SHONAN INSTITUTE OF
TECHNOLOGY JOURNAL
Vol. 54, No. 1, 2020

Applications of Graph Theory towards Data Storage Media

Akiko MANADA*

Abstract: Graph theory, which studies the theoretical properties of graphs, is an area of applied and discrete mathematics, and has various applications towards engineering. For example, network topologies can be characterized by graphs and those topologies can be used to find better solutions in smooth communications.

In this paper, we explain how graph theory has been applied in data storage media. In particular, we explain storing/reading schemes for typical data storage media, and then present how graph theory plays an important role in coding for data storage media, together with our contributions. More precisely, we review our results from constrained systems, which control the appearance of certain data sequences, and how the results work for reliable data storage media.

KEYWORDS: Graph theory, Constrained system, Presentation, Data storage media

1 Introduction

A graph is a way to represent connections of objects using points (vertices) and lines (edges). For example, people are considered to be vertices and friendships can be presented using edges. Also, network connections are depicted using graphs by considering users in the networks as vertices.

Graph theory is, of course, a research for theoretically studies on graphs. Vertex colorings and the Max-flow problem (see, for example, [1]) would be typical examples of problems in graph theory. Graph theory is also considered to be an area of applied and discrete mathematics, and has various applications towards engineering. For example, network topologies can be characterized by graphs and those topologies can be used to find better solutions in smooth communications.

The main agenda of this paper is to present how graph theory is applied in the coding for data storage media. Indeed, typical data storage media (such as CDs, DVDs, HDDs and USB memory sticks) apply coding schemes to reduce the likelihood of errors, and theoretical analysis on those coding schemes come from the study of graphs representing the schemes. In this paper, we review the theoretical analysis and how it works, together with our contributions. Our contributions are strongly based on the study of constrained systems (see, for example, [2, 3]), which control the appearance of certain sequences in data. Such a control, in fact, works to reduce the errors in reading data, and therefore, the control is a key point to make the data media more reliable.

The rest of the paper is organized as follows. We go over fundamental back grounds and notations in Section 2. In Section 3, we explain typical data storage media and typical errors in storage media, and the relation with constrained systems. We then review in Section 4 our main contributions on storage media from the perspective of coding schemes utilizing graph theory. We terminate this paper with conclusion and future works in Section 5.

2 Preliminaries and Basic Backgrounds

We first go over fundamental backgrounds, based on [1, 2, 3], that will be used throughout this paper.

2.1 Language

Let Σ be an *alphabet*, a finite set of symbols. Throughout this paper, we mainly assume that $\Sigma =$

*Lecturer at Department of Information Science, Shonan Institute of Technology

$\{0,1\}$ for simplicity, but results on binary case can be naturally extended to q-ary cases (*i.e.*, when an alphabet consists of q symbols with $q \geq 2$).

A *word* $\boldsymbol{w} = w_0 w_1 \dots w_{\ell-1}$ is a finite-length sequence over Σ, and denote its length (that is, ℓ in this case) by $|\boldsymbol{w}|$. The *empty word* ϵ is a unique word of length 0. Given two words $\boldsymbol{w} = w_0 w_1 \dots w_{\ell-1}$ and $\hat{\boldsymbol{w}} = \hat{w}_0 \hat{w}_1 \dots \hat{w}_{\hat{\ell}-1}$ such that $\hat{\ell} \leq \ell$, we say that $\hat{\boldsymbol{w}}$ is a *subword* of $\boldsymbol{w}$ if $\hat{\boldsymbol{w}} = w_i w_{i+1} \dots w_{i+\hat{\ell}-1}$ for some $0 \leq i \leq \ell - \hat{\ell}$. In particular, $\hat{\boldsymbol{w}}$ is a *prefix* of $\boldsymbol{w}$ when $i = 0$, and a *suffix* of $\boldsymbol{w}$ when $i = \ell - \hat{\ell}$. The notation $\boldsymbol{w}\boldsymbol{w}'$ represents the word generated by concatenating $\boldsymbol{w}$ followed by $\boldsymbol{w}'$. In particular, $\boldsymbol{w}^r$ for some integer $r \geq 0$ represents the word generated by concatenating r copies of $\boldsymbol{w}$. We assume by convention that for any word $\boldsymbol{w}$, $\epsilon\boldsymbol{w} = \boldsymbol{w}\epsilon = \boldsymbol{w}$ (so ϵ is a prefix and a suffix of $\boldsymbol{w}$) and $\boldsymbol{w}^0 = \epsilon$.

2.2 Graph Theory

A *graph* $G = (V, E)$ consists of the vertex set V and the edge set $E \subset V \times V$, where each edge $e = (u, v)$ in E can be characterized by two end points $u, v \in V$. If the direction of edges is not considered, then G is simply called a graph and $(u, v) = (v, u)$ holds. However, when the direction should be considered, vertices u and v are called the *starting vertex* and the *terminating vertex*, respectively, of edge e, and such a graph is called a *directed graph* or a *digraph* in short. Throughout this paper, we assume that graphs are always directed if not specified.

A *path* $\pi : p_0, p_1, \dots, p_k$ is a sequence of vertices such that $(p_t, p_{t+1}) \in E$ for each $0 \leq t \leq k$. A directed graph is called *irreducible* or *strongly-connected* if a path from vertex u to vertex v exits for each pair of vertices u, v.

Each graph can be represented in terms of a matrix. Given a graph G of vertex set $V = \{v_1, v_2, \dots, v_{|V|}\}$, the *adjacency matrix* $A = A_G$ of G is a $|V| \times |V|$ square matrix such that

- the i-th row and the i-th column correspond to v_i, $1 \leq i \leq |V|$; and
- the (i, j)-coordinate is the number of edges from vertex v_i to vertex v_j, $1 \leq i, j \leq |V|$.

A *labelled directed graph* $G = (V, E, L)$ is a digraph with a function $L : E \to \Sigma$ assigned to edges. More precisely, in a labelled directed graph, each edge e is labelled with a symbol $s = L(e) \in \Sigma$. A labelled graph is called *deterministic* if edges starting from the same vertex are labelled distinctly. In other words, for a deterministic graph, if the edges e, e' start from the same vertex, then $L(e) \neq L(e')$.

2.3 Constrained System

A *constrained system* $\mathcal{S}$ is a set of words generated by reading off labels along the paths in a labelled directed graph G. In this case, we say that $\mathcal{S}$ is *presented* by G or G is a *presentation* of $\mathcal{S}$. A constrained system $\mathcal{S}$ is called *irreducible* if there exists an irreducible presentation for $\mathcal{S}$, and is called *reducible* if not.

It has to be mentioned that any constrained system $\mathcal{S}$ can be characterized by a *forbidden set* $\mathcal{F}$, a set of *forbidden words*. In other words, a word $\boldsymbol{w}$ is in $\mathcal{S} = \mathcal{S}_{\mathcal{F}}$ if and only if $\boldsymbol{w}$ does not contain any forbidden word $\boldsymbol{f} \in \mathcal{F}$ as a subword. A typical example of constrained systems is the set of binary words satisfying the (d, k)-*Run-Length-Limited (RLL) constraint* such that

- there are at least d 0's between two consecutive 1's; and
- the run-length of 0's is at most k.

We call the constrained system the (d, k)-*RLL system*. The conditions above derive that (d, k)-RLL system is characterized by a forbidden set

$$\mathcal{F} = \{11, 101, \dots, 10^{d-1}1, 0^{k+1}\}.$$

Furthermore, we can easily observe that (d, k)-RLL system is indeed a constrained system since it has a presentation in Figure 1.

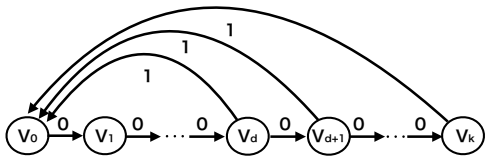

Figure 1: A presentation of the (d, k)-RLL system

Let $W_\ell(\mathcal{S})$ be the set of words of length ℓ in a constrained system $\mathcal{S}$. The *capacity* $C(\mathcal{S})$ of $\mathcal{S}$ is defined to be

$$C(\mathcal{S}) = \lim_{\ell \to \infty} \frac{\log_2 |W_\ell(\mathcal{S})|}{\ell}.$$

The capacity can be considered as an asymptotic growth rate. Indeed, the number of words of length ℓ in $\mathcal{S}(\mathcal{S})$ is estimated as

$$|W_\ell(\mathcal{S})| = 2^{\ell C(\mathcal{S})}.$$

In addition, the capacity is an important value when considering the coding scheme since the capacity gives the maximum coding rate under finite state encoders.

It is important to emphasize that each constrained system has a deterministic presentation, and that the capacity can be computed as

$$C(\mathcal{S}) = \log_2 \lambda_A, \tag{1}$$

where λ_A is the largest eigenvalue for the adjacency matrix of a deterministic presentation (see, for example, [2, Section 4]).

3 Data Storage Media and Typical Errors

There are various types of data storage media that have been used these days. In this section, we focus on, as examples, Compact Discs (CDs) and Digital Versatile Discs (DVDs) in detail together with their typical errors. We also explain how constrained systems work to reduce the likelihood of errors.

3.1 Compact Disc and Digital Versatile Disc

Compact Discs (CDs) and Digital Versatile Discs (DVDs) are classic storage media that have been often used in real life. For those media, the surface is covered with a "land", a shiny layer which reflects the laser. When data is stored in the media (where data is again considered to be a binary word), "pits" whose ends are at the positions of 1's are created on the land. The land and pits have different reflectances, which is a key for reading data. Indeed, when data is read form the media, a laser is irradiated to the land and determine when reflectance changes occur. The positions of these reflectance changes are read as the positions of 1's, and the number of 0's between 1's is counted based on the time course after the last reflectance change (see Figure 2).

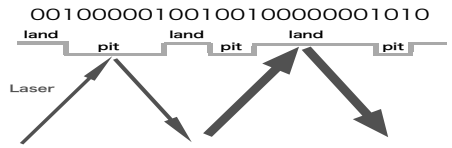

Figure 2: An image on how to store and read data for CDs and DVDs

For these media, typical errors occur when reading data. For example, let us suppose that two 1's are located pretty close. Then the reflectance detector might not be able to catch the reflectance changes. Furthermore, if two 1's are located very far, then the error in measuring time after the last reflectance change (*i.e.* clock drift) might occur. Those errors cause misreads in data, so some robustness against these errors should be applied.

3.2 Relation with Constrained Systems

To avoid these typical errors discussed above, encoding data to binary words so that the run-length of 0's are suitably adjusted would be effective. This way

of thinking is strongly related to the (d, k)-RLL constraint described in Subsection 2.3, which restricts the run-length of 0's. Indeed, CDs and DVDs utilize RLL constraints to avoid such anticipatable errors.

4 Main Contributions

In this section, we review our main contributions for data storage media. In particular, we summarize our results for constrained systems.

4.1 Main Contributions for Shannon Covers

For a constrained system $\mathcal{S}$, let us consider the set of deterministic presentations G of $\mathcal{S}$. Amongst all deterministic presentations for $\mathcal{S}$, we can find one with the smallest number of vertices. Such a presentation is called a *Shannon cover* of $\mathcal{S}$, which we denote by $G_{\mathcal{S}}$. For example, the presentation in Figure 1 is the Shannon cover of the (d, k)-RLL system.

Shannon covers play important roles in the study of constrained systems. Indeed, the study on constrained systems is strongly based on the study on Shannon covers of constrained systems. For example, Shannon covers are considered to be canonical presentations for constrained systems. Furthermore, the capacity of a constrained system (computed as (1)) is easily derived using its Shannon cover, since the complexity of computing the largest eigenvalue of the adjacency matrix of graph G depends on the number of vertices in G.

It is known that a Shannon cover turns out to be unique when a constrained system is irreducible (see, for example, [2, Theorem 3.3.18]), and an algorithm to find the Shannon cover is also well known. However, when a constrained system is reducible (not irreducible), then there can be two or more Shannon covers for the system, and there does not exist an algorithm to find a Shannon cover, up to this point.

We presented some properties regarding Shannon covers, based on the presentation generated under an algorithm introduced by Chrochemore, Mignosi and Restivo [5], which we call the *CMR presentation*. In this case, we always argue based on a natural assumption that a finite forbidden set $\mathcal{F}$ of a constrained system is *non-redundant*; that is, each forbidden word $\boldsymbol{f} \in \mathcal{F}$ is not a subword any other forbidden word $\boldsymbol{f}' \in \mathcal{F}$ (see [4] for details).

Definition 1 (Chrochemore, Mignosi, Restivo [5]). *Given a non-redundant finite forbidden set $\mathcal{F}$ of a constrained system $\mathcal{S} = \mathcal{S}_{\mathcal{F}}$, the CMR presentation of $\mathcal{S}$ is generated as follows;*

- *the vertex set $V := \{\boldsymbol{v} : \boldsymbol{v}$ is a proper prefix of some $\boldsymbol{f} \in \mathcal{F}\}$ and $Q := V \cup \mathcal{F}$.*
- *for each $\boldsymbol{v} \in \mathcal{F}$ and each symbol $a \in \Sigma$, let $\boldsymbol{u}$ be the longest suffix of $\boldsymbol{v}a$ in Q.*
  - *if $\boldsymbol{u} \notin \mathcal{F}$, draw an edge labeled a from $\boldsymbol{v}$ to $\boldsymbol{u}$.*
  - *if $\boldsymbol{u} \in \mathcal{F}$, draw no edge labeled a from $\boldsymbol{v}$.*

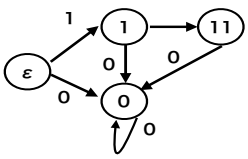

Figure 3: The CMR presentation when $\mathcal{F} = \{01, 111\}$ and $\Sigma = \{0, 1\}$. Observe that elements in $V = \{\epsilon, 1, 0, 11\}$ are proper prefixes of 01 or 110.

Theorem 1 shows that the CMR presentation derives the Shannon cover when a constrained system is characterized by a unique forbidden word.

Theorem 1 (Theorem 4.1 in [6] and Corollary 3.8 in [4]). *Let $\mathcal{F}$ be a singleton set. If a constrained system $\mathcal{S}_{\mathcal{F}}$ characterized by $\mathcal{F}$ is irreducible, then the CMR presentation is the Shannon cover of $\mathcal{S}_{\mathcal{F}}$.*

We next focused on the class of *standard* constrained systems, which are defined as follows.

Definition 2. *We call a constrained system $\mathcal{S}_{\mathcal{F}}$ standard if its forbidden set $\mathcal{F}$ has one of the following forms.*

1. **Prefix Matching (PM):** *for a fixed symbol $a \in \Sigma$, the forbidden set $\mathcal{F}$ consists of $t = |\Sigma| - 1$ words*

$$\mathcal{F} = \{a^{d_1}b_1, a^{d_2}b_2, \ldots, a^{d_s}b_t\}$$

for distinct t symbols $b_1, b_2, \ldots, b_t \in \Sigma \setminus \{a\}$.

2. **Suffix Matching (SM):** *for a fixed symbol $a \in \Sigma$, the forbidden set $\mathcal{F}$ consists of $t = |\Sigma| - 1$ words*

$$\mathcal{F} = \{b_1a^{h_1}, b_2a^{h_2}, \ldots, b_sa^{h_t}\}$$

for distinct t symbols $b_1, b_2, \ldots, b_t \in \Sigma \setminus \{a\}$.

It can be shown that standard constrained systems are always reducible, so we cannot apply the results on Shannon covers to the standard constrained systems. The following results (Theorem 2 and Corollary 1) show the properties of Shannon covers for standard constrained systems.

Theorem 2 (Lemmas IV.5 and IV.7 in [7])**.** *A Shannon cover for a standard constrained system is easily obtained from the CMR presentation. In particular, the CMR presentation turns out to be a Shannon cover when the standard constrained system is characterized by a forbidden set of a prefix matching.*

Corollary 1 (Corollary IV.6 in [7])**.** *Shannon covers for standard constrained systems are always unique.*

Thus, our contributions on Shannon covers imply that the CMR presentations can be a good starting point for finding the Shannon covers. Indeed, the CMR presentations have smaller number of vertices compared with other well-known graphs (*e.g.* De Bruijn graphs), so they can be easily constructed with lower complexity.

4.2 Main Contributions for Irreducibility

As we have observed so far, the irreducibility is an important characteristic when we study Shannon covers. We therefore considered some measurements to determine whether a given constrained system is irreducible or not, and presented some results on the irreducibility in [8].

Our first result is related to the *antidictionary* $\mathcal{A}(\boldsymbol{w})$ of $\boldsymbol{w}$, which is defined as follows.

Definition 3. *For a word $\boldsymbol{w}$, let $\boldsymbol{x}$ be a word such that*

- *$\boldsymbol{x}$ is not a subword of $\boldsymbol{w}$; and*
- *any proper subword of $\boldsymbol{x}$ is a subword of $\boldsymbol{w}$.*

We call such an $\boldsymbol{x}$ a minimal forbidden word of $\boldsymbol{w}$, and the antitictionary $\mathcal{A}(\boldsymbol{w})$ of $\boldsymbol{w}$ is a set of minimal forbidden words of $\boldsymbol{w}$.

We first showed the irreducibility from the perspective of the antidictionary as follows.

Theorem 3 (Theorem 3.2 in [8])**.** *Let $\mathcal{A}(\boldsymbol{w})$ be the antidictionary of a word $\boldsymbol{w}$, and let $\mathcal{S} = \mathcal{S}_{\mathcal{F}}$ be a constrained system characterized by a nonempty (proper) subset $\mathcal{F}$ of $\mathcal{A}(\boldsymbol{w})$. If $\boldsymbol{f}$ is not a subword of $\boldsymbol{w}^2$ for any $\boldsymbol{f} \in \mathcal{F}$, then $\mathcal{S}$ is irreducible.*

We also derived the irreducibility by focusing on the size of the alphabet as follows.

Theorem 4 (Theorem 3.11 in [8])**.** *If a forbidden set $\mathcal{F}$ of a constrained system $\mathcal{S}$ satisfies $|\mathcal{F}| \leq |\Sigma| - 2$, then $\mathcal{S}$ is irreducible.*

Theorem 5 (Theorem 3.14 in [8])**.** *Let $\mathcal{F}$ be a forbidden set of a constrained system $\mathcal{S}$ such that $|\mathcal{F}| = |\Sigma| - 1$. Then $\mathcal{S}$ is reducible if and only if $\mathcal{F}$ is standard.*

5 Conclusion

In this paper, we presented applications of graph theory in the area of engineering. In particular, we first described the summaries on typical data storage media, and then introduced how the notions of graph theory can be applied to the data storage media. Results based on graph theory (or constrained systems to be more precise) derive important contributions in coding theory for data storage media.

As a future work, we aim to apply the notion of graph theory towards DNA storage media, new data storage media with dramatically high density and long-lasting (see, for example, [9]). For example, proper coding schemes for DNA storage media should be considered based on the notion of forbidden words, so the study of constrained systems will be a promising approach. Furthermore, since the delay in reading data is one of the serious problems in DNA storage media, a data compression scheme based on graph theory will be useful.

References

[1] D. B. West, *Introduction to Graph Theory (Second Edition)*. Pearson Education, 2002.

[2] D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding*. Cambridge University Press, 1995.

[3] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," unpublished Lecture Note. [Online]. Available: http://ronny.cswp.cs.technion.ac.il/wp-content/uploads/sites/54/2016/05/chapters1-9.pdf

[4] A. Manada, "Minimal presentations of sofic shifts and properties of periodic-finite-type shifts," Ph.D. dissertation, Queen's University, 2009.

[5] M. Crochemore, F. Mignosi, and A. Restivo, "Automata and forbidden words," *Information Processing Letters*, vol. 67, no. 3, pp. 111–117, 1998.

[6] A. Manada and N. Kashyap, "On the shannon covers of certain irreducible constrained systems of finite type," in *Proc. 2006 IEEE Int. Symp. Inform. Theory*, 2006, pp. 1477–1481.

[7] A. Manada, "On a shannon cover of certain reducible shift of finite type," in *Proc. 2012 Int. Symp. Inform. Theory and Its Applications*, 2012, pp. 606–610.

[8] T. Kobayashi, A. Manada, T. Ota, and H. Morita, "On the irreducibility of certain shifts of finite type," *IEICE Trans. Fundamentals*, vol. E96-A, no. 12, pp. 1024–1031, June 2013.

[9] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "Dna-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, September 2015.