Binary Matrix Completion Under Diameter Constraints

Tomohiro Koana
Algorithmics and Computational Complexity, Faculty IV, Technische Universität Berlin, Germany

Vincent Froese
Algorithmics and Computational Complexity, Faculty IV, Technische Universität Berlin, Germany

Rolf Niedermeier
Algorithmics and Computational Complexity, Faculty IV, Technische Universität Berlin, Germany

Abstract

We thoroughly study a novel but basic combinatorial matrix completion problem: Given a binary incomplete matrix, fill in the missing entries so that the resulting matrix has a specified maximum diameter (that is, upper-bounding the maximum Hamming distance between any two rows of the completed matrix) as well as a specified minimum Hamming distance between any two of the matrix rows. This scenario is closely related to consensus string problems as well as to recently studied clustering problems on incomplete data.

We obtain an almost complete picture concerning the complexity landscape (P vs NP) regarding the diameter constraints and regarding the number of missing entries per row of the incomplete matrix. We develop polynomial-time algorithms for maximum diameter three, which are based on Deza's theorem [Discret. Math. 1973, J. Comb. Theory, Ser. B 1974] from extremal set theory. In this way, we also provide one of the rare links between sunflower techniques and stringology. On the negative side, we prove NP-hardness for diameter at least four. For the number of missing entries per row, we show polynomial-time solvability when there is only one missing entry and NP-hardness when there can be at least two missing entries. In general, our algorithms heavily rely on Deza's theorem and the correspondingly identified sunflower structures pave the way towards solutions based on computing graph factors and solving 2-SAT instances.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Mathematics of computing → Discrete mathematics

Keywords and phrases sunflowers, binary matrices, Hamming distance, stringology, consensus problems, complexity dichotomy, combinatorial algorithms, graph factors, 2-Sat, Hamming radius

Digital Object Identifier 10.4230/LIPIcs.STACS.2021.47

Related Version *Full Version*: `https://arxiv.org/abs/2002.05068`

Funding *Tomohiro Koana*: Partially supported by the DFG project FPTinP (NI 369/16).

Acknowledgements We are grateful to Christian Komusiewicz for helpful feedback on an earlier version of this work and to Stefan Szeider for pointing us to the work on clustering incomplete data [9]. We also thank Curtis Bright for mentioning the connection to the Hadamard matrix completion problem.

1 Introduction

In combinatorial matrix completion problems, given an incomplete matrix over a fixed alphabet with some missing entries, the goal is to fill in the missing entries such that the resulting "completed matrix" (over the same alphabet) fulfills a desired property. Performing a parameterized complexity analysis, Ganian et al. [14, 13] and Eiben et al. [9] recently contributed to this growing field by studying various desirable properties. More specifically, Ganian et al. [14] studied the two properties of minimizing the rank or of minimizing the

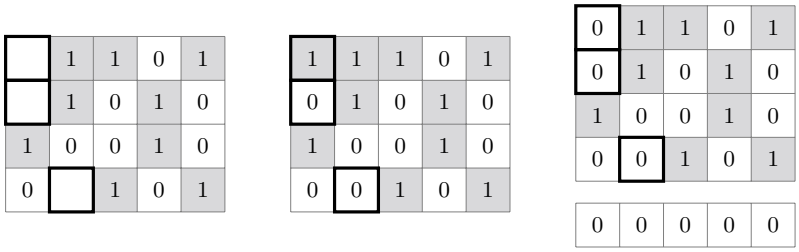


Figure 1 An illustration of matrix completion problems with the input matrix (left). Missing entries (and their completions) are framed by thick lines. The middle matrix is a completion of diameter four and the right matrix is a completion of radius three with the center vector below. Note that missing entries in the same column might be filled with different values to meet the diameter constraint, whereas this is never necessary for the radius constraint.

number of distinct rows of the completed matrix. Ganian et al. [13] analyzed the complexity of completing an incomplete matrix so that it fulfills certain constraints and can be partitioned into subspaces of small rank. Eiben et al. [9] investigated clustering problems where one wants to partition the rows of the completed matrix into a given number of clusters of small radius or of small diameter. Finally, Koana et al. [20] studied two cases of completing the matrix into one which has small (local) radius. The latter two papers [9, 20] rely on Hamming distance as a distance measure; in general, all considered matrix completion problems are NP-hard and thus the above papers [9, 14, 13, 20] mostly focused on parameterized complexity studies. In this work, we focus on a desirable property closely related to small radius, namely diameter bounds. Doing so, we further focus on the case of binary alphabet. For a matrix $\mathbf{T} \in \{0,1\}^{n\times\ell}$, let $\gamma(\mathbf{T}) := \min_{i\neq i'\in[n]} d(\mathbf{T}[i],\mathbf{T}[i'])$ and $\delta(\mathbf{T}) := \max_{i\neq i'\in[n]} d(\mathbf{T}[i],\mathbf{T}[i'])$, where d denotes the Hamming distance and $\mathbf{T}[i]$ denotes the i-th row of $\mathbf{T}$. We use the special symbol $\square$ to represent a missing entry. Specifically, we study the following problem.

Diameter Matrix Completion (DMC)

Input: An incomplete matrix $\mathbf{S} \in \{0,1,\square\}^{n\times\ell}$ and $\alpha \leq \beta \in \mathbb{N}$.
Question: Is there a completion $\mathbf{T} \in \{0,1\}^{n\times\ell}$ of $\mathbf{S}$ with $\alpha \leq \gamma(\mathbf{T})$ and $\delta(\mathbf{T}) \leq \beta$?

Before motivating the study of DMC, we refer to the example in Figure 1 that also illustrates significant differences between radius minimization [20] and diameter minimization (the latter referring to $\delta(\mathbf{T}) \leq \beta$ above).

Compare DMC with Constraint Radius Matrix Completion as studied by Koana et al. [20]:

Constraint Radius Matrix Completion (ConRMC)

Input: An incomplete matrix $\mathbf{S} \in \{0,1,\square\}^{n\times\ell}$ and $r \in \mathbb{N}^n$.
Question: Is there a completion $\mathbf{T} \in \{0,1\}^{n\times\ell}$ of $\mathbf{S}$ and a row vector $v \in \{0,1\}^{\ell}$ such that $d(v,\mathbf{T}[i]) \leq r[i]$ for all $i \in [n]$?

An important difference between DMC and ConRMC is that in DMC we basically have to compare all rows against each other, but in ConRMC we have to compare one "center row" against all others. Indeed, this makes these two similarly defined problems quite different in many computational complexity aspects.

Now, let us consider potential application scenarios where DMC may be relevant. It is a natural combinatorial matrix problems which may appear in the following contexts:

- In coding theory, one may want to "design" (by filling in the missing entries) codewords that are pairwise neither too close (parameter α in DMC) nor too far (parameter β in DMC) from each other. One prime example is the completion into a Hadamard matrix [18]. This is a special case of DMC with $n = \ell$ and $\alpha = \beta = n/2$.
- In computational biology, one may want to minimize the maximum distance of sequences in order to determine their degree of relatedness (thus minimizing β); missing entries refer to missing data points.[1]
- In data science, each row may represent an entity with its attributes, and solving the DMC problem may fulfill some constraints with respect to the pairwise (dis)similarity of the completed entities.
- In stringology, DMC seems to constitute a new and natural problem, closely related to several intensively studied consensus problems (many of which are NP-hard for binary alphabets) [1, 4, 5, 6, 16, 17, 21, 23].

Somewhat surprisingly, although simple to define and well-motivated, in the literature there seems to be no systematic study of DMC and its computational complexity. The two closest studies are the work of Eiben et al. [9] and Koana et al. [20]. Eiben et al. [9] focus on clustering while we focus on only finding one cluster (that is, the whole resulting matrix with small diameter). Another crucial difference from the work of Eiben et al. [9] is that we also model the aspect of achieving a minimum pairwise distance (not only a maximum diameter); actually, one may say that we essentially combine their "dispersion" and diameter clustering problems (for the special case of a single cluster). In this sense the problems are incomparable.

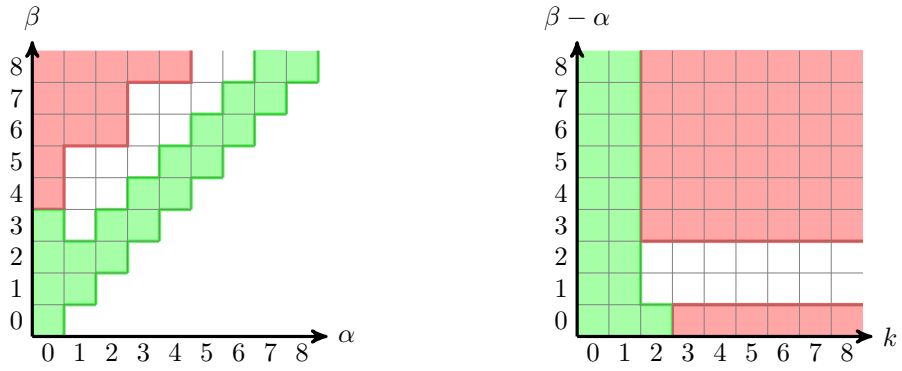
We perform a more fine-grained complexity study in terms of diameter bounds α, β and the maximum number k of missing entries in any row. Note that in bioinformatics applications matrix rows may represent sequences with few corrupted data points, thus resulting in small values for k. In fact, computational complexity with respect to this kind of parameters has been studied in the context of computational biology [1, 5, 17]. We identify polynomial-time cases as well as NP-hard cases, taking significant steps towards a computational complexity dichotomy (polynomial-time solvable versus NP-hard), leaving fairly few cases open. While the focus of the previous works [9, 20] is on parameterized complexity studies, in this work we settle more basic algorithmic questions on the DMC problem, relying on several combinatorial insights, including results from (extremal) combinatorics (most prominently, Deza's theorem [8]). Indeed, we believe that exploiting sunflowers based on Deza's theorem in combination with corresponding use of algorithms for 2-SAT and graph factors is our most interesting technical contribution. In this context, we also observe the phenomenon that the running time bounds that we can prove for odd values of α (the "lower bound for dissimilarity") are significantly better than the ones for even values of α – indeed, for even values of α the running time exponentially depends on α while it is independent of α for odd values of α. We survey our results in Figure 2 which also depicts remaining open cases.

2 Preliminaries

For $m \leq n \in \mathbb{N}$, let $[m, n] := \{m, \dots, n\}$ and let $[n] := [1, n]$.

For a matrix $\mathbf{T} \in \{0,1\}^{n \times \ell}$, we denote by $\mathbf{T}[i, j]$ the entry in the i-th row and j-th column ($i \in [n]$ and $j \in [\ell]$) of $\mathbf{T}$. We use $\mathbf{T}[i, :]$ (or $\mathbf{T}[i]$ in short) to denote the *row vector* $(\mathbf{T}[i, 1], \dots, \mathbf{T}[i, \ell])$ and $\mathbf{T}[:, j]$ to denote the *column vector* $(\mathbf{T}[1, j], \dots, \mathbf{T}[n, j])^T$. For subsets

[1] Here, it would be particularly natural to also study the case of non-binary alphabets; however, most of our positive results probably only hold for binary alphabet.



(a) Complexity of DMC with respect to combinations of constant values for α and β.

(b) Complexity of DMC with respect to combinations of the maximum number k of missing entries in any row and $\beta - \alpha$.

Figure 2 Overview of our results. Green denotes polynomial-time solvability and red denotes NP-hardness. White cells indicate open cases.

$I \subseteq [n]$ and $J \subseteq [\ell]$, we write $\mathbf{T}[I, J]$ to denote the submatrix containing only the rows in I and the columns in J. We abbreviate $\mathbf{T}[I, [\ell]]$ and $\mathbf{T}[[n], J]$ as $\mathbf{T}[I, :]$ (or $\mathbf{T}[I]$ for short) and $\mathbf{T}[:, J]$, respectively. We use the special character $\square$ for a *missing* entry. A matrix $\mathbf{S} \in \{0, 1, \square\}^{n \times \ell}$ is called *incomplete* if it contains a missing entry, and it is called *complete* otherwise. We say that $\mathbf{T} \in \{0, 1\}^{n \times \ell}$ is a *completion* of $\mathbf{S} \in \{0, 1, \square\}^{n \times \ell}$ if either $\mathbf{S}[i, j] = \square$ or $\mathbf{S}[i, j] = \mathbf{T}[i, j]$ holds for all $i \in [n]$ and $j \in [\ell]$.

Let $u, w \in \{0, 1, \square\}^{\ell}$ be row vectors. Let $D(u, w) := \{j \in [\ell] \mid u[j] \neq w[j] \wedge u[j] \neq \square \wedge w[j] \neq \square\}$ be the set of column indices where u and v disagree (not considering positions with missing entries). The *Hamming distance* between u and w is $d(u, w) := |D(u, w)|$. Note that the Hamming distance obeys the triangle inequality $d(u, w) \leq d(u, v) + d(v, w)$ for a complete vector $v \in \{0, 1\}^{\ell}$. For a subset $J \subseteq [\ell]$, we also define $d_J(u, w) := d(u[J], w[J])$. Let $u', v', w' \in \{0, 1\}^{\ell}$ be complete row vectors. Then, it holds that $d(u', w') = |D(u', v') \triangle D(v', w')| = |D(u', v')| + |D(v', w')| - 2|D(u', v') \cap D(v', w')|$. The binary operation $u \oplus v$ replaces the missing entries of u with the corresponding entries in v for $v \in \{0, 1\}^{\ell}$. We sometimes use string notation to represent row vectors, such as 001 for $(0, 0, 1)$.

3 Constant Diameter Bounds α and β

In this section we consider the special case (α, β)-DMC of DMC, where $\alpha \leq \beta$ are some fixed constants. We prove the results depicted in Figure 2a. To start with, we show the following simple linear-time special case which will subsequently be used several times.

▶ **Lemma 1.** DMC *can be solved in linear time for a constant number ℓ of columns.*

Proof. If $\alpha > 0$ and $n > 2^{\ell}$, then there is no completion $\mathbf{T}$ of $\mathbf{S}$ with $\gamma(\mathbf{T}) \geq \alpha > 0$. Thus, we can assume that the input matrix comprises of at most $n\ell \leq 2^{\ell} \cdot \ell$ (that is, constantly many) entries for the case $\alpha > 0$. Suppose that $\alpha = 0$. Consider a set $\mathcal{V} \subseteq \{0, 1\}^{\ell}$ in which the pairwise Hamming distances are at most β. We simply check whether each row vector in the input matrix can be completed to some row vector in $\mathcal{V}$ in $O(n \cdot 2^{\ell}) = O(n)$ time. Since there are at most $2^{2^{\ell}}$ choices for $\mathcal{V}$, this procedure can be done in linear time. ◀

3.1 Polynomial time for $\alpha = 0$ and $\beta \leq 3$

As an entry point, we show that $(0,1)$-DMC is easily solvable. To this end, we call a column vector *dirty* if it contains both 0 and 1. Clearly, for $\alpha = 0$, we can ignore columns that are not dirty since they can always be completed without increasing the Hamming distances between rows. Hence, throughout this subsection, we assume that the input matrix contains only dirty columns. Now, any $(0,1)$-DMC instance is a **Yes**-instance if and only if there is at most one dirty column in the input matrix:

▶ **Lemma 2.** *A matrix $\mathbf{S} \in \{0,1,\square\}^{n\times\ell}$ admits a completion $\mathbf{T} \in \{0,1\}^{n\times\ell}$ with $\delta(\mathbf{T}) \leq 1$ if and only if $\mathbf{S}$ contains at most one dirty column.*

Proof. Suppose that $\mathbf{S}$ contains two dirty columns $\mathbf{S}[:,j_0]$ and $\mathbf{S}[:,j_1]$ for $j_0 \neq j_1 \in [\ell]$. We claim that $\delta(\mathbf{T}) \geq 2$ holds for any completion $\mathbf{T}$ of $\mathbf{S}$. Let $i \in [n]$. Then, there exist $i_0, i_1 \in [n]$ with $\mathbf{T}[i,j_0] \neq \mathbf{T}[i_0,j_0]$ and $\mathbf{T}[i,j_1] \neq \mathbf{T}[i_1,j_1]$. If $\delta(\mathbf{T}) \leq 1$, then we obtain $\mathbf{T}[i_0,j_1] = \mathbf{T}[i,j_1]$ and $\mathbf{T}[i_1,j_0] = \mathbf{T}[i,j_0]$. Now we have $d(\mathbf{T}[i_0],\mathbf{T}[i_1]) \geq 2$ because $\mathbf{T}[i_0,j_0] \neq \mathbf{T}[i_1,j_0]$ and $\mathbf{T}[i_0,j_1] \neq \mathbf{T}[i_1,j_1]$. The reverse direction follows easily. ◀

Lemma 2 implies that one can solve $(0,1)$-DMC in linear time. In the following, we extend this to a linear-time algorithm for $(0,2)$-DMC (Theorem 12) and a polynomial-time algorithm for $(0,3)$-DMC (Theorem 13).

For these algorithms, we make use of a concept from extremal set theory, known as Δ-systems [19]. We therefore consider matrices as certain set systems.

▶ **Definition 3.** *For a matrix $\mathbf{T} \in \{0,1\}^{n\times\ell}$, let $\mathcal{T}$ denote the set system $\{D(\mathbf{T}[i],\mathbf{T}[n]) \mid i \in [n-1]\}$. Moreover, for $x \in \mathbb{N}$, let $\mathcal{T}_x$ denote the set system $\{D(\mathbf{T}[i],\mathbf{T}[n]) \mid i \in [n-1], d(\mathbf{T}[i],\mathbf{T}[n]) = x\}$.*

The set system $\mathcal{T}$ contains the subsets (without duplicates) of column indices corresponding to the columns where the row vectors $\mathbf{T}[1],\ldots,\mathbf{T}[n-1]$ differ from $\mathbf{T}[n]$. For given $\mathbf{T}[n]$, all the rows of $\mathbf{T}$ can be determined from $\mathcal{T}$, as we have binary alphabet.

The concept of Δ-systems has previously been used to obtain efficient algorithms [9, 10, 11]. They are defined as follows:

▶ **Definition 4** (Weak Δ-system)**.** *A set family $\mathcal{F} = \{S_1,\ldots,S_m\}$ is a* weak Δ-system *if there exists an integer $\lambda \in \mathbb{N}$ such that $|S_i \cap S_j| = \lambda$ for any pair of distinct sets $S_i, S_j \in \mathcal{F}$. The integer λ is called the* intersection size *of $\mathcal{F}$.*

▶ **Definition 5** (Strong Δ-system, Sunflower)**.** *A set family $\mathcal{F} = \{S_1,\ldots,S_m\}$ is a* strong Δ-system *(or* sunflower*) if there exists a subset $C \subseteq S_1 \cup \cdots \cup S_m$ such that $S_i \cap S_j = C$ for any pair of distinct sets $S_i, S_j \in \mathcal{F}$. We call the set C the* core *and the sets $P_i = S_i \setminus C$ the* petals *of $\mathcal{F}$.*

Clearly, every strong Δ-system is a weak Δ-system.

Our algorithms employ the combinatorial property that under certain conditions the set system $\mathcal{T}$ of a matrix $\mathbf{T}$ with bounded diameter forms a strong Δ-system (which can be algorithmically exploited). We say that a family $\mathcal{F}$ of sets is h-uniform if $|S| = h$ holds for each $S \in \mathcal{F}$. Deza [8] showed that an h-uniform weak Δ-system is a strong Δ-system if its cardinality is sufficiently large (more precisely, if $|\mathcal{F}| \geq h^2 - h + 2$). Moreover, Deza [7] also proved a stronger lower bound for uniform weak Δ-systems in which the intersection size is exactly half of the cardinality of each set.

▶ **Lemma 6** ([7, Théorème 1.1])**.** *Let $\mathcal{F}$ be a (2μ)-uniform weak Δ-system with intersection size μ. If $|\mathcal{F}| \geq \mu^2 + \mu + 2$, then $\mathcal{F}$ is a strong Δ-system.*

We extend this result to the case in which the set size is odd in the full version.

▶ **Lemma 7.** *Let $\mathcal{F}$ be a $(2\mu+1)$-uniform weak Δ-system.*
(i) *If the intersection size of $\mathcal{F}$ is $\mu+1$ and $|\mathcal{F}| \geq \mu^2+\mu+3$, then $\mathcal{F}$ is a strong Δ-system.*
(ii) *If the intersection size of $\mathcal{F}$ is μ and $|\mathcal{F}| \geq (\mu+1)^2+\mu+3$, then $\mathcal{F}$ is a strong Δ-system.*

In order to obtain a linear-time algorithm for $(0,2)$-DMC, we will prove that for $\mathbf{T} \in \{0,1\}^{n\times\ell}$ with $\delta(\mathbf{T}) \leq 2$ and sufficiently large ℓ, the set system $\mathcal{T}$ is a sunflower. This yields a linear-time algorithm via a reduction to a linear-time solvable special case of ConRMC. We start with a simple observation on matrices of diameter two, which will be helpful in the subsequent proofs.

▶ **Observation 8.** *Let $\mathbf{T} \in \{0,1\}^{n\times\ell}$ be a matrix with $\delta(\mathbf{T}) \leq 2$. For each $T_1 \in \mathcal{T}_1$ and $T_2, T_2' \in \mathcal{T}_2$, it holds that $T_1 \subseteq T_2$ and that $|T_2 \cap T_2'| \geq 1$ (otherwise there exists a pair of rows with Hamming distance three).*

The next lemma states that $|\mathcal{T}_2|$ restricts the number of columns.

▶ **Lemma 9.** *Let $\mathbf{T} \in \{0,1\}^{n\times\ell}$ be a matrix consisting of only dirty columns with $\delta(\mathbf{T}) \leq 2$. If $\mathcal{T}_2 \neq \emptyset$, then $\ell \leq |\mathcal{T}_2|+1$.*

Proof. First, observe that $\ell = |\bigcup_{T_1\in\mathcal{T}_1} T_1 \cup \bigcup_{T_2\in\mathcal{T}_2} T_2|$ because each column of $\mathbf{T}$ is dirty. Thus, it follows from Observation 8 that $\ell = |\bigcup_{T_2\in\mathcal{T}_2} T_2|$. We prove the lemma by induction on $|\mathcal{T}_2|$. Clearly, we have at most two columns if $|\mathcal{T}_2| = 1$. Suppose that $|\mathcal{T}_2| \geq 2$. For $T_2 \in \mathcal{T}_2$, we claim that

$$\ell = \left|\bigcup_{T_2'\in\mathcal{T}_2} T_2'\right| = \left|\bigcup_{T_2'\in\mathcal{T}_2\setminus\{T_2\}} T_2'\right| + \left|T_2 \setminus \bigcup_{T_2'\in\mathcal{T}_2\setminus\{T_2\}} T_2'\right| \leq |\mathcal{T}_2|+1.$$

The induction hypothesis gives us that $|\bigcup_{T_2'\in\mathcal{T}_2\setminus\{T_2\}} T_2'| \leq |\mathcal{T}_2|$. For the other term, observe that $|T_2 \setminus \bigcup_{T_2'\in\mathcal{T}_2\setminus\{T_2\}} T_2'| \leq |T_2 \setminus T_2''| = |T_2| - |T_2 \cap T_2''|$ for $T_2'' \in \mathcal{T}_2 \setminus \{T_2\}$. Hence, it follows from Observation 8 that the second term is at most 1. ◀

Next, we show that a matrix with diameter at most two has radius at most one as long as it has at least five columns. Thus, we can solve DMC by solving ConRMC with radius one, which can be done in linear time via a reduction to 2-SAT [20]. We use the following lemma concerning certain intersections of a set with elements of a sunflower.

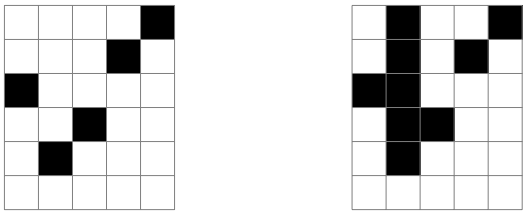
▶ **Lemma 10** ([11, Lemma 8])**.** *Let $\lambda \in \mathbb{N}$, let $\mathcal{F}$ be a sunflower with core C, and let X be a set such that $|X \cap S| \geq \lambda$ for all $S \in \mathcal{F}$. If $|\mathcal{F}| > |X|$, then $\lambda \leq |C|$ and $|X \cap C| \geq \lambda$.*

▶ **Lemma 11.** *Let $\mathbf{T} \in \{0,1\}^{n\times\ell}$ be a matrix with $\delta(\mathbf{T}) \leq 2$. If $\ell \geq 5$, then there exists a vector $v \in \{0,1\}^\ell$ such that $d(v, \mathbf{T}[i]) \leq 1$ for all $i \in [n]$.*

Proof. If $\mathcal{T}_2 = \emptyset$, then we are immediately done by definition, because $d(\mathbf{T}[n], \mathbf{T}[i]) \leq 1$ for all $i \in [n]$ (see Figure 3a for an illustration). Since $\ell \geq 5$, Lemma 9 implies $|\mathcal{T}_2| \geq 4$.

It follows from Observation 8 that $\mathcal{T}_2$ is a 2-uniform weak Δ-system with intersection size one (see Figure 3b). Thus, $\mathcal{T}_2$ is a sunflower by Lemma 6. Let $\{j_{\text{core}}\}$ denote the core of $\mathcal{T}_2$. Note that $|T_1 \cap T_2| \geq 1$ holds for each $T_1 \in \mathcal{T}_1$ and $T_2 \in \mathcal{T}_2$ by Observation 8. Now we can infer from Lemma 10 (let $X = T_1$, $\lambda = 1$, and $\mathcal{F} = \mathcal{T}_2$) that $\mathcal{T} \subseteq \{T_1\}$, where $T_1 = \{j_{\text{core}}\}$.

Hence, it holds that $d(v, \mathbf{T}[i]) \leq 1$ for all $i \in [n]$, where $v \in \{0,1\}^\ell$ is a row vector such that $v[j_{\text{core}}] = 1 - \mathbf{T}[n, j_{\text{core}}]$ and $v[j] = \mathbf{T}[n, j]$ for each $j \in [\ell] \setminus \{j_{\text{core}}\}$. ◀



(a) The case $\mathcal{T}_2 = \emptyset$. **(b)** The case $|\mathcal{T}_2| \geq 4$.

Figure 3 Illustration of Lemma 11 with $n = 6$. A black cell denotes a value different from row $\mathbf{T}[6]$. In (b) the set system $\mathcal{T}_2$ forms a sunflower with core $\{2\}$. In both cases the radius is one.

▶ **Theorem 12.** $(0, 2)$-DMC *can be solved in* $O(n\ell)$ *time.*

Proof. Let $\mathbf{S} \in \{0, 1, \square\}^{n \times \ell}$ be an input matrix of $(0, 2)$-DMC. If $\ell \leq 4$, then we use the linear-time algorithm of Lemma 1. Henceforth, we assume that $\ell \geq 5$.

We claim that $\mathbf{S}$ is a **Yes**-instance if and only if the ConRMC instance $I = (\mathbf{S}, 1^n)$ is a **Yes**-instance.

($\Rightarrow$) Let $\mathbf{T}$ be a completion of $\mathbf{S}$ with $\delta(\mathbf{T}) \leq 2$. Since $\ell \geq 5$, there exists a vector v such that $d(v, \mathbf{T}[i]) \leq 1$ for all $i \in [n]$ by Lemma 11. It follows that I is a **Yes**-instance.

($\Leftarrow$) Let v be a solution of I. Let $\mathbf{T}$ be the matrix such that for each $i \in [n]$, $\mathbf{T}[i] = \mathbf{S}[i] \oplus v$ (recall that $u \oplus v$ denotes the vector obtained from u by replacing all missing entries of u with the entries of v in the corresponding positions). Then, we have $d(v, \mathbf{T}[i]) \leq 1$ for each $i \in [n]$. By the triangle inequality, we obtain $d(\mathbf{T}[i], \mathbf{T}[i']) \leq d(v, \mathbf{T}[i]) + d(v, \mathbf{T}[i']) \leq 2$ for each $i, i' \in [n]$.

Since ConRMC can be solved in linear time when $\max_{i \in [n]} r[i] = 1$ [20, Theorem 1], it follows that $(0, 2)$-DMC can be solved in linear time. ◀

In the full version, we show polynomial-time solvability of $(0, 3)$-DMC. The overall idea is, albeit technically more involved, similar to $(0, 2)$-DMC. We first show that the set family $\mathcal{T}$ of a matrix $\mathbf{T}$ with $\delta(\mathbf{T}) = 3$ contains a sunflower by Lemma 7. We then show that such a matrix has a certain structure which again allows us to reduce the problem to the linear-time solvable special case of ConRMC with radius one.

▶ **Theorem 13.** $(0, 3)$-DMC *can be solved in* $O(n\ell^4)$ *time.*

Our algorithms work via reductions to ConRMC. Although ConRMC imposes an upper bound on the diameter implicitly by the triangle inequality, it is seemingly difficult to enforce any lower bounds (that is, $\alpha > 0$). In the next subsection, we will see polynomial-time algorithms for $\alpha > 0$, based on reductions to the graph factorization problem.

3.2 Polynomial time for $\beta = \alpha + 1$

We now give polynomial-time algorithms for (α, β)-DMC with constant $\alpha > 0$ given that $\beta \leq \alpha + 1$. As in Section 3.1, our algorithms exploit combinatorial structures revealed by Deza's theorem (Lemmas 6 and 7). Recall that $\mathcal{T}$ denotes a set system obtained from a complete matrix $\mathbf{T}$ (Definition 3). We show that $\mathcal{T}$ essentially is a sunflower when $\gamma(\mathbf{T}) \geq \alpha$ and $\delta(\mathbf{T}) \leq \alpha + 1$. For the completion into such a sunflower, it suffices to solve the following matrix completion problem, which we call Sunflower Matrix Completion.

Sunflower Matrix Completion (SMC)

Input: An incomplete matrix $\mathbf{S} \in \{0,1,\square\}^{n\times\ell}$ and $s, m \in \mathbb{N}$.

Question: Is there a completion $\mathbf{T} \in \{0,1\}^{n\times\ell}$ of $\mathbf{S}$ such that $D(\mathbf{T}[1],\mathbf{T}[n]),\ldots,D(\mathbf{T}[n-1],\mathbf{T}[n])$ are pairwise disjoint sets each of size at most s and $\sum_{i\in[n-1]} |D(\mathbf{T}[i],\mathbf{T}[n])| \geq m$.

Intuitively speaking, the problem asks for a completion into a sunflower with empty core and bounded petal sizes. All algorithms presented in this subsection are via reductions to SMC. First, we show that SMC is indeed polynomial-time solvable. We prove this using a well-known polynomial-time algorithm for the graph problem (g,f)-Factor [12].

(g,f)-Factor

Input: A graph $G=(V,E)$, functions $f,g\colon V\to\mathbb{N}$, and $m'\in\mathbb{N}$.

Question: Does G contain a subgraph $G'=(V,E')$ such that $|E'|\geq m'$ and $g(v)\leq \deg_{G'}(v)\leq f(v)$ for all $v\in V$?

▶ **Lemma 14.** *For constant $s>0$,* SMC *can be solved in $O(n\ell\sqrt{n+\ell})$ time.*

Proof. Let $(\mathbf{S},s,m)$ be an SMC instance. Let a_j^x be the number of occurrences of $x\in\{0,1\}$ in $\mathbf{S}[:,j]$ for each $j\in[\ell]$. We can assume that $a_j^0\geq a_j^1$ for each $j\in[\ell]$ (otherwise swap the occurrences of 0's and 1's in the column). If $a_j^0\geq 2$ and $\mathbf{S}[n,j]=1$ for some $j\in[\ell]$, then we can return **No** since there will be two intersecting sets. Also, if $a_j^1\geq 2$, then we return **No**.

We construct an instance of (g,f)-Factor as follows. We introduce a vertex u_i for each $i\in[n-1]$ and a vertex v_j for each $j\in[\ell]$. The resulting graph G will be a bipartite graph with one vertex subset $\{u_1,\ldots,u_{n-1}\}$ representing rows and the other $\{v_1,\ldots,v_\ell\}$ representing columns. Essentially, we add an edge between u_i and v_j if the column $\mathbf{S}[:,j]$ can be completed such that the i-th entry differs from all other entries on $\mathbf{S}[:,j]$ (see Figure 4 for an illustration). Intuitively, such an edge encodes the information that column index j can be contained in a petal of the sought sunflower. Formally, there is an edge $\{u_i,v_j\}$ if and only if there is a completion $t_j\in\{0,1\}^n$ of $\mathbf{S}[:,j]$ in which $t_j[i]=1-t_j[n]$ and $t_j[h]=t_j[n]$ for all $h\in[n-1]\setminus\{i\}$. We set $g(u_i):=0$ and $f(u_i):=s$ for each $i\in[n-1]$, $g(v_j):=a_j^1$ and $f(v_j):=1$ for each $j\in[\ell]$, and $m':=m$. This construction can be done in $O(n\ell)$ time. To see this, note that the existence of an edge $\{u_i,v_j\}$ only depends on a_j^0, a_j^1, and $\mathbf{S}[i,j]$.

- If $a_j^0\leq 1$ and $a_j^1=0$, then add the edge $\{u_i,v_j\}$. The corresponding completion t_j can be seen as follows:
  - If $\mathbf{S}[h,j]=\square$ for all $h\in[n-1]$, then let $t_j[i]:=1$ and let $t_j[h]:=0$ for all $h\in[n]\setminus\{i\}$.
  - If $\mathbf{S}'[h,j]=0$ for some $h\in[n-1]$, then $\mathbf{S}'[h',j]=\square$ for all $h'\in[n]\setminus\{h\}$. If $h\neq i$, then let $t_j[i]:=1$ and let $t_j[h]:=0$ for all $h\in[n]\setminus\{i\}$. Otherwise, let $t_j[h]:=1$ for all $h\in[n]\setminus\{i\}$.
- If $a_j^0=1$ and $a_j^1=1$, then add the edge $\{u_i,v_j\}$ if $\mathbf{S}[i,j]\neq\square$.
- If $a_j^0\geq 2$ and $a_j^1=0$, then add the edge $\{u_i,v_j\}$ if $\mathbf{S}[i,j]=\square$.
- If $a_j^0\geq 2$ and $a_j^1=1$, then add the edge $\{u_i,v_j\}$ if $\mathbf{S}[i,j]=1$ (because $\mathbf{S}[n,j]$ must be completed with 0).

The correctness of the reduction easily follows from the definition of an edge: If $\mathbf{T}$ is a solution for $(\mathbf{S},s,m)$, then the corresponding subgraph of G contains the edge $\{u_i,v_j\}$ for each $i\in[n-1]$ and each $j\in D(\mathbf{T}[i],\mathbf{T}[n])$. Conversely, a completion of $\mathbf{S}$ is obtained from a subgraph G' by taking for each edge $\{u_i,v_j\}$ the corresponding completion t_j as the j-th column. Note that no vertex v_j can have two incident edges since $f(v_j)=1$. Moreover, if v_j has no incident edges, then this implies that $g(v_j)=a_j^1=0$. Hence, we can complete all missing entries in column j by 0.

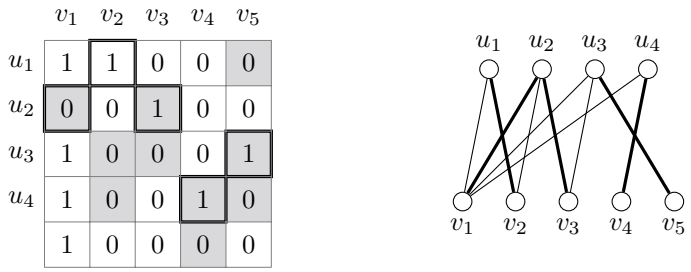


Figure 4 A completion of a 5×5 incomplete matrix (left). The known entries are highlighted in gray. A bipartite graph as constructed in the reduction (right). Note that the entries framed by thick lines (which differ from all others in the same column) correspond to the subgraph represented by the thick lines.

Regarding the running time, note that the constructed graph G has at most $n\ell$ edges and $\sum_{i\in[n-1]} f(u_i) \in O(n)$ and $\sum_{j\in[\ell]} f(v_j) \in O(\ell)$. Since (g,f)-Factor can be solved in $O(|E|\sqrt{f(V)})$ time [12] for $f(V) = \sum_{v\in V} f(v)$, SMC can be solved in $O(n\ell\sqrt{n+\ell})$ time. ◀

Using Lemma 14, we first show that (α,α)-DMC can be solved in polynomial time.

▶ **Theorem 15.** (α,α)-DMC *can be solved in* $O(n\ell\sqrt{n+\ell})$ *time.*

Proof. We first show that (α,α)-DMC can easily be solved if α is odd. Consider row vectors $u,v,w \in \{0,1\}^\ell$ and let $U := D(u,v)$ and $W := D(v,w)$. Then, $d(u,v)+d(v,w)+d(w,u) = |U|+|W|+(|U|+|W|-2|U\cap W|) = 2(|U|+|W|-|U\cap W|)$ and hence $d(u,v)+d(v,w)+d(w,u)$ is even. Thus, we can immediately answer **No** if $n \geq 3$. It is also easy to see that DMC can be solved in linear time if $n \leq 2$.

We henceforth assume that α is even. Eiben et al. [9, Theorem 34] provided a linear-time algorithm for $(0,\alpha)$-DMC with constant n (and arbitrary α) using reductions to integer linear programming (ILP). It is straightforward to adapt their ILP formulation to show that (α,α)-DMC can also be solved in linear time for constant n (basically, we just need the additional constraint that each pairwise distance is at least α). So we can assume that $n \geq (\alpha/2)^2 + (\alpha/2) + 3$ (otherwise (α,α)-DMC can be solved in linear time). We claim that there is a completion $\mathbf{T}$ of $\mathbf{S}$ with $\gamma(\mathbf{T}) = \delta(\mathbf{T}) = \alpha$ if and only if the SMC instance $(\mathbf{S}', \alpha/2, \alpha n/2)$ is a **Yes**-instance for the matrix $\mathbf{S}' \in \{0,1,\square\}^{(n+1)\times\ell}$ obtained from $\mathbf{S}$ with an additional row vector $\square^\ell$.

($\Rightarrow$) Let $\mathbf{T}$ be a completion of $\mathbf{S}$ with $\gamma(\mathbf{T}) = \delta(\mathbf{T}) = \alpha$. Then, $\mathcal{T}$ is a weak Δ-system with intersection size $\alpha/2$. Since $|\mathcal{T}| \geq (\alpha/2)^2 + (\alpha/2) + 2$, Lemma 6 tells us that $\mathcal{T}$ is a sunflower. Let C be the core of $\mathcal{T}$. Consider the completion $\mathbf{T}'$ of $\mathbf{S}'$ such that

- $\mathbf{T}'[[n],:] = \mathbf{T}$,
- $\mathbf{T}'[n+1,j] = 1 - \mathbf{T}[n,j]$ for each $j \in C$, and
- $\mathbf{T}'[n+1,j] = \mathbf{T}[n,j]$ for each $j \in [\ell] \setminus C$.

Note that $D(\mathbf{T}'[i],\mathbf{T}'[n+1]) = D(\mathbf{T}'[i],\mathbf{T}'[n]) \setminus C$ for each $i \in [n-1]$. Note also that $D(\mathbf{T}'[n],\mathbf{T}'[n+1]) = C$. Hence, $D(\mathbf{T}'[1],\mathbf{T}'[n+1]),\ldots,D(\mathbf{T}'[n],\mathbf{T}'[n+1])$ are pairwise disjoint sets of size $\alpha/2$.

($\Leftarrow$) Let $\mathbf{T}'$ be a completion of $\mathbf{S}'$ such that $D(\mathbf{T}'[1],\mathbf{T}'[n+1]),\ldots,D(\mathbf{T}'[n],\mathbf{T}'[n+1])$ are pairwise disjoint sets of size $\alpha/2$. For the completion $\mathbf{T} = \mathbf{T}'[[n],:]$ of $\mathbf{S}$, it holds that $d(\mathbf{T}[i],\mathbf{T}[i']) = |D(\mathbf{T}'[i],\mathbf{T}'[n+1]) \triangle D(\mathbf{T}'[i'],\mathbf{T}'[n+1])| = |D(\mathbf{T}'[i],\mathbf{T}'[n+1])| + |D(\mathbf{T}'[i'],\mathbf{T}'[n+1])| = \alpha$ for each $i,i' \in [n]$. ◀

Now we proceed to develop polynomial-time algorithms for the case $\alpha + 1 = \beta$. We will make use of the following observation made by Froese et al. [11, Proof of Theorem 9].

▶ **Observation 16.** *Let* $\mathbf{T} \in \{0,1\}^{n \times \ell}$ *with* $\gamma(\mathbf{T}) \geq \alpha$ *and* $\delta(\mathbf{T}) \leq \beta = \alpha + 1$. *For* $T_\alpha \neq T'_\alpha \in \mathcal{T}_\alpha$ *and* $T_\beta \neq T'_\beta \in \mathcal{T}_\beta$, *it holds that* $|T_\alpha \cap T'_\alpha| = \lfloor \alpha/2 \rfloor$, $|T_\alpha \cap T_\beta| = \lceil \alpha/2 \rceil = \lfloor \beta/2 \rfloor$, *and* $|T_\beta \cap T'_\beta| = \lceil \beta/2 \rceil$.

Surprisingly, odd α seems to allow for significantly more efficient algorithms than even α.

▶ **Theorem 17.** (α, β)-DMC *with* $\beta = \alpha + 1$ *can be solved in*
(i) $O(n\ell\sqrt{n+\ell})$ *time for odd* α, *and*
(ii) $(n\ell)^{O(\alpha^3)}$ *time for even* α.

Proof. (i) We can assume that $n \geq \beta^2/2 + \beta + 7$ holds since otherwise the problem is linear-time solvable via a reduction to ILP as in the proof of Theorem 15. Suppose that $\mathbf{S}$ admits a completion $\mathbf{T}$ with $\gamma(\mathbf{T}) \geq \alpha$ and $\delta(\mathbf{T}) \leq \beta$. Since $\mathcal{T} = \mathcal{T}_\alpha \cup \mathcal{T}_\beta$ and $|\mathcal{T}| \geq \beta^2/2 + \beta + 6$, it follows that $\max\{|\mathcal{T}_\alpha|, |\mathcal{T}_\beta|\} \geq c := (\beta/2)^2 + (\beta/2) + 3$. We consider two cases depending on the size of $\mathcal{T}_\alpha$ and $\mathcal{T}_\beta$.

1. Suppose that $|\mathcal{T}_\alpha| \geq c$. Since $\mathcal{T}_\alpha$ is a weak Δ-system with intersection size $(\alpha-1)/2$, $\mathcal{T}_\alpha$ is a sunflower with a core of size $(\alpha-1)/2$ and petals of size $(\alpha+1)/2$ by Lemma 7 (ii). We claim that $\mathcal{T}_\beta = \emptyset$. Suppose not and let $T_\beta \in \mathcal{T}_\beta$. Consequently, we obtain $|T_\alpha \cap T_\beta| = (\alpha+1)/2$ for all $T_\alpha \in \mathcal{T}_\alpha$ by Observation 16, which contradicts Lemma 10.
2. Suppose that $|\mathcal{T}_\beta| \geq c$. Again, $\mathcal{T}_\beta$ is a sunflower whose core C has size $\beta/2$ by Lemma 6. By Observation 16 and Lemma 10, $T_\alpha \supseteq C$ holds for each $T_\alpha \in \mathcal{T}_\alpha$. Now suppose that there exist $T_\alpha \neq T'_\alpha \in \mathcal{T}_\alpha$. Since $C \subseteq T_\alpha$ and $C \subseteq T'_\alpha$, it follows that $|T_\alpha \cap T'_\alpha| \geq \beta/2$, thereby contradicting Observation 16. Hence, we have $|\mathcal{T}_\alpha| \leq 1$.

We construct an instance I of SMC covering both cases above, as in Theorem 15. We use the matrix $\mathbf{S}'$ obtained from $\mathbf{S}$ by appending a row vector $\square^\ell$, and we set $s := \beta/2$ and $m := ns - 1$. Basically, we allow at most one "petal" to have size $s - 1$. We return **Yes** if and only if I is a **Yes**-instance. The correctness can be shown analogously to the proof of Theorem 15.

(ii) Suppose that there is a completion $\mathbf{T}$ of $\mathbf{S}$ with $\gamma(\mathbf{T}) \geq \alpha$ and $\delta(\mathbf{T}) \leq \beta$. Again, we can assume that $n > 2c$ for $c := (\beta/2)^2 + (\beta/2) + 4$, and consider a case distinction regarding the size of $\mathcal{T}_\alpha$ and $\mathcal{T}_\beta$.

1. Suppose that $|\mathcal{T}_\alpha| \geq c$ and $|\mathcal{T}_\beta| \geq c$. It follows from Observation 16 and Lemmas 6 and 7 that $\mathcal{T}_\alpha$ and $\mathcal{T}_\beta$ are sunflowers. Let C_α and C_β be the cores of $\mathcal{T}_\alpha$ and $\mathcal{T}_\beta$, respectively. Note that $|C_\alpha| = \alpha/2$ and $|C_\beta| = \alpha/2 + 1$, and hence $C_\alpha \subsetneq C_\beta$ holds by Observation 16 and Lemma 10. Let $j \in [\ell]$ be such that $C_\alpha \cup \{j\} = C_\beta$ and let $\mathbf{T}' := \mathbf{T}[:, [\ell] \setminus \{j\}]$. Then, the set family $\mathcal{T}'$ is a sunflower with a core of size $\alpha/2$ and petals of size $\alpha/2$. Hence, there exists $j \in [\ell]$ such that the (α, α)-DMC instance $\mathbf{S}[:, [\ell] \setminus \{j\}]$ is a **Yes**-instance. On the other hand, if there is a completion $\mathbf{T}'$ of $\mathbf{S}[:, [\ell] \setminus \{j\}]$ with $\gamma(\mathbf{T}') = \delta(\mathbf{T}') = \alpha$, then $\gamma(\mathbf{T}) \geq \alpha$ and $\delta(\mathbf{T}) \leq \alpha + 1$ hold for any completion $\mathbf{T}$ of $\mathbf{S}$ with $\mathbf{T}[:, [\ell] \setminus \{j\}] = \mathbf{T}'$.
2. Suppose that $|\mathcal{T}_\alpha| \geq c$ and $|\mathcal{T}_\beta| < c$. The same argument as above shows that $T_\alpha \cap T_\beta = C$ holds for each $T_\alpha \in \mathcal{T}_\alpha$ and $T_\beta \in \mathcal{T}_\beta$, where C is the size-$\alpha/2$ core of sunflower $\mathcal{T}_\alpha$. Let $I_\beta = \{i \in [n-1] \mid d(\mathbf{T}[i], \mathbf{T}[n]) = \beta\}$ be the row indices that induce the sets in $\mathcal{T}_\beta$ and let $J_\beta = \bigcup_{T_\beta \in \mathcal{T}_\beta} T_\beta$. Consider $\mathbf{T}' = \mathbf{S}[[n] \setminus I_\beta, [\ell] \setminus (C \cup J_\beta)]$ and note that the family $\mathcal{T}'$ consists of pairwise disjoint sets, each of size $\alpha/2$. We use this observation to obtain a reduction to SMC. The idea is to test all possible choices for $\mathbf{T}'$, that is, we simply try out all possibilities to choose the following sets:
   - $C \subseteq [\ell]$ of size exactly $\alpha/2$.

- $I_\beta \subseteq [n-1]$ of size at most c.
- $J_\beta \subseteq [\ell] \setminus C$ of size at most $\beta \cdot c$ such that $d_{[\ell]\setminus(C\cup J_\beta)}(\mathbf{S}[i_\beta], \mathbf{S}[n]) = 0$ for all $i_\beta \in I_\beta$.

For each possible choice, we check whether it allows for a valid completion. Formally, it is necessary that the following exist:

- A completion t_C of $\mathbf{S}[n, C]$ such that $\mathbf{S}[i, j] \neq t_C[j]$ for all $i \in [n-1]$ and $j \in C$.
- A completion t_{J_β} of $\mathbf{S}[n, J_\beta]$ such that $d(t_{J_\beta}, \mathbf{S}[i, J_\beta]) = 0$ for all $i \in [n-1] \setminus I_\beta$.
- A completion t_{i_β} of $\mathbf{S}[i_\beta, J_\beta]$ for each $i_\beta \in I_\beta$ such that $d(t_{i_\beta}, t_{J_\beta}) = \alpha/2 + 1$ for each $i_\beta \in I_\beta$ and $d(t_{i_\beta}, t_{i'_\beta}) = \alpha$ for each $i_\beta \neq i'_\beta \in I_\beta$.

The existence of the above completions can be checked in $O(n)$ time. We then construct an SMC instance $(\mathbf{S}', \alpha/2, (n - |I_\beta| - 1) \cdot \alpha/2)$, where $\mathbf{S}'$ is an incomplete matrix with $n' = n - |I_\beta|$ rows and $\ell - |C| - |J_\beta|$ columns defined as follows:

- $\mathbf{S}'[[n'-1]] = \mathbf{S}[[n-1] \setminus I_\beta, [\ell] \setminus (C \cup J_\beta)]$.
- $\mathbf{S}'[n', j] = \square$ for each $j \in [\ell] \setminus (C \cup J_\beta)$ such that $\mathbf{S}[i_\beta, j] = \square$ for all $i_\beta \in I_\beta \cup \{n\}$.
- $\mathbf{S}'[n', j] = \mathbf{S}[i_\beta, j]$ for each $j \in [\ell]\setminus(C\cup J_\beta)$ such that $\mathbf{S}[i_\beta, j] \neq \square$ for some $i_\beta \in I_\beta \cup \{n\}$.

Overall, we solve at most $(n\ell)^{O(\alpha^3)}$ SMC instances and hence it requires $(n\ell)^{O(\alpha^3)}$ time.

3. Suppose that $|\mathcal{T}_\alpha| < c$ and $|\mathcal{T}_\beta| \geq c$. Let $i \in [n-1]$ be such that $d(\mathbf{T}[i], \mathbf{T}[n]) = \beta$. Then, $d(\mathbf{T}[i], \mathbf{T}[i']) = \alpha$ holds for each $i' \in [n-1] \setminus \{i\}$ with $d(\mathbf{T}[i'], \mathbf{T}[n]) = \beta$. Since there are at least $c - 1 = (\beta/2)^2 + (\beta/2) + 3$ such row indices, it follows that this case is essentially equivalent to the previous case (by considering row i as the last row). ◀

A natural question is whether one can extend our approach above to the case $\beta = \alpha + 2$ (particularly $\alpha = 1$ and $\beta = 3$). The problem is that the petals of the sunflowers $\mathcal{T}_2$ and $\mathcal{T}_3$ may have nonempty intersections. Thus, reducing to SMC to obtain a polynomial-time algorithm is probably hopeless.

3.3 NP-hardness

Hermelin and Rozenberg [17, Theorem 5] proved that ConRMC (under the name Closest String with Wildcards) is NP-hard even if $r[i] = 2$ for all $i \in [n]$. Using this result, we prove the following (the proof is in the full version).

▶ **Theorem 18.** (α, β)-DMC *is NP-hard if* $\beta \geq 2\lceil \alpha/2 \rceil + 4$.

It remains open whether NP-hardness also holds for $(\alpha, \alpha + 3)$-DMC with $\alpha \geq 1$ (recall that $(0, 3)$-DMC is polynomial-time solvable). In Section 4, however, we show NP-hardness for $\beta = \alpha + 3$ when α and β are part of the input.

4 Bounded number k of missing entries per row

In this section, we consider DMC with α and β being part of the input, hence not necessarily being constants. We consider the maximum number k of missing entries in any row as a parameter (DMC is clearly trivial for $k = 0$). We obtain two polynomial-time algorithms and two NP-hardness results. Our polynomial-time algorithms are based on reductions to 2-SAT (see the full version for the proof).

▶ **Theorem 19.** DMC *can be solved in* $O(n^2\ell)$ *time*

(i) *for* $k = 1$, *and*

(ii) *for* $k = 2$ *and* $\alpha = \beta$.

$$\mathbf{T} = \left[\begin{array}{c|c|c|c} 001 & 111 & 001 & 000000000 \\ 111 & 111 & 001 & 000000000 \\ \hline 111 & 111 & 010 & 111111111 \\ 111 & 010 & 111 & 111111111 \end{array}\right]$$

Figure 5 An illustration of the reduction from Orthogonal Vectors, where $\mathcal{U} = \{010, 110\}$ and $\mathcal{V} = \{110, 101\}$.

To complement this result, we show that the quadratic dependence on n in the running time of Theorem 19 is inevitable under Orthogonal Vectors Conjecture (OVC), which states that Orthogonal Vectors cannot be solved in $O(n^{2-\epsilon} \cdot \ell^c)$ time for any $\epsilon, c > 0$ (it is known that Strong Exponential Time Hypothesis implies OVC [24]).

Orthogonal Vectors
Input: Sets $\mathcal{U}, \mathcal{V}$ of row vectors in $\{0,1\}^\ell$ with $|\mathcal{U}| = |\mathcal{V}| = n$.
Question: Are there row vectors $u \in \mathcal{U}$ and $v \in \mathcal{V}$ such that $u[j] \cdot v[j] = 0$ holds for all $j \in [\ell]$?

▶ **Theorem 20.** DMC *cannot be solved in* $O(n^{2-\epsilon} \cdot \ell^c)$ *time for any* $c, \epsilon > 0$*, unless* OVC *breaks.*

Proof. We reduce from Orthogonal Vectors. Let $u_1, \ldots, u_n, v_1, \ldots, v_n \in \{0,1\}^\ell$ be row vectors. Consider the matrix $\mathbf{T} \in \{0,1\}^{2n \times 6\ell}$ where

$$\mathbf{T}[i, [3j-2, 3j]] = \begin{cases} 001 & \text{if } u_i[j] = 0, \\ 111 & \text{if } u_i[j] = 1, \end{cases} \qquad \mathbf{T}[i, [3\ell+3j-2, 3\ell+3j]] = 000,$$

$$\mathbf{T}[n+i, [3j-2, 3j]] = \begin{cases} 010 & \text{if } v_i[j] = 0, \\ 111 & \text{if } v_i[j] = 1, \end{cases} \qquad \mathbf{T}[n+i, [3\ell+3j-2, 3\ell+3j]] = 111,$$

for each $i \in [n]$ and $j \in [\ell]$ (see Figure 5 for an illustration). We show that $\delta(\mathbf{T}) = 5\ell$ if and only if there are $i, i' \in [n]$ such that u_i and $v_{i'}$ are orthogonal. By construction, we have

$$d(\mathbf{T}[i, [3j-2, 3j]], \mathbf{T}[n+i', [3j-2, 3j]]) = \begin{cases} 2 & \text{if } u_i[j] = 0 \text{ or } v_i[j] = 0, \\ 0 & \text{otherwise.} \end{cases}$$

for any $i, i' \in [n]$ and $j \in [\ell]$. Thus, it holds for any orthogonal vectors u_i and v'_i that $d(\mathbf{T}[i], \mathbf{T}[n+i']) = 5\ell$. Conversely, suppose that there exist $i < i' \in [2n]$ such that $d(\mathbf{T}[i], \mathbf{T}[i']) = 5\ell$. It is easy to see that $i \in [n]$, $i' \in [n+1, 2n]$ (otherwise $d(\mathbf{T}[i], \mathbf{T}[i']) \leq 3\ell$). Then, the vectors u_i and $v_{i'-n}$ are orthogonal. ◀

Finally, we prove the following two NP-hardness results (the proofs are in the full version).

▶ **Theorem 21.** DMC *is NP-hard*
(i) *for* $k = 2$ *and* $\alpha + 3 \leq \beta$*, and*
(ii) *for* $k = 3$ *and* $\alpha = \beta$*.*

The proof for Theorem 21 is based on rather technical reductions from $(3, \text{B2})$-SAT [2] and Cubic Monotone 1-in-3 SAT [22]. The challenge here is to ensure the bounds on the Hamming distances between all row pairs. To overcome this challenge, we adjust pairwise row distances by making heavy use of the matrix, in which one pair of rows has distance exactly two greater than any other:

▶ **Lemma 22.** *For each $n \geq 3$ and $i < i' \in [n]$, one can construct in $n^{O(1)}$ time, a matrix $\mathbf{B}^n_{i,i'} \in \{0,1\}^{n \times \ell}$ with n rows and $\ell := (\binom{n}{2} - 1)(2n-1)$ columns such that for all $h \neq h' \in [n]$,*

$$d(\mathbf{B}^n_{i,i'}[h], \mathbf{B}^n_{i,i'}[h']) = \begin{cases} \gamma(\mathbf{B}^n_{i,i'}) + 2 & \text{if } (h,h') = (i,i'), \\ \gamma(\mathbf{B}^n_{i,i'}) & \text{otherwise.} \end{cases}$$

The problem of deciding whether an incomplete matrix $\mathbf{S} \in \{1, -1, \square\}^{n \times n}$ can be completed into a Hadamard matrix as seen in Johnson [18], is equivalent to DMC with $n = \ell$ and $\alpha = \beta = n/2$. We conjecture that one can adapt the proof of Theorem 21 (ii) to show the NP-hardness of this problem. We also conjecture that DMC with $k = 3$ is actually NP-hard for every value of $\beta - \alpha$. Similar reductions might work here as well. By contrast, we believe the case $k = 2$ and $\beta - \alpha = 1$ to be polynomial-time solvable, again by reducing to 2-SAT.

5 Conclusion

Together with the recent work of Eiben et al. [9], we are seemingly among the first in the context of stringology that make extensive use of Deza's theorem and sunflowers. While Eiben et al. [9] achieved classification results in terms of parameterized (in)tractability, we conducted a detailed complexity analysis in terms of polynomial-time solvable versus NP-hard cases. Figure 2 provides a visual overview on our results for Diameter Matrix Completion (DMC), also spotting concrete open questions.

Going beyond open questions directly arising from Figure 2, we remark that it is known that the clustering variant of DMC can be solved in polynomial time when the number of clusters is two and the matrix is complete [15]. Hence, it is natural to ask whether our tractability results can be extended to this matrix completion clustering problem as well. Furthermore, we proved that there are polynomial-time algorithms solving DMC when $\beta \leq 3$ and $\alpha = 0$ (Theorems 12 and 13). This leads to the question whether these algorithms can be extended to matrices with arbitrary alphabet size. Next, we are curious whether the phenomenon we observed in Theorem 17 concerning the exponential dependence of the running time for $(\alpha, \alpha+1)$-DMC when α is even but independence of α when it is odd can be further substantiated or whether one can get rid of the "α-dependence" in the even case. In terms of standard parameterized complexity analysis, we wonder whether DMC is fixed-parameter tractable with respect to $\beta + k$ (in our NP-hardness proof for the case $\beta = 4$ (Theorem 18) we have $k \in \theta(\ell)$). Finally, performing a multivariate fine-grained complexity analysis in the same spirit as in recent work for Longest Common Subsequence [3] would be another natural next step.

References

1 Vineet Bafna, Sorin Istrail, Giuseppe Lancia, and Romeo Rizzi. Polynomial and APX-hard cases of the individual haplotyping problem. *Theoretical Computer Science*, 335(1):109–125, 2005.

2 Piotr Berman, Marek Karpinski, and Alex D. Scott. Approximation hardness of short symmetric instances of MAX-3SAT. *Electronic Colloquium on Computational Complexity (ECCC)*, 049, 2003.

3 Karl Bringmann and Marvin Künnemann. Multivariate fine-grained complexity of longest common subsequence. In *29th Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA '18)*, pages 1216–1235, 2018.

4 Laurent Bulteau, Vincent Froese, and Rolf Niedermeier. Tight hardness results for consensus problems on circular strings and time series. *SIAM Journal on Discrete Mathematics*, 34(3):1854–1883, 2020.

5 Laurent Bulteau, Falk Hüffner, Christian Komusiewicz, and Rolf Niedermeier. Multivariate algorithmics for NP-hard string problems. *Bulletin of the EATCS*, 114, 2014.

6 Laurent Bulteau and Markus L. Schmid. Consensus strings with small maximum distance and small distance sum. *Algorithmica*, 82(5):1378–1409, 2020.

7 Michel Deza. Une propriété extrémale des plans projectifs finis dans une classe de codes équidistants. *Discrete Mathematics*, 6(4):343–352, 1973.

8 Michel Deza. Solution d'un problème de Erdős-Lovász. *Journal of Combinatorial Theory, Series B*, 16(4):166–167, 1974.

9 Eduard Eiben, Robert Ganian, Iyad Kanj, Sebastian Ordyniak, and Stefan Szeider. On clustering incomplete data. *CoRR*, abs/1911.01465, 2019. `arXiv:1911.01465`.

10 Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2006.

11 Vincent Froese, René van Bevern, Rolf Niedermeier, and Manuel Sorge. Exploiting hidden structure in selecting dimensions that distinguish vectors. *Journal of Computer and System Sciences*, 82(3):521–535, 2016.

12 Harold N. Gabow. An efficient reduction technique for degree-constrained subgraph and bidirected network flow problems. In *15th Annual ACM Symposium on Theory of Computing, (STOC '83)*, pages 448–456, 1983.

13 Robert Ganian, Iyad Kanj, Sebastian Ordyniak, and Stefan Szeider. On the parameterized complexity of clustering incomplete data into subspaces of small rank. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, (AAAI '20)*, pages 3906–3913, 2020.

14 Robert Ganian, Iyad A. Kanj, Sebastian Ordyniak, and Stefan Szeider. Parameterized algorithms for the matrix completion problem. In *35th International Conference on Machine Learning, (ICML '18)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1642–1651. PMLR, 2018.

15 Leszek Gąsieniec, Jesper Jansson, and Andrzej Lingas. Approximation algorithms for Hamming clustering problems. *Journal of Discrete Algorithms*, 2(2):289–301, 2004.

16 Jens Gramm, Rolf Niedermeier, and Peter Rossmanith. Fixed-parameter algorithms for Closest String and related problems. *Algorithmica*, 37(1):25–42, 2003.

17 Danny Hermelin and Liat Rozenberg. Parameterized complexity analysis for the closest string with wildcards problem. *Theoretical Computer Science*, 600:11–18, 2015.

18 Charles R Johnson. Matrix completion problems: a survey. In *Matrix Theory and Applications*, volume 40 of *Proceedings of Symposia in Applied Mathematics*, pages 171–198. American Mathematical Society, 1990.

19 Stasys Jukna. *Extremal Combinatorics: With Applications in Computer Science*. Springer Science & Business Media, 2011.

20 Tomohiro Koana, Vincent Froese, and Rolf Niedermeier. Parameterized algorithms for matrix completion with radius constraints. In *31st Annual Symposium on Combinatorial Pattern Matching, (CPM '20)*, pages 20:1–20:14, 2020.

21 Ross Lippert, Russell Schwartz, Giuseppe Lancia, and Sorin Istrail. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, 3(1):23–31, 2002.

22 Cristopher Moore and J. M. Robson. Hard tiling problems with simple tiles. *Discrete & Computational Geometry*, 26(4):573–590, 2001.

23 Markus L. Schmid. Finding consensus strings with small length difference between input and solution strings. *ACM Transactions on Computation Theory*, 9(3):13:1–13:18, 2017.

24 Ryan Williams. A new algorithm for optimal 2-constraint satisfaction and its implications. *Theoretical Computer Science*, 348(2-3):357–365, 2005.