Reliable Community Search in Dynamic Networks

Yifu Tang
Deakin University
Geelong, Australia
tangyif@deakin.edu.au

Jianxin Li∗
Deakin University
Geelong, Australia
jianxin.li@deakin.edu.au

Nur Al Hasan Haldar
The University of Western Australia
Perth, Australia
nur.haldar@uwa.edu.au

Ziyu Guan
Xidian University
Xi’an, China
zyguan@xidian.edu.cn

Jiajie Xu
Soochow University
Suzhou, China
xujj@suda.edu.cn

Chengfei Liu
Swinburne University of Technology
Melbourne, Australia
cliu@swin.edu.au

1 INTRODUCTION

Local community search has attracted much attention in recent years and has shown its great success in different applications, e.g., personalized recommendation [19, 20], destination marketing [34]. In general, local community search aims to identify a densely connected structure with regard to a query vertex. Majority of the existing works [9, 29, 30] on local community search consider static network structure. For instance, Clauset et al. [9] and Luo et al. [30] proposed local modularity to measure the quality of the community in the static network. A random walk based community detection model from multiple static networks is proposed in [29]. Some existing works [7, 12] take into consideration that network structure may change over time and propose local community search in dynamic networks. For example, Bu et al. [7] provides a modularity-based criterion to find local communities in a dynamic network and update them in an incremental manner by monitoring the changes. In another work, DiTursi et al. [12] discovered the dynamic communities with optimal time intervals by minimizing temporal conductance, a well-known metric to measure the quality of a community. However, the above-mentioned works assess the community quality using their aggregated structural cohesiveness at independent timestamps and ignore the evolving structure of a community over time. In the dynamic network, the continuity of the community cohesiveness is an important factor in determining whether a community is reliable. For example, it is desirable to hire a team that continuously delivers high-quality outputs together over time in the collaboration network. Finding the user groups with a longer duration of reacting to the social event can help better analyze user behavior on social media. Existing works largely ignore the continuity of the community cohesiveness. In addition, they did not consider the edge weight, e.g., connection strength between a node pair, which incurs the new computational challenge to solve local community search in dynamic networks.

To fill this research gap, we propose a novel community model of $\theta, k$-core reliable community in dynamic networks where: (i) the community is a $k$-core with each edge weight no less than the weight threshold $\theta$, and (ii) spans over a period of time. The most reliable local community search aims to find the community with the maximum reliability score, which is defined by coupling temporal continuity and member engagement. In other words, this work jointly models the three important properties, i.e., connection strength, cohesiveness continuity, and member engagement, of a community in a dynamic network.

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Cyril-Tang/CRC-query.
vertex exceeds a given threshold. The proposed searching algorithm takes exponential complexity. However, the persistent community in [22, 24] did not consider the weight of the edge and its time complexity is too high for dealing with large-scale networks. The frequency-based subgraph in [25, 35] ignored the continuity of the cohesive structure and failed to maximize member engagement.

To solve the proposed problem of the most reliable local community search, a naive idea is to enumerate all the possible community candidates and select the satisfied results by checking their edge weight and duration. However, this may incur an exponential time cost. To address the computational challenge, in this paper, we firstly propose an efficient eligible edge filtering online search algorithm that utilizes the minimum edge requirement of k-core to compute the reliability upper bound to prune a large number of edge sets without probing their corresponding (\(\theta,k\))-core community candidates. To further accelerate the query processing, we develop a weighted core forest index by maintaining the standard \(\theta\)-threshold values and the \((\theta,k)\)-core structural information of vertices, which supports efficient retrieval of \(k\)-core with regard to different thresholds and timestamps. Following this, we design an index-based dynamic programming algorithm, and derive the reliability upper bound of communities w.r.t. the time interval during the dynamic programming procedure to avoid probing the unsatisfactory community candidates. Besides, the index construction, maintenance, and compression are well presented in this paper.

The main contributions of this work are as below:

- We propose a novel problem of the most reliable community search that jointly considers community continuity, community size, and connection strength for online network analysis services.
- We develop an efficient online search algorithm by deriving and applying the properties of pruning the ineligible edges w.r.t. the given query conditions.
- We further present a weighted core forest index and develop an index-based dynamic programming algorithm to solve the most reliable community search problem in a more efficient way.
- We conduct extensive experiments to show the efficiency and effectiveness of the proposed algorithms and community model by using eight real-world datasets and comparing with three existing studies.

The remainder of this paper is organized as follows. First, we formalize the most reliable local community search problem in Section 2 and develop the online search algorithm in Section 3. Then, we introduce our index structure and the detailed index-based search algorithm in Section 4. The procedures of index construction, maintenance and compression are shown in Section 5. Experimental evaluation and results are discussed in Section 6. Finally, we discuss the related work in Section 7 and conclude the work in Section 8.

2 PRELIMINARIES AND PROBLEM DEFINITION

In this section, we first present the preliminaries and then formalize the problem of the most reliable local community search.

Definition 2.1 (Dynamic Networks). A dynamic network \(G = \{G_t, \ldots, G_T\}\) is a sequence of time-variant weighted graph instances
\{G_t_1, ..., G_T\} \text{ s.t. } t_1 < t_2 < ... < T, \text{ where each timestamped instance } G_t = (V_t, E_t, W_t) \text{ contains a set of vertices } V_t, \text{ a set of edges } E_t \text{ with the weights } W_t(e) \in (0, 1) \text{ for } e \in E_t.

In this work, we ignore the isolated vertices, so that vertex updates can be supported by edge insertions and deletions. The edge insertion with a new endpoint can represent the vertex addition and edge deletion isolated an endpoint reflects the vertex deletion. For simplicity, we assume all graph instances share a fixed set of vertex \(V\), i.e. \(G = (V, E_t, W_t)\). The edge weight is a widely-used network feature to represent the interaction frequency, similarity, or connection strength between vertices. In this work, we normalize the edge weight to be in \((0, 1]\).

For a graph instance \(G_t = (V, E_t, W_t)\), \(deg(u, G_t)\) denotes the degree of a vertex \(u\) in \(G_t\), which is the number of neighbors of \(u\) in \(G_t\). Like [37], we consider \(k\)-core in \(G_t\) as a connected subgraph \(G_t^{k} = (V_t^k, E_t^k, W_t^k)\) where each vertex has the degree no less than \(k\), i.e. \(\forall u \in V_t^k, \text{deg}(u, G_t^k) \geq k\).

**Definition 2.2** \((\theta,k)\)-core. Given a graph instance \(G_t = (V, E_t, W_t)\), an integer \(k\), and a threshold \(\theta\), a connected subgraph \(G_t^{\theta,k} = (V_t^{\theta,k}, E_t^{\theta,k}, W_t^{\theta,k})\) is called a \((\theta,k)\)-core of \(G_t\) if \(G_t^{\theta,k}\) is a \(k\)-core and each edge has the weight no less than \(\theta\), i.e. \(\forall u \in V_t^{\theta,k}, \text{deg}(u, G_t^{\theta,k}) \geq k \text{ and } \forall e \in E_t^{\theta,k}, W_t^{\theta,k}(e) \geq \theta\).

**Definition 2.3** (Time Interval based \((\theta,k)\) Reliable Community (CRC)). Given a dynamic network \(\mathcal{G} = \{G_t_1, ..., G_T\}\), an integer \(k\), a threshold \(\theta\), and a time interval \(T_C = [t_s, t_e]\), a \((\theta,k)\)-reliable community is a subgraph \(C = (V_C, E_C)\) that spans continuously from \(t_s\) to \(t_e\) and for each timestamp \(t_n \in T_C\), the subgraph induced by \(E_C\) from the graph instance \(G_t_n\) is a \((\theta,k)\)-core, i.e. \(\forall t_n \in T_C, G_{t_n}[E_{t_n}]\) is a \((\theta,k)\)-core of \(G_{t_n}\). In the remainder of this work, \((\theta,k)\)-core reliable community is called CRC for brevity.

Based on Definition 2.3, a CRC maintains a cohesive structure with the required connection strength of a time interval in the dynamic network. Its reliability score can be measured by coupling the continuity and size of the community as below:

**Definition 2.4** (Reliability Score of CRC) Given a dynamic network \(\mathcal{G} = \{G_t_1, ..., G_T\}\), a query vertex \(q\), a threshold \(\theta\), and a structural constraint integer \(k\), and a query time interval \(T_Q = [t_i, t_j]\), \(k\) query vertex \(q\), a query vertex \(q\), a threshold \(\theta\), a structural constraint integer \(k\), and a query time interval \(T_Q = [t_i, t_j]\), the problem of the Most Reliable Local Community Search is to find the CRC \(C = (V_C, E_C)\) and its continuous time interval \(T_C = [t_s, t_e]\), satisfying

\[
\arg \max_{C \subseteq \mathcal{V}} \frac{\sum_{t_s \leq t \leq t_e} \text{deg}(q, C)}{|C|} \geq \gamma \\forall e \in E_C, \forall t \in T_C, W_t(e) \geq \theta, \text{ and } T_C \subseteq T_Q.
\]

As shown in Figure 1, when the query time interval is \([t_1, t_3]\), we can obtain two reliable communities \(C_1\) and \(C_2\) w.r.t. the query input \((v_0, 0.5, 2)\). The maximal 2-core is composed of 10 vertices. When \(\alpha = 1\), we have \(S_{\text{rel}}(C_1) = S_{\text{rel}}(C_2) = 0.57\). When \(\alpha\) increases to 2, we have \(S_{\text{rel}}(C_1) = 0.63 < S_{\text{rel}}(C_2) = 0.77\), and \(C_2\) with longer duration becomes the optimal result.

To solve the most reliable community search problem, a naive solution is to compute all the \((\theta,k)\)-cores at each timestamp, and then verify their longest duration in the dynamic network. After that, their reliability scores can be obtained by multiplying their size and the number of continuous timestamps. Finally, the most reliable community can be returned by selecting the ones with the maximum reliability scores. However, the operation of finding all the \((\theta,k)\)-cores needs to probe all the combinations of edges that form a connected subgraph with no less than \(k(k + 1)/2\) edges and \(k + 1\) vertices, i.e., satisfying the conditions of minimal \(k\)-core component. Thus, we can remark that the computational cost of finding the most reliable community is in exponential complexity.

### 3 ONLINE RELIABLE COMMUNITY SEARCH

To efficiently solve the problem of reliable local community search, in this section, we will present a novel Eligible Edge Filtering (EEF) based Online CRC Search algorithm. Different from the naive idea discussed in Section 2, EEF does not need to generate all the \((\theta,k)\)-core candidates, which can greatly reduce the query time cost.

Given a graph instance \(G_{t_n} = (V, E_{t_n})\) and a weight threshold \(\theta\), we can filter out “ineligible” edges whose weights are less than \(\theta\) and maintain only the “eligible” edges. We identify the set of eligible edges, denoted by \(E_{t_n,\theta}\). i.e. \(E_{t_n,\theta} = \{e \in E_{t_n} | W_{t_n}(e) \geq \theta\}\) for the CRC construction.

**Definition 3.1** (Eligible Lasting Time of Edge). Given a graph instance \(G_{t_n} = (V, E_{t_n})\) at timestamp \(t_n \in [t_i, t_j]\), and a threshold \(\theta\), for an edge \(e \in E_{t_n,\theta}\), its eligible lasting time \(\lambda_{t_n,\theta}(e)\) is measured by the length of the longest time interval \([t_m, t_n]\) \((t_m \leq t \leq t_n)\) when \(W_{t_n}(e) \geq \theta\) for all \(t \in [t_m, t_n]\).

Eligible lasting time calculates the number of continuous timestamps that the edge is “eligible” until the current timestamp. For example, in Figure 1, for an edge \(e = (v_0, v_1)\), we have \(\lambda_{t_0,0.5}(e) = 1\), \(\lambda_{t_0,0.6}(e) = 2\) and \(\lambda_{t_0,0.6}(e) = 0\). We can easily derive that the eligible lasting time of edges can be incrementally computed by accessing the dynamic network chronologically.

Based on the eligible time, we can easily identify common edges of multiple continuous graph instances, that can be utilized to construct CRC with varying duration. Different from the vertex-induced subgraph, the subgraph induced by an eligible edge set
provides the guarantee to meet the requirement of edge weight \( \theta \). Therefore, the eligible edge set can be used to prune the unqualified \( k \)-core candidates by using the following property.

**Property 3.1 (Minimum \( k \)-core).** Given an edge set \( E_{t_n, \theta} \) at timestamp \( t_n \) with regards to a threshold \( \theta \), \( E_{t_n, \theta} \) can be pruned without probing its induced communities if \( |E_{t_n, \theta}| < k(k + 1)/2 \), i.e., the number of edges does not meet the density requirement of \( k \)-core.

In addition, we can calculate the upper bound of the reliability score of CRCs constructed using \( E_{t_n, \theta} \). Given \( G' = (V', E') \) as a \( k \)-core in the induced subgraph \( G_{t_n}[E_{t_n, \theta}], |E_{t_n, \theta}| \geq |E'| \geq (k - \lambda)|V'|/2 \) must hold because there are at least \( k \) edges for a vertex in the \( k \)-core. Therefore, if \( G' \) can form a CRC with its duration as \( d \), then its vertex size satisfies that \( |V'| \leq 2|E_{t_n, \theta}|/k \). Thus, we have \( N(V') \leq 2|E_{t_n, \theta}|/k \) and \( N(T) = d/|T_Q| \) and we can determine the upper bound of reliability score of the CRC constructed by the given eligible edge set.

For simplicity of presentation, we omit the normalizers \( |V|_{\mathit{max}} \) of the maximum community size and \( |T_Q| \) of the query time interval in the following equations.

**Property 3.2 (\( S_r \) Upper Bound of CRC w.r.t. \( E_{t_n, \theta} \)).** Given an eligible edge set \( E_{t_n, \theta} \), an integer \( k \), the upper bound reliability score \( \mathit{UBR}_{t_n}^d \) of the CRC constructed using \( E_{t_n, \theta} \) whose duration is \( d \) is calculated as:

\[
\mathit{UBR}_{t_n}^d = (1 + \alpha^2) \cdot \frac{|E_{t_n, \theta}| |k \cdot d}{(\alpha^2 \cdot 2|E_{t_n, \theta}|/k) + d}
\]

The key idea of \( \mathit{EEF} \)-based Online CRC Search is to filter out edges in each graph instance \( G_{t_n} \) using the given threshold \( \theta \) while maintaining the lasting time of each edge by a timestamp. As shown in Algorithm 1, we first initialize \( C_{\mathit{opt}} \) and \( \mathit{maxS} \) to store the most reliable community and its reliability score (line 1). Then, for each timestamp \( t_n \), we traverse the edges of \( G_{t_n} \) starting from the query vertex \( q \) in Breath-First Search manner. In the meantime, vertices and edges that violate the degree and weight constraints are pruned. During the traversing, the eligible time of edges is updated incrementally, and the eligible edges are added to the edge set \( E_{t_n, \theta} \).

Having \( E_{t_n, \theta} \), we calculate its upper bound \( \mathit{UBR}_{t_n}^d \) of potential CRC whose duration is \( 1 \) (lines 2-10). Then, we visit each timestamp \( t_n \) in the descending order of \( \mathit{UBR}_{t_n}^d \), which provides a best-first search strategy to exploit the CRCs. We utilize \( E_{t_n, \theta} \) to construct CRC with duration \( d \) iterating from 1 to \( |[t_i, t_n]| |(|E_{t_n, \theta}||E_{t_n, \theta}|) \) (lines 11-19). At each iteration, we select the edge set \( E_{t_n, \theta} \) where each edge has the eligible time no less than \( d \) and update the upper bound w.r.t. \( d \). Then we adopt Property 3.1 and Property 3.2 to prune the CRC construction if \( |E'| \) is too small or \( \mathit{UBR}_{t_n}^d \) cannot exceed \( \mathit{maxS} \). After that, we can extract the CRC \( C \) (i.e., local maximal \( k \)-core) from the induced subgraph \( G_{t_n}[E'] \) by finding the connected component containing \( q \) after the core decomposition, and then update \( C_{\mathit{opt}} \) and \( \mathit{maxS} \). Finally, the algorithm returns \( C_{\mathit{opt}} \) as the optimal result.

The time complexity of Algorithm 1 can be analyzed as below. For each graph instance \( G_{t_n} = (V, E_{t_n}) \), it takes \( O(|V| + |E_{t_n}|) \) to run the Breath-First Search that requires to visit very vertex and edge once. At the same time, the eligible time of each edge is obtained (lines 2-18). Then, it needs \( O(|[t_i, t_n]| \cdot |E_{t_n}|) \) to compute \(|[t_i, t_n]| \cdot |E_{t_n}|\) number of CRCs where each CRC is obtained by a core decomposition process that needs to consume \( O(|E_{t_n}|) \) [4] (lines 20-27). Therefore, for the query interval of \( |T_Q| \) timestamps, Algorithm 1 takes \( O(\sum_{t_n \in T_Q} (|V| + |E_{t_n}|) + ([t_i, t_n] \cdot |E_{t_n}|)) \) in total, which can be rewritten as \( O(|T_Q| \cdot (|V| + |E_{t_n}|) + (|T_Q| \cdot |E_{t_n}|)) \), i.e., \( O(|T_Q|^2 \cdot |V| + |E_{t_n}| \cdot |T_Q| \cdot (|V| + |E_{t_n}|)) \), where \(|E_{t_n}|\) denotes the average number of edges of the graph instances.

4 INDEX BASED RELIABLE COMMUNITY SEARCH

To further accelerate the query processing, in this section, we first propose a forest index structure, called \textit{Weighted Core Forest Index (WCF-Index)}, to maintain the \((\theta, k)\)-core vertices for each graph instance \( G_{t_n} \). Then, we develop an index-based dynamic programming algorithm by using the proposed index and derive the reliable score upper bound with great pruning power to accelerate the query algorithm.

4.1 WCF-Index

The general idea of this index is to maintain the vertex candidates of the \((\theta, k)\)-core with regards to the given \( \theta \) and \( k \) at each timestamp, from which we can work out the satisfied CRCs containing \( q \) with the different continuous time intervals.

**Definition 4.1 (\( \theta \)-threshold of a Vertex).** Given a graph instance \( G_t = (V, E_t) \) at a timestamp \( t \) and an integer \( k \), for a vertex \( u \in V \), it may have a set of \( \theta \) values and their corresponding \((\theta, k)\)-core
Figure 2: θ-threshold of $G_{k_1}$  
Figure 3: θ-tree of $G_{k_1}$, $k=2$

subgraphs containing $u$. Thus, we take the largest $θ$ value in the $θ$
set as the θ-threshold of $u$, denoted as $θ$-thres$_{G_1}(u, G_{k_1})$.

Example 4.1. Figure 2 shows the θ-threshold of vertices in $G_{k_1}$ in
Figure 1 (a) with regards to different $k$ values, e.g., $θ$-thres$_{G_1}(v_1, G_{k_1})$
0.5 because $(0.5, 2)$-core (i.e., $\{v_0, v_1, v_2, v_3, v_4\}$) exists in $G_{k_1}$, but
no one $(θ'$, 2)-core containing $v_1$ exists if $θ' > 0.5$.

According to the above definition and the example, we are able to
justify whether a vertex $v$ is contained in a $(θ, k)$-core for given $θ$
and $k$ if the θ-threshold values of vertices are maintained. However,
the θ-threshold only implies the vertex candidates of a $(θ, k)$-core,
but fails to reflect the structural connectivity of vertices. Therefore,
it is highly desirable to design an index structure for maintaining the
θ-threshold and the structure information of vertices together.

Yang et al. in [44] proposed a forest-based index to query $(k, η)$-
core in a static uncertain graph where $k$ implies the degree con-
straint of the vertex and $η$ implies the probability of the vertex to
appear in the subgraph. By maintaining $η$-tree$_k$ for each $k$, it can
accelerate the search of the all $(k, η)$-core with custom $η$ require-
ments. Motivated by $η$-tree$_k$, in this work, we extend the concept
of θ-tree$_k$ to the dynamic weighted network to construct the θ-
tree$_{G_1}$ for each $k$ at time $t$, which can support quick retrieval of
local maximal $(θ, k)$-core from the indexed graph instance $G_{k_1}$.

Definition 4.2 ($θ$-tree$_{G_1}$). Given a graph instance $G_{k_1}$, an integer
$k$, $θ$-thres$_{G_1}$ index is a tree structure, satisfying

1. **Node**: each tree node $V$ is a set of maximal connected
vertices in $G_{k_1}$ with same $θ$-threshold value, denoted as $V, θ$,
i.e., $∀v \in V$, $θ$-thres$_{G_1}(v, G_{k_1}) = θ, V$.

2. **Parent-child relationship**: for a node $W$, $N_W(G_{k_1})$

denotes the tree nodes that are connected to $W$ in $G_{k_1}$
with $θ$-threshold smaller than $W, θ$. The parent node $V$ of $W$ is
the node with the largest $θ$-threshold in $N_W(G_{k_1})$, i.e.
$V = argmax _{v \in N_W(G_{k_1})} V, θ$.

Example 4.2. Figure 3 presents the constructed $θ$-tree$_{G_1}$, $G_{k_1}$
from threshold of $G_{k_1}$ in Figure 2. If we search $(0.5, 2)$-core on
θ-tree$_{G_1}$, three tree nodes will be returned, i.e., $\{v_0, v_1, v_2, v_3, v_4\}$,
and $\{v_7, v_8, v_9\}$. These tree nodes can induce two $(0.5, 2)$-core, i.e.,
g1 and g2.

$θ$-tree can be composed of several trees where each tree repre-
sents a connected component in the graph instance. We denote $T$
as the WCF-Index where $T[k][t]$ represents the $θ$-tree$_{G_1}$ of each $k$
and $t$ in the dynamic network.

Remark 1. In this work, we set the $θ$-threshold as the standard
values $\{0, 0.1, 0.2, ..., 0.9, 1\}$. If the $θ$-threshold of a vertex is not
in the standard set, we will round it down to the nearest stan-
dard value. Accordingly, fetching $(θ, k)$-core with non-standard $θ$
value will also be processed as the nearest rounded down standard
value. For instance, to fetch $(θ, k)$-core with $θ = 0.55$, the index
accesses the tree nodes from $θ$-threshold of 0.5 and then examines the
$θ$-threshold of vertices in the tree node $V$ if $V, θ < 0.55$. In the
following discussion, we skip this process for simplicity.

### 4.2 Dynamic Programming based CRC search

To solve the most reliable community search problem, we need to
compare CRC with different duration. In this section, we develop a
dynamic programming algorithm based on the recursive relation of
CRCs ending in consecutive timestamps and utilize the WCF-Index
to search CRC with varying duration efficiently.

Assume that the lasting time interval of the CRC is fixed (so does
the duration), then we only need to extract the CRC with the largest
size. Given the duration of the CRC is $d$ and the last timestamp
it spans is $t_n$, we denote the maximal CRC w.r.t. the query input
as $(d, t_n)$. We can easily derive the following recursive relation
between CRCs:

$$C(d, t_n) \subseteq C(d - 1, t_{n-1}) \cap C(d - 1, t_n)$$

The base situation is $C(1, \cdot)$ that can be retrieved from WCF-Index.

Based on Eq. 4, we can devise a DP algorithm to compute $C(d, t_n)$.
More specifically, to get $C(d, t_n)$, we simply compute the intersection
of $C(d - 1, t_{n-1})$ and $C(d - 1, t_n)$ and extract the local maximal
$(θ, k)$-core using core decomposition. The intermediate result of
$C(d, t_n)$ with varying $d$ is maintained to support the adoption of
the dynamic programming.

If at a timestamp $t_n$, the maximal community $C(1, t_n)$ does not exist,
i.e., for a given query, there is no such subgraph satisfying $(θ, k)$-
core constraint at time $t_n$, then it implies that further calculations
depending on $C(1, t_n)$ are unnecessary. In this work, these kinds of
timestamps like $t_n$ are called anchored timestamps of a query time
interval. The anchored timestamps split the query interval $[t_i, t_j]$
into several non-overlapping time intervals $T_k = \{T_1, T_2, ...\}$. For
each interval $T_i \in T_k$, we can compute the upper bound of the reliabil-
ity score of the communities.

Property 4.1 ($S_{rel}$ Upper Bound of CRC w.r.t. $T_i$). Given a time
interval $T_i = [t_i, t_j]$ where $C(1, t_n)$ exists for every $t_n \in [t_i, t_j]$
we can construct an array $M = \{\mu_1, \mu_2, ..., \mu_n\}$ to store the size of $C(1, \cdot)$,
where $\mu_n$ denotes the size of $C(1, t_n)$. The upper bound reliability
score (UBR) in this time interval can be calculated by:

$$UBR_{T_i} = max _{\mu_n \in M} \left((1 + \alpha^2) \cdot \frac{\mu_n \cdot LCT(\mu_n, M)}{\alpha^2 \cdot \mu_n + LCT(\mu_n, M)}\right)$$

where $LCT(\mu_n, M)$ stands for the length of the Longest Consecutive
Timestamps $S$ in array $M$ such that $i \in S[\mu_i \leq \mu]$. 

### Figure 4: UBR of $[t_1, t_5]$
Example 4.3. For a community \( C(1, t_n) \) of size \( \mu_n \), the largest reliability score of a CRC constructed by \( C(1, t_n) \) is determined by the longest possible duration that \( \mu_n \) can remain. Hence, the upper bound score of a time interval is the maximum value among all the largest possible scores for each \( C(1, \cdot) \) size. Figure 4 shows an illustrative example of calculating the UBR of interval \([t_1, t_5]\) where we assume \( |V_k|_\text{max} = 10 \). The size of \( C(1, t_5) \) is 4 and the longest continuous timestamps for this size is \( (t_3, t_5) \), so the largest possible score of the CRC constructed by \( C(1, t_5) \) is \( 0.5/10 + 4 + 0.5/5 + 3 = 0.5 \). The largest possible reliability score is obtained by a CRC constructed by \( C(1, t_2) \) that contains three vertices and spans for five timestamps.

Remark 2. Similar to Property 2, we use \( UBR_{opt}^d \) to denote the upper bound calculated by the size of \( C(d, \cdot) \). However, it can determine the maximum reliability score of the community with duration longer than \( d \). We can derive that

\[
UBR_{opt}^d = \max_{\mu \in M} \left( \frac{1 + \alpha^2}{\alpha^2} \cdot \mu \cdot (d + LCT(\mu, M) - 1) \right) \tag{6}
\]

where \( \mu \) denotes the size of \( C(d, t_n) \) and \( LCT(\mu, M) - 1 \) represents the additional lasting timestamps of size \( \mu \) on top of \( d \). In the process of community search, \( UBR_{opt}^d \) can be updated with different duration of CRC and provide sustainable pruning power. Having \( UBR_{opt}^d \) calculated for each interval \( t_i \) and updated during the community exploration, we can skip exploring CRCs if \( UBR_{opt}^d \) is no larger than the reliability score of intermediate community candidates we have obtained during the query processing.

Algorithm 2 presents the detailed dynamic programming procedure of the WCF-Index based CRC Search. We first initialize a table \( L_C \), \( maxS \), and \( C_{opt} \) to store the extracted communities, the maximum reliability score, and the most reliable community (line 1). For each timestamp \( t_n \in [t_1, t_j] \), we can obtain \( C(1, t_n) \) from the \( \theta\)-tree,\( t_n \) index, and store it in \( L_C[1][|n|] \) (lines 2-6). In addition, we also determine whether \( t_n \) is an anchored timestamp. Then we split \( T_Q \) into several non-overlapping time intervals \( T_Q = \{T_1, T_2, \ldots\} \) with valid \( C(1, \cdot) \) by the anchored timestamps, and calculate their upper bound reliability score (lines 7-8). For each individual time interval \( T_i = [t_s, t_e] \), if its upper bound is no larger than \( maxS \), then the time interval is pruned (line 10). Otherwise, we initialize an array \( Q \) to store the size of CRC and compute the CRC with various duration \( d \) based on Eq. 4, and store the intermediate CRC \( C(d, t_s) \) in \( L_C[d][x] \) (lines 11-16). Then, we update \( maxS \) and \( C_{opt} \) and add the size of \( C(d, t_s) \) to \( Q \) for upper bound calculation (lines 17-20). Once \( C(d, \cdot) \) has been explored for every \( t_s \in [t_1, t_e] \), we can update the \( UBR \) and determine whether it is necessary to explore communities with longer duration in this interval (line 21-22).

Finally, the algorithm returns the most reliable community \( C_{opt} \) whose reliability score is the largest.

The time complexity of Algorithm 2 is dominated by the operation of finding CRCs with various duration (lines 10-23) as the \( C(1, \cdot) \) community can be queried from the index in constant time. In the worst case, there are up to \( |T_Q|^2 \) subgraphs to be explored and the community construction takes \( O(|E_i|) \) complexity, where \( T_Q \) is the query interval and \( |E_i| \) is the average number of edges of graph instance \( G_i \). The total complexity is \( O(|T_Q|^2 \cdot |E_i|) \). Compared with the EEF-based Online CRC Search Algorithm, WCF-Index can avoid searching a large number of edges, which helps to reduce the time cost of computing \( C(1, \cdot) \).

5 WCF INDEX CONSTRUCTION, MAINTENANCE, AND COMPRESSION

In this section, we describe the procedure of index construction, and propose index maintenance and compression strategies to support efficient query processing over dynamic weighted networks and reduce the time and space cost of the index.

5.1 WCF-Index Construction

The main idea of constructing WCF-Index is to build the \( \theta\)-tree,\( t \) for each graph instance \( G_t \) for \( k \in [1, k_{\text{max}}] \), where \( k_{\text{max}} \) denotes the maximum core number of the vertex in \( G_t \). To obtain the \( \theta\)-tree,\( t \), we group the vertices by their \( \theta\)-threshold value and add the vertex groups as tree nodes into the \( \theta\)-tree according to the \( \theta\)-threshold and connectivity of the tree nodes, i.e., the vertex groups.
Algorithm 3 presents the procedure of building the WCF-Index $I$. For each graph instance $G_t$, we build $\theta$-tree$_{k,t}$ for each available $k$ by first computing $\theta$-threshold of vertices and then construct and insert tree nodes to the $\theta$-tree index. We first initialize two graphs $G_{pre}$ and $G_{cur}$ to store intermediate states of edge filtering (line 3). Then we iteratively pick $\theta^* \in \Theta$ in descending order. For each $\theta^*$, we get the edge set $E_{t,\theta'}$ whose weights are no less than $\theta'$ and obtain the induced subgraph $G_t[E_{t,\theta'}]$ as the current state $G_{cur}$. We can obtain a set of vertices $V_{th}$ whose core number in $G_{cur}$ is increased with regards to the core number in the last state $G_{pre}$. This implies that for a vertex $w \in V_{th}$, the $\theta$-threshold of $w$ is $\theta^*$ w.r.t. its increased core number (lines 4-7). After that, we set the previous state $G_{pre}$ to be $G_{cur}$ and obtain the distinct values of the newly increased core numbers $K$ (lines 8-9). Then, according to the newly identified $\theta$-threshold, we can construct and add tree nodes to the $\theta$-tree$_k$, for $k^*$ in $K$ (lines 10-21). To do that, we get $\theta$-tree$_{k,t}$ from $I[k^*][t]$, then we identify the groups of connected vertices whose $\theta$-threshold at $k^*$ as the tree node $X$ (lines 11-14). To determine the position of $X$, we find each tree node $Y$ that contains any neighbor $v$ of $G_t[X]$ and its root $Z$, so that $X, Y, Z$ are connected (lines 15-18). If $Z.\theta > X.\theta$, $X$ is assigned as the parent of $Z$, otherwise, their $\theta$-threshold are the same because the smaller $\theta^*$ has not been visited yet, so we need to merge $X$ to $Z$ (lines 19-21). After iterating all the standard threshold values, we can construct all the $\theta$-tree$_k$ completely for each possible $k$ of each $G_t$ and return the WCF-Index $I$. The space cost of WCF-Index is $O(\sum_{t \in T} \sum_{u \in V} \text{core}(u, G_t))$ as each vertex $u$ appears $\text{core}(u, G_t)$ times in each graph instance.

Example 5.1. Consider the graph instance $G_t$ in Figure 1, we calculate the $\theta$-threshold of its vertices by inducing $G_{t_i}[E_{t_i,\theta}]$ with increasing $\theta$. Upon inducing $G_{t_i}[E_{t_i,0.5}]$, comparing to $G_{t_i}[E_{t_i,0.6}]$, we can observe the core number increase of $v_0$ and $v_1$ from 1 to 2, which implies $\theta$-res$_3(0.6, G_{t_i}) = \theta$-res$_3(0.5, G_{t_i}) = 0.5$. So we add tree node to $\theta$-tree$_2$ because the core number is increased to 2. One node $X = \{v_0, v_1\}$ is constructed as $v_0$ and $v_1$ are connected. $\theta$-tree$_{2,t}$ has two nodes $Y_1 = \{v_2, v_3, v_4\}$ and $Y_2 = \{v_7, v_8, v_9\}$ from previous steps and the node that contains neighbors of $X$ is $Y_1$, whose root is itself. $X$ can thus be added as the parent of $Y_1$ since $Y_1.\theta > X.\theta$. By now, $\theta$-tree$_2$ contains three tree nodes. After inducing $G_{t_i}[E_{t_i,0.7}]$ we can construct $\theta$-tree$_{2,t}$ as Figure 3. $\theta$-tree$_{1,t}$ and $\theta$-tree$_{3,t}$ will also be obtained.

5.2 WCF-Index Maintenance

In general, dynamic networks might have subtle changes, i.e., a small percentage of edges and vertices change or update, in two consecutive timestamps. It is time-consuming to simply compute the $\theta$-threshold for all vertices and re-construct the index. Therefore, in this section, it is highly desirable to develop an index maintenance strategy and update the index using the small number of changed edges and vertices only.

To explore the relationships of vertices with regards to different core numbers, we are motivated by the work in [36] that proposed an incremental core number update method for an evolving graph. It supports to locate a small set of vertices whose core number will be affected by using the below two concepts. The other works [15, 26, 45] also follow the similar concepts of subcore and purecore in [36] to maintain the core numbers.

Definition 5.1 (subcore in [36]). Given a graph $G=(V, E)$ and a vertex $u \in V$, the subcore of $u$ denoted as $S_u$, is a set of vertices having the same core number as $u$ and connected with $u$ via a path, where each vertex on the path has the same core number as $u$.

Definition 5.2 (purecore in [36]). Given a graph $G=(V, E)$ and a vertex $u \in V$, the purecore of $u$ denoted as $P_u$, is a set of vertices where each vertex $w \in P_u$ satisfies:

1. Condition 1: the core number $\text{core}(w, G)$ of $w$ is equal to the core number $\text{core}(u, G)$ of $u$.
2. Condition 2: $w$ has a set $W$ of neighbors whose core numbers are no less than $\text{core}(w, G)$, and $|W|$ is larger than $\text{core}(u, G)$.
3. Condition 3: $w$ is connected to $u$ via a path, where each vertex on the path satisfies the conditions (1) and (2).

Specifically, given two vertices $u$ and $v$ in a graph $G=(V, E)$, and $\text{core}(u, G) \leq \text{core}(v, G)$, if an edge $(u, v)$ is removed from $G$, then only the vertices in the subcore set $S_u$ may have their core number decreased; if an edge $(u, v)$ is added to $G$, then only the vertices in the purecore set $P_u$ may have their core number increased. Thus, we can extend the rules to the weighted graph, in which the updates include edge insertion, edge deletion, and edge weight change.

Considering that $G_{t_m}$ is obtained by inserting an edge $(u, v)$ with weight $\theta^*$ to $G_{t_n} = (V, E_{t_n}, W_{t_n})$. For a vertex $w \in V$, if
$\theta$-thres$_k (w, G_{t_n}) = \theta''$ and $\theta'' \geq \theta'$, then $\theta$-thres$_k (w, G_{t_n})$ will remain unchanged.

**Property 5.1 (Insertion of an Edge).** Given a graph $G_{t_n}$ and $\theta$-tree$_{k,t_n}$ for each $k \in [1, k_{\text{max}}]$, and two vertices $u$ and $v$ such that $\theta$-thres$_k (u, G_{t_n}) \leq \theta$-thres$_k (v, G_{t_n})$, if an edge $(u, v)$ is inserted with weight $\theta'$, then only the vertices $\{w \in P_u | \theta$-thres$_k (w, G_{t_n}) < \theta'\}$ may have their $\theta$-threshold increased.

**Property 5.2 (Deletion of an Edge).** Given a graph $G_{t_n}$ and $\theta$-tree$_{k,t_n}$ for each $k \in [1, k_{\text{max}}]$, and two vertices $u$ and $v$ such that $\theta$-thres$_k (u, G_{t_n}) \leq \theta$-thres$_k (v, G_{t_n})$, if an edge $(u, v)$ is removed with weight $\theta''$, then only the vertices $\{w \in S_u | \theta$-thres$_k (w, G_{t_n}) \leq \theta''\}$ may have their $\theta$-threshold decreased.

The edge insertion and edge deletion can be treated as the update of edge weight.

**Property 5.3 (Update of Edge Weight).** Given a graph $G_{t_n}$ and $\theta$-tree$_{k,t_n}$ for each $k \in [1, k_{\text{max}}]$, two vertices $u$ and $v$ such that $\theta$-thres$_k (u, G_{t_n}) \leq \theta$-thres$_k (v, G_{t_n})$, we have (1) if the weight of edge $(u, v)$ increases from $\theta_1$ to $\theta_2$ ($\theta_1 < \theta_2$), then only the vertices $\{w \in P_u | \theta$-thres$_k (w, G_{t_n}) \in (\theta_1, \theta_2)\}$ may have their $\theta$-threshold increased; (2) if the weight of edge $(u, v)$ decreases from $\theta_2$ to $\theta_1$ ($\theta_1 < \theta_2$), then only the vertices $\{w \in S_u | \theta$-thres$_k (w, G_{t_n}) \in (\theta_1, \theta_2)\}$ may have their $\theta$-threshold decreased.

**Example 5.2.** Figure 5(a) shows an updated graph instance of $G'_t$, by adding an edge $(v_5, v_9)$ with weight 0.3 to $G_{t_1}$. We can identify $P_{v_5} = \{v_0, v_2, v_3, v_5, v_6, v_7, v_9\}$ in $G_{t_1}$. In addition, the $\theta$-threshold of $v_5$ or $v_9$ is less than 0.3 and other vertices’ $\theta$-threshold is no less than 0.3. According to Property 5.1, only $v_5$ and $v_9$ may have their $\theta$-threshold increased. After recalculating $\theta$-thres$_2 (v_5, G'_{t})$ and $\theta$-thres$_2 (v_9, G'_{t})$, we can update the tree index from Figure 3 to Figure 5(b).

### 5.3 WCF-Index Compression

Sometimes, the graph instances of some consecutive timestamps may be similar because the edge weight and graph structure change progressively over time. Besides that, one tree node usually contains multiple vertices as it gathers many connected vertices with the same threshold. It is likely to have much duplicate information across $\theta$-tree indices. Thus, we need to develop an index compression strategy in order to reduce the redundancy. The key idea is to utilize a virtual node to replace the tree node that contains multiple vertices and appears frequently. The actual vertices of the virtual nodes are stored in an auxiliary table.

#### 6 EXPERIMENT

We conduct extensive experiments to evaluate the performance of our proposed algorithms, including EEF-based Online CRC Search in Algorithm 1, WCF-Index based CRC Search in Algorithm 2 and $\theta$-Tree Construction in Algorithm 3, denoted as EEF-CRC, WCF-CRC and WCF-Construct, respectively. We implement a baseline method based on maximal spanning core (SpanCore) [14], which calculates all the k-core subgraphs with different k value in various time intervals. We additionally remove the edges and calculate the reliability score of each candidate subgraph to get the optimal results. We compare the effectiveness of our proposed community model with PC [22] and SC [35]. We also evaluate the effectiveness of index maintenance (WCF-Maintain) and index compression. All the experiments are conducted on a Windows machine with an Intel i9-10900F CPU @ 2.80GHz and 32.0 GB DDR4-RAM.

**6.1 Experimental Setup**

Datasets. We conduct the experiments on eight real-world dynamic network datasets collected from SNAP\(^1\) and Network Data Repository\(^2\). In BitcoinAlpha (BA) and BitcoinOTC (BO) datasets,

---

\(^1\)https://snap.stanford.edu/data/

\(^2\)https://networkrepository.com
the edge weight represents the rating between two users. In the remaining datasets, the edge weight is calculated from the interaction frequency. The edge weight of all the datasets is normalized to [0, 1] by min-max normalization. The statistics of the dataset are shown in Table 1. The number of vertices and edges are denoted as $|V|$ and $|E|$, respectively. For each dataset, we first sort the edges by chronological order, and then divide them into $|T|$ partitions, i.e., $|E|/|T|$ edges, where $|T|$ is the target number of graph instances. It guarantees each graph instance contains meaningful $k$-core components. For example, the largest dataset StackOverflow (SOF) is divided into $|T| = 100$ snapshots and the medium-sized datasets, e.g., TechAsTopology (TAT), Retweet, etc. are divided into $|T| = 30$ instances. We denote $|V|$, $|E|$, density and $k_{\text{max}}$ as the average of vertex numbers, edge numbers, density, and the largest core numbers of the $|T|$ graph instances, respectively.

Parameters. Table 2 shows the detailed setting of the parameters used in the experiments. To better fit the model and cover more meaningful situations, we vary the query parameter $k$ as 20%, 40%, 60%, 80% of $k_{\text{max}}$ for each dataset with the default value 40%. The threshold value varies from 0.0 to 0.8 with the default value of 0.4. The length of the query time interval was specified as 4, 8, 12, 16, 20 with the default value of 12. Their default values are marked in bold font. We also vary the $\alpha$ parameter from 0 to 6 to show its effect on the returned community. We sample 100 query vertices whose core numbers are uniformly distributed in $[1, k_{\text{max}}]$ for each dataset and report their average running time as the time cost. The average core number of the 100 query vertices is shown in Table 1 as $\bar{k}_{\text{query}}$. In general, $\bar{k}_{\text{query}}$ is around 50% of $k_{\text{max}}$, which reflects the common scenario of query vertex.

6.2 Evaluation of Query Efficiency

In this section, we present the performance of SpanCore, EEF-CRC, and WCF-CRC under the default parameter settings. Figure 7 demonstrates the time cost when we run the three algorithms over eight datasets. Both of the proposed algorithms outperform the baseline algorithm SpanCore. EEF-CRC is slightly faster than SpanCore, as they both need to determine the core number of the vertices but EEF-CRC only requires the local information of the query vertex.

WCF-CRC runs much faster than EEF-CRC. For instance, WCF-CRC reduces the time cost of EEF-CRC by about 89 times.

To show the impact of each parameter, we also evaluate the efficiency of the proposed algorithms by varying the values of parameters $k$, $\theta$, and $t$, respectively. We utilize two representative datasets Reddit and SOF to demonstrate the experimental results.

Varying $k$. Figure 8 shows the average time cost of our proposed algorithms when $k$ varies from 20% to 80% of the corresponding $k_{\text{max}}$ values. WCF-CRC is significantly more efficient than the other two algorithms, and SpanCore consumes the most time in all settings. For instance, SpanCore takes 2.46s, EEF-CRC takes 1.76s, while WCF-CRC only needs 0.03s to complete the query processing in Reddit dataset where $k$ is 40% of the average large core number (i.e., $k = 5$). With the increase of $k$, all the algorithms consume decreasing time. But SpanCore and EEF-CRC are less sensitive to $k$ than WCF-CRC because they need to scan all the edges and compute the core numbers of the vertices induced by the edges, while WCF-CRC can directly retrieve the core numbers using WCF-Index.

Varying threshold $\theta$. Figure 9 shows the average time cost of our proposed algorithms when the threshold $\theta$ varies from 0 to 0.8. WCF-CRC outperforms SpanCore and EEF-CRC significantly. For instance, in Reddit, when $\theta = 0.2$, SpanCore takes 3.33s, EEF-CRC takes 2.41s while WCF-CRC takes 0.12s. WCF-CRC is faster than the other two by more than 20 times. When $\theta$ is given as 0.8, the three algorithms take 1.62s, 1.17s and 0.002s, respectively, i.e.,
WCF-CRC can reduce the time cost of EEF-CRC by about 500 times. With the increase of $\theta$, the speedup trend of WCF-CRC becomes significant because there are small number of tree nodes in WCF-Index when $\theta$ is set as a large value, i.e., more vertices can be pruned. Similarly, SpanCore and EEF-CRC also consume less time because the significant number edges can be pruned with the higher theta threshold. However, scanning all the edges in the locally connected subgraph for each query vertex is inevitable.

Varying time span $|T_Q|$, Figure 10 shows the average time cost of the proposed algorithms when the time span $|T_Q|$ varies from 4 to 20. All the algorithms consume higher time cost when the query time span $|T_Q|$ increases. For instance, in Reddit dataset, SpanCore takes 0.96s, 1.91s, 2.98s, 4.24s, and 5.44s. EEF-CRC takes 0.63s, 1.35s, 2.04s, 2.73s and 3.55s, respectively. But, WCF-CRC only takes 0.0012s, 0.02s, 0.027s, 0.029s and 0.037s at the same settings. The growth of runtime is consistent with the time complexity in Section 3 and Section 4.2.

Efficiency of Upper Bound. Figure 11 shows the efficiency of the proposed algorithms with (i.e., EEF-UB, WCF-UB) and without (i.e., EEF-Base, WCF-Base) using the upper bound pruning strategy on the Reddit dataset with varying $|T_Q|$. We can find that the pruning capability of the upper bound can be accelerated with the increase of the query time interval. For instance, EEF-UB is faster than EEF-Base by 5% when $|T_Q| = 4$, but the acceleration can achieve by 10% when $|T_Q| = 20$. Compared to WCF-Base, the improvement of WCF-UB is not significant. Since the nature of the upper bound is to estimate the maximum reliability score of the community candidates for each potential time interval, there are more chances to prune more intermediate community candidates when query time interval is large.

6.3 Evaluation of Index Construction and Maintenance

In Figure 12, we report the runtime of WCF-Construct for all the graph instances of eight datasets. Generally, it takes around 10s for datasets with small number of vertices like BA, BO, and Email. For the largest dataset like SOF, it needs 25h to complete the index construction.

We also evaluate the effectiveness of the index maintenance method. Taking the first graph instance of Reddit as the base, we randomly sample 100, 200, 300, 500, 1000 edges and mix the operation of edge insertion, deletion, and weight update to generate a synthetic instance. Figure 13 shows the time cost of WCF-Construct and WCF-Maintain on the synthetic graph, where the speed up of WCF-Maintain is significant. For instance, reconstructing the index takes 8s, but WCF-Maintain only takes 4s when 1000 edges are updated.

6.4 Evaluation of Scalability

We evaluate the scalability of proposed algorithms including EEF-CRC, WCF-CRC and WCF-Construct by using five graph instances from two datasets Reddit and HepPh. For each dataset, we generate four new datasets with different sizes by randomly sampling 20%, 40%, 60%, 80% edges from the dataset, respectively. The dataset itself is considered with the 100% data size. Figure 14 shows the time cost of WCF-CRC and EEF-CRC on the size-varying datasets. With the increase of the data size, we can find that the running time of WCF-CRC and EEF-CRC grow in a gentle trend, which implies that both algorithms are easily applicable to large-scale networks.

Furthermore, we also show the scalability of index construction in Figure 15. From this, we can observe a linear increasing trend of the construction time. For instance, the index constructing time of Reddit is 4.7s, 12.8s, 22.3s, 34.0s and 47.9s when the size of the dataset increases as 20%, 40%, 60%, 80% and 100%, respectively.

6.5 Evaluation of Index Size with Compression

We show the WCF-Index size of the eight datasets in Table 3. For each dataset, we take ten graph instances. The largest dataset Stack-Overflow takes 874,931kb and the smallest dataset BitcoinAlpha takes 333kb. The compressed size is the sum of the compressed
### Table 3: Index Size & Compression (kb)

<table>
<thead>
<tr>
<th></th>
<th>BitcoinAlpha</th>
<th>BitcoinOtc</th>
<th>Retweet</th>
<th>TAT</th>
<th>Email</th>
<th>Reddit</th>
<th>HepPh</th>
<th>StackOverflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data</td>
<td>1,108</td>
<td>1,648</td>
<td>4,935</td>
<td>13,619</td>
<td>5,324</td>
<td>37,376</td>
<td>263,168</td>
<td>1,153,024</td>
</tr>
<tr>
<td>Original Index</td>
<td>333</td>
<td>645</td>
<td>1,569</td>
<td>7,047</td>
<td>1,139</td>
<td>45,216</td>
<td>252,561</td>
<td>874,931</td>
</tr>
<tr>
<td>Compressed</td>
<td>328</td>
<td>616</td>
<td>1,418</td>
<td>6,205</td>
<td>1,135</td>
<td>39,924</td>
<td>158,734</td>
<td>835,721</td>
</tr>
<tr>
<td>Auxiliary Table</td>
<td>3</td>
<td>11</td>
<td>120</td>
<td>430</td>
<td>3</td>
<td>834</td>
<td>11,406</td>
<td>5,361</td>
</tr>
</tbody>
</table>

### 6.6 Evaluation of Query Effectiveness

To show the effectiveness of finding communities in dynamic or temporal networks, we compare our (θ, k)-core reliable community (CRC) with the Persistent Community (PC) and Stable Community (SC) proposed by Li [22] and Qin [35], respectively. To do this, we select five graph instances of Reddit dataset and return the largest community C obtained by SC, PC, and our CRC. To show the quality of returned communities, we utilize three community quality metrics:

- **Average Snapshot Density (ASD)** measures how dense is the community and captures the intuition that a good community should be closely connected inside. The larger is the density, the closer the community is connected. Average snapshot density is calculated as the average density of the community in each snapshot: \( ASD = \frac{\sum_{t=1}^{T} \text{density}(G_t[C])}{t} \).

- **Average Snapshot Core (ASCORE)** captures the degree information of vertices and evaluates the closeness of the community. The larger is the core number, the more interactions each vertex will keep with others in the community. ASCore calculates the average value of the average core number of each vertex in each snapshot: \( ASCore = \frac{\sum_{t=1}^{T} \text{core}(s, G_t[C])/|V|}{t} \).

- **Average Snapshot Conductance (ASCCond)** measures how “well-knit” the graph is. The higher is the conductance, the easier the community can communicate with the vertices outside the community. In the local community detection task, the smaller conductance is desired as it implies the community is tightly self-capsulated. Here, the average snapshot conductance is calculated as the average conductance of the community in each snapshot: \( ASD = \frac{\sum_{t=1}^{T} \text{conductance}(G_t[C])}{t} \).

Table 4 shows the experimental results of evaluating the community quality on the Reddit dataset. The community is obtained with the same structural cohesiveness constraint (core number or number of neighbors equals 8). We vary the duration or frequency of the community (\( r \) in SC and PC, \( d \) in CRC), denoted by \( t \) in Table 4, to compare the community quality with different temporal features. The size of each community is also provided as the average snapshot size ASD in Table 4. It can be observed that CRC performs best in all the measurements. For example, when \( t = 4 \), CRC finds a community with the highest ASD of 0.39, the highest ASCore of 10.4 and the lowest ASCCond of 0.83. When \( t = 4 \) and 5, PC and CRC have similar ASCCond score. PC generally finds the largest community at the cost of lower density and cohesiveness. CRC outperforms SC with the larger community size and the closer connection.

Community with varying \( r \). Figure 16 shows an example of obtained reliable community by querying the vertex funny in the Reddit dataset where \( x \) varies from 0 to 6. With the increase of \( a \), the duration of the optimal CRC increases and the community size decreases. The progressive change of community duration shows that parameter \( x \) is able to smoothy adjust the balance between community size and duration. Figure 17 shows the trending change of the selected three quality metrics ASD, ASCore and ASCCond when \( x \) increases. All the scores increase significant when \( x \) varies from 0 to 2. After that, their trends become steady relatively.

### 7 RELATED WORKS

Local Community Search in Static and Time-varying Networks. Local community search has been studied in many existing
works. In static networks, existing methods can be classified into two categories. The first method is based on random walk, which aims to assign scores to the vertex from the query vertex and identify the local community based on the scores. Wu et al. [43] used a single random walker, and Bian et al. [5] introduced multiple walkers to assign vertex scores based on the hitting probability. Bian [6] further proposed memory based multiple walker that records the entire visiting history and supports multiple local communities w.r.t. different query vertices simultaneously. Another method is based on capturing cohesiveness structures [17] such as k-truss [2, 16, 27], k-core [3, 11, 21] and k-clique [10, 39]. To deal with the changes of network data over time, Takañoffa et al. [41] explored local community mining in the dynamic social network by extending L-metric [8] to an incremental version. Luo et al. [31] divided the formation of the local community into three stages and designed different dynamical membership functions to construct the local community with better cohesiveness. DiTuri et al. [12] proposed PHASR method to detect local community in the dynamic networks, and Papadopoulos [33] expanded PHASR to fit the distributed processing standard of Apache Spark engine.

**K-core Community Search.** In this work, we consider the community size (number of vertices) and duration (continuity of the vertex engagement) as two important factors, so we use k-core model which is defined on vertex attributes rather than other classic models like k-truss that defined on edge attributes. K-core was firstly introduced by Seidman et al. [37] and becomes one of the most widely used measurements of graph cohesiveness. Seidman et al. [40] developed a greedy algorithm to discover the dense subgraph by iteratively removing the vertices with the minimum degrees. Batagelj et al. [4] proposed a linear core decomposition algorithm to compute the core number of all the vertices. Cui et al. [11] developed a local community search algorithm that starts from a query vertex q and spans iteratively to include the local optimal vertex into the community. Barbieri et al [3] proposed an index structure based on the nested feature of the core number and improved the community search significantly. Following [3], Fang et al. [13] improved the efficiency of index construction. To include network dynamics, Li et al. [23] devised a core maintenance algorithm in large dynamic networks. Wu [42] proposed distributed algorithms based on the block-centric model to compute cores in the temporal graph. Galimberti et al. [14] identified all the maximal k-core with various time span and k value. Based on [14], Hung and Tseng [18] extended to the maximal lasting k-core. However, these works mainly focused on core decomposition in dynamic graphs and didn’t consider the edge weight and query vertex like this work.

**Cohesive Subgraph Mining in Dynamic Networks.** Our work also relates to dense subgraph mining, which aims to identify the densely connected vertices in temporal or dynamic networks. Abdelhamid et al. [1] proposed an incremental approach called InCGM+ to extend the traditional Frequent Subgraph Mining (FSM) method into dynamic networks by only updating the “fringe” patterns. Ma et al. [32] proposed a fast computation algorithm to identify dense subgraphs in temporal graphs where edges have positive or negative weights. However, their method relies on the “evolving convergence phenomenon” that assumes weights of all edges are increasing or decreasing in the same direction which is too strict for the real world. And they didn’t consider the community continuity. Semertzidis and Pitoura [38] proposed the problem of querying the frequent subgraph patterns in the directed dynamic networks and returns the top-k durable matches. Liu et al. [28] considered the duration of the found dense subgraphs using an expectation-maximization method. They are unable to deal with edge weight, which is limited in real applications.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we first discussed the necessity of reliable local community in dynamic networks and proposed the novel most reliable community search problem. Then, we developed an online (θ, k)-core reliable community search approach by pruning the ineligible edges based on the given threshold and their lasting times. After that, we designed an effective WCF-Index to maintain the vertex candidates of (θ, k)-core subgraphs, and developed an efficient index-based dynamic programming approach. Finally, the empirical evaluations on a variety of datasets and parameter settings illustrate the efficiency and effectiveness of the proposed approaches. In this work, we mainly focus on single query vertex situation. However, our proposed algorithms can be extended to support querying a set of vertices. For EEF algorithm, we can start the edge search from all query nodes simultaneously, and maintain a visited edge set to avoid repeat traverse. Then we can follow the EEF algorithm to return the valid local k-core. For WCF-Index query algorithm, multiple query nodes can be easily supported by filtering the intermediate result that does not contain all the query nodes. To explore the significance of multiple query nodes in community discovery, one potential research direction is to investigate the local engagement of the query nodes and identify the meaningful communities.

ACKNOWLEDGMENTS

This work was mainly supported by the ARC Linkage Project under Grant No. LP180100750.
REFERENCES


