# On Extremal Rates of Storage over Graphs

Zhou Li and Hua Sun
EE Department, University of North Texas
*Email: zhouli@my.unt.edu, hua.sun@unt.edu*

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206619

***Abstract*—A storage code over a graph maps $K$ independent source symbols, each of $L_w$ bits, to $N$ coded symbols, each of $L_v$ bits, such that each coded symbol is stored in a node of the graph and each edge of the graph is associated with one source symbol. From a pair of nodes connected by an edge, the source symbol that is associated with the edge can be decoded. The ratio $L_w/L_v$ is called the symbol rate of a storage code and the highest symbol rate is called the capacity. We show that the three highest capacity values of storage codes over graphs are $2, 3/2, 4/3$. We characterize all graphs over which the storage code capacity is $2$ and $3/2$, and for capacity value of $4/3$, necessary condition and sufficient condition (that do not match) on the graphs are given.**

## I. Introduction

Motivated by the heterogeneity of modern distributed storage systems, a storage code problem over graphs is introduced in [1], [2], where a storage code maps $K$ independent source symbols, $W_1, \cdots, W_K$ to $N$ coded symbols, $V_1, \cdots, V_N$, and the coded symbols are stored in the node set of a graph $\{V_1, \cdots, V_N\}$ (so that $V_n$ denotes both the coded symbol and the node). The heterogeneous data recovery pattern is captured by the edges of the graph, where each edge $\{V_i, V_j\}$ is associated with one source symbol $W_k$ and from $(V_i, V_j)$, we can decode $W_k$. As the structure of the graph can be very diverse, versatile distributed storage and data access requirements can be accommodated. An example of the storage code problem over a graph is given in Fig. 1. The metric of pursuit is the capacity $C$ of a storage code over a graph, i.e., the highest possible symbol rate, defined as $L_w/L_v$, where $L_w(L_v)$ is the number of bits contained in each source (coded) symbol and $L_w/L_v$ represents the number of source symbol bits reliably stored in each coded symbol bit.

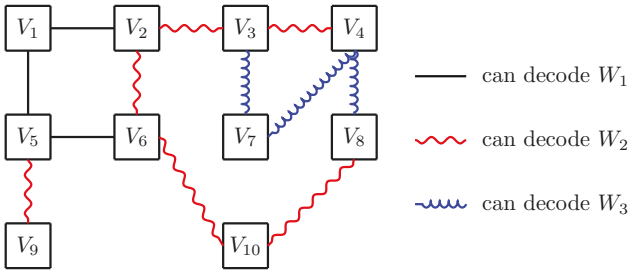


Fig. 1: An example graph of a storage code problem with $K = 3$ source symbols and $N = 10$ coded symbols.

The graph based storage code problem is not new in the sense that it can be equivalently transformed to a network coding problem [1]–[4] and adding further security constraints (i.e., beyond desired data decodability, leakage about other source symbols is prevented), it is intimately related to conditional disclosure of secrets [5]–[8] and secret sharing [9], [10]. What is new is the view brought by [1] - finding extremal networks/graphs. Instead of first fixing the network/graph and then finding its highest rate, we focus on the extremal (highest) capacity values and aim to find the networks/graphs whose capacity is equal to the extremal values (see Fig. 2). This complementary view is useful in identifying critical combinatorial graph structures that limit the rate and in separating more tractable graph classes in terms of capacity characterization. Considering that networks are becoming more and more heterogeneous and solving each network instance becomes infeasible and impossible (as hard instances that require non-linear codes for achievability or non-Shannon information inequalities for converse are well known [11]–[13]), this extremal rate (network) approach might be a fruitful direction to produce new results and insights.

In this work, we start from the highest possible capacity values and for the two highest rates - 2 and $3/2$, all extremal graphs with corresponding extremal capacity values are easily characterized. For extremal rate of 2, absolute no interference is allowed as $L_w = 2L_v$, i.e., a pair of nodes can just store the desired source symbols. As long as there exists interference, the maximal capacity value drops to $3/2$, the next extremal rate, and all storage code instances with capacity $3/2$ only require intra-source symbol coding, i.e., mixing of symbols from the same source symbol. When rate of $3/2$ cannot be achieved, the next highest capacity value is shown to be $4/3$, which is our main focus and the corresponding graphs turn out to be highly technical. We identify necessary condition (converse required) and sufficient condition (achievability provided) for graphs with storage code capacity $4/3$ (see Fig. 2). The converse is based on delicate arguments on the intimate relation between the maximum amount of interference (undesired source symbols) allowed and the minimum amount of desired source symbols needed. The achievable scheme uses vector linear codes that carefully control the alignment of interfering source symbols and the independence of desired source symbols. The conditions are stated in terms of the presence (absence) of critical nodes (edges) of the graph, whose combinatorial structure constrains the code rate.

## II. Problem Statement and Definitions

Consider $K$ independent uniform source symbols $W_1, \cdots, W_K$ of size $L_w$ bits each.

$$H(W_1, \cdots, W_K) = H(W_1) + \cdots + H(W_K),$$

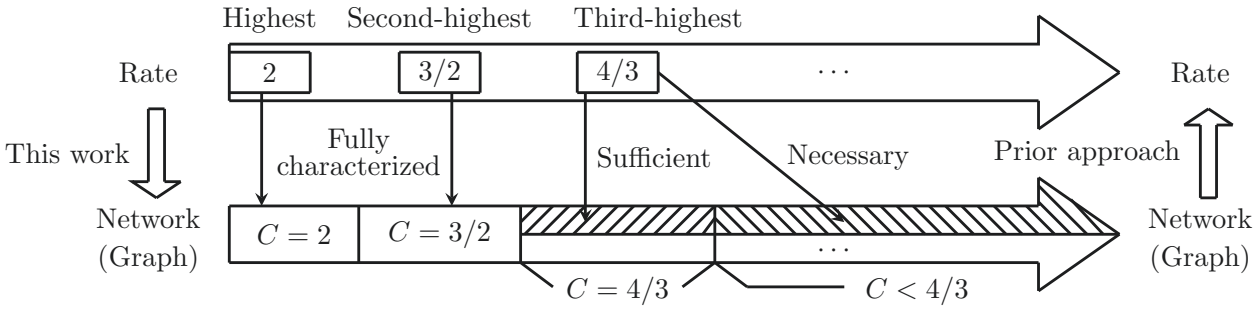


Fig. 2: The extremal rate and network approach of this work and results obtained.

$$L_w = H(W_1) = \cdots = H(W_K). \tag{1}$$

Consider $N$ coded symbols $V_1, \cdots, V_N$, each of $L_v$ bits. Our interest lies in the relative size of $L_w, L_v$ (see (3)) and coding over arbitrary finite fields is allowed, so $L_w, L_v$ can take arbitrarily large values (that are not necessarily integers).

The source symbol recoverability constraint on the coded symbols is specified by a graph $G = (\mathcal{V}, \mathcal{E}, t)$, where the node[1] set $\mathcal{V} = \{V_1, \cdots, V_N\}$, the edge set $\mathcal{E}$ is a set of unordered pairs from $\mathcal{V}$, and the function $t$ associates each edge $\{V_i, V_j\} \in \mathcal{E}$ with a source symbol $W_k, k \in \{1, 2, \cdots, K\} \triangleq [K]$, i.e., $t(\{V_i, V_j\}) = W_k$. For each edge $\{V_i, V_j\} \in \mathcal{E}$ such that $t(\{V_i, V_j\}) = W_k$, we can decode $W_k$ with no error, i.e.,

$$H(W_k|V_i, V_j) = 0 \text{ if } t(\{V_i, V_j\}) = W_k. \tag{2}$$

Isolated nodes are trivial as they are not connected to any edges and thus involve no constraints. Without loss of generality, we assume in this work that any graph contains no isolated nodes.

A mapping from the source symbols $W_1, \cdots, W_K$ to the coded symbols $V_1, \cdots, V_N$ that satisfies the decoding constraint (2) specified by a graph $G = (\mathcal{V}, \mathcal{E}, t)$ is called a storage code. The (achievable) symbol rate is defined as

$$R \triangleq \frac{L_w}{L_v} \tag{3}$$

whose supremum is called the capacity, $C \triangleq \sup_{L_w} L_w/L_v = \lim_{L_w \to \infty} L_w/L_v$, as block codes are allowed.

Next we introduce some graph definitions to facilitate the presentation of our results.

### A. Graph Definitions

*Definition 1 ($W_k$-Edge, $W_k$-Path, and $W_k$-Component):* An edge that is associated with $W_k$ is called a $W_k$-edge. A sequence of distinct connecting $W_k$-edges is called a $W_k$-path. A $W_k$-component is a maximal subgraph wherein every edge is a $W_k$-edge and every two nodes are connected by a $W_k$-path (an isolated node is defined as a trivial component).

For example, in Fig. 1, $\{V_1, V_2\}$ (also all solid black edges) is a $W_1$-edge; the sequence of $W_1$-edges $(\{V_2, V_1\}, \{V_1, V_5\}, \{V_5, V_6\})$ is a $W_1$-path and also a $W_1$-component.

[1]Note that we abuse the notation by using $V_n$ to denote both a coded symbol and a node of the graph, which will not cause confusion.

*Definition 2 (Internal Edge and Residing Path):* A $W_k$-edge that connects two nodes (say $V_i, V_j$) in a $W_{k'}$-path, $k' \neq k$ is said to be internal and the $W_{k'}$-path with end nodes $V_i, V_j$ is called the residing path of the internal $W_k$-edge $\{V_i, V_j\}$.

For example, in Fig. 1, the $W_2$-edge $\{V_2, V_6\}$ is an internal edge as it connects two nodes $V_2, V_6$ in the $W_1$-path $(\{V_2, V_1\}, \{V_1, V_5\}, \{V_5, V_6\})$, which is then its residing path.

*Definition 3 ($M$-Color Node):* A node whose connected edges are associated with $M$ different source symbols is called an $M$-color node.

For example, in Fig. 1, $V_1, V_9$ are 1-color nodes and $V_5, V_6$ are 2-color nodes.

We need to further distinguish two types of 2-color nodes, defined as follows.

*Definition 4 (Normal 2-Color Node and $W_k$-Special 2-Color Node):* For a 2-color node $V$ that is connected to $W_k$-edges and $W_{k'}$-edges, $k \neq k'$, if the nodes connected to $V$ through $W_k$-edges are all 1-color, then $V$ is called a $W_k$-special 2-color node (or just a special 2-color node when $W_k$ does not need to be highlighted). A 2-color node that is not special is said to be normal.

For example, in Fig. 1, the 2-color node $V_5$ is $W_2$-special as $V_9$ is the only node that is connected to $V_5$ through $W_2$-edges and $V_9$ is 1-color; the 2-color node $V_6$ is normal as it is connected to a 2-color node $V_2$ through a $W_2$-edge and is connected to a 2-color node $V_5$ through a $W_1$-edge.

*Definition 5 (Graph Class $\mathcal{G}_{C=R^*}, \mathcal{G}_{C\geq R^*}, \mathcal{G}_{C<R^*}$):* The set of graphs whose storage code capacity is equal to\no smaller than\strictly smaller than $R^*$ is denoted by $\mathcal{G}_{C=R^*} \backslash \mathcal{G}_{C\geq R^*} \backslash \mathcal{G}_{C<R^*}$.

## III. Results

Our results are presented in this section, along with illustrative examples and observations.

### A. Extremal Graphs with Storage Code Capacity $2, 3/2$: $\mathcal{G}_{C=2}$, $\mathcal{G}_{C=3/2}$

The three highest extremal capacity values and the full extremal graph characterization for the two highest extremal capacity values are established in the following theorem.

*Theorem 1:* $[\mathcal{G}_{C=2}, \mathcal{G}_{C=3/2}]$ The three highest storage code capacity values are $2, 3/2, 4/3$. The storage code capacity of a graph is equal to 2 ($G \in \mathcal{G}_{C=2}$) if and only if every node

is 1-color. The storage code capacity of a graph is equal to $3/2$ ($G \in \mathcal{G}_{C=3/2}$) if and only if $(a)$ there exists a 2-color node and all nodes are 1-color or 2-color and $(b)$ there are no connected 2-color nodes.

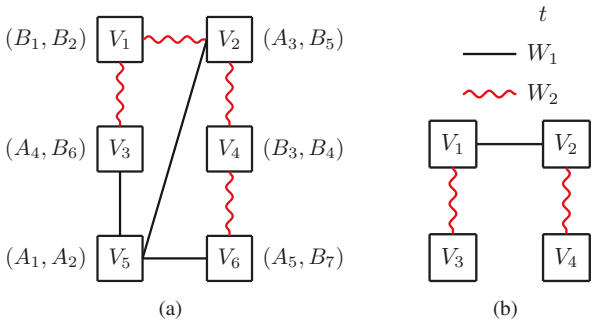


Fig. 3: (a) An example graph $G \in \mathcal{G}_{C=3/2}$. $W_1 = (a_1, a_2, a_3), W_2 = (b_1, b_2, b_3)$, and each $A_i \backslash B_j$ is a generic linear combination of $(a_1, a_2, a_3) \backslash (b_1, b_2, b_3)$. (b) An example graph $G \in \mathcal{G}_{C<3/2}$ where two 2-color nodes $V_1, V_2$ are connected.

The proof of Theorem 1 is fairly straightforward and is deferred to the full version of this work [14]. An example of the achievable scheme (code construction) for $G \in \mathcal{G}_{C=3/2}$ is shown in Fig. 3.(a). An example graph that does not belong to $\mathcal{G}_{C=2} \cup \mathcal{G}_{C=3/2}$ is shown in Fig. 3.(b). An intuitive explanation on why the rate is upper bounded by $4/3$ is as follows. $V_3$ can at most contribute $L_v$ bits of information about $W_2$. $\{V_1, V_3\}$ is a $W_2$-edge so that $V_1$ has to provide at least the remaining $L_w - L_v$ bits of information about $W_2$, leaving at most $L_v - (L_w - L_v) = 2L_v - L_w$ bits of room for $W_1$. The same reasoning applies to $V_2$. Finally, $\{V_1, V_2\}$ is a $W_1$-edge so that the size of the remaining room must accommodate the $L_w$ bits of $W_1$, i.e., $2(2L_v - L_w) \geq L_w$ so that $R = L_w/L_v \leq 4/3$.

### *B. Extremal Graphs with Capacity $4/3$: $\mathcal{G}_{C=4/3}$ with $K = 2$ Source Symbols*

Next we focus on the storage code capacity value of $4/3$, whose extremal graph characterization turns out to be highly non-trivial. In this work, we exclusively consider the cases where there are $K = 2$ source symbols to illustrate the results in a simpler setting while noting that generalizations to more than 2 source symbols are possible and deferred to the full version of this work [14].

The obtained necessary and sufficient conditions are rather involved. To make the results more clear we give a summarizing chart in Fig. 4.

*1) Sufficient Condition: Internal Edge and 1-Color Node:* A crucial graphic structure for the achievability of rate $4/3$ is the absence of internal edges (or when they exist, the presence of 1-color nodes in their residing paths).

*Theorem 2:* [Sufficient Condition of $\mathcal{G}_{C=4/3}$] With $K = 2$ source symbols, a graph $G \in \mathcal{G}_{C \geq 4/3}$ if $G$ contains no internal edge or for any internal edge, its residing path contains a 1-color node.

The proof of Theorem 2 is presented in [14]. To illustrate the idea, two examples are shown in Fig. 5, where Example $(a)$

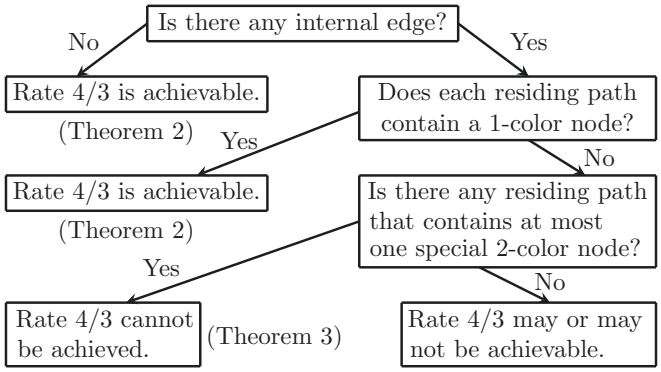


Fig. 4: A summary of sufficient and necessary conditions of $\mathcal{G}_{C=4/3}$ with $K = 2$.

contains no internal edge; Example $(b)$ contains two internal edges $\{V_2, V_3\}$ and $\{V_3, V_5\}$. Internal $W_2$-edge $\{V_2, V_3\}$ resides in $W_1$-path $(\{V_2, V_1\}, \{V_1, V_3\})$, which contains 1-color node $V_1$ and internal $W_1$-edge $\{V_3, V_5\}$ resides in $W_2$-path $(\{V_3, V_2\}, \{V_2, V_4\}, \{V_4, V_5\})$, which contains 1-color node $V_4$. So the condition of Theorem 2 is satisfied and rate $4/3$ is achievable. We next explain how to construct the code.

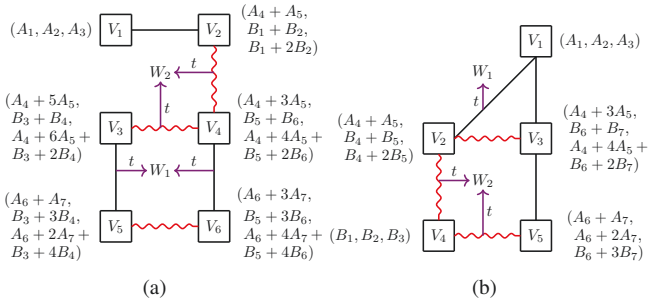


Fig. 5: Two example graphs $G \in \mathcal{G}_{C \geq 4/3}$ and code constructions for rate $4/3$. $W_1 = (a_1, a_2, a_3, a_4), W_2 = (b_1, b_2, b_3, b_4)$ and each $A_i \backslash B_j$ is a generic linear combination of $(a_1, \cdots, a_4) \backslash (b_1, \cdots, b_4)$.

We are targeting at rate $L_w/L_v = 4/3$ so that any pair of nodes connected by an edge contain $2L_v = 3L_w/2$ bits. Except from $L_w$ bits from the desired source, at most we can tolerate $2L_v - L_w = L_w/2$ undesired bits (i.e., interference). Then the key is to guarantee for any $W_k$-edge, $k \in \{1, 2\}$, the interference from $W_{3-k}$ has at most half source size. That is, $W_{3-k}$ symbols shall be assigned according to $W_k$-edges ($W_k$-components). When there is no internal edge (or residing path contains 1-color nodes), such interference based assignment automatically ensures the independence (thus decodability) of desired source symbols. We now come back to the examples in Fig. 5 to see how to implement the above code design idea.

Consider Example $(a)$ first and Example $(b)$ will follow similarly. We set $L_w / \log_2 p = 4$ so that $W_1 = (a_1, a_2, a_3, a_4)$ and $W_2 = (b_1, b_2, b_3, b_4)$, where each symbol is from a sufficiently large finite field $\mathbb{F}_p$ (the exact field size can be found in the general proof in [14]). To achieve rate $R = L_w/L_v = 4/3$, we set $L_v = 3 \log_2 p$, i.e., each $V_n$ contains three symbols

from the same field. We generate a number of generic linear combinations of $(a_1, \cdots, a_4) \backslash (b_1, \cdots, b_4)$ and denote them as $(A_1, A_2, \cdots) \backslash (B_1, B_2, \cdots)$. For now, it suffices to view each $A_i \backslash B_j$ as a random linear combination of symbols from $W_1 \backslash W_2$ and if we can collect four linearly independent combinations of $A_i \backslash B_j$, then we can recover $W_1 \backslash W_2$. The detailed randomized construction is again deferred to the general proof. Each one of the three symbols in $V_n$ will be a linear combination of some $A_i$ and $B_j$ symbols. We first assign the $A_i$ and $B_j$ symbols in each $V_n$ and then linearly combine them to produce the final three symbols in $V_n$.

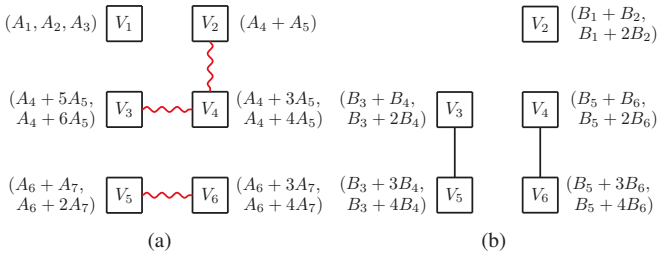


Fig. 6: (a) $W_2$-component decomposition of $W_1$-connected nodes in Fig. 5.(a), according to which $A_i$ symbols are assigned. (b) $W_1$-component decomposition of $W_2$-connected nodes in Fig. 5.(a), according to which $B_j$ symbols are assigned.

Consider nodes that are connected to $W_1$-edges so that some $A_i$ symbols need to be assigned, i.e., all nodes $V_1, \cdots, V_6$. The 1-color nodes are trivial (i.e., $V_1$), and we just assign three distinct $A_i$ symbols. Next, consider the remaining 2-color nodes $V_2, \cdots, V_6$ for which the $A_i$ symbols are assigned according to $W_2$-components (see Fig. 6.(a)). $V_2, \cdots, V_6$ form two $W_2$-components - one consists of $V_2, V_3, V_4$ and the other consists of $V_5, V_6$. For each $W_2$-component, we assign generic linear combinations of the *same* $2 = \frac{1}{2} L_w / \log_2 p$ $A_i$ symbols (say $A_{i_1}, A_{i_2}$) so that the interference dimension is limited to two. Further, a normal 2-color node and a $W_2$-special 2-color node will get two generic linear combinations of $(A_{i_1}, A_{i_2})$ and a $W_1$-special 2-color node will get one generic linear combination of $(A_{i_1}, A_{i_2})$. For example, consider $W_2$-component with nodes $V_2, V_3, V_4$, where the $A_i$ symbols appeared are limited to $A_4, A_5$; $V_2$, as a $W_1$-special 2-color node, gets one combination $A_4 + A_5$ and $V_3, V_4$, as normal 2-color nodes, each gets two generic combinations (e.g., $V_3$ gets $A_4 + 5A_5, A_4 + 6A_5$). The other $W_2$-component with nodes $V_5, V_6$ is assigned similarly - the $A_i$ symbols are limited to $A_6, A_7$.

The assignment for nodes connected to $W_2$-edges is exactly the same (see Fig. 6.(b)). Nodes $V_2, \cdots, V_6$ are connected to $W_2$-edges and they are all 2-color. The $B_j$ symbols are assigned according to $W_1$-components, i.e., $V_2$ (as a single-node component) gets generic linear combinations of $B_1, B_2$; $V_3, V_5$ form a $W_1$-component and get generic linear combinations of $B_3, B_4$; $V_4, V_6$ form a $W_1$-component and the $B_j$ symbols are limited to $B_5, B_6$.

The last step is to combine the $A_i, B_j$ symbols so that each $V_n$ has only three symbols. This step is simple, if a node gets at most three $A_i, B_j$ symbols, then just set them as $V_n$ (e.g., $V_1, V_2$); otherwise the node must be normal 2-color, which gets two generic combinations of $A_i$ and two generic combinations of $B_j$ and we just add one arbitrary combination (say the last) of $A_i$ and $B_j$ together to reduce the total number of symbols to three (e.g., $V_3, V_4, V_5, V_6$).

Finally, let us verify why the decoding constraints (2) are satisfied. An edge that contains 1-color node is straightforward, e.g., from $W_1$-edge $\{V_1, V_2\}$, we have $A_1, A_2, A_3, A_4 + A_5$, so as long as the $A_i$ combinations are generic we can recover $W_1 = (a_1, \cdots, a_4)$. For edges that connect two 2-color nodes (e.g., $W_2$-edge $\{V_3, V_4\}$), we have 1) the interference dimension is limited to two as our assignment is based on components of interfering sources (e.g., we may decode $A_4, A_5$ and remove them, leaving us with only $B_j$ symbols); 2) the four symbols from the desired source have full rank (e.g., $B_3, B_4, B_5, B_6$ are generic combinations) so that we can recover the desired source symbol. Note that because there is no internal edge, for any $W_k$-edge, the two nodes obtain distinct desired $W_k$ symbols, e.g., for $W_2$-edge $\{V_3, V_4\}$, $V_3$ is assigned $B_3, B_4$ symbols and $V_4$ is assigned $B_5, B_6$ symbols as $V_3, V_4$ belong to distinct $W_1$-components (refer to Fig. 6.(b)). If $V_3, V_4$ belong to the same $W_1$-component, then the $W_2$-edge $\{V_3, V_4\}$ will be internal).

The code construction for Example $(b)$ in Fig. 5 follows from the same procedure as that of Example $(a)$. That is, first consider 1-color nodes and assign generic combinations (e.g., $V_1, V_4$); for remaining 2-color nodes, assign $W_k$ symbols according to $W_{3-k}$-components (e.g., the $W_1$ space of the $W_2$-edge $\{V_2, V_3\}$ is spanned by $A_4, A_5$, and the $W_2$ space of the $W_1$-edge $\{V_3, V_5\}$ is spanned by $B_6, B_7$); finally combine the four symbols to three for normal 2-color nodes (e.g., $V_3$). The decoding constraints (2) are easily verified as the interference dimension is strictly controlled and desired source symbols are sufficiently generic because after removing 1-color nodes, there no longer exist internal edges.

*2) Necessary Condition: Residing Path and Special 2-Color Node:* The sufficient condition of the achievability of rate $4/3$ in Theorem 2 requires the absence of internal edges or the presence of 1-color node in residing paths. Considering the complementary cases, we identify a crucial graphic structure for the unachievability of rate $4/3$ - the presence of at most one special 2-color node in a residing path.

*Theorem 3:* [Necessary Condition of $\mathcal{G}_{C=4/3}$] With $K = 2$ source symbols, a graph $G \in \mathcal{G}_{C<4/3}$ if $G$ has a residing path which contains no 1-color node and at most one special 2-color node.

The proof of Theorem 3 is presented in [14]. To illustrate the idea, an example is shown in Fig. 7, where the internal $W_2$-edge $\{V_1, V_2\}$ resides in the $W_1$-path $(\{V_1, V_3\}, \{V_3, V_4\}, \{V_4, V_2\})$ and this residing path contains only one special 2-color node $V_3$ and no 1-color node. So the condition of Theorem 3 is satisfied and rate $4/3$ cannot be achieved. To see why, we next give an intuitive explanation.

Suppose rate $4/3$ is achievable, i.e., $L_w/L_v = 4/3$. Then we can show that for any 2-color node (e.g., $V_3$), it must

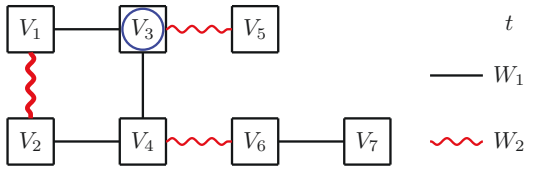


Fig. 7: An example graph $G \in \mathcal{G}_{C<4/3}$ where the internal edge $\{V_1, V_2\}$ is highlighted and the only special 2-color node $V_3$ in its residing path is highlighted.

contain at least $L_w/4$ bits of information about each of $W_1$ and $W_2$ (captured through conditional entropy. See Lemma 1 of [14]). This is because the connecting node can provide at most $L_v = 3L_w/4$ bits of information about the desired source symbol (e.g., $V_5$ can contribute $L_v = 3L_w/4$ bits on $W_2$ at most and the remaining $L_w - L_v = L_w/4$ bits must come from $V_3$). Further, if the 2-color node is normal (e.g., $V_4$), it must contain exactly $L_w/2$ bits of information about each of $W_1$ and $W_2$ (see Lemma 2 of [14]). The reason is that for two connecting 2-color nodes, the amount of interference allowed is at most $2L_v - L_w = L_w/2$ bits and a pair of nodes must contribute $L_w$ bits of information about the desired source symbol (thus $L_w/2$ from each node). For example, consider $W_1$-edge $\{V_2, V_4\}$, where from an interference view, $V_2$ can contain at most $L_w/2$ bits on $W_2$; from the desired source view, $V_2$ must also contribute at least $L_w/2$ bits on $W_2$ because of the $W_2$-edge $\{V_1, V_2\}$.

We now consider the propagation of interference through the residing $W_1$-path $(\{V_2, V_4\}, \{V_4, V_3\}, \{V_3, V_1\})$. Start from the normal 2-color node $V_2$, which contains $L_w/2$ bits on $W_2$ and as a $W_1$-edge can tolerate at most $L_w/2$ bits on $W_2$, then the normal 2-color node $V_4$ must contain the same $L_w/2$ bits on $W_2$ (see Lemma 3 of [14]). We are now at $V_4$ and continue the $W_1$-path through edge $\{V_3, V_4\}$, where $V_3$ is special so that $V_3$ contains at least $L_w/4$ bits on $W_2$ and this $L_w/4$ bits are contained in the total $L_w/2$ interference bits in $V_4$. Continue further the $W_1$-path through edge $\{V_3, V_1\}$, where the $L_w/4$ bits on $W_2$ in $V_3$ must be contained in the $L_w/2$ bits on $W_2$ in $V_1$. This in turn means that the $L_w/2$ bits on $W_2$ in $V_1$ must overlap with the $L_w/2$ bits on $W_2$ in $V_2$ (in the $L_w/4$ bits on $W_2$ in $V_3$), thus the internal $W_2$-edge $\{V_1, V_2\}$ cannot contribute $L_w/2+L_w/2 = L_w$ independent bits for the desired $W_2$ source and we have arrived at a contradiction.

From the above reasoning, we can now illuminate the role of special and normal 2-color nodes in a residing path. For an internal $W_k$-edge, its residing $W_{3-k}$-path made up of 2-color nodes must have two normal 2-color end nodes, each of which contains $L_w/2$ independent bits of information about the desired source $W_k$ (e.g., $V_1, V_2$ about $W_2$). In the residing $W_{3-k}$-path, a normal 2-color node will keep the interference on $W_k$ to the same $L_w/2$ dimensions (e.g., $V_2, V_4$ have the same $L_w/2$ dimensions about $W_2$ and $V_1, V_3$ have the same $L_w/2$ dimensions about $W_2$) while a special 2-color node will inherit at least $L_w/4$ interference dimensions on $W_k$ (e.g., $V_3$ gets at least $L_w/4$ dimensions of $W_2$ from $V_3$). Conversely, a special 2-color node in a residing $W_{3-k}$-path can change at most $L_w/4$ dimensions of the interference on $W_k$ (which is the desired source for the internal $W_k$-edge), so to ensure the independence of the desired source at the internal edge we need at least two special 2-color nodes in the residing path. Along this line, we can also see the role of 1-color node in a residing path, i.e., it completely stops the propagation of interference (see $V_4$ in the residing $W_2$-path $(\{V_3, V_2\}, \{V_2, V_4\}, \{V_4, V_5\})$ of Fig. 6.(b), where $V_3, V_5$ can hold independent $W_1$ bits although $\{V_3, V_5\}$ is internal).

## IV. Discussion

An extremal rate perspective is taken to study the storage code problem over graphs. For the highest capacity values, we have identified a number of combinatorial structures that have significant impact on the code rate - $M$-color code (i.e., the number of sources associated with a node), internal edge (which captures a direct conflict between alignment of undesired source symbols and independence of desired source symbols), normal 2-color node\special 2-color node (for rate $4/3$, which keeps the same interference\which could change interference up to the extent of $1/4$ source size). Both the achievability and converse results are guided by a linear dimension counting view. The sufficient and necessary conditions presented are not the largest that our proof technique can lead to, i.e., we can solve more graph instances, but a systematic description is still out of current reach. It is not clear which rates will turn out to have hard capacity instances. Specifically, all extremal graphs with storage code capacity $4/3$ appear to go beyond the techniques of this work. Regarding generalizations, we note that our model is the most elementary, where we have focused on the highest capacity values, i.e., best rate scenarios instead of lowest capacity values, i.e., worst rate scenarios, or other physically meaningful rates; decoding constraints are placed on a pair of nodes in this work instead of an arbitrary set of nodes, i.e., we may have a hypergraph rather than a graph [2]; each edge is associated with only one source symbols instead of multiple source symbols where the decoding structure can be more diverse [1]; each source (coded) symbol is assumed to have equal size instead of arbitrarily different sizes so that in this asymmetric (fully heterogeneous) setting, we may generalize extremal rate to extremal rate tuple (region). Finally, from an extremal rate and network perspective, we may view combinatorial objects using the metric of capacity and study further extremal (largest, densest, most (linearly) independent) graphs, set families, vector spaces etc. along the line of extremal combinatorics [15].

## Acknowledgement

This work is supported in part by NSF under Grant CCF-2007108 and Grant CCF-2045656.

## References

[1] Z. Li and H. Sun, "On Extremal Rates of Secure Storage over Graphs," *arXiv preprint arXiv:2204.06511*, 2022.

[2] S. Sahraei and M. Gastpar, "GDSP: A graphical perspective on the distributed storage systems," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2218–2222.
[3] R. W. Yeung, *Information Theory and Network Coding*. Springer, 2008.
[4] C. K. Ngai and R. W. Yeung, "Network coding gain of combination networks," in *Information Theory Workshop*. IEEE, 2004, pp. 283–287.
[5] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 151–160.
[6] Z. Li and H. Sun, "Conditional disclosure of secrets: A noise and signal alignment approach," *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 4052–4062, 2022.
[7] ——, "On the linear capacity of conditional disclosure of secrets," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 3202–3207.
[8] Z. Wang and S. Ulukus, "Communication cost of two-database symmetric private information retrieval: A conditional disclosure of multiple secrets perspective," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 402–407.
[9] E. F. Brickell and D. M. Davenport, "On the classification of ideal secret sharing schemes," *Journal of Cryptology*, vol. 4, no. 2, pp. 123–134, 1991.
[10] H.-M. Sun and S.-P. Shieh, "Secret sharing in graph-based prohibited structures," in *Proceedings of INFOCOM'97*, vol. 2. IEEE, 1997, pp. 718–724.
[11] R. Dougherty, C. Freiling, and K. Zeger, "Network coding and matroid theory," *Proc. IEEE*, vol. 99, no. 3, pp. 388 – 405, Mar. 2011.
[12] S. Kamath, V. Anantharam, D. Tse, and C.-C. Wang, "The two-unicast problem," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3865–3882, 2016.
[13] H. Sun and S. A. Jafar, "Index Coding Capacity: How far can one go with only Shannon Inequalities?" *IEEE Trans. on Inf. Theory*, vol. 61, no. 6, pp. 3041–3055, 2015.
[14] Z. Li and H. Sun, "On Extremal Rates of Storage over Graphs," *arXiv preprint arXiv:2210.06363*, 2022.
[15] S. Jukna, *Extremal combinatorics: with applications in computer science*. Springer, 2011, vol. 571.