SDREGION: Fast Spotting of Changing Communities in Biological Networks

Serene W.H. Wong
University Health Network
Toronto, Ontario, Canada
swong@cse.yorku.ca

Chiara Pastrello
University Health Network
Toronto, Ontario, Canada
chiara.pastre@gmail.com

Max Kotlyar
University Health Network
Toronto, Ontario, Canada
maxk.email@gmail.com

Christos Faloutsos
Department of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
christos@cs.cmu.edu

Igor Jurisica
University Health Network
Medical Biophysics and Computer Science
University of Toronto
Toronto, Ontario, Canada
juris@ai.utoronto.ca

ABSTRACT

Given a large, dynamic graph, how can we trace the activities of groups of vertices over time? Given a dynamic biological graph modeling a given disease progression, which genes interact closely at the early stage of the disease, and their interactions are being disrupted in the latter stage of the disease? Which genes interact sparsely at the early stage of the disease, and their interactions increase as the disease progresses? Knowing the answers to these questions is important as they give insights to the underlying molecular mechanism to disease progression, and potential treatments that target these mechanisms can be developed.

There are three main contributions to this paper. First, we designed a novel algorithm, *SDREGION*, that identifies subgraphs that decrease or increase in density monotonically over time, referred to as d-regions or i-regions, respectively. We introduced the objective function, $\Delta density$, for identifying d-(i-)regions. Second, *SDREGION* is a generic algorithm, applicable across several real datasets. In this manuscript, we showed its effectiveness, and made observations in the modeling of the progression of lung cancer. In particular, we observed that *SDREGION* identified d-(i-)regions that capture mechanisms that align with literature. Importantly, findings that were identified but were not retrospectively validated by literature may provide novel mechanisms in tumor progression that will guide future biological experiments. Third, *SDREGION* is scalable with a time complexity of $O(mlogn + nlogn)$ where m is the number of edges, and n is the number of vertices in a given dynamic graph.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219854

KEYWORDS

dynamic graphs, decreasing density subgraph detection, increasing density subgraph detection, temporal data

ACM Reference Format:
Serene W.H. Wong, Chiara Pastrello, Max Kotlyar, Christos Faloutsos, and Igor Jurisica. 2018. SDREGION: Fast Spotting of Changing Communities in Biological Networks. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219854

1 INTRODUCTION

Given data with time stamps, entities and the relationship between them, how would you trace the behavior of a group of entities over time? Given a dynamic biological graph modeling a given disease progression, which genes should new treatments be targeting at? What are the underlying molecular mechanisms that contribute to disease progression? Can the connections between a group of genes give insights to these questions? If we can identify subgraphs that interact closely at the early stage of the disease, and their interactions decrease as the disease progresses, then it may suggest that mechanisms needed for more normal biological functions are being disrupted by the advances of the disease. Conversely, if we can identify subgraphs that interact sparsely at the early stage of the disease, and their interactions increase as the disease progresses, then it may suggest that the disease is causing interaction between genes that are not present in normal conditions in order to involve processes needed by the disease. Potential novel treatments can then be developed to target these mechanisms. In this paper, we proposed a novel algorithm, *SDREGION* (sparsifying or densifying region), that identifies subgraphs such that their density monotonically decreases or increases over time, referred to as d-regions or i-regions respectively. Fig. 1 depicts a d-region identified by *SDREGION*, and Fig. 2 depicts an i-region detected by *SDREGION*. Both of them are biologically meaningful, and will be discussed in detail in Section 5. In fact, an effective and scalable algorithm that focuses on this problem would be useful not only in the biomedical

domain, but in many other domains as well. Large, dynamic complex graphs are used to model data with time information in many fields. Some examples of such complex graphs include biological networks, social networks, co-authorship networks and phone-call networks.

There are three main contributions to this paper:

(1) **Novel Algorithm** - We designed *SDREGION*, a scalable and effective algorithm to identify subgraphs that decrease or increase in density monotonically over time.
(2) **Effectiveness** - *SDREGION* is effective in real data, and observations were made. While it can be applied in many domains, we showed its effectiveness on data that models the tumor progression of non-small cell lung cancer (NSCLC).
(3) **Scalable** - *SDREGION* is scalable with a time complexity of $O(mlogn + nlogn)$ where m is the number of edges, and n is the number of vertices in a given dynamic graph.

Reproducibility: Our code and the datasets that were used are available at http://www.cs.utoronto.ca/~juris/data/KDD2018/.

Our proposed algorithm, *SDREGION* was carefully designed such that it overcomes the problem of the naive approach. The naive approach uses any algorithm that finds dense subgraphs designed for static graphs, then checks the vertices in these subgraphs across time points to verify whether the number of edges are decreasing among them. This naive approach suffers from the overshadowing of dense subgraphs. Furthermore, we introduced the objective function, $\Delta density$, that is used in *SDREGION*.

While *SDREGION* is generic and can be used in many domains, we showed its effectiveness in the progression of tumor in NSCLC using 4 NSCLC gene expression datasets. The effectiveness of *SDREGION* was shown through gene signature (a set of genes such that its gene expression levels have a unique association with a specific biological condition) enrichment (over-representation) analysis and pathway (a group of genes working together to achieve a biological function) enrichment (over-representation) analysis. From gene signature enrichment analysis, all d-regions and i-regions were significantly enriched ($p < 0.05$) in biologically relevant gene signatures. Importantly, d-(i-)regions that were not enriched in gene signatures related to lung were still enriched in biologically relevant gene signatures. This means that the d-(i)regions identified at least one biologically relevant signature that has been described in another condition, and provides a novel hypothesis about its relevance in NSCLC. This observation may lead to possible novel lung cancer gene signatures. Moreover, pathway enrichment analysis on both the d-regions and i-regions showed that *SDREGION* identified subgraphs that represent tumor progression mechanisms. In particular, d-(i-)regions capture mechanisms related to RB1 related pathways, integrins, focal adhesions, ECM, neuronal system and neurotransmission that align with literature. Importantly, findings that were identified but were not highlighted in our observations may suggest potential novel mechanisms in tumor progression.

Recently, dynamic graphs have been explored with interest across multiple domains as they provide useful framework for modeling and analysis. Many studies have focused on finding dense blocks, subgraphs or communities in dynamic graphs, e.g., [23], [24], [25] and [11]. Other studies directed their attention to detecting changes in dynamic graphs. For example, identifying changes on the appearing, disappearing, splitting or merging communities over time, e.g., [12], [2]; detecting when a change occurs [25]; dynamic model for group evolution [26]; providing a descriptive summary of changes, e.g., [21]; and to detect unstable communities in network ensembles [17]. In this manuscript, we proposed an algorithm to identify subgraphs such that their density monotonically decreases or increases over time. This increasing or decreasing pattern enforced over all time points is important in various fields. For example in cancer biology, it is important to know how connections change between a set of genes from the initial stage all the way to the advance stage as this may provide insights to the progression of cancer. To the best of our knowledge, no current method can identify subgraphs that decrease or increase in density monotonically over time; thus, our results provide the first baseline data for future comparisons.

Table 1 compares the features of *SDREGION* with other related work on dynamic graphs.

2 RELATED WORK

Many studies have focused on detecting dense blocks or subgraphs in dynamic graphs or tensors. For example, M-Zoom [23] and D-Cube [24] detect dense blocks in tensors. Epasto et al. [11] addressed the problem of finding densest subgraphs in dynamic graphs.

Community detection in dynamic networks has also been studied. For example, DiTursi et al. [7] found local temporal communities in dynamic graphs where communities have high interactions within and low interactions outside over a period of time. GraphScope [25] discovers community in time-evolving graphs, and determines the points of change in time.

Other studies have focused on detecting changes in dynamic graphs. TimeFall [12] uses Minimum Description Length (MDL) to find communities that evolve over time (e.g., appear, disappear, split, merge), and selects cut-points in time when there are abrupt structural changes in communities. HOCTracker [2] is a framework to detect evolutionary events in hierarchical and overlapping communities where evolutionary events include birth, death, growth, contract, merge, split of communities. Moreover, dynamic model for group evolution has also gained attention. COMENGO [26] is a dynamic model for group evolution that includes the joining and quitting mechanisms.

Some research groups concentrated on characterizing and summarizing large dynamic graphs. For example, TimeCrunch [21] describes the underlying behavior in dynamic graphs using concise summaries such as the number of cliques that persist throughout all time points, or the number of stars that appear periodically.

While many methods have been developed to identify frequent, dense subgraphs, Rahman et al. [17] proposed the region detection of graphs such that their variability is maximized. They introduced the definition of unstable communities. An unstable community has several subgraphs induced by it, and their frequency distributions are nearly uniform.

3 PRELIMINARIES AND PROBLEM DEFINITION

In this section, we provide problem definition and introduce notation used throughout the paper. Section 3.1 provides notations

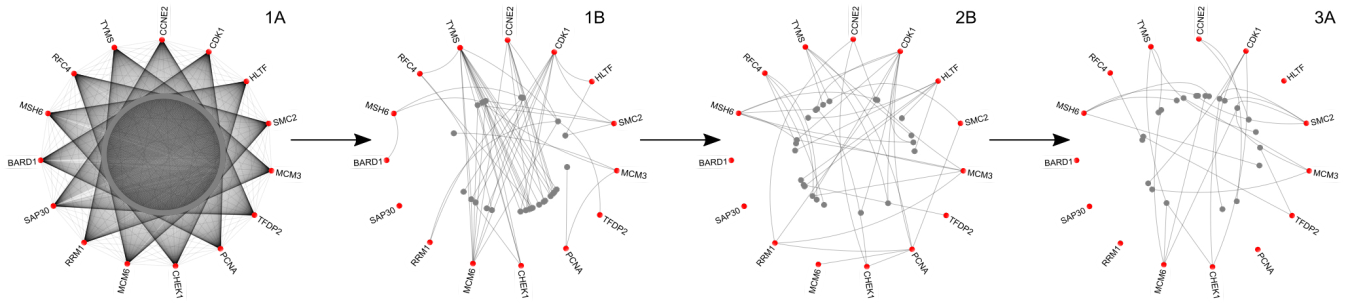

Figure 1: Result by *SDREGION*: d-region 0 in the *ChitaleMA1* dataset. Each graph represent a stage (1A, 1B, 2B, 3A), and stages increase progressively from left to right. Red vertices belong to Wikipathways "Retinoblastoma (RB) in cancer" pathway. The disruption of *RB1* related pathways is important to cancer progression showing that our results captured relevant cancer mechanisms.

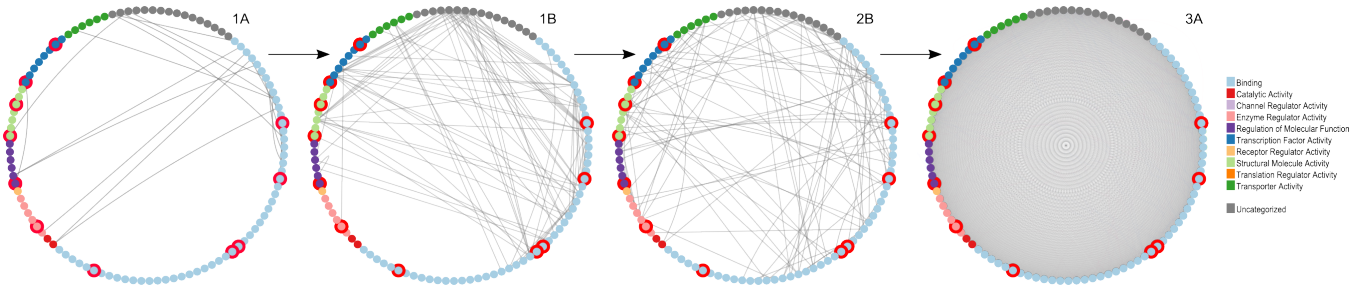

Figure 2: Result by *SDREGION*: i-region 0 in the *ChitaleMA1* dataset. Each graph represents a stage (1A, 1B, 2B, 3A) of lung cancer, and cancer progresses from left to right. Node colors are based on Gene Ontology. Highlighted in red are vertices belonging to a stem cell signature, and the activation of these genes have been observed in cancers; thus, this i-region is pointing to a known cancer initiating mechanism.

Table 1: *SDREGION* fulfills all specifications.

	SDREGION	M-Zoom [23]	D-Cube [24]	TimeCrunch [21]	Graphscope [25]
Pattern enforced across all time points	✓			✓	
Increasing density	✓				
Decreasing density	✓				
Scalable	✓	✓	✓	✓	✓
Effective on real data	✓	✓	✓	✓	✓

and definitions used in this paper. Sections 3.2, 3.3 introduce the objective function and the sparsifying condition used in *SDREGION*. Section 3.4 states our problem definition.

3.1 Notations and definitions

Let $G(V, E)$ denote a dynamic graph with T time steps, where V is the set of vertices, and E is the set of edges in G. $G = \bigcup_t G_t(V, E_t)$, where G_t, E_t corresponds to the graph and edge set for the t^{th} time step. Let $m = \sum_1^T (|E_t|)$ denote the number of edges in G, and $n = |V|$ denote the number of nodes in G. Let $deg_{G_t}(v)$ denote the degree of vertex v in G_t. Let $N_{G_t}(v)$ denote the set of neighbors of v in G_t. Let $density(G_t)$ denote the density for graph G_t where $density = |E_t|/|V|$.

3.2 Objective function

In order for *SDREGION* to return subgraphs that have their density monotonically decrease over time, we introduce an objective function, $\Delta density$. Let $\Delta density(G)$ denote the Δ density for G where

$$\Delta density = \sum_1^{T-1} (density(G_t) - density(G_{t+1})).$$

Lemma 3.1.

$$\Delta density = \sum_1^{T-1} (density(G_t) - density(G_{t+1})) = (|E_1| - |E_T|)/|V|$$

Proof. Expand the summation, and cancel the middle terms. □

3.3 Sparsifying condition

SDREGION is to return subgraphs that monotonically decrease in density, throughout all time points, not just the first and the last. Therefore, we introduce the sparsifying condition:

$$density(G'_t) \geq density(G'_{t+1}) \forall t \in [1, T-1]$$

To obtain subgraphs that monotonically increase in density throughout all time points, reverse the time sequence in the input dynamic graph, G.

3.4 Problem definition

Given a dynamic graph, G, and the number of subgraph, k, find k subgraphs that

(1) maximize the objective function, Δ density
(2) the sparsifying condition is satisfied.

These subgraphs are referred to as d-regions.

To obtain i-regions, reverse the time sequence in the input dynamic graph, G.

4 PROPOSED METHOD

4.1 Algorithm

Our proposed algorithm, *SDREGION* is a heuristic algorithm. Naive enumeration of all possible subgraphs will be combinatorial, therefore, we proposed a heuristic algorithm.

SDREGION begins with G, and greedily searches for a d-(i-) region. *SDREGION* removes a vertex at a time, greedily selecting a vertex according to Lemma 4.1. *SDREGION* then returns a graph configuration that has the maximum Δ density value, and satisfies the sparsifying condition.

Lemma 4.1. *The removal of $v \in V(G)$ with the minimum difference in degree in G_1 and G_T results in the highest Δ density in G' where $V(G') = V(G) \setminus \{v\}$.*

Proof. Recall that Δ density for G is $(|E_1| - |E_T|)/n$, and in undirected graphs, $\sum_{v \in V} deg_{G_t}(v) = 2|E_t|$. Thus, $|E_1| = \frac{1}{2}[deg_{G_1}(v_1) + ... + deg_{G_1}(v_i) + ... + deg_{G_1}(v_n)]$, and $|E_T| = \frac{1}{2}[deg_{G_T}(v_1) + ... + deg_{G_T}(v_i) + ... + deg_{G_T}(v_n)]$

Thus, the Δ density for G:

$$\frac{1}{2n}[\sum_{v \in V} deg_{G_1}(v_i) - deg_{G_T}(v_i)]$$

From the above equation, and that $\frac{1}{2(n-1)}$ will stay constant regardless of which v_i to remove, removing v_i such that $deg_{G_1}(v_i) - deg_{G_T}(v_i)$ has the minimum value will result in the highest Δ density for G'. □

Algorithm 1 gives a top level description of the algorithm. *SDREGION* begins with the entire graph G, finds one d-(i-)region in each iteration, removes the d-(i-)region from G and searches again for another d-(i-)region.

Note that for any given graph, a d-(i-)region may not exist. Furthermore, given the greedy nature of the algorithm, it is possible

Algorithm 1: *SDREGION*

Input: Dynamic graph:G, no. of d-(i-)region:k, [no. of restart:k'], [% of vertices to remain:p]
Output: $\leq k$ d-(i-)regions

```
1  ResultSet = ∅;
2  while No. of d-(i-)regions found are < k do
3      result ← FindARegion (G);
4      if No d-(i-)region is found ∧ restart < k' then
5          V(G) ← V(G) \ {v} where v ∈ V(G) is an arbitrary
             vertex;
6      else
7          ResultSet ← ResultSet ∪ {result} ;
8          V(G) ← V(G) \ U where
             |U| = (1 − p)|V(result)| ∧ v ∈ U are arbitrary
             vertices in V(result)
9  end
10 return ResultSet;
```

that no subgraph that satisfies the sparsifying condition is found. Therefore, lines 4 − 5 provide an option for restarts. Restarting allows for a different ordering for the removal of vertices, and a d-(i-)region may be found. The ordering of the removal of vertices is affected as for each vertex that is removed, its neighbors will be updated. In fact, different sets of neighbors from different time graphs will be updated. Thus, the calculation for which vertex to be removed first may be changed.

Since *SDREGION* can be applied in different domains, it is desirable for some applications that a vertex can be included in more than one d-(i-)region. Thus, line 8 provides an option to not remove all vertices of the identified d-(i-)region from G before the next d-(i-)region is searched.

Algorithm 2: FindARegion (G)

Input: Current graph G
Output: 0 or 1 d-(i-)region

```
1  Configuration = ∅;
2  initialize min-heap MH, hash H;
3  while |V(G)| ≠ ∅ do
4      v ← RemoveNode (G, MH, H);
5      V(G) ← V(G) \ {v};
6      Calculate Δ density for G;
7      if G satisfies density(G_t) ≥ density(G_{t+1}) ∀t ∈ [1, T − 1]
         then
8          Configuration ← Configuration ∪ {v} ;
9  end
10 return C ∈ Configuration with the max Δ density;
```

Algorithm 2 finds a single d-(i-)region. Again, note that it is not the case that all graphs will have a d-(i-)region, and that given the greedy nature of *SDREGION*, it is possible that no subgraph satisfies the sparsifying condition. H is a hash to store vertices that have been removed, and MH is a min-heap that stores the degree difference between G_1 and G_T for each vertex. *SDREGION* is

designed such that variables in lines 6, 7, 10 are updated, and computations in these lines are not calculated from scratch. Furthermore, *SDREGION* does not store the entire graph for each configuration, but the vertices that are removed, the maximum Δ density so far, and indices to keep track of configurations.

Algorithm 3: RemoveNode (G, MH, H)

Input: Current graph G, min-heap MH, hash H
Output: $v \in V(G)$ to be removed, according to Lemma 4.1

1 $v \leftarrow v \in V(G)$ with the min value in MH;
2 $V(G) \leftarrow V(G) \setminus \{v\}$;
3 $H(v) = 1$;
// update MH
4 **foreach** $t \in \{1, T\}$ **do**
5 **foreach** $u \in N_{G_t}(v)\, s.t.\, H(u) = 0$ **do**
6 update u in MH;
7 **end**
8 **end**
9 **return** v;

Algorithm 3 removes a vertex based on Lemma 4.1, and updates MH. In order for line 6 to be executed in $O(logn)$ (Recall that the decrease key operation in MH is $O(logn)$), the computation for the value to be updated has to be in constant time. We implemented a hash to compute the value to be updated in MH for constant time operation as the value to be updated is dependent on both $N_{G_1}(v)$ and $N_{G_T}(v)$.

4.2 Time complexity

In this section, we show that *SDREGION* has a time complexity of $O(mlogn + nlogn)$.

Lemma 4.2. *SDREGION has a time complexity of* $O(mlogn + nlogn)$.

Proof. MH is a min-heap implemented in a hash indexed minimum heap. Recall that both the decrease key operation and the delete minimum operation in a min-heap is $O(logn)$.

In FindARegion(G), the bottleneck is line 4. Line 2 in FindARegion(G) takes $O(n)$, lines 6, 7, 10 take constant time if variables are updated, and computations in these lines are not calculated from scratch. In RemoveNode(G, MH, H), line 1 is in $O(logn)$. Lines $1-3$ in RemoveNode(G, MH, H) execute n times as all vertices will be removed once, and lines $4-7$ in RemoveNode(G, MH, H) execute m times as every edge will be updated once. Therefore, the time complexity for FindARegion(G) is $O(mlogn + nlogn)$.

The bottleneck for the *SDREGION* algorithm is line 3, FindARegion(G). FindARegion(G) is to be executed $k + k'$ times, and $k + k'$ is a constant. Therefore, the time complexity of the proposed *SDREGION* algorithm is $O(mlogn + nlogn)$.

□

4.3 Scalability

SDREGION is carefully designed to be quasilinear time $O(mlogn)$, refer to Fig. 3. Starting off with a real dataset, *chitaleMA*1, used in this manuscript, random graphs were generated by randomly removing edges from it. We measured the time taken to return 3 d-regions with no restart and no % of vertices to remain. All experiments were performed on an machine with Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz and 256GB RAM. *SDREGION* was implemented in Java.

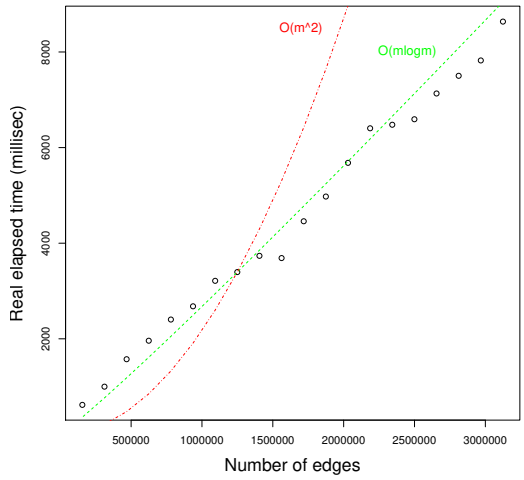

Figure 3: Runtime for *SDREGION* is in black circles, $O(mlogm)$ is in green, and $O(m^2)$ is in red.

5 TUMOR PROGRESSION BY SDREGION IN REAL DATA

In this Section, we demonstrate the effectiveness on the d-(i-)regions identified by our proposed algorithm *SDREGION*.

While *SDREGION* can be applied to many different domains, we applied it to gene expression data across disease stages to model lung cancer progression.

The input to *SDREGION* and the datasets used are discussed in Sections 5.1 and 5.2. Section 5.3 presents the results of the first three d-regions and i-regions with no restart and no % of vertices to remain from *SDREGION*. For these identified d-(i-)regions, we then used gene signature and pathway enrichment (over-representation) analyses to show that they are biologically meaningful and are also meaningful in terms of tumor progression. Importantly, potential novel understanding to tumor progression is identified. We explained the methods used for gene signature and pathway enrichment analyses in Sections 5.4 and 5.5, respectively, and presented the results in Sections 5.6 and 5.7.

5.1 Input to SDREGION

The input to *SDREGION* is a dynamic graph consisting of co-expression graphs generated by lung cancer gene expression data where edges represent pairs of genes that have high absolute correlation values. The input graph G has T time steps, corresponding to T stages of lung cancer that a dataset has, and G_t is the co-expression graph for stage t. We assume that the dynamic graph has already

Table 2: Co-expression graph size

Dataset	Stage	Nodes	Edges
ChitaleMA1	1A	12500	781188
ChitaleMA1	1B	12039	781188
ChitaleMA1	2B	12500	781188
ChitaleMA1	3A	12500	781188
ChitaleMA2	IA	12494	781188
ChitaleMA2	IB	12500	781188
ChitaleMA2	IIIA	12500	781188
Okayama	IA	17944	2031221
Okayama	IB	18574	2031221
Okayama	II	20009	2031221
Raponi	Ia	12491	781188
Raponi	Ib	12494	781188
Raponi	IIb	12500	781188
Raponi	IIIa	12500	781188

been constructed, and *SDREGION* takes it as input. Refer to Table 2 for more information on the input graphs to *SDREGION*.

5.2 Datasets

We used 4 NSCLC microarray gene expression datasets to demonstrate that *SDREGION* provides meaningful results to the progression of tumor. We obtained the datasets from the Gene Expression Omnibus database [9], and the Memorial Sloan-Kettering Cancer Center. The 4 datasets used [5], [16] and [19] are referred to as *ChitaleMA1*, *ChitaleMA2* (both at http://cbio.mskcc.org/public/lung_array_data/), *Okayama* (GSE31210) and *Raponi* (GSE4573), respectively in this paper. All 4 datasets are NSCLC datasets, and more specifically, *ChitaleMA1*, *ChitaleMA2*, *Okayama* are adenocarcinoma datasets, and *Raponi* is a squamous cell lung carcinoma dataset. Datasets were chosen based on the number of stages, the number of samples they have, and the platform used.

5.3 SDREGION results

Tables 3 and 4 summarize the results from *SDREGION*. For d-regions, for a given set of vertices, the number of edges decreases from $|E'_1|$ to $|E'_T|$. For example, for d-region 0 for *ChitaleMA1*, for the same group of 221 vertices, the number of edges decreases from 23,860 to 566.

For i-regions, for a given set of vertices, the number of edges increases from $|E'_1|$ to $|E'_T|$. For example, for i-region 0 for *ChitaleMA1*, for the same group of 106 vertices, the number of edges increases from 21 to 5,556. Notice that G'_T is almost a full clique as a full clique would have 5,565 edges. Figure 2 depicts this i-region.

Importantly, our proposed algorithm *SDREGION* ensures that the sparsifying condition holds, i.e., all time graphs are decreasing or increasing monotonically in density, not just the first time graph and the last time graph. Figures 4a and 4b depict that the sparsifying condition is satisfied.

Table 3: Output of *SDREGION*: d-regions

Dataset	d-region	n	$\|E'_1\|$	$\|E'_T\|$	Lung	All
ChitaleMA1	0	221	23860	566	13	196
ChitaleMA1	1	3341	329170	70759	6	152
ChitaleMA1	2	52	1261	42	5	11
ChitaleMA2	0	510	58480	6630	20	338
ChitaleMA2	1	393	34750	640	0	5
ChitaleMA2	2	763	72728	14949	23	516
Okayama	0	1873	506852	165274	36	615
Okayama	1	1393	208127	75357	5	97
Okayama	2	380	30747	8453	17	290
Raponi	0	711	153185	13055	0	2
Raponi	1	283	32347	784	2	25
Raponi	2	233	22848	408	6	78

Table 4: Output of *SDREGION*: i-regions

Dataset	i-region	n	$\|E'_1\|$	$\|E'_T\|$	Lung	All
ChitaleMA1	0	106	21	5556	0	1
ChitaleMA1	1	99	65	4797	2	28
ChitaleMA1	2	7739	178910	498883	0	3
ChitaleMA2	0	792	5209	94714	1	81
ChitaleMA2	1	240	395	22181	1	5
ChitaleMA2	2	345	1049	30213	0	9
Okayama	0	595	1714	127050	4	31
Okayama	1	1875	94518	283491	15	318
Okayama	2	690	3739	55606	0	16
Raponi	0	5133	117338	450155	1	57
Raponi	1	85	65	2522	0	2
Raponi	2	406	834	11525	4	60

5.4 Gene signature enrichment (over-representation) analysis

We used gene signature enrichment analysis as part of our evaluation of the meaningfulness of our results from *SDREGION*. A gene signature is a set of genes such that its gene expression levels have a unique association with a specific biological condition. GeneSigDB [6] is a gene signature database that contains manually curated gene signatures from PubMed indexed publications. GeneSigDB focuses on gene signatures of, for examples, cancer, lung disease, development and stem cell biology.

Two sets of gene signatures were downloaded in January 2018 from GeneSigDB Release 4. One set involves all human gene signatures in GeneSigDB, and the second set was the result of a search using the keyword "lung". In this paper, we refer to the first set as *ALL*, and the second set as *LUNG* gene signatures.

For each set of gene signatures, and for each identified d-(i-) region, gene signature enrichment was performed using the hypergeometric test. The set of background genes used were the genes

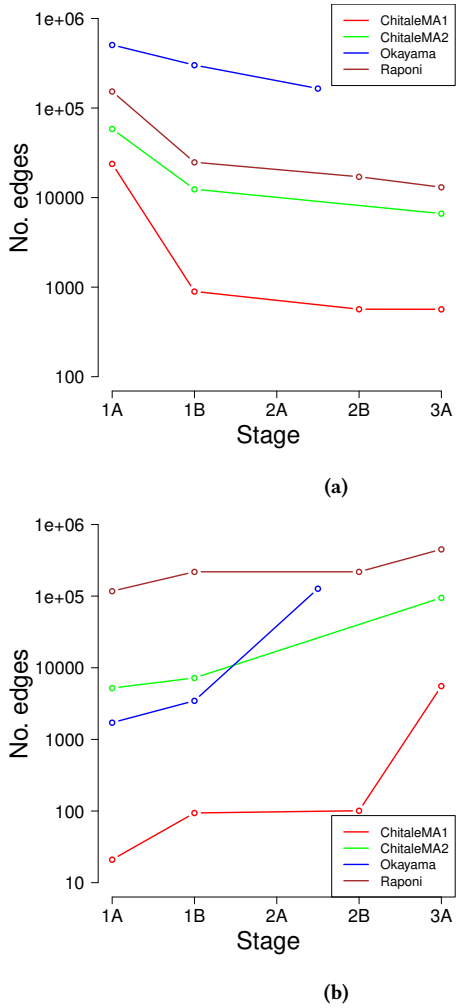

Figure 4: (a) Number of edges in a d-region identified by *SDREGION* across cancer stages.
(b) Number of edges in an i-region identified by *SDREGION*, across cancer stages.

in the chipset for each dataset. P values were adjusted using false discovery rate (FDR) for multiple testing.

5.5 Pathway enrichment (over-representation) analysis

We used pathway enrichment analysis as part of our evaluation to the meaningfulness of our results from *SDREGION*. A pathway is a group of genes working together to achieve a biological function.

Known pathways from pathDIP [18] v2.5 were used. For each pathway, and for each identified d-(i-)regions, pathway enrichment was performed using the hypergeometric test. As in gene signature enrichment, the set of background genes used were the genes in the chipset for each dataset. P values were adjusted using false discovery rate (FDR) for multiple testing.

5.6 Observations of SDREGION's results in gene signature analysis

Observation 1: All 24/24 *d-(i-)regions identified were significantly enriched in ALL showing that they are biologically meaningful.*

Observation 2: 17/24 *d-(i-)regions identified were significantly enriched in LUNG showing that they are meaningful to biological conditions related to lung. Recall that our datasets are lung cancer datasets.*

Observation 3: Our findings may point to possible novel lung gene signatures. The 7 *d-(i-)regions identified that were not significantly enriched in LUNG but were significantly enriched in ALL may be potential novel lung gene signatures.*

Results for gene signature enrichment analysis are in Tables 3 and 4. The Lung and All columns in these tables indicate the number of gene signatures that a given d-(i-)-region is enriched in (adjusted p-value < 0.05) in their respective categories, *LUNG* and *ALL*.

10/12 identified d-regions were enriched in *LUNG*, and 12/12 were enriched in *ALL*. 7/12 identified i-regions were enriched in *LUNG*, and again, 12/12 were enriched in *ALL*. This result shows that the identified d-(i-)-regions are biologically meaningful. Importantly, all the identified d-(i-)regions that were not enriched in *LUNG* were enriched in *ALL* indicating that they are potentially novel gene signatures in lung. For example, *Raponi* d-region 0 is enriched in two signatures, one related to human embryonic stem cells and one that identifies targets of CREB in myeloid leukemia cells. Interestingly, CREB regulates diverse cellular processes, among which are cell differentiation and proliferation, and its overexpression has been associated with negative prognosis in NSCLC patients never smoker[20]. *Raponi* i-region 1, on the other hand, is enriched in a signature of mismatch repair (MMR) deficiency and one related to serum cholesterol in heart diseases. MMR deficiency is frequently present in colorectal and ovarian cancer, but in a murine model K-ras mutant it has been demonstrated to accelerate lung tumorigenesis[8]. *ChitaleMA1* i-region 0 is enriched in only one signature, linked to human embryonic stem cells (cells whose gene expression pattern leads to their continuous self-renewal and pluripotency). Interestingly, activation of the human embryonic stem cell signature genes has been observed in several cancers, but, even more importantly, in stem cells of smokers airways - where the expression pattern is similar to that of lung cancer [22]. Recall that our datasets are lung cancer datasets. This result highlights once again that the d-(i-)regions identified with *SDREGION* are biologically meaningful and identify subnetworks characteristic of cancer development mechanism.

5.7 Observations of SDREGION's results in pathway analysis

Observation 1: Our findings on d-regions are meaningful to the progression of NSCLC. d-regions captured biological functions such that their disruption is important to cancer progression. Thus, there were fewer edges among these gene groups as cancer progresses.

Observation 2: Our findings on i-regions showed meaningful

results in tumor progression. They captured biological mechanisms related to tumor progression that are aligned with literature.

Pathway enrichment of d-regions identified 36 pathways shared among all three adenocarcinoma datasets (shown in Table 5). Interestingly, cell cycle and several of its phases and checkpoints are present in this list. It is well known that cell cycle's tight regulation is fundamental for normal cells' development, and that cancer cells need deregulation of the cell cycle and its checkpoints to grow - so much as to be known as one of the hallmarks of cancer [13]. Specifically, two key regulators of cell cycle known for their role in cancer are present in this list: aurora kinases and retinoblastoma (*RB1*), and the latter, in this analysis, was found enriched only among d-regions.

RB1 is the first identified tumor suppressor gene and the functional inactivation of pRb (phosphorylated RB protein) or its related pathways are events shared by nearly all human cancers. pRb role in tumorigenesis has first been linked to its ability to regulate the cell cycle by repression of proliferation-related genes. pRb regulation mechanisms have been well studied, and the best known happens through the interactions with the E2F family of transcription factors involved in DNA replication and repair, and G2/M progression [14]. pRb role in tumorigenesis has been subsequently linked to other important functions, such as regulation of immune functions and control of cell adhesion. The latter is particularly important for cancer metastasis (together with epithelial-mesenchymal transition (EMT)), and it has been shown that pRb deficiency not only correlates with cancer aggressiveness but also leads to faster metastasis (by definition, an event occurring at later cancer stages). In particular, pRb has been shown to stabilize adherens junctions, and its loss to disrupt these structures (leading to EMT). Among the cell adhesion proteins regulated by pRb are cadherins and integrins (as well as pathways involved in integrin-mediated cell-to-extracellular matrix (ECM) adhesion) [10].

Our findings on d-regions in *RB1* related pathways as well as integrins, focal adhesions and ECM capture the importance of the disruption of these pathways for cancer progression and metastasis, which are aligned with known molecular mechanism in literature. Furthermore, this result also suggests that, even when *RB1* is not frequently mutated, as it is the case of NSCLC [4], *RB1* and downstream pathways are impaired. Figure 1 shows how edges connecting proteins belonging to pRb pathway decrease with cancer stage in d-region 0 of the *ChitaleMA1* dataset. Networks were created using NAViGaTOR 3.0 (ophid.utoronto.ca/navigator/ [3]), and node colors are based on Gene Ontology, as per legend (http://www.geneontology.org/[1]).

Pathway enrichment on i-regions also shows meaningful results in tumor progression. The *Okayama* dataset returned the highest number (72) of enriched pathways, many of which are related to neuronal system and neurotransmission. Tumors can stimulate the formation of new nerve endings (neoneurogenesis) secreting neurogenic factors and axon guidance molecules. Moreover, neurotransmitters have been shown to play an important role in tumorigenesis, suppressing the immune response, affecting tumor vascularization and increasing cells migratory activity, and they have been related to tumor progression [15]. Finding such pathways enriched in our i-regions further confirms the relevance of our results.

Table 5: Pathways significantly enriched in d-regions, overlap of all 3 adenocarcinoma datasets

Database	Pathway	Lowest p-value
KEGG	**Cell cycle**	1.86E-13
KEGG	DNA replication	3.06E-08
PID	PLK1 signaling events	4.96E-12
PID	**E2F transcription factor network**	7.38E-11
PID	Aurora B signaling	3.34E-08
PID	FOXM1 transcription factor network	2.45E-07
Reactome	**Cell cycle**	2.24E-41
Reactome	**Cell cycle, mitotic**	1.12E-39
Reactome	Mitotic prometaphase	3.26E-22
Reactome	Resolution of sister chromatid cohesion	2.56E-21
Reactome	Mitotic metaphase and anaphase	1.29E-20
Reactome	m phase	2.97E-20
Reactome	Mitotic anaphase	4.89E-20
Reactome	Separation of sister chromatids	1.04E-19
Reactome	Rho GTPases activate formins	7.10E-16
Reactome	**Cell cycle checkpoints**	3.02E-14
Reactome	DNA replication	6.53E-14
Reactome	**g2/m checkpoints**	3.28E-12
Reactome	s phase	3.79E-12
Reactome	Synthesis of DNA	2.43E-11
Reactome	DNA strand elongation	3.87E-10
Reactome	Activation of ATR in response to replication stress	5.70E-10
Reactome	Unwinding of DNA	3.39E-08
Reactome	Activation of the pre-replicative complex	6.67E-08
Reactome	Phosphorylation of Emi1	2.13E-04
SPIKE	**g2/m phase of the cell cycle**	2.97E-20
SPIKE	DNA damage induced g1/s checkpoint	1.90E-12
SPIKE	g1/s phase of the cell cycle	1.93E-11
SPIKE	Repair of interstrand crosslinks	9.78E-07
SPIKE	**ATM signaling network**	2.48E-04
SPIKE	Mismatch repair	2.45E-03
Wikipathways	**Retinoblastoma (RB) in cancer**	2.38E-24
Wikipathways	**Cell cycle**	9.63E-17
Wikipathways	DNA replication	1.28E-12
Wikipathways	g1 to s cell cycle control	3.74E-07
Wikipathways	Gastric cancer network	5.38E-05

Bold are pathways related to *RB1*; p-values are the lowest p-values for the given pathway across d-(i-)regions.

6 CONCLUSIONS

There were three main contributions to this paper.

(1) **Novel Algorithm** - We designed *SDREGION*, a scalable and effective algorithm to identify subgraphs that decrease or increase in density monotonically over time.
(2) **Effectiveness** - *SDREGION* is effective in real data, and observations were made.
(3) **Scalable** - *SDREGION* is scalable with a time complexity of $O(mlogn + nlogn)$.

We designed a novel algorithm, *SDREGION*, that identifies subgraphs that decrease or increase in density monotonically over time. We introduced the objective function, $\Delta density$, for identifying d-(i-)regions. *SDREGION* is effective in real data. While *SDREGION* is generic and can be used in many domains, we showed its effectiveness in NSCLC. From gene signature enrichment analysis, all d-(i-)regions were significantly enriched in biologically relevant gene signatures, indicating the effectiveness of *SDREGION*. Importantly, d-(i-)regions identified may point to possible novel lung gene signatures. Moreover, pathway enrichment analysis on both the d-regions and i-regions showed that *SDREGION* identified subgraphs characterizing tumor progression. In particular, d-(i-)regions capture mechanisms that align with literature. Importantly, findings that were identified but were not highlighted in our observations may suggest potential novel mechanisms in the progression of tumor.

ACKNOWLEDGMENTS

This work was supported in part by the Ontario Research Fund No. GL2-01-030 to I.J.; the Natural Sciences and Engineering Research Council of Canada No. NSERC PDF-487917-2016 to S.W.H.W., NSERC 104105 to I.J.; the Canada Foundation for Innovation No. CFI 12301, 203373, 29272, 225404 to I.J.; the Canada Research Chair Program No. 203373, 225404 to I.J.. This material is based upon work supported by the National Science Foundation No. CNS-1314632, IIS-1408924 to C.F.. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. We are thankful to the authors of M-Zoom [23] for the code.

REFERENCES

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25.

[2] Sajid Yousuf Bhat and Muhammad Abulaish. 2015. HOCTracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks. *IEEE Transactions on Knowledge and Data engineering* 27, 4 (2015), 1019–1013.

[3] Kevin R Brown, David Otasek, Muhammad Ali, Michael J McGuffin, Wing Xie, Baiju Devani, Ian Lawson van Toch, and Igor Jurisica. 2009. NAViGaTOR: network analysis, visualization and graphing Toronto. *Bioinformatics* 25, 24 (2009), 3327–3329.

[4] Deborah L Burkhart and Julien Sage. 2008. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature Reviews Cancer* 8, 9 (2008), 671.

[5] D Chitale, Y Gong, B S Taylor, S Broderick, C Brennan, R Somwar, B Golas, L Wang, N Motoi, J Szoke, J M Reinersman, J Major, C Sander, V E Seshan, M F Zakowski, V Rusch, W Pao, W Gerald, and M Ladanyi. 2009. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. *Oncogene* 28, 31 (8 2009), 2773–83. https://doi.org/10.1038/onc.2009.135

[6] AedÃŋn C. Culhane, Markus S. SchrÃűder, Razvan Sultana, Shaita C. Picard, Enzo N. Martinelli, Caroline Kelly, Benjamin Haibe-Kains, Misha Kapushesky, Anne-Alyssa St Pierre, William Flahive, Kermshlise C. Picard, Daniel Gusenleitner, Gerald Papenhausen, Niall O'Connor, Mick Correll, and John Quackenbush. 2012. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research* 40, D1 (2012), D1060–D1066. https://doi.org/10.1093/nar/gkr901

[7] Daniel J DiTursi, Gaurav Ghosh, and Petko Bogdanov. 2017. Local Community Detection in Dynamic Networks. *arXiv preprint arXiv:1709.04033* (2017).

[8] Charlene M Downey and Frank R Jirik. 2015. DNA mismatch repair deficiency accelerates lung neoplasm development in K-rasLA1/+ mice: a brief report. *Cancer medicine* 4, 6 (2015), 897–902.

[9] R Edgar, M Domrachev, and A E Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 1 (1 2002), 207–210.

[10] Brienne E Engel, W Douglas Cress, and Pedro G Santiago-Cardona. 2015. The retinoblastoma protein: A master tumor suppressor acts as a link between cell cycle and cell adhesion. *Cell health and cytoskeleton* 7 (2015), 1.

[11] Alessandro Epasto, Silvio Lattanzi, and Mauro Sozio. 2015. Efficient densest subgraph computation in evolving graphs. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 300–310.

[12] Jure Ferlez, Christos Faloutsos, Jure Leskovec, Dunja Mladenic, and Marko Grobelnik. 2008. Monitoring network evolution using MDL. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 1328–1330.

[13] Yousef Ahmed Fouad and Carmen Aanei. 2017. Revisiting the hallmarks of cancer. *American journal of cancer research* 7, 5 (2017), 1016.

[14] Jack Hutcheson, Agnieszka K Witkiewicz, and Erik S Knudsen. 2015. The RB tumor suppressor at the intersection of proliferation and immunity: relevance to disease immune evasion and immunotherapy. *Cell Cycle* 14, 24 (2015), 3812–3819.

[15] Mario Mancino, Elisabet Ametller, Pedro Gascón, and Vanessa Almendro. 2011. The neuronal influence on tumor progression. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1816, 2 (2011), 105–118.

[16] H Okayama, T Kohno, Y Ishii, Y Shimada, K Shiraishi, R Iwakawa, K Furuta, K Tsuta, T Shibata, S Yamamoto, S Watanabe, H Sakamoto, K Kumamoto, S Takenoshita, N Gotoh, H Mizuno, A Sarai, S Kawano, R Yamaguchi, S Miyano, and J Yokota. 2012. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res* 72, 1 (1 2012), 100–111.

[17] Ahsanur Rahman, Steve TK Jan, Hyunju Kim, B Aditya Prakash, and TM Murali. 2016. Unstable Communities in Network Ensembles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 504–512.

[18] Sara Rahmati, Mark Abovsky, Chiara Pastrello, and Igor Jurisica. 2016. pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic acids research* 45, D1 (2016), D419–D426.

[19] M Raponi, Y Zhang, J Yu, G Chen, G Lee, J M Taylor, J Macdonald, D Thomas, C Moskaluk, Y Wang, and D G Beer. 2006. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.* 66, 15 (8 2006), 7466–7472.

[20] Hye-Sook Seo, Diane D Liu, B Nebiyou Bekele, Mi-Kyoung Kim, Katherine Pisters, Scott M Lippman, Ignacio I Wistuba, and Ja Seok Koo. 2008. Cyclic AMP response element-binding protein overexpression: a feature associated with negative prognosis in never smokers with non-small cell lung cancer. *Cancer research* 68, 15 (August 2008), 6065âĂŤ6073. https://doi.org/10.1158/0008-5472.can-07-5376

[21] Neil Shah, Danai Koutra, Tianmin Zou, Brian Gallagher, and Christos Faloutsos. 2015. Timecrunch: Interpretable dynamic graph summarization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1055–1064.

[22] Renat Shaykhiev, Rui Wang, Rachel K Zwick, Neil R Hackett, Roland Leung, Malcolm AS Moore, Camelia S Sima, Ion Wa Chao, Robert J Downey, Yael Strulovici-Barel, et al. 2013. Airway basal cells of healthy smokers express an embryonic stem cell signature relevant to lung cancer. *Stem cells* 31, 9 (2013), 1992–2002.

[23] Kijung Shin, Bryan Hooi, and Christos Faloutsos. 2016. M-zoom: Fast dense-block detection in tensors with quality guarantees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 264–280.

[24] Kijung Shin, Bryan Hooi, Jisu Kim, and Christos Faloutsos. 2017. D-cube: Dense-block detection in terabyte-scale tensors. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 681–689.

[25] Jimeng Sun, Christos Faloutsos, Spiros Papadimitriou, and Philip S Yu. 2007. Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 687–696.

[26] Tianyang Zhang, Peng Cui, Christos Faloutsos, Yunfei Lu, Hao Ye, Wenwu Zhu, and Shiqiang Yang. 2017. come N go: A Dynamic Model for Social Group Evolution. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 4 (2017), 41.