

Storage and Repair Bandwidth Tradeoff for Distributed Storage Systems with Clusters and Separate Nodes

Jingzhao Wang^{*}, Tinghan Wang[†], Yuan Luo[‡]

^{*}Email: wangzhe.90@sjtu.edu.cn [†]Email: wth19941018@sjtu.edu.cn [‡]Email: yuanluo@sjtu.edu.cn

Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai 200240, China

Abstract

The optimal tradeoff between node storage and repair bandwidth is an important issue for distributed storage systems (DSSs). As for realistic DSSs with clusters, when repairing a failed node, it is more efficient to download more data from intra-cluster nodes than from cross-cluster nodes. Therefore, it is meaningful to differentiate the repair bandwidth from intra-cluster and cross-cluster. For cluster DSSs the tradeoff has been considered with special repair assumptions where all the alive nodes are utilized to repair a failed node. In this paper, we investigate the optimal tradeoff for cluster DSSs under more general storage/repair parameters. Furthermore, a regenerating code construction strategy achieving the points in the optimal tradeoff curve is proposed for cluster DSSs with specific parameters as a numerical example. Moreover, the influence of separate nodes for the tradeoff is also considered for DSSs with clusters and separated nodes.

I. INTRODUCTION

As data center storage expands at scale, storage node failures are more prevalent [1], where distributed storage systems (DSSs) with erasure coding are widely utilized to ensure data reliability [2], [3], [4]. When a storage node in DSSs has failed, to recover the failed node, a new node will download data from others which are called helper nodes. The amount of data to download is called the repair bandwidth. In [5], the tradeoff between node storage and bandwidth to repair one node is investigated for homogeneous DSSs [6] where all the nodes (hard disks or other storage devices) have the same parameters (storage per node, repair bandwidth, etc.). Meanwhile, regenerating codes are proposed based on the tradeoff to reduce the repair bandwidth of DSSs.

Contrast to homogeneous DSSs, in heterogeneous DSSs [6], [7], nodes can have different storage and repair bandwidths. In [8], the communication cost among nodes is taken into consideration, where the storage of each node is equal, but the repair bandwidth varies based on the location of failed nodes. In realistic storage systems, nodes in the same cluster (rack) may be connected to each other with cheaper and faster networks (i.e. local area networks) [9], where downloading data from each other may be faster and cheaper. For the sake of reducing

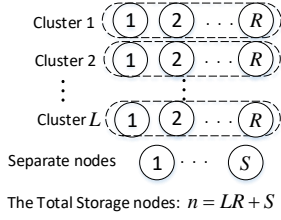


Fig. 1: System model for CSN-DSS

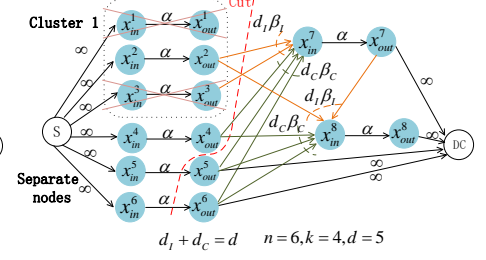
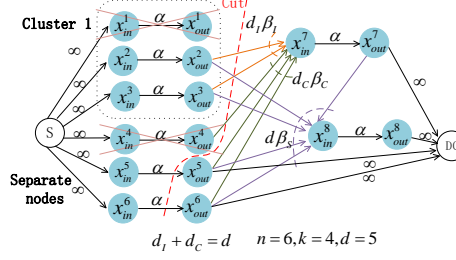


Fig. 2: The IFGs of one cluster and separate nodes distributed storage system

communication cost, it is more efficient to download more data from intra-cluster nodes and less from cross-cluster nodes, rather than downloading the same amount of data from others.

This paper investigates a type of heterogeneous DSSs with clusters and separate nodes, where the tradeoff between node storage and repair bandwidth is characterized on more flexible parameter settings. The tradeoff for DSSs with only two clusters (racks) is analysed in [10]. In [11], the authors consider cluster DSSs with one relay node in each cluster. The relay node collects data in its cluster and transmits to nodes in other clusters. In [12], the properties of DSS with multiple clusters are considered under a specific assumption that the new node will download data from all the other nodes when one node has failed. In current paper, the optimal tradeoff for cluster DSSs is investigated under more general settings that the new node does not need to download data from all the other nodes. The traditional homogeneous DSS and model in [10] and [12] can be obtained by specializing parameters of our general model. On the other hand, the tradeoff for DSSs with clusters and separate nodes is also analysed. Moreover, a regenerating code construction strategy is investigated for cluster DSSs with specific parameters.

The rest of this paper is organized as follows. The model of DSS with clusters and separate node (CSN-DSS) is introduced and the problem is formulated in Section II. In Section III, the properties of cluster DSSs are analysed in two parts, which are proved in Theorem 1 and Theorem 2, respectively. As a general case, the DSS with clusters and one separate node is analysed in Theorem 3. Afterward, the tradeoff between node storage and repair bandwidth is characterized. Some numerical results are illustrated in Section IV, where a regenerating code construction strategy is investigated for cluster DSSs. Finally, Section V presents the conclusion and future work.

II. PRELIMINARIES AND PROBLEM FORMULATION

Subsection II-A defines main parameters of CSN-DSSs. As an efficient tool to analyse DSSs, the information flow graph is introduced in Subsection II-B. The problem investigated in this paper is formulated in Subsection II-C.

A. Distributed Storage System with Cluster and Separate Nodes (CSN-DSS)

CSN-DSS Model: The cluster and separate nodes distributed storage system, illustrated in Figure 1, consists of S separate nodes and L clusters each with R nodes. The total number of the storage nodes is $n = LR + S$. A data

file of size \mathcal{M} symbols is divided into k fragments, each of size $\alpha = \mathcal{M}/k$ symbols. The k fragments are encoded into n fragments of size α and stored at n nodes. Any k encoded fragments out of n suffice to recover the original data file, which is called the (n, k) MDS¹ property.

When repairing a failed node, a newcomer locating in the same cluster will download data from other alive nodes and repair the failed one. If the failed node is a separate node, the newcomer is still a separate node.

If a **node in cluster** has failed, the newcomer downloads β_I symbols each from d_I intra-cluster nodes and β_C symbols each from d_C cross-cluster nodes (including nodes from other clusters and separate nodes). Let

$$d \triangleq d_I + d_C$$

and $d \geq k$ based on the (n, k) MDS property [13].

In realistic distributed storage systems, the storage servers may connect to each other with local area networks or external networks. The communication between servers in the same local area network is cheaper than servers in different local area networks and connected with external networks. The servers connected in the same local area network can be seen as being in the same cluster. The separate nodes connect to cluster nodes by external networks. In order to reduce the bandwidth cost, it is better to download more data from nodes in the same cluster and less data from outer nodes, namely, $\beta_I \geq \beta_C$. On the other hand, the intra-cluster nodes are used preferentially in general case, when repairing failed nodes. Therefore, it is reasonable to assume that all the intra-cluster alive nodes are used for repair, namely $d_I = R - 1$ in current paper.

If a **separate node** has failed, the newcomer downloads β_S symbols each from d other nodes including nodes in clusters and separate nodes, which means that the newcomer needs only d helper nodes to repair a failed node, no matter where the failed node is located.

In a CSN-DSS, we dub (n, k, L, R, S) the **node parameters** and $(\alpha, d_I, \beta_I, d_C, \beta_C, \beta_S)$ the **storage/repair parameters** for simplicity. The intra-cluster and cross-cluster bandwidth of repairing a cluster node is defined as

$$\gamma_I \triangleq d_I \beta_I \quad \text{and} \quad \gamma_C \triangleq d_C \beta_C,$$

respectively. The bandwidth of repairing a separate node is $\gamma_S \triangleq d \beta_S$. When $\beta_I = \beta_C = \beta_S$, the traditional homogeneous DSS in [5] is obtained.

B. Information Flow Graph (IFG)

To analyse the performance of distributed storage systems, the information flow graph is proposed in [5], which consists of three kinds of nodes: a single data source S , storage nodes x_{in}^i , x_{out}^i , and a data collector DC as shown in Figure 2 (a). A physical storage node (hard disk or other storage device) is represented by a storage input node x_{in}^i and an output node x_{out}^i , where pre-computing is permitted when transmitting data. x_{in}^i and x_{out}^i are connected by a directed edge with capacity identical to the storage size α of the node. Throughout this paper, x^i is used to present x_{in}^i and x_{out}^i as a storage node.

¹An (n, k) maximum distance separate (MDS) code encodes k information symbols to n symbols such that any k symbols of n suffice to recover the original information symbols

As is mentioned before, the original data file is divided into k fragments and encoded into n fragments stored at n nodes, which is represented by n edges from node S to $\{x_{in}^i\}_{i=1}^n$ with infinite capacity.

At the initial time, the source node S stores data to the n storage nodes. Then S becomes inactive and the n storage nodes become active. When a node x^j fails, it becomes inactive. The repair procedure creates an active newcomer x^{n+1} to the graph as a substitute by connecting edges each with capacity β from $d(\geq k)$ surviving active nodes, which means the newcomer downloads β symbols from each of d alive nodes. Note that the values of β vary in heterogeneous DSSs. The total number of active nodes remains n after each repair. Based on the (n, k) MDS, a data collector DC connects to arbitrary k active nodes with direct edges of infinite capacity, which means any k nodes suffice to reconstruct the original data file. There may be many DC s connecting different sets of k active nodes, but only one DC is drawn visually.

An example of information flow graph for CSN-DSSs is illustrated in Figure 2 (a), where $n = 6(L = 1, R = 3, S = 1)$, $k = 4, d = 5$. Two nodes x^1 and x^6 have failed successively. When node x^1 (a node in cluster 1) has failed, a newcomer x^7 downloads β_I symbols each from x^2 and x^3 (two intra-cluster nodes) and β_C symbols each from x^4, x^5 and x^6 (three cross-cluster nodes). When node x^6 has failed, a second newcomer x^8 is added by downloading β_S symbols from five alive nodes (nodes in clusters or separate nodes). This model only handles one node failure at a time, downloading data from d helper nodes, a subset of the $n - 1$ alive nodes.

An important notion associated with the information flow graph is that of minimum cuts: In an IFG, a (direct) cut between S and DC is defined as a subset \mathcal{C} of edges such that every directed path from S to DC contains at least one edge in \mathcal{C} . The min-cut is the cut between S and DC in which the total sum of the edge capacities is smallest.

C. Problem Formulation

As is proved in [5], the reconstruction problem for every DC reduces exactly to multicasting the original data from a single source S to every DC . With relative works on network coding [14], [15], a tradeoff between node storage and repair bandwidth can be maintained by analysing the min-cuts between S and all possible DC s.

Let \mathcal{G} be the set of all possible information flow graphs of a CSN-DSS with node parameters (n, k, L, R, S) and storage/repair parameters $(\alpha, d_I, \beta_I, d_C, \beta_C, \beta_S)$. Consider any given finite IFG $G \in \mathcal{G}$, with a finite set of data collectors. If the minimum of the min-cuts separating the source with each data collector is larger than or equal to the data object size \mathcal{M} , then there exists a linear network code such that all data collectors can recover the data object (see Proposition 1 in [5]). Denote the graph with minimum min-cut by G^* . The **capacity** of a CSN-DSS is defined as

$$\mathbf{C}(\mathcal{G}) \triangleq \text{min-cut of } G^*.$$

In order to send data of size \mathcal{M} from the source to any data collectors,

$$\mathbf{C}(\mathcal{G}) \geq \mathcal{M} \tag{1}$$

should be satisfied. As $\mathbf{C}(\mathcal{G})$ depends on node parameters and storage/repair parameters, when node parameters (n, k, L, R, S) are fixed, a tradeoff between node storage α and repair bandwidth parameters $(d_I, \beta_I, d_C, \beta_C, \beta_S)$

will be characterized. The set of points $(\alpha, d_I, \beta_I, d_C, \beta_C, \beta_S)$ which satisfies $\mathbf{C}(\mathcal{G}) \geq \mathcal{M}$ is feasible in the sense of reliably storing the original file of size \mathcal{M} .

Therefore, to analyse the capacity or tradeoff properties of CSN-DSSs is to analyse the min-cuts of IFGs for given node parameters, which is illuminated in the following sections.

III. ANALYSIS OF CSN-DSSs

In this section, we investigate the min-cuts of IFGs for CSN-DSSs and prove the algorithms to generate the IFG achieving the capacity of given CSN-DSS. Some useful terms and notations are defined in Subsection III-A. Subsection III-B considers the min-cuts of IFGs with no separate selected nodes. With similar methods, the influence of one separate selected node is investigated in Subsection III-C.

A. Terminologies and Min-cut Calculation

For a given IFG $G \in \mathcal{G}$, the main problem is to find the min-cuts between source S and each DC . Because there are no paths among different DC s, the min-cuts between S and each different DC can be analysed in the same way. For simplicity, assume the IFGs in the following parts only contain one single DC and the min-cut only indicates the min-cut separating S and DC . This subsection introduces some important terminologies such as repair sequence, selected node distribution, cluster order and relative location, with which the method for calculating min-cuts of IFGs is illuminated.

Topological order: Note that every directed acyclic graph has a topological order (see [16], Chapter 3), which is an ordering of its vertices such that the existence of a path from v_i to v_j implies $i < j$. The k output nodes connected by every DC can be topologically sorted.

Min-cut between S and DC : Let $\{x_{out}^{t_i}\}_{i=1}^k$ be the set of output nodes connected by the data collector, which are topologically ordered. The min-cut between S and DC can be calculated out by cutting $\{x_{out}^{t_i}\}_{i=1}^k$ one by one in the topological order, which is proved in [5] Lemma 2. Each time cutting a node, a part of the min-cut is determined, called a **part-cut value**. So the min-cut between S and DC is the summation of the k part-cut values. For example, in Figure 2 (a), The DC connects to four output nodes $x_{out}^6, x_{out}^5, x_{out}^7, x_{out}^8$, which are topologically ordered. Cut the five nodes one by one, as is shown by the red dashed line, we then get the four part-cut values $\alpha, \alpha, (\beta_C + 2\beta_I)(\leq \alpha), (2\beta_C)(\leq \alpha)$ respectively. Note that if $\beta_C + 2\beta_I \geq \alpha$, the cut line will be between x_{in}^7 and x_{out}^7 . As a result, the third part-cut value will change to α .

Repair sequence and selected nodes: It is obvious that when a DC connects to a newcomer instead of connecting to an original node, the part-cut value may be smaller than α . So smaller min-cuts can be derived as the DC connects to more newcomers. Based on the MDS property mentioned in Subsection II-A, a DC connects to $k(\leq n)$ nodes, it is possible to find an IFG with a DC only connecting to k newcomers. Note that each newcomer corresponds to an original node failure and completes the repair procedure. Clearly, the topological order of k output nodes $\{x_{out}^{t_i}\}_{i=1}^k$ corresponds to a **repair sequence** of original nodes. These original nodes contained in a repair sequence are called **selected nodes**.

In homogeneous distributed storage systems, all the storage/repair parameters (α, β, d) are the same for different nodes. The repair sequence won't affect the minimum min-cut. As is proved in [5], when the DC connects to k newcomers and the newcomer x^{t_i} downloads data from all the former newcomers $\{x^{t_j}\}_{j=1}^{i-1}$, the minimum min-cut is reached.

However, in a CSN-DSS, the storage/repair parameters are different for nodes in cluster and separate nodes. Different repair sequences result in different min-cuts of the IFGs. As is illustrated in Figure 2, there two different repair sequences (x^1, x^4) in (a) and (x^1, x^3) in (b). The corresponding min-cuts are differently $2\alpha + 2\beta_I + \beta_C$ and $2\alpha + 2\beta_I + \beta_I$ respectively, as are shown by the red dash cut lines. Here we only consider two newcomer for the simplicity of the figures and assume $2\beta_I, \beta_C$ are less than α . Consequently, the repair sequence determines the minimum min-cut directly.

Selected node distribution: For any given k selected nodes, without loss of generality, assume the clusters are relabeled by the number of selected nodes in descending order. In the other words, cluster 1 contains the most selected nodes, and cluster L contains the least selected nodes. Define the **selected node distribution** as $\mathbf{s} = (s_0, s_1, s_2, \dots, s_L)$, where $s_i (1 \leq i \leq L)$ is the number of selected nodes in cluster i , and the first component s_0 is the number of selected separate nodes. Meanwhile, the set of all possible selected node distributions is defined as follows.

$$\mathcal{S} = \left\{ \mathbf{s} = (s_0, s_1, s_2, \dots, s_L) : s_{i+1} \leq s_i, 0 \leq s_i \leq R, \text{ for } 1 \leq i \leq L; 0 \leq s_0 \leq S; \sum_{i=0}^L s_i = k \right\}$$

Note that the selected node distribution describes the total number of selected nodes in each cluster and separate nodes. Moreover, we need to represent the topological order of k selected output nodes $\{x_{out}^{t_i}\}_{i=1}^k$ corresponding to k selected original nodes, called cluster order.

Cluster order: Let the **cluster order** $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_k)$ denote the repair sequence, where $\pi_i (1 \leq i \leq k)$ is the index of the cluster which contains the newcomer x^{t_i} corresponding to the failed node. If the i th node is a separate node, π_i equals to 0. Note that the cluster index is enough to define the repair sequence, because the storage/repair parameters for each node in the same cluster are the same.

For a certain selected node distribution $\mathbf{s} = (s_0, s_1, s_2, \dots, s_L)$, there are different cluster orders. The set of possible cluster orders is defined as

$$\Pi(\mathbf{s}) = \left\{ \boldsymbol{\pi} = (\pi_1, \dots, \pi_k) : \sum_{j=1}^k \mathbb{I}(\pi_j = i) = s_i, i \in \{0, 1, \dots, L\} \right\},$$

where $\mathbb{I}(\pi_j = i)$ is an indicator function which equals 1 if $\pi_j = i$, and 0 otherwise.

The relationship between selected node distribution and cluster order is illustrated in Figure 3, where the selected nodes are numbered. The selected node distribution $\mathbf{s} = (1, 4, 3, 1)$ means that, in the IFG of this CSN-DSS, the DC connects 1 separate node, 4 nodes from cluster 1, 3 nodes from cluster 2 and 1 node from cluster 3. The cluster order $\boldsymbol{\pi}(\mathbf{s}) = (1, 2, 3, 1, 2, 3, 1, 2, 1, 0)$ is a possible repair sequence for \mathbf{s} .

The selected nodes are labeled from 1 to k as Figure 3 shows, although it's enough to record the cluster number in the cluster order as the nodes in one cluster are undifferentiated. For the nodes in a cluster order $\boldsymbol{\pi}$, it's also

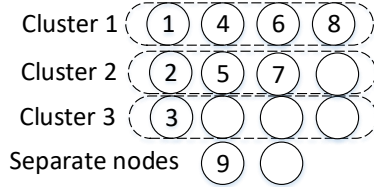


Fig. 3: The numbered nodes are selected nodes and the selected node distribution is $\mathbf{s} = (1, 4, 3, 1)$. A corresponding cluster order is $\boldsymbol{\pi} = (1, 2, 3, 1, 2, 1, 2, 1, 0)$ for the CSN-DSS with $n = 15$ nodes, $k = 9$ selected nodes.

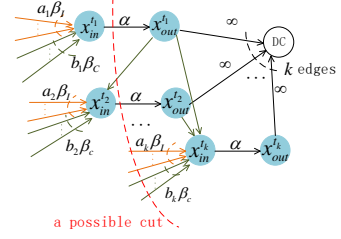


Fig. 4: The min-cut IFG G with no separate nodes in selected nodes

needed to identify the precedence of selected nodes in each cluster. Assume the i -th node in cluster order $\boldsymbol{\pi}$ is the $h_{\boldsymbol{\pi}}(i)$ -th node in its cluster. We called $h_{\boldsymbol{\pi}}(i)$ the **relative location** of the i -th node and

$$h_{\boldsymbol{\pi}}(i) = \sum_{j=1}^i \mathbb{I}(\pi_j = \pi_i), \quad (2)$$

where $1 \leq i \leq k$. For an example, in Figure 3, the cluster order is $\boldsymbol{\pi} = (1, 2, 3, 1, 2, 1, 2, 1, 0)$ and $h_{\boldsymbol{\pi}}(4) = \mathbb{I}(\pi_1 = \pi_4) + \mathbb{I}(\pi_2 = \pi_4) + \mathbb{I}(\pi_3 = \pi_4) + \mathbb{I}(\pi_4 = \pi_4) = 1 + 0 + 0 + 1 = 2$ where $\pi_4 = 1$. The corresponding sequence of $h_{\boldsymbol{\pi}}(i)$ is then $(1, 1, 1, 2, 2, 3, 3, 4, 1)$.

Calculating the min-cut between S and DC of an IFG:

When calculating the min-cut of IFG G , it is to cut the output nodes contacted by the DC k times in topological order. Consider two disjoint sets U and \bar{U} of the nodes in G . Assume S and the original nodes $x^i (1 \leq i \leq n)$ are contained in U and DC is contained in \bar{U} at the beginning. Every time we cut G , some nodes are added into U and \bar{U} respectively. When G^* is cut k times, all the nodes of G^* are contained in U or \bar{U} and the set of edges emanating from U to \bar{U} is a cut between S and DC . Let \mathcal{C} denote the edges in the cut set, i.e., the set of edges going from U to \bar{U} .

As is illustrated in Figure 4, the k selected nodes $\{x^{t_i}\}_{i=1}^k$ are in topological order. When cutting node x^{t_1} , there are two possible cases.

- If $x_{in}^{t_1}$ is in U , the edge $(x_{in}^{t_1}, x_{out}^{t_1})$ is contained in \mathcal{C} . The part-cut value equals α .
- In case of $x_{in}^{t_1}$ is in \bar{U} , since $x_{in}^{t_1}$ has an in-degree of $d = d_I + d_C$ and it is the topologically first newcomer in \bar{U} , all the incoming edges of $x_{in}^{t_1}$ must be in \mathcal{C} , which consists of d_I edges from intra-cluster nodes and d_C edges from cross-cluster nodes. The part-cut value equals $d_I \beta_I + d_C \beta_C$.

When cutting node x^{t_2} , the first case is similar to x^{t_1} and the part-cut value is also α . For the second case, if x^{t_2} is in the same cluster with x^{t_1} , the incoming edges of $x_{in}^{t_2}$ consist of $d_I - 1$ edges from intra-cluster nodes and d_C edges from cross-cluster nodes, then the part-cut value equals to $(d_I - 1) \beta_I + d_C \beta_C$. On the other hand, if x^{t_2} is in different clusters with x^{t_1} , the incoming edges of $x_{in}^{t_2}$ still contain d_I edges from intra-cluster nodes but $d_C - 1$ edges from cross-cluster nodes, and the part-cut value equals $d_I \beta_I + (d_C - 1) \beta_C$.

Now consider node $x^{t_i}(1 \leq i \leq k)$, the i th newcomer:

- If $x_{in}^{t_i} \in U$, the edge $(x_{in}^{t_i}, x_{out}^{t_i})$ must be in \mathcal{C} .
- If $x_{in}^{t_i} \in \bar{U}$, node $x_{in}^{t_i}$ has $d = d_I + d_C$ incoming edges consisting of two parts: edges from nodes in U and edges from nodes in \bar{U} . Cut set \mathcal{C} only includes the first part of incoming edges among which let a_i denote the number of **edges from intra-cluster nodes** and b_i denote **edges from cross-cluster nodes**. It's obvious that $0 \leq a_i \leq d_I$ and $0 \leq b_i \leq d_C$. Note that when i increases by 1, either a_i or b_i will decrease by 1 and will not decrease when a_i or b_i equals 0. Since at most $i - 1$ incoming edges of $x_{in}^{t_i}$ can be from $x_{out}^{t_j}(1 \leq j \leq i - 1)$ already contained in \bar{U} ,

$$a_i + b_i \geq d - (i - 1), \quad (3)$$

for $1 \leq i \leq k$. Equality holds if a_i and b_i will not decrease to 0 as i increases.

If the i th selected node is a separate node, let c_i denote the number of incoming edges of $x_{in}^{t_i}$ from U and

$$c_i = d - (i - 1).$$

The respective values of a_i and b_i depend on the repair sequence of original nodes, namely, the selected node distribution \mathbf{s} and cluster order $\boldsymbol{\pi}$. The sum of the capacity of these edges is called the i th **part incoming weight**

$$w_i(\mathbf{s}, \boldsymbol{\pi}) = \begin{cases} a_i\beta_I + b_i\beta_C & \text{if the } i\text{th selected node is a cluster node,} \\ c_i\beta_S & \text{if the } i\text{th selected node is a separate node} \end{cases}. \quad (4)$$

If the selected node distribution \mathbf{s} or cluster order $\boldsymbol{\pi}$ is fixed beforehand, $w_i(\mathbf{s}, \boldsymbol{\pi})$ can be written as $w_i(\boldsymbol{\pi})$ or w_i for simplicity. On the other hand, a_i, b_i, c_i can be written as $a_i(\boldsymbol{\pi}), b_i(\boldsymbol{\pi}), c_i(\boldsymbol{\pi})$ for specific $\boldsymbol{\pi}$, respectively.

For a fixed selected node distribution \mathbf{s} , the min-cut varies for different cluster orders $\boldsymbol{\pi} \in \Pi(\mathbf{s})$. The min-cut for $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_k)$ is defined as

$$MC(\mathbf{s}, \boldsymbol{\pi}) \triangleq \sum_{i=1}^k \min\{w_i(\boldsymbol{\pi}), \alpha\}. \quad (5)$$

With the above definitions, for an IFG G with specified selected node distribution and cluster order, the min-cut can be figured out. In the following subsection, the min-cuts of cluster DSSs without separate nodes will be analysed.

B. The min-cuts of IFGs with no separate selected nodes

In this subsection, we assume the selected nodes are all cluster nodes, namely, $\mathbf{s} = (s_0 = 0, s_1, \dots, s_L)$, in which case, our CSN-DSS model can be seen as cluster DSS model. For given node parameters (n, k, L, R, S) , to find the IFG G^* with the minimum min-cut among all possible IFGs is equivalent to find the corresponding selected node distribution \mathbf{s} and cluster order $\boldsymbol{\pi}$. As \mathbf{s} and $\boldsymbol{\pi}$ both influence the min-cuts, the analysis comprises two steps:

1. Fix the selected node distribution \mathbf{s} and analyse the min-cuts for different cluster orders $\boldsymbol{\pi}$ (see the proof of vertical order algorithm in Theorem 1).
2. Fix the cluster order generating algorithm and analyse the min-cuts for different \mathbf{s} (see the proof of horizontal selection algorithm in Theorem 2).

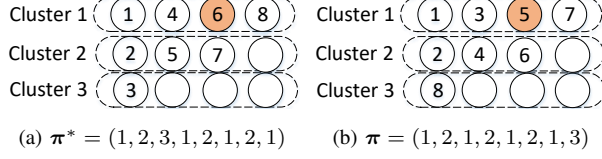


Fig. 5: The numbered nodes are selected nodes. There are two cluster orders π^* and π for a selected node distribution $\mathbf{s} = (0, 4, 3, 1)$.

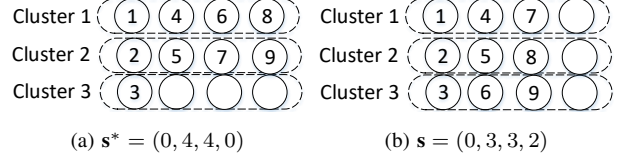


Fig. 6: The numbered nodes are selected nodes. There are two selected node distributions \mathbf{s}^* and \mathbf{s} for the CSN-DSS ($n = 12, k = 8, L = 3, R = 4, S = 0$).

Vertical order algorithm for $d_I = R - 1$

When the selected node distribution $\mathbf{s} = (s_0 = 0, s_1, \dots, s_L)$ is fixed, the cluster order $\pi^* = (\pi_1^*, \pi_2^*, \dots, \pi_k^*)$ generated by the vertical order algorithm achieves the minimum min-cut among all the possible IFGs, which is proved in Theorem 1. In [12], the above conclusion is considered based on the special assumption that all the alive nodes are used to repair the failed node, namely, $d_I = R - 1$ and $d_C = n - R$. We will investigate and prove this problem in more general settings. The number of helper nodes from cross-cluster, d_C , varies from $k - R + 1$ to $n - R$ and does not need to be $n - R$ of [12], following from the condition that $d_I + d_C \geq k$.

Algorithm 1 Vertical order algorithm

Input: $\mathbf{s} = (s_0 = 0, s_1, \dots, s_L)$. Initial cluster label $j \leftarrow 1$;

Output: $\pi^* = (\pi_1^*, \dots, \pi_k^*)$.

```

1: for  $i = 1$  to  $k$  do
2:   if the  $i$ -th selected node is a separate node then  $\pi_i^* \leftarrow 0$ ; continue;
3:   end if
4:   if  $s_j = 0$  then  $j = 1$ ;
5:   else  $\pi_i^* \leftarrow j$ ;  $s_j \leftarrow s_i - 1$ ;  $j \leftarrow (j \bmod L) + 1$ ;
6:   end if
7: end for

```

An example of Algorithm 1 is illustrated in Figure 5 (a), where $\mathbf{s} = (0, 4, 3, 1)$. After three iterations, s_3 equals to 0 and $\pi_4^* = 1$ in the next iteration. The final output of the algorithm is $\pi^* = (1, 2, 3, 1, 2, 1, 2, 1)$. If the selected node distribution \mathbf{s} is fixed, a cluster order determines an IFG and the min-cut $MC(\mathbf{s}, \pi)$ can be calculated. Note that $MC(\mathbf{s}, \pi)$ depends on the i -th part incoming weight $w_i(\pi) = a_i(\pi)\beta_I + b_i(\pi)\beta_C$ ($1 \leq i \leq k$) and a useful property for $a_i(\pi)$ (the coefficient of β_I) is proved in Lemma 1.

Lemma 1. For a given selected node distribution $\mathbf{s} = (0, s_1, s_2, \dots, s_L)$ of the system model in Figure 1, the multi-set² $[a_i(\pi)]_{i=1}^k = [a_1(\pi), a_2(\pi), \dots, a_k(\pi)]$ consists of the same elements for all the different cluster orders $\pi \in \Pi(\mathbf{s})$ and $a_i(\pi) = d_I + 1 - h_\pi(i)$ for $1 \leq i \leq k$.

²A multi-set is a generalization of the concept of a set that, unlike a set, allows multiple instances of the multi-set's elements.

Proof. Assume $\pi^* = (\pi_1^*, \dots, \pi_k^*)$ and $\pi = (\pi_1, \dots, \pi_k)$ are two different cluster orders for the same selected node distribution $\mathbf{s} = (s_1, \dots, s_L)$. For example, in Figure 5, the selected node distribution is $\mathbf{s} = (4, 3, 1)$ and the corresponding two different cluster orders are $\pi^* = (1, 2, 3, 1, 2, 1, 2, 1)$ and $\pi = (1, 2, 1, 2, 1, 2, 1, 3)$. Note that the coloured node 6 in Figure 5 (a) is the third selected node in Cluster 1. When cutting node 6, two intra-cluster nodes are contained in \bar{U} (node 1 and node 3), which won't be counted in the part-cut value. Then $a_6(\pi^*) = d_I - 2 = R - 1 - 2 = 1$. Now consider cluster order π in Figure 5 (b). Although the coloured node 5 is the 5th node in π , it is the third selected node in Cluster 1 and $a_5(\pi) = d_I - 2 = R - 1 - 2 = 1$. It's easy to see that the value of $a_i(\pi^*)$ only depends on the number of selected nodes in the same cluster before repairing the current node, which is defined by the relative location $h_{\pi^*}(i)$, namely, $a_i(\pi^*) = d_I + 1 - h_{\pi^*}(i)$ for $1 \leq i \leq k$.

If $\mathbf{s} = (0, s_1, s_2, \dots, s_L)$ is fixed, the set of $h_{\pi}(i)$ for nodes in Cluster l is $\{1, 2, \dots, s_l\}$ for $1 \leq l \leq L$, no matter where the nodes locate in the cluster orders. For all the cluster orders $\pi \in \Pi(\mathbf{s})$, the multi-set $[a_i(\pi)]_{i=1}^k$ consists of L sets $\{d_I + 1 - 1, d_I + 1 - 2, \dots, d_I + 1 - s_l\}$ for $1 \leq l \leq L$. \square

When the selected node distribution \mathbf{s} is fixed, a property of the min-cuts of IFGs for different cluster orders is proved in Theorem 1.

Theorem 1. *For the given node parameters (n, k, L, R, S) and any given selected node distribution $\mathbf{s} \in \mathcal{S}$ with $s_0 = 0$, the vertical cluster order π^* obtained by the vertical order algorithm achieves the minimum min-cut among all the possible IFGs with \mathbf{s} . In other words,*

$$MC(\mathbf{s}, \pi^*) \leq MC(\mathbf{s}, \pi),$$

holds for arbitrary $\pi \in \Pi(\mathbf{s})$. $MC(\mathbf{s}, \pi)$ is defined by (5).

Proof. Assume $(w_{u_1}(\pi), \dots, w_{u_k}(\pi))$ is a non-increasing order of elements in multi-set $[w_i(\pi)]_{i=1}^k$, namely, $w_{u_1}(\pi) \geq \dots \geq w_{u_k}(\pi)$. This proof consists of two parts. In Part 1, we will prove that

$$\sum_{i=k-t+1}^k w_{u_i}(\pi^*) \leq \sum_{i=k-t+1}^k w_{u_i}(\pi), \quad (6)$$

for any $1 \leq t \leq k$. Note that (6) means the sum of the minimum t elements in multi-set $[w_i(\pi^*)]_{i=1}^k$ is no more than the sum of the minimum t elements in $[w_i(\pi)]_{i=1}^k$ for any $1 \leq t \leq k$. With the help of (6), Part 2 completes the proof by considering the relationship between α and $w_i(\pi^*)$ in formula (5).

Part 1: As $w_i(\pi) = a_i(\pi)\beta_I + b_i(\pi)\beta_C$ (see equation (4)), the coefficient of β_I and β_C are considered respectively, and let

$$\phi_i(\pi) \triangleq a_i(\pi) + b_i(\pi),$$

for simplicity. In the following part, we will compare $\phi_i(\pi^*)$ and $\phi_i(\pi)$ one by one and prove inequality (7) which is important to prove (6).

For any cluster order $\pi \in \Pi(\mathbf{s})$, let sequence $(\phi_{t_1}(\pi), \dots, \phi_{t_k}(\pi))$ denote the non-increasing order of the elements of multi-set $[\phi_i(\pi)]_{i=1}^k$, namely, $\phi_{t_1}(\pi) \geq \dots \geq \phi_{t_k}(\pi)$. Based on Algorithm 1, it's easy to verify that the sequences

$(\phi_1(\pi^*), \dots, \phi_k(\pi^*))$, $(b_1(\pi^*), \dots, b_k(\pi^*))$ and $(w_1(\pi^*), \dots, w_k(\pi^*))$ are all non-increasing. Assume p ($1 \leq p \leq k$) is the integer that satisfies the following conditions:

$$b_p(\pi^*) = 0 \text{ and } b_{p-1}(\pi^*) > 0,$$

meaning that exactly d_C selected cross-cluster nodes are already cut when cutting the p -th node by the cluster order π^* . Then $\phi_i(\pi^*) = a_i(\pi^*) + b_i(\pi^*) = d - i + 1$ for $1 \leq i \leq p$. For the remaining nodes in π^* , $b_i(\pi^*) = 0$ and $\phi_i(\pi^*) = a_i(\pi^*)$ ($i > p$).

- When $1 \leq i \leq p$, as $\phi_i(\pi) \geq d - i + 1$ (see (3)), $\phi_{t_i}(\pi) \geq \phi_i(\pi) \geq d - i + 1 = \phi_i(\pi^*)$.
- When $p + 1 \leq i \leq k$, as $a_i(\pi^*) \geq a_{i+1}(\pi^*) \geq \dots \geq a_k(\pi^*)$, **at most** $k - i$ elements of multi-set $[a_i(\pi^*)]_{i=1}^k$ are less than $a_i(\pi^*)$.

► If $a_{t_i}(\pi) < a_i(\pi^*)$, we first assume $a_{t_j}(\pi) < a_j(\pi^*)$ for all $i + 1 \leq j \leq k$ and will derive a contradiction. It's obvious that **at least** $k - i + 1$ elements of multi-set $[a_{t_i}(\pi)]_{i=1}^k$ are less than $a_i(\pi^*)$. As is proved in Lemma 1, $[a_{t_i}(\pi)]_{i=1}^k$ and $[a_i(\pi)^*]_{i=1}^k$ contain the same elements, $[a_i(\pi)^*]_{i=1}^k$ then contains **at least** $k - i + 1$ elements of multi-set $[a_{t_i}(\pi)]_{i=1}^k$ are less than $a_i(\pi^*)$, which a contradiction. There then exists at least one $a_{t_j}(\pi)$ ($i + 1 \leq j \leq k$) not less than $a_i(\pi^*)$, and $\phi_{t_i}(\pi) \geq \phi_{t_j}(\pi) \geq a_{t_j}(\pi) \geq a_i(\pi^*) = \phi_i(\pi)$.

► If $a_{t_i}(\pi) \geq a_i(\pi^*)$, $\phi_{t_i}(\pi) = a_{t_i}(\pi) + b_{t_i}(\pi) \geq a_{t_i}(\pi) > a_i(\pi^*) = \phi_i(\pi^*)$.

Then it can be proved that

$$\sum_{i=k-t+1}^k (a_i(\pi^*) + b_i(\pi^*)) = \sum_{i=k-t+1}^k \phi_i(\pi^*) \leq \sum_{i=k-t+1}^k \phi_{t_i}(\pi) \stackrel{(a)}{\leq} \sum_{i=k-t+1}^k (a_{u_i}(\pi) + b_{u_i}(\pi)) \quad (7)$$

Note that $\sum_{i=k-t+1}^k a_{u_i}(\pi) + b_{u_i}(\pi) = \sum_{i=k-t+1}^k \phi_{u_i}(\pi)$ is the sum of t elements of multi-set $[\phi_i(\pi)]_{i=1}^k$. Inequality (a) is based on the fact that $\sum_{i=k-t+1}^k \phi_{t_i}(\pi)$ is the sum of the minimum t elements of $[\phi_i(\pi)]_{i=1}^k$, not greater than the sum of any t elements of $[\phi_i(\pi)]_{i=1}^k$.

With the above consequence, it can be proved that

$$\begin{aligned} \sum_{i=k-t+1}^k w_{u_i}(\pi^*) &= \sum_{i=k-t+1}^k w_i(\pi^*) = \sum_{i=k-t+1}^k (a_i(\pi^*) * \beta_I + b_i(\pi^*) * \beta_C) \\ &= \sum_{i=k-t+1}^k a_i(\pi^*) * \beta_I + \left(\sum_{i=k-t+1}^k (a_i(\pi^*) + b_i(\pi^*)) - \sum_{i=k-t+1}^k a_i(\pi^*) \right) * \beta_C \\ &\stackrel{(b)}{\leq} \sum_{i=k-t+1}^k a_i(\pi^*) * \beta_I + \left(\sum_{i=k-t+1}^k (a_{u_i}(\pi) + b_{u_i}(\pi)) - \sum_{i=k-t+1}^k a_i(\pi^*) \right) * \beta_C \\ &\stackrel{(c)}{\leq} \sum_{i=k-t+1}^k a_i(\pi^*) * \beta_I + \left(\sum_{i=k-t+1}^k a_{u_i}(\pi) - \sum_{i=k-t+1}^k a_i(\pi^*) \right) * \beta_I + \sum_{i=k-t+1}^k b_{u_i}(\pi) * \beta_C \\ &= \sum_{i=k-t+1}^k a_{u_i}(\pi) * \beta_I + \sum_{i=k-t+1}^k b_{u_i}(\pi) * \beta_C = \sum_{i=k-t+1}^k w_{u_i}(\pi), \end{aligned}$$

where (b) is based on inequality (7) and (c) is because of $\beta_I \geq \beta_C$.

Part 2: Assume there are t_1 elements in $[w_i(\pi^*)]_{i=1}^k$ and t_2 elements in $[w_i(\pi)]_{i=1}^k$ greater than α .

- If $t_1 < t_2$, $MC(\mathbf{s}, \pi^*) = t_1\alpha + \sum_{i=t_1+1}^{t_2} w_i(\pi^*) + \sum_{i=t_2+1}^k w_i(\pi^*) \leq t_2\alpha + \sum_{i=t_2+1}^k w_{u_i}(\pi) = MC(\mathbf{s}, \pi)$.
- If $t_1 = t_2$, it's easy to prove $MC(\mathbf{s}, \pi^*) \leq MC(\mathbf{s}, \pi)$, using (6).

- If $t_1 > t_2$, $MC(\mathbf{s}, \boldsymbol{\pi}^*) = t_2\alpha + \sum_{i=t_2+1}^{t_1} \min\{w_i(\boldsymbol{\pi}^*), \alpha\} + \sum_{i=t_1+1}^k w_i(\boldsymbol{\pi}^*) \leq t_2\alpha + \sum_{i=t_2+1}^k w_i(\boldsymbol{\pi}^*) \leq t_2\alpha + \sum_{i=t_2+1}^k w_{u_i}(\boldsymbol{\pi}) = MC(\mathbf{s}, \boldsymbol{\pi})$.

□

The consequence of Theorem 1 can be verified by Algorithm 1 and the numerical examples illustrated in Figure 5 (a) and (b). To analyse the influence of selected node distribution, for any input \mathbf{s} , let

$$\boldsymbol{\pi}^*(\mathbf{s}) = (\pi^*(\mathbf{s})_1, \pi^*(\mathbf{s})_2, \dots, \pi^*(\mathbf{s})_k) \quad (8)$$

denote **the unique cluster order** generated by the vertical order algorithm. In the following part, we investigate the min-cuts for different selected node distributions \mathbf{s} , where the cluster orders are $\boldsymbol{\pi}^*(\mathbf{s})$.

Horizontal selection algorithm for $d_I = R - 1$

The vertical order algorithm generates the cluster order achieving the minimum min-cut for any given selected node distribution \mathbf{s} . In this part, we assume all the cluster orders are generated by the vertical order algorithm and analyse the min-cuts for different selected node distributions. In Theorem 2, it is proved that the minimum min-cut among possible IFGs is achieved by the selected node distribution $\mathbf{s}^* = (s_0^*, s_1^*, s_2^*, \dots, s_L^*)$ generated by the horizontal selection algorithm and the cluster order $\boldsymbol{\pi}^*(\mathbf{s}^*)$ generated by the vertical order algorithm.

Algorithm 2: *Horizontal selection algorithm:*

The horizontal selected node distribution is $\mathbf{s}^* = (s_0^*, s_1^*, s_2^*, \dots, s_L^*)$ ($\sum_{i=0}^L s_i^* = k$), where

$$s_i^* = \begin{cases} R, & i \leq \lfloor \frac{k-s_0^*}{R} \rfloor \\ k - \lfloor \frac{k-s_0^*}{R} \rfloor R, & i = \lfloor \frac{k-s_0^*}{R} \rfloor + 1 \\ 0, & i > \lfloor \frac{k-s_0^*}{R} \rfloor + 1 \end{cases}$$

In this section, the situation without separate nodes is considered, namely, $s_0^* = 0$. An example of this algorithm is illustrated in Figure 6 (a), where $k = 8, R = 4$. Based on the horizontal algorithm, $s_1^* = R = 4$, $s_2^* = R = 4$ and $s_3 = k - 2R = 0$. Another property of $a_i(\boldsymbol{\pi})$, the coefficients of β_I , is proved in the following lemma, when the horizontal selected algorithm is used.

Lemma 2. *For the given node parameters (n, k, L, R, S) , the coefficients of β_I satisfies that*

$$a_i(\boldsymbol{\pi}^*(\mathbf{s}^*)) \leq a_i(\boldsymbol{\pi}^*(\mathbf{s}))$$

for $1 \leq i \leq k$, where \mathbf{s}^* is the selected node distribution generated by the horizontal selection algorithm and $\mathbf{s} \in \mathcal{S}$ with $s_0 = 0$. Note that $\boldsymbol{\pi}^*(\cdot)$ is defined by (8).

Proof. As is proved in Lemma 1, the coefficient of β_I , $a_i(\boldsymbol{\pi}^*(\mathbf{s})) = d_I + 1 - h_{\boldsymbol{\pi}^*(\mathbf{s})}(i)$. We will analyse the relative location $h_{\boldsymbol{\pi}^*(\mathbf{s})}(i)$ with jumping points defined in (9) and prove that $h_{\boldsymbol{\pi}^*(\mathbf{s}^*)}(i) \geq h_{\boldsymbol{\pi}^*(\mathbf{s})}(i)$ for $1 \leq i \leq k$.

Based on the vertical order algorithm, it's easy to verify the following two properties of $h_{\boldsymbol{\pi}^*(\mathbf{s})}(i)$:

- $1 \leq h_{\boldsymbol{\pi}^*(\mathbf{s})}(i) \leq R$ for $1 \leq i \leq k$,
- $0 \leq h_{\boldsymbol{\pi}^*(\mathbf{s})}(i+1) - h_{\boldsymbol{\pi}^*(\mathbf{s})}(i) \leq 1$ for $1 \leq i \leq k-1$,

meaning that $h_{\pi^*(s)}(i)$ is non-decreasing and will increase one time at most by 1. For an example, the sequence of relative location $h_{\pi^*(s)}(i)$ is (1, 1, 1, 2, 2, 3, 3, 4, 4) in Figure 6 (a) and when $i = 3, 5$ or 7 , the value of $h_{\pi^*(s)}(i)$ will increase by 1 for the next time. These values of i are called **jumping points**, denoted by

$$J(\pi^*(s)) = (j_0(\pi^*(s)), j_1(\pi^*(s)), \dots, j_{s_1-1}(\pi^*(s)), j_{s_1}(\pi^*(s))), \quad (9)$$

which depends on $s = (s_0, s_1, \dots, s_k)$. We set $j_0(\pi^*(s)) = 0$ and $j_{s_1}(\pi^*(s)) = k$ as the beginning and ending of the jumping point vector, then

$$j_i(\pi^*(s)) - j_{i-1}(\pi^*(s)) = \#\{t | h_{\pi^*(s)}(t) = i, 1 \leq t \leq k\}$$

for $1 \leq i \leq s_1$. Based on the definition of cluster order, it's obvious that

$$j_i(\pi^*(s)) - j_{i-1}(\pi^*(s)) \geq j_{i+1}(\pi^*(s)) - j_i(\pi^*(s)) \quad (10)$$

for $1 \leq i \leq s_1 - 1$.

We will use induction method to prove $j_i(\pi^*(s^*)) \leq j_i(\pi^*(s))$ for $1 \leq i \leq s_1 - 1$. Based on the vertical order algorithm, $j_1(\pi^*(s^*)) \leq j_1(\pi^*(s))$. Assume $j_t(\pi^*(s^*)) \leq j_t(\pi^*(s))$, it's needed to prove $j_{t+1}(\pi^*(s^*)) \leq j_{t+1}(\pi^*(s))$.

There are $k - j_{t+1}(\pi^*(s^*))$ nodes remaining after jumping point $j_{t+1}(\pi^*(s^*))$. Based on the horizontal selection algorithm,

$$j_{i+1}(\pi^*(s^*)) - j_i(\pi^*(s^*)) = j_1(\pi^*(s^*)) \text{ or } j_1(\pi^*(s^*)) - 1$$

for $1 \leq i \leq R - 1$. Then

$$k - j_{t+1}(\pi^*(s^*)) \geq (R - t - 1)(j_{t+1}(\pi^*(s^*)) - j_t(\pi^*(s^*)) - 1). \quad (11)$$

Assume

$$j_{t+1}(\pi^*(s^*)) > j_{t+1}(\pi^*(s)). \quad (12)$$

As $j_t(\pi^*(s^*)) \leq j_t(\pi^*(s))$, then

$$\begin{aligned} j_{t+1}(\pi^*(s)) - j_t(\pi^*(s)) &< j_{t+1}(\pi^*(s^*)) - j_t(\pi^*(s^*)) \\ \Rightarrow j_{t+1}(\pi^*(s)) - j_t(\pi^*(s)) &\leq j_{t+1}(\pi^*(s^*)) - j_t(\pi^*(s^*)) - 1. \end{aligned} \quad (13)$$

Then

$$\begin{aligned} k - j_{t+1}(\pi^*(s)) &\stackrel{(a)}{\leq} (s_1 - t - 1)(j_{t+1}(\pi^*(s)) - j_t(\pi^*(s))) \\ &\stackrel{(b)}{\leq} (s_1 - t - 1)(j_{t+1}(\pi^*(s^*)) - j_t(\pi^*(s^*)) - 1) \\ &\leq (R - t - 1)(j_{t+1}(\pi^*(s^*)) - j_t(\pi^*(s^*)) - 1) \\ &\stackrel{(c)}{\leq} k - j_{t+1}(\pi^*(s^*)), \end{aligned}$$

where (a) is based on (10), (b) is because of (13) and (c) results from (11). Hence,

$$j_{t+1}(\pi^*(s^*)) \leq j_{t+1}(\pi^*(s)),$$

contradicting assumption (12), and it can be proved that $j_{t+1}(\pi^*(\mathbf{s}^*)) \leq j_{t+1}(\pi^*(\mathbf{s}))$. Since $h_{\pi^*(\mathbf{s})}(i) = t$ for $j_{t-1}(\pi^*(\mathbf{s})) \leq i \leq j_t(\pi^*(\mathbf{s}))$ ($t = 1, 2, \dots, s_1$) and $j_i(\pi^*(\mathbf{s}^*)) \leq j_i(\pi^*(\mathbf{s}))$ for $1 \leq i \leq s_1 - 1$, it can be proved that

$$h_{\pi^*(\mathbf{s}^*)}(i) \geq h_{\pi^*(\mathbf{s})}(i), \quad (14)$$

for $1 \leq i \leq k$. Hence, $a_i(\pi^*(\mathbf{s}^*)) \leq a_i(\pi^*(\mathbf{s}))$ for $1 \leq i \leq k$. \square

Theorem 2. For the given node parameters (n, k, L, R, S) , when the selected node distribution \mathbf{s}^* is generated by the horizontal selection algorithm and the corresponding cluster order is generated by the vertical order algorithm, the min-cut of this IFG isn't greater than any IFGs. In other words,

$$MC(\mathbf{s}^*, \pi^*(\mathbf{s}^*)) \leq MC(\mathbf{s}, \pi^*(\mathbf{s})),$$

for all $\mathbf{s} \in \mathcal{S}$ with $s_0 = 0$. Note that $\pi^*(\cdot)$ is defined by (8). $MC(\mathbf{s}, \pi)$ is defined by (5).

Proof. Based on the vertical order algorithm, sequence $(w_1(\pi^*(\mathbf{s})), \dots, w_k(\pi^*(\mathbf{s})))$ is non-increasing for all $\mathbf{s} \in \mathcal{S}$ and $s_0 = 0$. Similarly to the proof of Theorem 1, it is only needed to prove that

$$w_i(\pi^*(\mathbf{s}^*)) \leq w_i(\pi^*(\mathbf{s}))$$

for $1 \leq i \leq k$. From the definition of $a_i(\pi^*(\mathbf{s}))$, $b_i(\pi^*(\mathbf{s}))$ and $h_{\pi^*(\mathbf{s})}(i)$ (see (2)), it is known that

$$a_i(\pi^*(\mathbf{s})) = d_I + 1 - h_{\pi^*(\mathbf{s})}(i) \quad (15)$$

for $1 \leq i \leq k$. When $d_C - (i - h_{\pi^*(\mathbf{s})}(i)) \geq 0$,

$$b_i(\pi^*(\mathbf{s})) = d_C - (i - h_{\pi^*(\mathbf{s})}(i)). \quad (16)$$

We assume $b_i(\pi^*(\mathbf{s}))$ decreases to 0 when $i = i^*(\mathbf{s})$, where $i^*(\mathbf{s})$ is a function of \mathbf{s} . It will not decrease anymore and $w_i(\pi^*(\mathbf{s})) = a_i(\pi^*(\mathbf{s}))\beta_I$ for $i \geq i^*(\mathbf{s})$, where

$$i^*(\mathbf{s}) - h_{\pi^*(\mathbf{s})}(i^*(\mathbf{s})) = d_C. \quad (17)$$

As is proved in Lemma 2 (14), $h_{\pi^*(\mathbf{s})}(i) \leq h_{\pi^*(\mathbf{s}^*)}(i)$ for $1 \leq i \leq k$, then $i^*(\mathbf{s}^*) \geq i^*(\mathbf{s})$ based on (17).

• When $1 \leq i \leq i^*(\mathbf{s})$,

$$\begin{aligned} w_i(\pi^*(\mathbf{s})) - w_i(\pi^*(\mathbf{s}^*)) &= (a_i(\pi^*(\mathbf{s})) - a_i(\pi^*(\mathbf{s}^*)))\beta_I + (b_i(\pi^*(\mathbf{s})) - b_i(\pi^*(\mathbf{s}^*)))\beta_C \\ &\stackrel{(a)}{=} (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))\beta_I - (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))\beta_C \\ &= (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))(\beta_I - \beta_C) \stackrel{(b)}{\geq} 0. \end{aligned}$$

Note that (a) is based on (15) and (16). (b) comes from (14) and $\beta_I \geq \beta_C$. Then $w_i(\pi^*(\mathbf{s})) \geq w_i(\pi^*(\mathbf{s}^*))$.

• When $i^*(\mathbf{s}) + 1 \leq i \leq i^*(\mathbf{s}^*)$, $b_i(\pi^*(\mathbf{s}))$ equals to 0 and will not decrease with i increasing, but

$$d_C - (i - h_{\pi^*(\mathbf{s})}(i)) \leq 0. \quad (18)$$

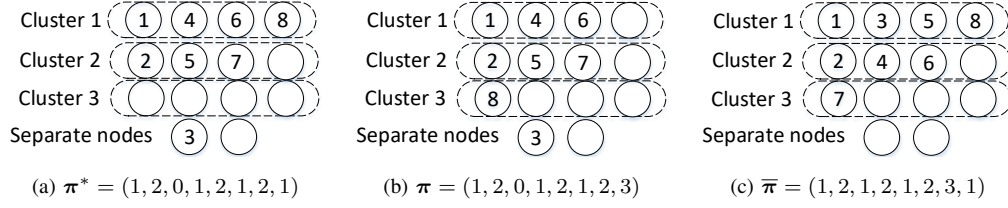


Fig. 7: The numbered nodes are selected nodes. There are three selected node distributions $\mathbf{s}^* = (1, 4, 3, 0)$, $\mathbf{s} = (1, 3, 3, 1)$ and $\bar{\mathbf{s}} = (0, 4, 3, 1)$, with cluster orders π^* , π and $\bar{\pi}$ respectively for CSN-DSS with node parameters $(n = 12, k = 9, L = 3, R = 4, S = 2)$.

Hence,

$$\begin{aligned}
 w_i(\pi^*(\mathbf{s})) - w_i(\pi^*(\mathbf{s}^*)) &= (a_i(\pi^*(\mathbf{s})) - a_i(\pi^*(\mathbf{s}^*)))\beta_I - b_i(\pi^*(\mathbf{s}^*))\beta_C \\
 &\stackrel{(c)}{\geq} (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))\beta_I - b_i(\pi^*(\mathbf{s}^*))\beta_C - (d_C - (i - h_{\pi^*(\mathbf{s})}(i)))\beta_C \\
 &= (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))\beta_I - (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))\beta_C \\
 &= (h_{\pi^*(\mathbf{s}^*)}(i) - h_{\pi^*(\mathbf{s})}(i))(\beta_I - \beta_C) \geq 0,
 \end{aligned}$$

where (c) bases on (18).

- When $i^*(\mathbf{s}^*) + 1 \leq i \leq k$, $w_i(\pi^*(\mathbf{s})) = a_i(\pi^*(\mathbf{s}))\beta_I \geq a_i(\pi^*(\mathbf{s}^*))\beta_I = w_i(\pi^*(\mathbf{s}^*))$. \square

The consequence of Theorem 2 can be verified by Algorithm 2 and the numerical examples illustrated in Figure 6 (a) and (b). As is proved by Theorem 1 and Theorem 2, the capacity of a cluster DSS with given node parameters and storage/repair parameters is $MC(\mathbf{s}^*, \pi^*(\mathbf{s}^*))$. Based on (1), the tradeoff between node storage and repair bandwidth can be characterized, which is illustrated in Section IV for specific numerical parameters.

C. The Min-cuts of IFGs with separate selected nodes

As is proved in last subsection, the horizontal selection algorithm and the vertical order algorithm will generate the selected node distribution \mathbf{s}^* and the corresponding cluster order $\pi^*(\mathbf{s}^*)$, achieving the minimum min-cut of the IFGs without separate selected nodes. In this subsection, we analyse the min-cut of IFGs with one separate selected nodes in Theorem 3, namely $s_0 = 1$ in \mathbf{s} .

Note that Theorem 1 focuses on the influence of cluster order π where the selected node distribution \mathbf{s} is fixed. On the other hand, Theorem 2 analyses different selected node distributions while the cluster order generating algorithm is fixed. The following theorem combines these two aspects and investigates the situation where there is one separate selected node.

Theorem 3. For the given node parameters (n, k, L, R, S) , assume the j -th selected node is a separate node, then the selected node distribution \mathbf{s}^* generated by the horizontal selection algorithm with cluster order $\pi^*(\mathbf{s}^*)$ generated by the vertical order algorithm minimizes the min-cut of all the IFGs. In other words,

$$MC(\mathbf{s}^*, \pi^*(\mathbf{s}^*)) \leq MC(\mathbf{s}, \pi),$$

for all $s \in \mathcal{S}$ with $s_0 = 1$ and $\pi \in \Pi(s)$ with $\pi_j = 0$. Note that $\pi^*(\cdot)$ is defined by (8). $MC(s, \pi)$ is defined by (5).

Due to space limitation, here we just sketch the proof of Theorem 3 in the following part.

Cluster order assignment: Let $\bar{\pi} = (\bar{\pi}_1, \bar{\pi}_1, \dots, \bar{\pi}_k)$ denote a cluster order with no separate selected nodes in its corresponding selected node distribution \bar{s} . When the j -th selected node is a separate node, let π denote the new cluster order with

$$\pi_i = \begin{cases} \bar{\pi}_i, & \text{if } 1 \leq i < j \\ 0, & \text{if } i = j \\ \bar{\pi}_{i-1}, & \text{if } j < i \leq k \end{cases}.$$

As is illustrated in Figure 7 (b) (c), node 3 is a separate selected node in π , then $\pi_1 = \bar{\pi}_1$, $\pi_2 = \bar{\pi}_2$, $\pi_3 = 0$ and $\pi_i = \bar{\pi}_{i-1}$ ($i = 4, 5, 6, 7, 8$). In Figure 7 (a), the optimal selected node distribution is $s^* = (1, 4, 3, 0)$ and the optimal cluster order is $\pi^* = (1, 2, 0, 1, 2, 1, 2, 1)$ generated by the vertical order algorithm. Note that the 3rd node in the cluster order is a separate node, which is fixed beforehand.

Main idea of the proof: For any cluster order π whose j -th node is a separate node, there always exists a cluster order $\bar{\pi}$ with no separate selected nodes satisfying the definition of π . Through investigating the relationship between $\bar{\pi}$ and π , the proof of Theorem 3 can be reduced to similar problems of Theorem 1 and Theorem 2, which is omitted here.

To analyse the minimum min-cut of IFGs with one separate selected node, the location of the separate selected node (the value of j in Theorem 3) is also important. Moreover, the relation among β_I , β_C and β_S need to be taken into consideration. The situation with multiple separate selected nodes can be investigated using similar methods, which is not introduced here due to the space limitation.

IV. NUMERICAL RESULTS AND CODE CONSTRUCTIONS FOR CLUSTER DSSS

In this section, Figure 8 illuminates some numerical capacity bounds for cluster DSSs without separate nodes. As is mentioned in [13], [17], interference alignment is an important method in regenerating code constructions, which is also applicative in cluster DSSs. A code construction strategy with interference alignment is investigated for a cluster DSS with specific parameters as an example (see Figure 9 and Figure 10). The code constructions for general cases can utilize similar methods.

Model configurations: Assume the node parameters are $(n = 12, k = 8, L = 3, R = 4, S = 0)$ as Figure 6 shows. Note that different relations between β_I and β_C result in different tradeoffs between storage per node α and bandwidth to repair one node. We consider a specific situation that $\beta_I = 2\beta_C$. As is proved before, the cluster order illustrated in Figure 6 (a) achieves the capacity of this DSS. Based on the bound constraint in (1), for specific values of $k - R + 1 \leq d_C \leq n - k$, the tradeoff between α and β_C is illuminated in Figure 8, where the file size \mathcal{M} is set to be 32 for simplicity.

As is illuminated in Figure 8, the tradeoff curve moves left as the cross-cluster helper nodes increase. Since α is the storage per node and β_C corresponds to the repair bandwidth, when the storage per node is fixed, the more helper nodes are utilized, the less bandwidth will be, which is consistent with the consequences of DSS without

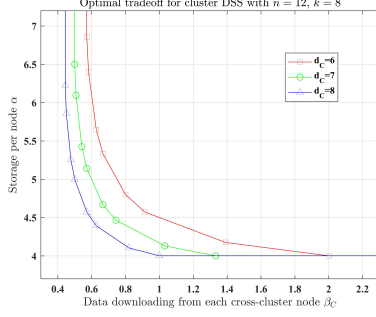


Fig. 8: Optimal tradeoff curves between storage per node α and data downloading from each cross-cluster node β_C , for cluster DSS with $n = 12, k = 8$ and $\beta_I = 2\beta_C$. There are three curves for different number of cross-cluster helper nodes, namely, $d_C = 6, 7, 8$, respectively. The original file size is $\mathcal{M} = 32$.

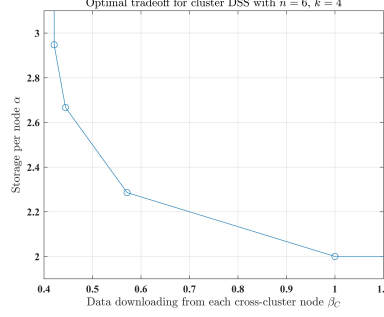


Fig. 9: Optimal tradeoff curves between storage per node α and data downloading from each cross-cluster node β_C , for cluster DSS with $n = 6, k = 4$ and $\beta_I = 2\beta_C, d_C = 3$. The original file size is $\mathcal{M} = 8$.

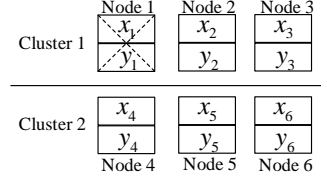


Fig. 10: MSR construction for cluster DSS with parameters $n = 6, k = 4, \beta_I = 2\beta_C, d_C = 3$ and $\mathcal{M} = 8$.

clusters in [5]. When $d_C = 8$, the point $(\alpha = 4, \beta_C = 2)$ achieves the minimum storage and the corresponding code constructions is called minimum-storage regenerating (MSR) codes, as is defined in [5]. We will investigate an MSR construction for cluster DSS with less nodes as an example.

Consider the cluster DSS with $n = 6, k = 4$ in Figure 10. It can be verified in Figure 9 that the MSR point of this system is $(\alpha = 2, \beta_C = 1)$, hence the amount of data downloading from each intra-cluster node is $\beta_I = 2\beta_C = 2$. **Encoding procedure:** As the original file size is $\mathcal{M} = 8$, we assume $x_i (1 \leq i \leq 4)$ and $y_i (1 \leq i \leq 4)$ are the original file symbols storing from Node 1 to Node 4 as Figure 10 illustrates. Two $(6,4)$ -MDS codes are used to encode symbols (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4) , respectively. Let

$$(x_1, x_2, x_3, x_4, x_5, x_6) = (x_1, x_2, x_3, x_4) [\mathbf{I}_{4 \times 4} | \mathbf{g} \ \mathbf{h}] \text{ and } (y_1, y_2, y_3, y_4, y_5, y_6) = (y_1, y_2, y_3, y_4) [\mathbf{I}_{4 \times 4} | \mathbf{g}' \ \mathbf{h}'], \quad (19)$$

where $\mathbf{I}_{4 \times 4}$ is an identity matrix and $\mathbf{g} = (g_1, g_2, g_3, g_4)^t$, $\mathbf{h} = (h_1, h_2, h_3, h_4)^t$, $\mathbf{g}' = (g'_1, g'_2, g'_3, g'_4)^t$, $\mathbf{h}' = (h'_1, h'_2, h'_3, h'_4)^t$. Then

$$\begin{aligned} x_5 &= g_1 x_1 + g_2 x_2 + g_3 x_3 + g_4 x_4, & y_5 &= g'_1 y_1 + g'_2 y_2 + g'_3 y_3 + g'_4 y_4, \\ x_6 &= h_1 x_1 + h_2 x_2 + h_3 x_3 + h_4 x_4, & y_6 &= h'_1 y_1 + h'_2 y_2 + h'_3 y_3 + h'_4 y_4. \end{aligned}$$

Repair procedure: Assume Node 1 has failed, based on the tradeoff in Figure 9 ($\alpha = 2$ and $\beta_C = 1$, $\beta_I = 2\beta_C = 2$), we will download 2 symbols each from Node 2 and Node 3 respectively and download 1 symbols each from Node 4 to Node 6. As the values of x_2, y_2, x_3, y_3 is known by downloading from Node 2 and Node 3, to calculate x_1 and y_1 , interference alignment can be used to eliminate x_4 and y_4 . For example, the symbols downloading from

Node 4, Node 5 and Node 6 are

$$symbol_4 = l_4x_4 + l'_4y_4,$$

$$symbol_5 = m_5x_5 + m'_5y_5 = m_5(g_1x_1 + g_2x_2 + g_3x_3 + g_4x_4) + m'_5(g'_1y_1 + g'_2y_2 + g'_3y_3 + g'_4y_4),$$

$$symbol_6 = n_6x_6 + n'_6y_6 = n_6(h_1x_1 + h_2x_2 + h_3x_3 + h_4x_4) + n'_6(h'_1y_1 + h'_2y_2 + h'_3y_3 + h'_4y_4),$$

respectively. Hence,

$$symbol_4 = l_4x_4 + l'_4y_4, \quad (20)$$

$$symbol_5 - m_5(g_2x_2 + g_3x_3) - m'_5(g'_2y_2 + g'_3y_3) = m_5g_1x_1 + m'_5g'_1y_1 + m_5g_4x_4 + m'_5g'_4y_4, \quad (21)$$

$$symbol_6 - n_6(h_2x_2 + h_3x_3) - n'_6(h'_2y_2 + h'_3y_3) = n_6h_1x_1 + n'_6h'_1y_1 + n_6h_4x_4 + n'_6h'_4y_4. \quad (22)$$

Note that the left parts of Eq. 20, Eq. 21 and Eq. 22 are real values. If the coefficients of x_1, y_1 and x_4, y_4 satisfy that

$$\text{rank} \left(\begin{bmatrix} m_5g_1 & m'_5g'_1 \\ n_6h_1 & n'_6h'_1 \end{bmatrix} \right) = 2 \text{ and } \text{rank} \left(\begin{bmatrix} l_4 & l'_4 \\ m_5g_4 & m'_5g'_4 \\ n_6h_4 & n'_6h'_4 \end{bmatrix} \right) = 1, \quad (23)$$

x_4 and y_4 can be eliminated, meanwhile, x_1 and y_1 can be solved, finishing the repair of Node 1.

As is proved in [13], there exists MDS codes satisfying the condition (23). When other nodes have failed, similar methods can be utilized to generate the parameters of corresponding MDS codes. The construction of MDS codes adaptive to all the node failures and cluster DSS with general parameters is more complicated and need more future work.

V. CONCLUSION AND FUTURE WORK

In this paper, DSSs with clusters and separate nodes are investigated, where the tradeoff between node storage and repair bandwidth is characterized on more flexible parameter settings. When a node in cluster DSSs has failed, the number of helper nodes varies based on the practical storage system demands. The influence of separate nodes is also analysed for DSSs with clusters and separated nodes. Moreover, a regenerating code construction strategy is proposed for cluster DSSs with specific parameters as a numerical example, achieving the points in the optimal tradeoff curve.

More general and practical regenerating codes for cluster DSSs need to be investigated further. On the other hand, more works are needed to characterize the influence of separate selected nodes for the min-cuts of CSN-DSSs more explicitly, when analysing the optimal tradeoff between node storage and repair bandwidth. Furthermore, the constructions of more flexible regenerating codes for CSN-DSSs are also meaningful for practical storage systems.

REFERENCES

- [1] W. Jiang, C. Hu, Y. Zhou, and A. Kanevsky, "Are disks the dominant contributor for storage failures?: A comprehensive study of storage subsystem failure characteristics," *Trans. Storage*, vol. 4, no. 3, pp. 7:1–7:25, Nov. 2008. [Online]. Available: <http://doi.acm.org/10.1145/1416944.1416946>
- [2] H. Zhang, H. Li, and S. Y. R. Li, "Repair tree: Fast repair for single failure in erasure-coded distributed storage systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 6, pp. 1728–1739, June 2017.

- [3] J. Li and B. Li, "Beehive: Erasure codes for fixing multiple failures in distributed storage systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 5, pp. 1257–1270, May 2017.
- [4] C. Huang, M. Chen, and J. Li, "Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems," *Trans. Storage*, vol. 9, no. 1, pp. 3:1–3:28, Mar. 2013. [Online]. Available: <http://doi.acm.org/10.1145/2435204.2435207>
- [5] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4539–4551, Sept 2010.
- [6] T. Ernvall, S. E. Rouayheb, C. Hollanti, and H. V. Poor, "Capacity and security of heterogeneous distributed storage systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 2701–2709, December 2013.
- [7] Q. Yu, K. W. Shum, and C. W. Sung, "Tradeoff between storage cost and repair cost in heterogeneous distributed storage systems," *Transactions on Emerging Telecommunications Technologies*, vol. 26, no. 10, pp. 1201–1211, 2015.
- [8] S. Akhlaghi, A. Kiani, and M. R. Ghanavati, "Cost-bandwidth tradeoff in distributed storage systems," *Computer Communications*, vol. 33, no. 17, pp. 2105 – 2115, 2010, special Issue: Applied sciences in communication technologies. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0140366410003506>
- [9] D. Ford, F. Labelle, F. I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan, "Availability in globally distributed storage systems," in *OsdI*, vol. 10, 2010, pp. 1–7.
- [10] J. Pernas, C. Yuen, B. Gastn, and J. Pujol, "Non-homogeneous two-rack model for distributed storage systems," in *2013 IEEE International Symposium on Information Theory*, July 2013, pp. 1237–1241.
- [11] N. Prakash, V. Abdrashitov, and M. Médard, "The storage vs repair-bandwidth trade-off for clustered storage systems," *CoRR*, vol. abs/1701.04909, 2017. [Online]. Available: <http://arxiv.org/abs/1701.04909>
- [12] J. y. Sohn, B. Choi, S. W. Yoon, and J. Moon, "Capacity of clustered distributed storage," in *2017 IEEE International Conference on Communications (ICC)*, May 2017.
- [13] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," *Proceedings of the IEEE*, vol. 99, no. 3, pp. 476–489, March 2011.
- [14] S. Y. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Trans. Inf. Theor.*, vol. 49, no. 2, pp. 371–381, Feb. 2003. [Online]. Available: <http://dx.doi.org/10.1109/TIT.2002.807285>
- [15] R. Ahlswede, N. Cai, S. Y. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theor.*, vol. 46, no. 4, pp. 1204–1216, Sep. 2006. [Online]. Available: <http://dx.doi.org/10.1109/18.850663>
- [16] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, 1st ed. Prentice Hall, 1993.
- [17] A. Fazeli, S. Goparaju, and A. Vardy, "Minimum storage regenerating codes for all parameters," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 76–80.