merging into one community. We consider the value of \( p_{u,u'} \) at which condition (18) detects a merge. This condition will hold with probability \( \geq \alpha \) if

\[
p_{u,u'} \geq p_{\text{in}} + \Phi^{-1}
\left(1 - \frac{1-\alpha}{2}\right) \sigma_m + \frac{d^2}{2n} - \frac{d^2}{n^2},
\]

(30)

where we use the the standard deviation of \( \hat{p}_{\text{in}} \pm \hat{p}_{u,u'} \),

\[
\sigma_m = \sqrt{\frac{2p_{\text{in}}(1-p_{\text{in}})}{\hat{q}(t)n'(n'-1)} + \frac{p_{u,u'}(1-p_{u,u'})}{(n')^2}}.
\]

(31)

However, it is also important to consider when reliably identifies the communities as being split. This occurs with probability \( \geq \alpha \) if

\[
p_{u,u'} \leq p_{\text{in}} - \Phi^{-1}
\left(1 - \frac{1-\alpha}{2}\right) \sigma_m + \frac{d^2}{2n} - \frac{d^2}{n^2},
\]

(32)

To illustrate the significance of bounds (30) and (32), Fig. 7(b) shows the detected number of communities for a pair of communities with \( n = 2500 \) nodes each (the detected number of communities is 1 if (18) holds, and 2 otherwise). The red line indicates the boundary of (30), and the yellow line indicates the boundary of (32), for \( \alpha = 0.9 \). When \( p_{u,u'} \) falls in the gap between these bounds, the detection tends to be unreliable. However, the size of this gap can be reduced by using larger sketch sizes to drive down the standard deviation \( \sigma_m \).

VIII. NUMERICAL RESULTS

We compare against four algorithms from the literature, each of which uses a different means for carrying forward information from one snapshot to the next. First, we use the classic Bayesian approach found in Yang et al. [17]. Second, we run the algorithm of Dinh et al. [14]. This algorithm uses a sketch-like concept by consolidating known communities into “supernodes” within a weighted graph. These supernodes are then incorporated into the next snapshot. Third, we use ESPRA (Evolutionary clustering based on Structural Perturbation and Resource Allocation similarity), which is based on structural perturbation theory [26]. This algorithm defines an objective function which explicitly balances two similarities: one which encourages temporal smoothness across snapshots, and one that takes into account only the community structure in the latest snapshot. Lastly, we independently cluster each snapshot as described in such works as [13] [27]. This algorithm, referred to here as (Independent), estimates the communities in the current snapshot using Static-Cluster, and then matches the estimates in adjacent snapshots using the Jaccard similarity coefficient [28]. Although Static-Cluster performs optimally in certain regimes, the main weakness of (Independent) is that it completely ignores information from the previous snapshot when clustering the current snapshot.

Further details regarding these algorithms are provided in Appendix D. Unless otherwise specified, all plots show an average over 20 independent runs. We set the initial sketch size to \( N' = \hat{q}(0)n' \).

A. Performance with small clusters

We first consider the performance of the proposed algorithm in the presence of small communities. We use normalized agreement to compare the planted communities \( \hat{C} = \{ \hat{C}_v \mid v \in \{1, \ldots, q\} \} \) and estimated communities \( C = \{ C_v \mid v \in \{1, \ldots, q\} \} \). Sets \( \hat{C} \) and \( C \) are padded with empty communities such that |\( \hat{C} | = | C | \). Then, normalized agreement is defined as [25]

\[
\hat{A} = \frac{1}{\hat{q}} \max \pi \sum_{v=1}^{\hat{q}} \frac{\left| \hat{C}_v \cap C_{\pi(v)} \right|}{| \hat{C}_v |},
\]

(33)

where \( \pi \) ranges over the permutations on \( \hat{q} \) elements. The normalized agreement for the snapshot at time \( t \) is denoted \( \hat{A}(t) \). Unlike the agreement metric defined earlier in [19], normalized agreement proves useful for quantifying performance in the presence of small clusters, since each community constitutes a fraction \( \frac{1}{\hat{q}(t)} \) of the normalized agreement, regardless of community size.

For summarizing the overall deviation in the actual and estimate communities for a snapshot sequence, we use the average-squared error

\[
E_{\hat{A}} = \frac{1}{T} \sum_{t=1}^{T} \left[ 1 - \hat{A}(t) \right]^2,
\]

(34)

where \( T \) is the total number of snapshots.

1. Grow-shrink benchmark

We use two concurrent instances of the grow-shrink benchmark with phase \( \phi = 0 \) for the first instance, and \( \phi = \pi/2 \) for the second instance. Figure 8 shows planted partitions for an example with \( f = 0.95 \). The community detection results are shown for all algorithms in Fig. 9(a), where the value of \( E_{\hat{A}} \) is plotted as a function of \( f \). The proposed algorithm has \( E_{\hat{A}} < 0.02 \) for all values of \( f \), whereas the other algorithms exhibit significantly larger values of \( E_{\hat{A}} \) especially for larger \( f \).

To gain further insight into the behavior of the algorithms, we plot a heat map of \( \hat{A}(t) \) for each algorithm in
Fig. 8. Planted partitions for a double-stacked version of the grow-shrink benchmark, with \( n = 250, f = 0.95, q = 4, p_n = 0.4, p_{out} = 0.1 \).

Table I shows the value of \( E_A \) as a function of \( n \), averaged over five trials. There is a small increase in \( E_A \) as \( n \) increases, due to a corresponding increase in the number of moving nodes (as described in Sec. VII C). Nonetheless, \( E_A \) remains below \( 5.5 \times 10^{-5} \) despite a dramatic increase in imbalance of the full graph, and despite the fact that the sketch size remains fixed.

Likewise, we run the birth-death benchmark with the parameters of Sec. VIII A 2, but with \( \gamma = 2 \tau_{\text{min}}/n \) such that \( \tau_{\text{min}} = 20 \) regardless of graph size. The smallest community size is attained immediately before death and after birth. The results are shown in Table I. Unlike the results for the grow-shrink benchmark, there is no increase in \( E_A \). This is consistent with the analysis in Sec. VII B, which showed no dependence on community size in the full graph.

For both benchmarks, Table I shows only a sub-linear increase in runtime as community size \( n \) increases, owing to the fixed sketch size. To expand on this result, we run all of the algorithms on the grow-shrink benchmark, and show the results as a function of \( n \) in Fig. 11. As expected, the proposed algorithm finishes very fast, in under two seconds for all cases. On the other hand, algorithms \[17\], ESPRA, and (Independent) all cluster the full graph, and therefore scale super-linearly with network size. Although \[14\] clusters a graph of reduced size at each step, nodes having changed edges are left as singleton nodes. In this example, the large number of edge changes forces a correspondingly large number of nodes to remain singletons, thus requiring the static clustering step to operate on large networks.

### B. Scalability

To demonstrate the scalability of the proposed algorithm, let us consider the minimum community size over all snapshots

\[
\tau_{\text{min}} = \min_{1 \leq t \leq \tau} \min_{1 \leq v \leq \delta(t)} \left| C_v(t) \right| .
\]

We run the grow-shrink benchmark using the same parameters as in Sec. VIII A 1, except with \( f = 1 - \tau_{\text{min}}/n \) such that the minimum cluster sizes are fixed at \( \tau_{\text{min}} = 200 \).

### C. Merge-split detection

We next execute the algorithms on the merge-split benchmark, with two concurrent instances as shown in Fig. 12(a). The plot in Fig. 12(b) shows the estimated number of communities as a function of time for the proposed algorithm, as well as for \[14\] and ESPRA. We omit (Independent) and ESPRA as they cannot handle the merge and split processes. Although the benchmark groundtruth undergoes an instantaneous transition between merged and split states, the network itself gradually interpolates between these states. This discrepancy
FIG. 9. Results for varying $f$ in the grow-shrink example in Fig. 8. For the proposed algorithm we set $n' = 50$. Plot of $E_{\tilde{A}}$ is shown in (a). Panels (b) through (f) show ensemble averages of $\tilde{A}(t)$ as a function of time along the horizontal axis, and $f$ along the vertical axis for each algorithm.

in timescales, along with the fact that the benchmark sets the transition at the theoretical detectability limit, means we cannot expect the estimated partitions to exactly match the planted partitions. Indeed, all three algorithms overestimate the span of time during which the communities are merged.

The estimates of the proposed algorithm are shown in Fig. 12(c), where we can see that nodes start being misclassified at $t = 13$. This is expected due to the shrinking gap between $p_0$ and the intercommunity edge densities, as described in Sec. VI. Nonetheless, the proposed algorithm detects the merge much closer to the benchmark’s merge time than the other two algorithms. We note that using larger values of $d$ in condition (18) will result in an earlier detection of the merge. In this way, $d$ can act as a tuning parameter to adjust the sensitivity of the merge detection.

For studying the performance of the algorithms’ split detection, we show the exact recovery limit for the sketch as a vertical white dashed line in Fig. 12(a). The proposed algorithm detects the split close to this limit, although we point out that the detection could be shifted earlier by increasing the sketch size. Despite clustering the full network, for which estimation should be easier, ESPRA does not exceed the performance of the proposed algorithm, and [14] fares even worst.

D. Mixed benchmark

So far, our results have considered individual benchmarks in isolation. We now run the proposed algorithm on the mixed benchmark from Fig. 2(a), which has concurrent birth-death, grow-shrink and merge-split processes. The partition estimates are shown in Fig. 13(a). Most of the mismatch occurs in the merge-split communities, which is consistent with our earlier results.

The sketches produced by the proposed algorithm are shown in Fig. 13(b). The sketch nodes are sorted vertically according to their planted communities, with the color indicating the estimated community of the corresponding node. The deviation from the ideal sketch in Fig. 2(b) lies mostly inside the merge-split communities, due to the errors present in the estimates of the full graph.

The estimated partitions for [14] are presented in Fig. 13(c). As with the earlier results, [14] encounters difficulties in correctly identifying the small clusters in the grow-shrink and birth-death communities.

IX. CONCLUSION

This paper concerned a sketch-based approach for community detection in time-evolving networks. We presented an SBM-based model along with possible evolutionary processes which may occur within this model. We then proposed sketch-based techniques for tracking these processes, as well as an algorithm incorporating these
techniques to produce community estimates for concurrent processes. We provided an analysis to guide the choice of sketch size, and generated numerical results comparing the proposed algorithm to full-scale community detection algorithms.

We conclude by briefly noting possible extensions. First, an arbitrary community detection algorithm may be used in place of Static-Cluster, provided that it can estimate the number of communities. Second, a straightforward extension to the network model and algorithm would allow the intracommunity edge density \( p_{in} \) to vary for each community. Third, our approach is extendable to other graph models as well, for example the Degree Corrected SBM (DCSBM) [29]. This can be accomplished by substituting a suitable sampling technique for constructing DCSBM sketches (e.g., the sampling method of [30]), a similarity definition between nodes in the full network and the sketch communities, and an appropriate technique for determining when clusters split or merge.

acknowledgments

This work was supported by NSF CAREER Award CCF-1552497 and NSF Award CCF-2106339. The University of Central Florida Advanced Research Computing Center provided computational resources that contributed to results reported herein.

appendix a: static clustering

We use the following algorithm to perform static clustering of graph \( G \) with \( N \) nodes.

\textbf{Static-Cluster}(G)

1. Construct \( B' \) from \( G \) using (12).
2. Calculate eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_N \) and corresponding eigenvectors of \( B' \).
3. Calculate \( q \) as the maximum value of \( i \) such that \( \lambda_i > \sqrt{\lambda_1} \).
4. for \( i = 1 \ldots q \) do
5. Build matrix \( M \) from the \( i \) normalized eigenvectors of \( B' \) corresponding to eigenvalues \( \lambda_1, \ldots, \lambda_i \). Apply k-means clustering to \( M \) to obtain community estimates \( C_i = \{ C_v \mid v \in \{1, \ldots, i\} \} \). We repeat 100 iterations with three random initializations and take the best result.
6. Calculate modularity \( Q_i \) of \( G \) with partition \( C_i \). Modularity is defined as in [31, Section IV].
7. end for
FIG. 12. (a) Planted partitions for a double-stacked version of the merge-split benchmark. The parameters of the model are $\tilde{q} = 4$, $n = 250$, $p_{in} = 0.5$, $p_{out} = 0.05$. We set $n' = 50$ for the proposed algorithm. This results in a sketch size of 100 in the merged state, and 200 in the split state. The exact recovery limit for the sketch, based on (20), is shown as a dashed white vertical line. (b) The estimated number of communities at each time step for the proposed algorithms, as well for [14] and ESPRA. (c) The estimated partitions for the proposed algorithm.

Steps 1-3 estimate the number of communities $q$, and are as described in [23]. We find that adding Steps 4-8 provides a more reliable estimate of the number of communities. These steps repeat k-means clustering, varying the number of clusters up to $q$, and then return the partition giving the highest modularity.

The additional steps do not increase the complexity of the algorithm $O(N^3)$.

Appendix B: Main algorithm: details

Two helper functions are needed. The first, \texttt{Sample}(G, N'), returns a set of $N'$ nodes sampled uniformly at random from $G$. The second infers community membership of nodes in the full graph $G$ based on the sketch community estimates $C'$, as described in Sec. V A. This function is defined as follows.

\textbf{Infer}(G, C')

(1) $C_v \leftarrow \emptyset$ for $v \in \{1, \ldots, q\}$
(2) for each node $i \in V$ do
(3) $w \leftarrow \arg\max_{v \in \{1, \ldots, q\}} \left| \left\{ (i,j) \in E | j \in C'_v \right\} \right| / |C'_v|$
(4) $C_w \leftarrow C_w \cup \{i\}$
(5) end for
(6) return partition $\mathcal{C} = \{ C_v \mid v = 1, \ldots, q \}$

We use the following definition: given a graph $G = (V, E)$ and node set $V' \subset V$, the subgraph of $G$
FIG. 14. Results showing improved performance of the proposed static clustering algorithm afforded by the modularity-based heuristic in Steps [13]. We generate a graph G with 800 nodes divided into q equal-sized communities, and p_{in} = 5p_{out}.

induced by V’ is denoted G[V’]. The complete definition of Main-Algorithm follows.

(1) G ← G(0)
(2) S’ ← Sample(G, N’)
(3) C(0) ← Infer(G, Static-Cluster(G[S’]))
(4) Build sketch set S(0) by sampling n’ nodes uniformly at random from each community C ∈ C(0).
   If n’ > |C(0)|, then include all nodes from C.
(5) r ← |C(0)|.
(6) for t = 1, 2, . . . do
(7) G ← G(t)
(8) {C_1, . . . , C_r}
   ← Infer(G[V(t) \ hat{\cal V}_{birth}], C(t−1))
(9) if hat{\cal V}_{birth} ≠ \emptyset then
(10) \cal G ← G[hat{\cal V}_{birth}]
(11) \cal S ← Sample(\cal G, n’)
(12) {C_{r+1}, . . . , C_{r+q}}
   ← Infer(\cal G, Static-Cluster(\cal G[\cal S]))
(13) r ← r + q
(14) end if
(15) for u ∈ {1, . . . , r}, where |C_u| > a do
(16) C’ ← C_u ∩ S(t−1)
(17) G’ ← G[C’]
(18) Let A be the adjacency matrix of G’. Calculate the eigenvalues \lambda_1, \lambda_2, . . . of B’ [defined in [12]].
(19) Calculate q as the maximum value of i such that \lambda_i > \sqrt{\lambda_1}.
(20) if q > 1 then
(21) G_u ← G[C_u]
(22) C'' ← Static-Cluster(G_u[C’])
(23) {C_u, C_{r+1}, . . . , C_{r+q−1}}
   ← Infer(G_u, C’’)
(24) r ← r + q − 1
(25) end if
(26) end for
(27) for community pairs u, u’ ∈ {1, . . . , r} do
(28) if \{18\} holds then
(29) C_u ← C_u ∪ C_{u’},
(30) C_{u’} ← \emptyset
(31) end if
(32) end for
(33) C(t) ← \{ C_v | v = 1, . . . , r \}
(34) V^- ← V(t−1) \setminus V(t)
(35) S(t) ← S(t−1) \setminus V^-
(36) Re-proportion sketch S(t) such that it contains min\{|n’, |C’|\} nodes from each community C ∈ C(t).
(37) end for

Steps [13] cluster the first graph snapshot. This estimate is used to construct a balanced sketch in step [4]. The remainder of the algorithm processes subsequent snapshots. Step [8] re-evaluates the community membership of existing nodes, as well as new nodes joining existing communities. Steps [11, 14] partition the set of newborn communities. Meanwhile, Steps [15, 26] handle splits within each community. Only communities with size greater than a are checked, as the spectral estimates become unreliable for small communities. We set a = 20. Steps [27, 32] handle merges among pairs of communities. Finally, steps [33, 36] generate the new sketch.

Appendix C: Derivations for analysis in Sec. VII

In this section, we denote a binomial random variable having n trials with probability of success p by Bin(n, p).

All of the binomial random variables found in this section indicate the number of edges between nodes and/or communities. Unless otherwise indicated, we assume that the number of edges is sufficient such that np ≥ 10, and that the network is sparse enough such that np ≥ 10. This justifies the use of Assumption [1, 24]. We denote a normal random variable with mean \mu and variance \sigma^2 by \cal N(\mu, \sigma^2).

1. Initial sketch

We comment here on the validity of the exact recovery analysis. The SBM model used for analyzing exact recovery in [25] does not have fixed community sizes, but rather defines a probability distribution over communities. The membership of each node is then sampled from this distribution. In fact, the initial sketch G’(0) adheres to this model, since the probability that a given node in the sketch belongs to community u is \frac{|C_u(0)|}{N'}.
2. Inferring community membership

Let \(m'_{(u \rightarrow)} = \sum_{w \neq u} m'_{u \rightarrow w}\) be the total number of sketch nodes moving out of community \(u\), and \(m'_{(v \rightarrow u)} = \sum_{w \neq u} m'_{w \rightarrow u}\) be the total number of sketch nodes moving into community \(u\). Then,

\[
s_{i,v}(t) = \frac{S'_{\text{out}} + S'_{\text{in}}}{n'} \tag{C1}
\]

where

\[
S'_{\text{in}} \sim \begin{cases} \text{Bin}(n' - m'_{(u \rightarrow)}, p_m), & v = u \\ \text{Bin}(m'_{v \rightarrow u}, p_m), & v \neq u \end{cases} \tag{C2a}
\]

\[
S'_{\text{out}} \sim \begin{cases} \text{Bin}(m'_{(u \rightarrow)}, p_{\text{out}}), & v = u \\ \text{Bin}(n' - m'_{v \rightarrow u}, p_{\text{out}}), & v \neq u \end{cases} \tag{C2b}
\]

Since the random variables in \((C2)\) are mutually independent, it follows from Assumption 1 that \(s_{i,u}(t) - s_{i,v}(t) \sim \mathcal{N}(\mu_{\Delta s}, \sigma_{\Delta s}^2)\) for any \(u \neq w\), where

\[
\mu_{\Delta s} = \frac{(p_m - p_{\text{out}})(n' - m'_{(u \rightarrow)} - m'_{v \rightarrow u})}{n'},
\]

\[
\sigma_{\Delta s}^2 = \frac{p_{\text{in}}(1 - p_{\text{in}}) + p_{\text{out}}(1 - p_{\text{out}})}{n'} - \frac{[p_{\text{in}}(1 - p_{\text{in}}) - p_{\text{out}}(1 - p_{\text{out}})](m'_{(u \rightarrow)} - m'_{v \rightarrow u})}{n'^2}. \tag{C3}
\]

We note that if there are few moving nodes or few edges between communities, then \(S'_{\text{in}}\) and \(S'_{\text{out}}\) may not be well-approximated by a normal random variable. Nonetheless, in these cases the expected values of \(S'_{\text{in}}\) and \(S'_{\text{out}}\) are small enough that they do not contribute significantly to the bound regardless.

Let \(f \sim \mathcal{N}(\mu, \frac{\sigma^2}{n'})\). If \((C4)\) holds, then \(\mu \geq \Phi^{-1}(1 - \frac{1 - \alpha}{2})\sqrt{\frac{\sigma^2}{n'}}\) and hence \(P(f < 0) \leq \frac{1 - \alpha}{2}\). Furthermore, from Assumption 2 we have \(\mu_{\Delta s} \geq \mu\) and \(\sigma_{\Delta s}^2 \leq \frac{\sigma^2}{n'}\), and thus \(P(s_{i,u}(t) < s_{i,v}(t)) \leq P(f < 0)\). Then, by applying the union bound, the probability of successfully inferring the community membership of node \(i\) is

\[
P\left(\bigcap_{u=1}^{n'} s_{i,u}(t) > s_{i,v}(t) \right) = 1 - P\left(\bigcap_{u=1}^{n'} s_{i,u}(t) < s_{i,v}(t)\right)\geq 1 - \sum_{u=1}^{n'} P(s_{i,u}(t) < s_{i,v}(t)) = \alpha. \tag{C4}
\]

3. Split detection

Following a similar line of reasoning as in Sec. VII A, exact recovery is efficiently solvable if

\[
\frac{2n'}{\ln(2n')} > 2(\sqrt{p_m} - \sqrt{p_{u,w}})^{-2}. \tag{C5}
\]

On the other hand, from \((C3)\), the split is asymptotically detectable in the spectrum of \(B'\) if

\[
n' \left(p_{\text{in}} - p_{u,w}\right)^2 > p_{\text{in}} + p_{u,w}, \tag{C6}
\]

which is equivalent to

\[
p_{u,w}^2 > p_{u,w} \left[2p_{\text{in}} + \frac{1}{n'}\right] + p_{\text{in}} \left(p_{\text{in}} - \frac{p_{\text{in}}}{n'}\right) \geq 0. \tag{C7}
\]

The bounds in \((C7)\) follow directly by solving \((C5)\) and \((C7)\) in terms of \(p_{u,w}\).

4. Merge detection

Define \(\hat{m} = \hat{q}(t)n'(n' - 1)/2\) as the maximum possible number of intracommunity edges at time \(t\). Then,

\[
\hat{p}_{\text{in}} = \frac{\hat{p}_{\text{in}}}{\hat{m}}, \tag{C8a}
\]

\[
\hat{p}_{u,w} = \frac{\hat{p}_{u,w}}{(n')^2}. \tag{C8b}
\]

where \(\hat{p}_{\text{in}} \sim \text{Bin}(\hat{m}, p_{\text{in}})\), and \(\hat{p}_{u,w} \sim \text{Bin}((n')^2, p_{u,w}(t))\). Since \(\hat{p}_{\text{in}}\) and \(\hat{p}_{u,w}\) are independent, it follows from Assumption 1 that \(\hat{p}_{\text{in}} - \hat{p}_{u,w} \sim \mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}}^2)\) and \(\hat{p}_{\text{in}} + \hat{p}_{u,w} \sim \mathcal{N}(\mu_{\text{in}}^+, \sigma_{\text{in}}^2)\), where \(\mu_{\text{in}}^+ = p_{\text{in}} + p_{u,w}\).

Let \(\beta' = 1 - \frac{1 - \alpha}{2}\). If condition \((C8)\) holds, then

\[
p_{u,w}^2 - p_{u,w}\left[2p_{\text{in}} + 2\Phi^{-1}(\beta')\sigma_m + \frac{d^2}{n}\right]
\]

\[
+ (p_{\text{in}} + \Phi^{-1}(\beta')\sigma_m)^2 - 2\Phi(p_{\text{in}} - \Phi^{-1}(\beta')\sigma_m) < 0, \tag{C9}
\]

which in turn implies that

\[
\mu^+ - \Phi^{-1}(\beta')\sigma_m < d\sqrt{\frac{\mu^+ - \Phi^{-1}(\beta')\sigma_m}{n}}. \tag{C10}
\]

Then, from \((C10)\) and the union bound, the probability that \((C10)\) will indicate a split state is

\[
P\left(\hat{p}_{\text{in}} - \hat{p}_{u,w} < d\sqrt{\frac{\hat{p}_{\text{in}} + \hat{p}_{u,w}}{n}}\right)
\]

\[
\geq P\left(\hat{p}_{\text{in}} - \hat{p}_{u,w} < \mu^+ + \Phi^{-1}(\beta')\sigma_m\right)
\]

\[
\bigg| \bigg| d\sqrt{\frac{\hat{p}_{\text{in}} + \hat{p}_{u,w}}{n}} \bigg| \bigg| > d\sqrt{\frac{\mu^+ - \Phi^{-1}(\beta')\sigma_m}{n}} \bigg| \bigg|
\]

\[
\geq 1 - P\left(\hat{p}_{\text{in}} - \hat{p}_{u,w} < \mu^+ + \Phi^{-1}(\beta')\sigma_m\right)
\]

\[
- P\left(\hat{p}_{\text{in}} + \hat{p}_{u,w} > \mu^+ - \Phi^{-1}(\beta')\sigma_m\right) = \alpha. \tag{C11}
\]

We can derive \((32)\) using a similar argument.
5. Detecting the birth event

From (C1) and (C8a) and under Assumption 1 with no moving nodes,
\[ s_{i,v}(t) \sim \mathcal{N}\left( p_{out}, \frac{p_{out}(1 - p_{out})}{n'} \right), \quad 1 \leq v \leq \hat{q}(t) \tag{C12} \]
\[ \hat{p}_{in} \sim \mathcal{N}\left( p_{in}, \frac{p_{in}(1 - p_{in})}{\hat{m}} \right) \tag{C13} \]

Then \( P(p^- \leq \hat{p}_{in} \leq p^+) = 1 - \beta \), and therefore
\[ P\left( \hat{p}_{in} - 3\sqrt{\frac{\hat{p}_{in}(1 - \hat{p}_{in})}{n'}} < p^- - 3\sqrt{\frac{p^+(1 - p^-)}{n'}} \right) \leq 1 - P(p^- \leq \hat{p}_{in} \leq p^+) = \beta. \tag{C14} \]

Furthermore, suppose that node \( i \) is joining community \( u \). For any community \( v \neq u \),
\[ P\left( s_{i,v}(t) \geq p_{out} + \Phi^{-1}(1 - \beta) \sqrt{\frac{p_{out}(1 - p_{out})}{n'}} \right) = \beta. \tag{C15} \]

Condition (21) is equivalent to
\[ p_{out} + \Phi^{-1}(1 - \beta) \sqrt{\frac{p_{out}(1 - p_{out})}{n'}} \leq p^- - 3\sqrt{\frac{p^+(1 - p^-)}{n'}}. \tag{C16} \]

Then, the probability that the condition inside (17) will fail for a particular \( v \) is
\[ P\left( s_{i,v}(t) \geq \hat{p}_{in} - 3\sigma \right) = P\left( s_{i,v}(t) \geq \hat{p}_{in} - 3\sqrt{\frac{\hat{p}_{in}(1 - \hat{p}_{in})}{n'}} \right) \leq P\left( s_{i,v}(t) \geq p_{out} + \Phi^{-1}(1 - \beta) \sqrt{\frac{p_{out}(1 - p_{out})}{n'}} \right) + P\left( \hat{p}_{in} - 3\sqrt{\frac{\hat{p}_{in}(1 - \hat{p}_{in})}{n'}} < p^- - 3\sqrt{\frac{p^+(1 - p^-)}{n'}} \right) \leq P\left( s_{i,v}(t) \geq p_{out} + \Phi^{-1}(1 - \beta) \sqrt{\frac{p_{out}(1 - p_{out})}{n'}} \right) \]
\[ + \ P\left( \hat{p}_{in} - 3\sqrt{\frac{\hat{p}_{in}(1 - \hat{p}_{in})}{n'}} < p^- - 3\sqrt{\frac{p^+(1 - p^-)}{n'}} \right) = 2\beta. \tag{C17} \]

Consequently, the probability that the condition will hold for all communities is
\[ P\left( \bigcap_{v=1}^{\hat{q}(t)} s_{i,v}(t) + 3\sigma < \hat{p}_{in} \right) \]
\[ = 1 - P\left( \bigcup_{v=1}^{\hat{q}(t)} s_{i,v}(t) + 3\sigma \geq \hat{p}_{in} \right) \]
\[ \geq 1 - \sum_{v=1}^{\hat{q}(t)} P\left( s_{i,v}(t) + 3\sigma \geq \hat{p}_{in} \right) \]
\[ = 1 - 2(\hat{q}(t) - 1) \beta = \alpha. \tag{C18} \]

Appendix D: Numerical results: details of algorithms used in comparison

For the algorithm of Yang et al. [17], we use the same tuning strategy as in the experimental results section of [17]. Specifically, we use the same temperature and iteration sequences, with \( \alpha = 0.8, \beta = 0.5, \gamma = 1, \mu_{kk} = 10 \). We run five instances of the algorithm with (1) \( \alpha_{kk} = 1, \beta_{kl} = 1 \), (2) \( \alpha_{kk} = 5, \beta_{kl} = 1 \), (3) \( \alpha_{kk} = 10, \beta_{kl} = 1 \), (4) \( \alpha_{kk} = 10^2, \beta_{kl} = 10 \), (5) \( \alpha_{kk} = 10^4, \beta_{kl} = 10 \), and then take the community assignments among the five trials yielding the highest average modularity (modularity is defined as in [17]). For the algorithm of Dinh et al. [14], we use the CNM algorithm [7] for the static clustering step, as in [14]. When running the ESPRA algorithm, we use the same parameters as used in the experimental results of [26]: \( \alpha = 0.8, \beta = 0.5 \). The algorithm (Independent) applies Static-Cluster to each snapshot to obtain community estimates \( \{\mathcal{C}_1, \ldots, \mathcal{C}_q\} \). To provide continuity in the community assignments of the nodes, community \( u \) in each snapshot at time \( t > 0 \) is matched to the community at time \( t-1 \) having the largest overlap according to the Jaccard coefficient. Specifically, for each community \( u \), we set the new estimate as \( C_u(t) = \mathcal{C}_{u'} \)
where \( u' = \arg \max_{1 \leq v \leq q} \frac{|C_u \cap C_{u'}(t-1)|}{|C_u \cup C_{u'}(t-1)|} \).

[1] C. Aggarwal and K. Subbian, “Evolutionary network analysis: A survey,” ACM Comput. Surv., vol. 47, no. 1, pp. 10:1–10:36, May 2014.
[2] D. Greene, D. Doyle, and P. Cunningham, “Tracking the evolution of communities in dynamic social networks,” in 2010 International Conference on Advances in Social Networks Analysis and Mining, Aug 2010, pp. 176–183.
[3] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, “Synopses for massive data: Samples, histograms, wavelets, sketches,” Foundations and Trends in
