Storage Capacity as an Information-Theoretic Analogue of Vertex Cover

Arya Mazumdar, Andrew McGregor, and Sofya Vorotnikova

***Abstract*—Motivated by applications in distributed storage, the storage capacity of a graph was recently defined to be the maximum amount of information that can be stored across the vertices of a graph such that the information at any vertex can be recovered from the information stored at the neighboring vertices. Computing the storage capacity is a fundamental problem in network coding and is related, or equivalent, to some well-studied problems such as index coding with side information and generalized guessing games. In this paper, we consider storage capacity as a natural information-theoretic analogue of the minimum vertex cover of a graph. Indeed, while it was known that storage capacity is upper bounded by minimum vertex cover, we show that by treating it as such we can get a $3/2$ approximation for planar graphs, and a $4/3$ approximation for triangle-free planar graphs. Since the storage capacity is closely related to the index coding rate, we get a 1.923 approximation of index coding rate for planar graphs and $3/2$ approximation for triangle-free planar graphs. Previously only an obvious 4 approximation of the index coding rate was known for planar graphs. We then develop a general method of "gadget covering" to upper bound the storage capacity in terms of the average of a set of vertex covers. This method is intuitive and leads to the exact characterization of storage capacity for various families of graphs, such as cycles with chords and certain Cartesian product graphs. Finally, we generalize the storage capacity notion to include recovery from partial failures in distributed storage. We show tight upper and lower bounds on this partial recovery capacity that scales nicely with the fraction of failure in a vertex.**

I. Introduction

The Shannon capacity of a graph [17] is a well studied parameter that quantifies the zero-error capacity of a noisy communication channel. Some other notions of graph capacity are also well known (see, e.g., [1]). In this paper, we focus on a recent definition of graph capacity, called the *storage capacity*, that we consider to be a natural information-theoretic analogue of the minimum vertex cover of a graph.

Suppose, every vertex of a graph can store a symbol (from any alphabet) with the criterion that the content of any vertex can be uniquely recovered from the contents of its neighborhood. Then the maximum information that can be stored in the graph is called the storage capacity of that graph [16]. This formulation is motivated by distributed storage, and generalizes the definition of *locally repairable codes* [13].

Formally, suppose we are given an n-vertex graph $G(V, E)$, where $V = [n] \equiv \{1, 2, \ldots, n\}$. Also, given a positive integer $q \geq 2$, let $H(X)$ be the Shannon entropy of the random variable X in q-ary units. Let $\{X_i\}_{i \in V}$, be random variables each with a finite sample space of size q. For any $I \subseteq [n]$, let $X_I \equiv \{X_i : i \in I\}$. Consider the solution of the following optimization problem:

$$\max H(X_1, \ldots, X_n) \tag{1}$$

such that $H(X_i | X_{N(i)}) = 0$, for all $i \in V$ where $N(i) = \{j \in V : (i, j) \in E\}$ is the set of neighbors of vertex i. This is the storage capacity of G and is denoted by $\mathrm{Cap}_q(G)$. Note that, although we hide the unit of entropy in the notation $H(\cdot)$, the unit should be clear from context, and the storage capacity should depend on it, as reflected in the subscript in the notation $\mathrm{Cap}_q(G)$. The absolute storage capacity is:

$$\mathrm{Cap}(G) = \lim_{q \to \infty} \mathrm{Cap}_q(G). \tag{2}$$

In [16], it was observed that the storage capacity is bounded by the size of the minimum vertex cover $\mathrm{VC}(G)$ of G.

$$\mathrm{Cap}(G) \leq |\,\mathrm{VC}(G)|. \tag{3}$$

This follows since the neighbors of $V \setminus \mathrm{VC}$ belong to VC and hence $H(X_V) = H(X_{\mathrm{VC}(G)}) + H(X_{V \setminus \mathrm{VC}(G)} | H(X_{\mathrm{VC}(G)})) = H(X_{\mathrm{VC}(G)}) \leq |\,\mathrm{VC}(G)|$. Since $H(X_V) = H(X_{\mathrm{VC}(G)})$, we think it is natural to view storage capacity as an *information theoretic analogue of vertex cover*. It was also shown in [16] that the storage capacity is at least the size $\mathrm{MM}(G)$ of the maximum matching of G. Since maximum matching and minimum vertex cover are two quantities within a factor of two of each other and maximum matching can be found in polynomial time, this fact gives a 2-approximation of the storage capacity. Improvement over maximum matching is unlikely to be achieved by simple means, since that would imply a better-than-2 approximation ratio for the minimum vertex cover problem [15].

This motivates us to look for natural families of graphs where minimum vertex cover has a better approximation. For example, for bipartite graphs maximum matching is equal to minimum vertex cover and hence storage capacity is exactly equal to the minimum vertex cover. Another obvious class, and our focus in Section III, is the family of planar graphs for which a PTAS for vertex cover is known [4], [5]. Note that many common network topologies are often planar (see, [10]).

In a related broadcast problem called *index coding* [6] planar topologies are of interest, and outerplanar topologies have already been studied [7]. It was shown in [16] that storage capacity is, in a coding-theoretic sense, dual to index coding and is equivalent to the *guessing game* problem of [12]. The index

College of Information and Computer Sciences, University of Massachusetts, Amherst. {arya,mcgregor,svorotni}@cs.umass.edu. This work was supported by NSF Awards CCF-0953754, IIS-1251110, CCF-1320719, CCF-1642658, CCF-BSF-1618512, and a Google Research Award.

coding rate for a graph G is defined to be the optimum value of the following minimization problem: $\min H(Y)$ where Y is such that $H(X_i|Y, X_{N(i)}) = 0$, for all $i \in V$. This is called the optimum index coding rate for the graph G, and we denote it as $\mathrm{Ind}_q(G)$. We can also define, $\mathrm{Ind}(G) = \lim_{q\to\infty} \mathrm{Ind}_q(G)$.

The index coding problem has been the subject of much recent attention. In particular it can be shown that any network coding problem can be reduced to an index coding problem [11]. It has been shown [16] that, $\mathrm{Cap}(G) = n - \mathrm{Ind}(G)$, and hence exact computation of $\mathrm{Ind}(G)$ and $\mathrm{Cap}(G)$ is equivalent although the approximation hardness could obviously differ. Note that, $\mathrm{Ind}(G) \geq \alpha(G)$, where $\alpha(G)$ is the independence number of G. Since, for planar graphs $\alpha(G) \geq n/4$, taking Y to be $X_{[n]}$ already gives a 4-approximation for index coding rate for planar graphs (since $H(Y) \leq n$). In this paper, we give a significantly better approximation algorithm for index coding rate of planar graphs. Not only that, due to the relation between index coding and storage capacity, we obtain an approximation factor significantly better than 2 for storage capacity.

Towards obtaining better approximation ratios for more general graphs, we then develop several upper bounding tools for storage capacity. Our approach revisits a linear program (LP) proposed by Blasiak, Kleinberg, and Lubetzky [8] that can be used to lower bound the optimum index coding rate or upper bound the storage capacity. We transform the problem of bounding this LP into the problem of constructing a family of vertex covers for the input graph. This in turn allows us to upper bound the storage capacity of any graph that admits a specific type of vertex partition. We then identify various graphs for which this upper bound is tight.

Since minimum vertex cover acts as an absolute upper bound on the rate of information storage in a graph, a natural question to ask is, if we store above the limit of minimum vertex cover in the graph, will any of the repair property be left? This is similar in philosophy to the rate-distortion theory of data compression, where one compresses beyond entropy limit and still can recover the data with some distortion. This question gives rise to the notion of recovery from partial failure. The *partial repair capacity* is a direct generalization in the context of distributed storage application to handle partial failure of vertices. In particular, suppose we lose $\delta \in [0, 1]$ proportion of the bits stored in a vertex. We still want to recover these bits by accessing the remaining $(1 - \delta)$-fraction of the bits in the vertex plus the contents of the neighborhood. What is the maximum amount of information that can be stored in the network with such restriction?

A summary of our results is as follows:

1) *Planar graphs.* We prove a $3/2$ approximation of storage capacity and 1.923 approximation for index coding rate for planar graphs. For triangle-free planar graphs, we get a $4/3$ approximation for storage capacity, and $3/2$ approximation for index coding rate.

2) *Tools for finding storage capacity upper bounds.* We develop an approach for bounding storage capacity in terms of multiple vertex covers. We use the approach to show a bound on any graph that admits a specific type of vertex partition. With this we derive capacities of a family of Cartesian product graphs and a family closely related to outerplanar graphs.

3) *Partial failure recovery.* We present upper and lower bounds on the capacity if recovery from neighbors is possible for up to δ-proportion failure of the bits stored in a server. These imply that the partial recovery capacity is same as the storage capacity when $\delta \geq \frac{1}{2}$. For an odd cycle with n nodes, we show that the partial recovery capacity is at most $\frac{n}{2}(1 + R_2(\delta))$, where $R_2(\delta)$ is the maximum rate of a binary error-correcting code with minimum distance δn. We also show that capacity of $\frac{n}{2}(2 - h_2(\delta))$ is polynomial time achievable, where $h_2(\delta)$ denotes the binary entropy function. Our bounds are tight, assuming the widely believed conjecture that $R_2(\delta) = 1 - h_2(\delta)$.

II. Preliminaries

Let $\mathrm{CP}(G)$ denote the fractional clique packing of a graph defined as follows: Let $\mathcal{C}$ be the set of all cliques in G. For every $C \in \mathcal{C}$ define a variable $0 \leq x_C \leq 1$. Then $\mathrm{CP}(G)$ is the maximum value of $\sum_{C\in\mathcal{C}} x_C(|C| - 1)$ subject to the constraint that $\sum_{C\in\mathcal{C}:u\in C} x_C \leq 1$ for all $u \in V$. Note that $\mathrm{CP}(G)$ can be computed in polynomial time in graphs where all cliques have constant size, such as planar graphs. Furthermore, $\mathrm{CP}(G)$ is at least the size of the maximum fractional matching $\mathrm{FM}(G)$ and they are obviously equal in triangle-free graphs since the only cliques are edges. $\mathrm{Cap}_q(G)$ is related to $\mathrm{CP}(G)$ as follows:

Lemma 1. $\mathrm{Cap}_q(G) \geq \mathrm{CP}(G)$ *for sufficiently large* q.

An equivalent result is known in the context of index coding [6]. The basic idea is that we can store $k-1$ units of information on a clique of size k by assigning $k - 1$ independent uniform random variables to $k - 1$ of the vertices and setting the final random variable to the sum (modulo q) of the first $k - 1$ variables. This idea can be extended to the fractional setting.

Below, let $G[S]$ denote the subgraph induced by $S \subseteq V$.

III. Approximation Algorithms for Planar Graphs

We next present approximation results for the storage capacity and optimal index coding rate of planar graphs. In our storage capacity result we use ideas introduced in [5] for the purpose of approximating the vertex cover of planar graphs. Specifically, they first considered a maximal set of vertex-disjoint triangles, reasoned about the vertex cover amongst these triangles, and then reasoned about the triangle-free induced subgraph on the remaining vertices. We consider a similar decomposition and reason about the integrality gap of vertex cover in each component. We parameterize our result in terms of the number of triangles; this will be essential in the subsequent result on optimal index coding rate.

Theorem 2. *Assume G is planar and let T be a set of $3t$ vertices corresponding to a maximal set of t vertex disjoint triangles. Then, $1 \leq \frac{\mathrm{Cap}(G)}{\mathrm{CP}(G)} \leq \frac{3t+k}{2t+3k/4}$ where k is the size of the minimum vertex cover of $G[V \setminus T]$. Hence $\mathrm{CP}(G)$ is a $3/2$ approximation for $\mathrm{Cap}(G)$ and $4/3$ approximation if G is triangle-free.*

Proof. Let $G' = G[V \setminus T]$. Partition the set of vertices into $T \cup C \cup I$ where C is the minimum vertex cover of G' and $I = V \setminus (T \cup C)$ is therefore an independent set. Let

X_V be the set of variables that achieve storage capacity. Therefore, $\mathrm{Cap}(G) = H(X_V) = H(X_T) + H(X_C|X_T) + H(X_I|X_C, X_T) = H(X_T) + H(X_C|X_T) \le 3t + k$ since for each $v \in I$, $H(X_v|X_C, X_T) = 0$ since $N(v) \subset C \cup T$.

Consider the fractional clique packing in which each of the t vertex-disjoint triangles in T receives weight 1. Then, $\mathrm{CP}(G) \ge 2t + \mathrm{CP}(G')$. Then it remains to show that $\mathrm{CP}(G') \ge 3k/4$. Note that since G' is triangle-free planar graph, it is 3-colorable by Grötzsch's theorem [14]. Furthermore, $\mathrm{CP}(G')$ is the maximum fractional matching which, by LP-duality, is the minimum fractional vertex cover. Hence it suffices to show that the size of the minimum fractional vertex cover of 3-colorable graph is at least $3/4$ of the size of the minimum (integral) vertex cover, i.e., $3k/4$. This can be shown as follows. Let $x_1, \ldots, x_n$ be an optimal fractional vertex cover, i.e., for all edges $uv \in G'$, $x_u + x_v \ge 1$. Since fractional vertex cover is $1/2$-integral, we may assume each $x_u \in \{0, 1/2, 1\}$. Let I_1, I_2, I_3 be a partition of $\{u \in [n] : x_u = 1/2\}$ corresponding to a 3-coloring where $\sum_{v\in I_1} x_v \ge \sum_{v\in I_2} x_v \ge \sum_{v\in I_3} x_v$. Then consider $y_1, \ldots, y_n$ where $y_u = 1$ iff $u \in I_2 \cup I_3$ or $x_u = 1$. Then $\sum_{u\in[n]} y_u \le \sum_{u\in I_2\cup I_3} y_u + \sum_{u\in[n]:x_u=1} y_u \le 2/3 \cdot 2 \cdot \sum_{u:x_u=1/2} x_u + \sum_{u:x_u=1} x_u \le 4/3 \cdot \mathrm{CP}(G')$, and $y_1, \ldots, y_n$ is a vertex cover because for every edge uv, at least one of $\{x_u, x_v\}$ is 1 or at least one of u and v is in $I_2 \cup I_3$. □

We next apply the previous theorem, together with the chromatic number of planar and triangle-free planar graphs to achieve a 1.923 approximation for $\mathrm{Ind}(G)$.

Theorem 3. *Assume G is planar and let T be a set of $3t$ vertices corresponding to t vertex disjoint triangles. Then,*

$$1 \le \frac{n - \mathrm{CP}(G)}{\mathrm{Ind}(G)} \le \begin{cases} \frac{3n+3t}{4n-3t} + \frac{3}{4} & \text{for } t \le \frac{179-\sqrt{16681}}{192} n \\ 4 - \frac{8t}{n} & \text{for } t \ge \frac{179-\sqrt{16681}}{192} n \end{cases}.$$

This is a $3/2$ approximation if G is triangle-free and maximizing over t implies a 1.923 approximation in general.

Proof. From Theorem 2, we know that $\mathrm{CP}(G) \ge 2t + 3/4|\,\mathrm{VC}(G')|$ where $G' = G[V \setminus T]$. Therefore,

$$\begin{aligned} n - \mathrm{CP}(G) &\le n - (2t + 3/4|\,\mathrm{VC}(G')|) \\ &= n - (2t + 3/4(n - 3t - \alpha(G'))) \\ &= (n+t)/4 + 3/4\,\alpha(G') \end{aligned}$$

On the other hand, $\mathrm{Ind}(G) \ge \alpha(G) \ge \alpha(G')$ where α denotes the size of the maximum independent set of the graph. Note that $\alpha(G) \ge n/4$ since G is planar and thus 4-colorable [2], [3]. Since G' has $n-3t$ vertices and is triangle-free and planar and thus 3-colorable [14], $n-3t \ge \alpha(G') \ge (n-3t)/3$. These inequalities imply $(n - \mathrm{CP}(G))/\,\mathrm{Ind}(G)$ is at most

$$\begin{aligned} &\max\left(\frac{(n+t)/4 + 3/4\,\alpha(G')}{\alpha(G)}, \frac{(n+t)/4 + 3/4\,\alpha(G')}{\alpha(G')}\right) \\ &\le \max\left(\frac{(n+t)/4 + 3/4(n-3t)}{n/4}, \frac{(n+t)/4}{(n-3t)/3} + 3/4\right) \\ &= \max\left(4 - \frac{8t}{n}, \frac{3n+3t}{4n-3t} + \frac{3}{4}\right). \end{aligned}$$

□

IV. Upper Bounds via Multiple Vertex Covers

In this section, we start by considering a linear program proposed by Blasiak, Kleinberg, and Lubetzky [8] that can be used to lower bound the optimum index coding rate. There are $\Omega(2^n)$ constraints but by carefully selecting a subset of constraints we can prove upper bounds on the storage capacity for a specific graph without solving the LP.

Our main goal in this section is to relate this linear program to finding a suitable family of vertex covers of the graph. In doing so, we propose a combinatorial approach to constructing good upper bounds that we think makes the process of proving strong upper bounds more intuitive. This allows us to prove a more general theorem that gives an upper bound on the storage capacity for a relatively large family of graphs. As an application of this theorem we show that a class of graphs closely related to the family of outerplanar graphs and another family of Cartesian product graphs have capacity exactly $n/2$. Proofs of this and the subsequent sections have been omitted for space constraints.

A. Upper Bound via the "Information Theoretic" LP

We first rewrite the index coding LP proposed by Blasiak, Kleinberg, and Lubetzky [8] for the purposes of upper-bounding storage capacity. We define a variable z_S for every $S \subseteq V$ that will correspond to an upper bound on $H(X_S)$. Let $\mathrm{cl}(S) = S \cup \{v : N(v) \subseteq S\}$ denote the *closure* of the set S consisting of vertices in S and vertices with all neighbors in S.

$$\begin{aligned} \text{maximize } z_V \quad \text{s.t.} \quad & z_\emptyset = 0 \\ & z_T - z_S \le |T \setminus \mathrm{cl}(S)| \quad \forall S \subseteq T \\ & z_S + z_T \ge z_{S\cap T} + z_{S\cup T} \quad \forall S, T \end{aligned}$$

The second constraint corresponds to $H(X_T) - H(X_S) = H(X_T|X_S) = H(X_T|X_{\mathrm{cl}(S)}) \le H(X_{T\setminus \mathrm{cl}(S)}) \le |T \setminus \mathrm{cl}(S)|$, whereas the last constraint follows from the sub-modularity of entropy. Hence, the optimal solution to the above LP is an upper bound on $\mathrm{Cap}(G)$. We henceforth refer to the above linear program as the *information theoretic* LP.

B. Upper Bound via Gadgets

k-cover by gadgets is a technique for proving upper bounds on the storage capacity of graph. The core idea is to construct a set of k vertex covers for the graph via the construction of various "gadgets". A gadget $g(A, B)$ is created as follows: take two sets of vertices A and B, take their closures $\mathrm{cl}(A)$ and $\mathrm{cl}(B)$, find $S = \mathrm{cl}(A) \cup \mathrm{cl}(B)$ and $T = \mathrm{cl}(A) \cap \mathrm{cl}(B)$. Then S and T form a gadget. Call S the outside of the gadget and T the inside. We note that by taking $A = \{v\}$ and $B = \emptyset$ we obtain a gadget with the outside $\{v\}$ and empty inside (assuming v has no neighbors of degree one); call such gadget *trivial*. Define the weight of a gadget to be $|A| + |B|$. If we color every inside and outside gadget set with one of k colors such that the union of all sets of the same color forms a vertex cover, the total weight of gadgets in such coloring provides an upper bound on $k\,\mathrm{Cap}(G)$. Note that for $k = 1$ using gadgets with non-empty inside can only increase the total weight, so the only gadgets we need to consider are the trivial ones corresponding

to individual vertices, and thus the construction is just a single vertex cover.

We can formulate the k-cover by gadgets (for fixed k) as the following linear program: Let $x_{S,c}$ be a variable where S is a set that is an outside or an inside of a gadget and c is one of k colors. Each variable has a corresponding weight $w_{S,c} = (|A| + |B|)/2$ where A and B are the 2 sets used to form the gadget that S is a part of. $x_{S,c} = 1$ if set S is colored with color c and 0 otherwise.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{k} \sum_{S,c} w_{S,c} x_{S,c} \\ \text{s.t.} \quad & \sum_{S:u \in S} x_{S,c} + \sum_{S:v \in S} x_{S,c} \geq 1 \quad \forall (u,v) \in E, \forall c \\ & \sum_{c} x_{H,c} = \sum_{c} x_{H',c} \quad \text{for all gadgets, where } H \text{ and } H' \text{ are the outside and inside of the gadget} \end{aligned}$$

The first condition states that every collection of sets of a fixed color is a vertex cover and the second states that the outside and inside of every gadget are used the same number of times.

Theorem 4. *Any feasible integral solution to the above k-cover by gadgets LP is an upper bound on* $\mathrm{Cap}(G)$.

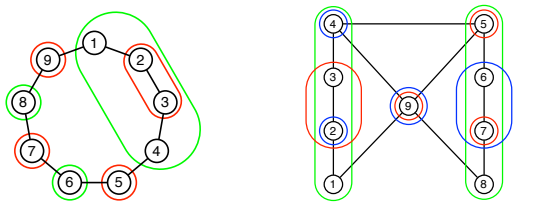
Proof. We prove this by showing that k-cover by gadgets follows from the fact that the optimal solution of the information theoretic LP is an upper bound. First, note which constraints correspond to the steps of forming a gadget:

$$\begin{aligned} z_A \leq |A| \ , \ z_B \leq |B| \quad & \text{take sets } A \text{ and } B \\ z_{\mathrm{cl}(A)} - z_A \leq 0 \quad & \text{find closure of } A \\ z_{\mathrm{cl}(B)} - z_B \leq 0 \quad & \text{find closure of } B \\ z_S + z_T \leq z_{\mathrm{cl}(A)} + z_{\mathrm{cl}(B)} \quad & \text{find } S = \mathrm{cl}(A) \cup \mathrm{cl}(B) \text{ and } T = \mathrm{cl}(A) \cap \mathrm{cl}(B) \end{aligned}$$

If we sum all the constraints, we obtain $z_S + z_T \leq |A| + |B|$. Assume, that we used g gadgets in the cover. By summing all corresponding constraints, we get $z_{S_1} + z_{T_1} + \cdots + z_{S_g} + z_{T_g} \leq |A_1| + |B_1| + \cdots + |A_g| + |B_g|$. Group the sets into color classes $C_1, \ldots, C_k$. Let $U_i = \cup_{S \in C_i} S$. The corresponding constraints are then, for all $i \in [k]$, $z_{U_i} - \sum_{S \in C_i} z_S \leq 0$ and $z_{cl(U_i)} - z_{U_i} \leq 0$. Note that $z_V = z_{cl(U_i)}$ since U_i is a vertex cover. By summing the $2k$ constraints and the one obtained from building gadgets, we get $k z_V \leq |A_1| + |B_1| + \cdots + |A_g| + |B_g|$. □

We next illustrate the use of the k-cover via gadgets approach with a couple of examples. First, we re-prove a result of Blasiak et al. [9] via a 2-cover by gadgets. Then we give an example of an outerplanar graph where it is necessary to consider a 3-cover by gadgets to establish a tight bound.

a) Odd Cycles: We prove that the storage capacity of an odd cycle of length n is $n/2$; see Figure 1(a) for an example where $n = 9$. $\mathrm{FM}(C_n) = n/2$, thus $\mathrm{Cap}(C_n) \geq n/2$. For the upper bound we form a gadget $g(A, B)$ by taking $A = \{v_1, v_3\}$, $B = \{v_2, v_4\}$ and obtaining outer set $S = \{v_1, v_2, v_3, v_4\}$ and inner set $T = \{v_2, v_3\}$. On the rest of the vertices we place trivial gadgets. Color S and trivial gadgets on $v_6, v_8, \ldots, v_{n-1}$ green, color T and trivial gadgets on $v_5, v_7, \ldots, v_n$ red. Green and red sets are then vertex covers and the total weight of all gadgets is n. Thus, $\mathrm{Cap}(C_n) \leq n/2$.



(a) An Odd Cycle. (b) An Outerplanar Graph.

Fig. 1. Two examples of k-cover upper bounds. See text for details.

b) An Outerplanar Graph: We prove that the storage capacity of the graph in Figure 1(b) is $14/3$. This capacity is achieved by the fractional clique cover. Create gadgets $g_1(A_1, B_1)$ and $g_2(A_2, B_2)$ from $A_1 = \{v_1, v_3\}$, $B_1 = \{v_2, v_4\}$, $A_2 = \{v_5, v_7\}$, and $B_2 = \{v_6, v_8\}$. Place a trivial gadget on each of the vertices v_2, v_4, v_5, v_7 and two trivial gadgets on v_9. Color the sets as:

- Red: v_5, v_7, v_9 and the inside of gadget g_1
- Blue: v_2, v_4, v_9 and the inside of gadget g_2
- Green: the outside sets of both gadgets

Each color corresponds to a vertex cover and the total weight of gadgets is 14.

C. $n/2$ Upper Bound via Vertex Partition

The next theorem uses a 2-cover by gadgets to prove that a certain family of graphs have capacity at most $n/2$. Subsequently, we will use this theorem to exactly characterize the capacity of various graph families of interest.

Theorem 5. *Suppose that the vertices of a graph G can be partitioned into sets X and Y such that:*

1) *$G[X]$ and $G[Y]$ are both bipartite.*
2) *S_X is an independent set in $G[X]$ and S_Y is an independent set in $G[Y]$*

where $S_X \subseteq X$ consists of all vertices in X with a neighbor in Y and $S_Y \subseteq Y$ consists of all vertices in Y with a neighbor in X. Then $\mathrm{Cap}(G) \leq n/2$.

We next apply Theorem 5 to prove that certain families of graphs have storage capacity exactly $n/2$.

1) Cartesian Product of a Cycle and a Bipartite Graph: The Cartesian product $G_1 \square G_2$ of graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ has vertex set $V_1 \times V_2$ and $(u, u')(v, v')$ is an edge iff $u = v$ and $u'v' \in E_2$ or $u' = v'$ and $uv \in E_1$.

Theorem 6. *Let C_k be a cycle with $k > 3$, B a bipartite graph, and $G = C_k \square B$. Then* $\mathrm{Cap}(G) = n/2$, *where n is the number of vertices in G.*

2) Cycles with Well-Separated Chords: Note that any (connected) outerplanar graph without cut vertices is a cycle with non-overlapping chords. The family of graphs we consider is more general as we permit the chords to overlap, but more restrictive since we require the endpoints of these chords to be far apart on the cycle. A natural open question is to characterize $\mathrm{Cap}(G)$ for all outerplanar graphs. All that was previously known is that if we assume each X_i is a linear combination of $\{X_j\}_{j \in N(i)}$, then $\mathrm{Cap}(G)$ equals *integral* clique packing [7].

Theorem 7. *Let G be a cycle with a chords whose endpoints are at least distance 4 apart on the cycle. Then* $\mathrm{Cap}(G) = n/2$.

V. Partial Recovery

We now extend the notion of storage capacity to cover for partial failures. This is a new generalization, that, as far as we understand, does not have a counterpart in index coding. As before, suppose we have a graph $G([n], E)$ and every vertex $i \in [n]$ stores a random vector $X_i \in \mathbb{F}_q^m$. We want the following repair criterion to be satisfied: if up to any $\delta \in [0, 1]$ proportion of the m coordinates of X_i are erased, they can be recovered by using the remaining content of the vertex i and $X_{N(i)}$, the contents in the neighbors of the vertex.

We define the *partial recovery capacity* of G to be the normalized asymptotic maximum total amount of information (in terms of q-ary unit) that can be stored in the graph G, $\mathrm{Cap}_q(G, \delta) = \lim_{m\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{m}$. It is easy to show $\mathrm{Cap}_q(G, 0) = n$ and $\mathrm{Cap}_q(G, 1) = \mathrm{Cap}(G)$. We will prove upper and lower bounds for $\mathrm{Cap}_q(G, \delta)$.

A. Impossibility bound

The partial recovery capacity can be defined via an entropy maximization problem, generalizing the storage capacity. Let $A_q(m, d)$ be the maximum possible size of a q-ary m-length error-correcting code with minimum distance d.

Theorem 8. *Let $H(X)$ be the entropy of X measured in q-ary units. Suppose, $X_i \in \mathbb{F}_q^m$, $i \in [n]$. For a graph $G([n], E)$, $\mathrm{Cap}_q(G, \delta)$ is upper bounded by the solution of the following optimization problem:* $\max \lim_{m\to\infty} H(X_1, \ldots, X_n)/m$, *such that,* $H(X_i \mid X_{N(i)}) \le \log_q A_q(m, \delta m + 1)$.

Define $R_q(\delta) \equiv \lim_{m\to\infty} \log_q A_q(m, \delta m + 1)/m$ assuming the limit exists. Since $R_q(\delta) = 0$ for $\delta \ge 1 - 1/q$, we deduce:

Corollary 9. *For any graph G,* $\mathrm{Cap}_q(G, \delta) = \mathrm{Cap}(G)$ *for* $\delta \ge 1 - \frac{1}{q}$. *In particular,* $\mathrm{Cap}_2(G, \delta) = \mathrm{Cap}(G)$ *for* $\delta \ge \frac{1}{2}$.

We next generalize the technique of upper bounding the storage capacity via an information theoretic linear program.

Theorem 10. *The optimal solution to the following LP is an upper bound on* $\mathrm{Cap}_q(G, \delta)$.

$$\begin{aligned} \textit{maximize}\ \ z_V \quad \textit{s.t.}\quad & z_\emptyset = 0 \\ & z_T - z_S \le f(T, S), \forall S \subseteq T \\ & z_S + z_T \ge z_{S\cap T} + z_{S\cup T} \quad \forall S, T \end{aligned}$$

where $f(T, S) = |T \setminus S| - (1 - R_q(\delta))|(\mathrm{cl}(S) \setminus S) \cap T|$.

Odd Cycles: Consider an odd cycle with n vertices (n is odd). The above LP implies, $n + 2R_q(\delta) + R_q(\delta)(n-2) \ge 2z_V - 2z_\emptyset$ and thus $\mathrm{Cap}_q(G, \delta) \le z_V \le \frac{n}{2}(1 + R_q(\delta))$.

B. Achievability bound

A naive achievability bound is $\mathrm{Cap}_q(G, \delta) \ge n(1 - h_q(\delta))$ for $\delta \le 1/2$, where $h_q(x) \equiv x \log_q(q-1) - x \log_q x - (1-x)\log_q(1-x)$. This follows by using an error-correcting code of length m, distance $\delta m + 1$, and rate $1 - h_q(\delta)$ in each of the vertices. Such codes exist, by the Gilbert-Varshamov bound. Also, $\mathrm{Cap}_q(G, \delta) \ge 0$, for $0 \le \delta \le 1 - \frac{1}{q}$. This simple bound can be improved by more carefully designing a code.

Theorem 11. *Given a graph G, let $\mathcal{C}$ be the set of all cliques of G. The generalized clique packing number* $\mathrm{CP}_\delta(G)$ *is defined to be the optimum of the following linear program. For* $0 \le x_C \le 1, \forall C \in \mathcal{C}$, $\max \sum_{C\in\mathcal{C}} x_C(|C| - h_q(\delta))$, *such that,* $\sum_{C\in\mathcal{C}: u\in C} x_C \le 1$. *Then,* $\mathrm{Cap}_q(G, \delta) \ge \mathrm{CP}_\delta(G)$ *if* $\delta \le 1 - 1/q$, *and* $\mathrm{Cap}_q(G, \delta) \ge \mathrm{CP}(G)$ *if* $\delta > 1 - 1/q$.

Odd Cycles: Consider an odd cycle on n nodes. Since there is a fractional matching of size $\frac{n}{2}$, we have $\mathrm{Cap}_q(G, \delta) \ge \frac{n}{2}(2 - h_q(\delta))$ for $\delta \le 1 - 1/q$, and $\mathrm{Cap}_q(G, \delta) \ge \frac{n}{2}$ when $\delta > 1 - \frac{1}{q}$. Compare this with the impossibility bound that we have, $\mathrm{Cap}_q(G, \delta) \le \frac{n}{2}(1 + R_q(\delta))$. It is widely conjectured that the optimal rate of an error-correcting code is given by $R_q(\delta) = 1 - h_q(\delta)$, for small q, which is also known as the Gilbert-Varshamov conjecture. If this conjecture is true, then our upper and lower bounds match exactly. In particular, for large q (i.e., $q \to \infty$), we have $h_q(\delta) \to \delta$ and $R_q(\delta) \to 1 - \delta$. Hence, our bounds match definitively in the regime of large q.

References

[1] N. Alon and A. Orlitsky. Source coding and graph entropies. *IEEE Transactions on Information Theory*, 42(5):1329–1339, 1996.

[2] K. Appel and W. Haken. Every planar map is four colorable. part i: Discharging. *Illinois J. Math.*, 21(3):429–490, 09 1977.

[3] K. Appel, W. Haken, and J. Koch. Every planar map is four colorable. part ii: Reducibility. *Illinois J. Math.*, 21(3):491–567, 09 1977.

[4] B. S. Baker. Approximation algorithms for np-complete problems on planar graphs. *J. ACM*, 41(1):153–180, 1994.

[5] R. Bar-Yehuda and S. Even. On approximating a vertex cover for planar graphs. In *Proceedings of the 14th Annual ACM Symposium on Theory of Computing, May 5-7, 1982, San Francisco, California, USA*, pages 303–309, 1982.

[6] Z. Bar-Yossef, Y. Birk, T. Jayram, and T. Kol. Index coding with side information. *IEEE Transactions on Information Theory*, 57(3):1479–1494, 2011.

[7] Y. Berliner and M. Langberg. Index coding with outerplanar side information. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 806–810. IEEE, 2011.

[8] A. Blasiak, R. Kleinberg, and E. Lubetzky. Lexicographic products and the power of non-linear network coding. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 609–618, 2011.

[9] A. Blasiak, R. D. Kleinberg, and E. Lubetzky. Index coding via linear programming. *CoRR*, abs/1004.1379, 2010.

[10] R. Bowden, H. X. Nguyen, N. Falkner, S. Knight, and M. Roughan. Planarity of data networks. In *Teletraffic Congress (ITC), 2011 23rd International*, pages 254–261. IEEE, 2011.

[11] M. Effros, S. El Rouayheb, and M. Langberg. An equivalence between network coding and index coding. *IEEE Transactions on Information Theory*, 61(5):2478–2487, 2015.

[12] M. Gadouleau, A. Richard, and S. Riis. Fixed points of boolean networks, guessing graphs, and coding theory. *SIAM Journal on Discrete Mathematics*, 29(4):2312–2335, 2015.

[13] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin. On the locality of codeword symbols. *IEEE Transactions on Information Theory*, 58(11):6925–6934, 2012.

[14] H. Grötzsch. Zur theorie der diskreten gebilde, vii: Ein dreifarbensatz fr dreikreisfreie netze auf der kugel. *Wiss. Z. Martin-Luther-U., Halle-Wittenberg, Math.-Nat. Reihe*, 8:109120, 1959.

[15] S. Khot and O. Regev. Vertex cover might be hard to approximate to within 2- ε. *Journal of Computer and System Sciences*, 74(3):335–349, 2008.

[16] A. Mazumdar. Storage capacity of repairable networks. *IEEE Transactions on Information Theory*, 61(11):5810–5821, 2015.

[17] C. Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):8–19, 1956.