This is a peer-reviewed, accepted author manuscript of the following research article: Knight, PA, le Gorrec, L, Mouysset, S & Ruiz, D 2023, 'Introducing the class of semidoubly stochastic matrices: a novel scaling approach for rectangular matrices', SIAM Journal on Matrix Analysis and Applications, vol. 44, no. 4, pp. 1731-1748. https://doi.org/10.1137/22M1519791

INTRODUCING THE CLASS OF SEMI-DOUBLY STOCHASTIC MATRICES: A NOVEL SCALING APPROACH FOR RECTANGULAR MATRICES

PHILIP A. KNIGHT*, LUCE LE GORREC *, SANDRINE MOUYSSET †, AND DANIEL RUIZ ‡

Abstract. It is easy to verify that if $\mathbf{A}$ is a doubly stochastic matrix then both its normal equations $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are also doubly stochastic; but the reciprocal is not true. In this paper, we introduce and analyse the complete class of nonnegative matrices whose normal equations are doubly stochastic. This class contains and extends the class of doubly stochastic matrices to the rectangular case. In particular, we characterise these matrices in terms of their row and column sums, and provide results regarding their nonzero structure. We then consider the diagonal equivalence of any rectangular nonnegative matrix to a matrix of this new class, and we identify the properties for such a diagonal equivalence to exist. To this end, we present a scaling algorithm, and establish the conditions for its convergence. We also provide numerical experiments to highlight the behaviour of the algorithm in the general case.

Key words. Matrix scaling; Rectangular sparse matrices; Combinatorial matrix theory.

AMS subject classifications. 15A48; 65F35

1. Introduction.

If $\mathbf{A} \geq 0$ is a square non-negative matrix with total support, then we can find a diagonal scaling so that $\mathbf{DAE}$ is doubly stochastic ($\mathbf{DAE}\mathbb{1} = \mathbf{E}\mathbf{A}^T\mathbf{D}\mathbb{1} = \mathbb{1}$), where $\mathbf{A} \geq 0$ means that $\mathbf{A}$ is nonnegative and $\mathbf{D}$ and $\mathbf{E}$ are diagonal matrices with positive diagonal. If $\mathbf{A} \geq 0$ is rectangular and has sufficient nonzeros, then it too can be scaled so that it has constant row and column sums (but no longer equal). Alternatively, one can prescribe arbitrary row and column sums, $\mathbf{r} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^n$ (so long as $\sum_{i=1}^m |r_i| = \sum_{j=1}^n |c_j|$), and scale $\mathbf{A}$ so that $\mathbf{DAE}\mathbb{1}_n = \mathbf{r}$ and $\mathbf{E}\mathbf{A}^T\mathbf{D}\mathbb{1}_m = \mathbf{c}$.

Note that a square matrix has support if it can be permuted so that it has a fully nonzero diagonal, and has total support if every nonzero entry can be permuted onto a fully nonzero diagonal. A generalisation of total support for rectangular matrices is the strong Hall property (see [2] for details). We use (and restate) a version of this property in Theorem 2.10.

The diagonal scaling problem has a long history in the mathematical literature, dating back to the 1930s [4], with applications in diverse areas outside linear algebra. Most recently it has emerged as being central to the solution of optimal transport problems associated with machine learning [10], as well as a key step in genome analysis [12].

In the general case, the precise conditions for existence of a scaling depend on $\mathbf{r}$, $\mathbf{c}$ and $\mathbf{A}$, and were set out by Brualdi [1] and Menon and Schneider [8], but they cannot be as neatly described as in the square case. A generic condition [7] for a given scaling to exist for $\mathbf{A}$ is that there exists a nonnegative matrix $\mathbf{B}$ with the same pattern as $\mathbf{A}$ for which $\mathbf{B}\mathbb{1}_n = \mathbf{r}$ and $\mathbf{B}^T\mathbb{1}_m = \mathbf{c}$. If a scaling exists, in both the square and

*Department of Mathematics and Statistics, University of Strathclyde (p.a.knight@strath.ac.uk, legorrecluce@gmail.com).

Funding: LlG was supported on this project by the Royal Academy of Engineering and the Office of the Chief Science Advisor for National Security under the UK Intelligence Community Postdoctoral Fellowship Programme.

†Université Paul Sabatier, Toulouse 3 (sandrine.mouysset@irit.fr).
‡Institut National Polytechnique de Toulouse (daniel.ruiz@toulouse-inp.fr).

rectangular cases, then it can be found using the Sinkhorn–Knopp algorithm [11, 13]. In fact, the existence of a scaling (particularly in the rectangular case) is confirmed by the convergence of this algorithm, although it may be more insightful to verify that Brualdi's conditions hold.

In this work we extend the class of doubly stochastic matrices to include a set of rectangular matrices. While it is impossible for a non-square nonnegative matrix to have row and column sums both equal to one since the sum of row sums must be equal to the sum of column sums, we may however extend a weaker condition satisfied by doubly stochastic matrices. We consider nonnegative matrices for which

$$\mathbf{A}\mathbf{A}^T\mathbb{1}_m = \mathbb{1}_m \quad \text{and} \quad \mathbf{A}^T\mathbf{A}\mathbb{1}_n = \mathbb{1}_n\,. \tag{1.1}$$

This trivially holds for doubly stochastic matrices since

$$\mathbf{A}\mathbf{A}^T\mathbb{1} = \mathbf{A}\mathbb{1} = \mathbb{1} \quad \text{and} \quad \mathbf{A}^T\mathbf{A}\mathbb{1} = \mathbf{A}^T\mathbb{1} = \mathbb{1}.$$

But it is a bigger class, even in the square case, as can be seen with

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1/\sqrt{2} \\ 0 & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{pmatrix}, \tag{1.2}$$

for which we have

$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which is doubly stochastic even though $\mathbf{A}$ is not. Notice that $\mathbf{A}$ does not have support, so it cannot even be scaled to doubly stochastic form.

We label as *semi-doubly stochastic* any nonnegative matrix, square or rectangular ($\mathbf{A} \in \mathbb{R}^{m\times n}$), for which (1.1) holds. We first show that such a matrix is essentially the direct sum of p connected rectangular sub-components $\mathbf{A}_i$, $i = 1, \ldots, p$, where $\mathbf{A}_i \in \mathbb{R}^{m_i \times n_i}$, each having constant row sums and constant column sums. A question that naturally arises is whether a given nonnegative matrix can be scaled to semi-doubly stochastic form. For the square case this is a very well studied problem and existence is conditional on the non-zero pattern of the matrix. It is also true in our generalisation. For example, consider $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ which is scalable to semi-doubly stochastic form if and only if $\mathbf{B} = \begin{bmatrix} \alpha\beta & \alpha\gamma \\ 0 & \delta\gamma \end{bmatrix}$ is semi-doubly stochastic, for some scalars $\alpha, \beta, \gamma, \delta > 0$. This can be recast as the system

$$\begin{cases} \overbrace{(\alpha\beta)^2}^{x} + \overbrace{(\alpha\gamma)^2}^{y} + \overbrace{\gamma^2\alpha\delta}^{z} & = 1, \qquad \text{(row sums of } \mathbf{B}\mathbf{B}^T\text{)} \\ \overbrace{(\gamma\delta)^2}^{u} + \gamma^2\alpha\delta & = 1, \\ (\alpha\beta)^2 + \overbrace{\alpha^2\beta\gamma}^{v} & = 1, \\ \alpha^2\beta\gamma + (\alpha\gamma)^2 + (\gamma\delta)^2 & = 1. \qquad \text{(row sums of } \mathbf{B}^T\mathbf{B}\text{)} \end{cases}$$

We immediately see that we require $y = (\alpha\gamma)^2 = 0$, and so $\mathbf{B}$ does not exist. We denote as *semi-scalable* any nonnegative matrix $\mathbf{A}$ for which we can find diagonal matrices $\mathbf{D}$ and $\mathbf{E}$ such that $\mathbf{B} = \mathbf{DAE}$ is semi-doubly stochastic (that is that $\mathbf{B}^T\mathbf{B}$ and $\mathbf{B}\mathbf{B}^T$ are doubly stochastic matrices) Note that throughout this article we assume that our matrices do not contain any zero row or column, as these are clearly not semi-scalable.

Our main motivation for investigating this type of matrices is its potential in co-clustering applications. Co-clustering is a data mining technique that extends clustering to uncover relationships between different features in a dataset. Connections between elements of the two features are represented in a rectangular data matrix and co-clustering aims to find row and column permutations to reveal consistent row and column blocks, the so-called co-clusters. Adapting doubly stochastic scaling to rectangular matrices can help in at least two different co-clustering approaches.

The first one is related to optimal transport [5]. It draws a parallel between scaling a rectangular matrix to one with (piecewise)[1] constant row and column sums, and finding the probability distributions of data and features random variables responsible for the observations stored in a data matrix. In the co-clustering context, these distributions are assumed to be mixtures of uniform distributions, with each component in the mixture corresponding to a co-cluster. Thus, permuting rows and columns of the data matrix according to the increasing order of the elements in the scaling factors can highlight the co-clustering structure. In the algorithm `CCOT` derived from these observations, the authors subsample the data matrix to get square matrices since there is no current algorithm to scale a general rectangular matrix to one with (piecewise) constant row and column sums. This, in turn, requires that they apply a majority vote over the co-clusterings uncovered using the sampled square matrices, which increases both the algorithm complexity, and the risk of co-clustering mistakes. We believe that the results we highlight in the current work may help improving the `CCOT` algorithm proposed in [5].

The second approach in which semi-doubly stochastic matrices clearly have a role is the spectral algorithm used to uncover block structures in matrices scaled into doubly stochastic form, proposed in [6]. In this work, permuted singular vectors of a doubly stochastic matrix are shown to have a piecewise constant shape when the matrix has a block structure, and permuting the matrix according to the size of the vectors entries highlights the underlying block structure. The results from [6] can be easily extended to semi-doubly stochastic matrices, thus enabling one to extend the spectral approach to rectangular matrices. As an example, in Figure 1 we show two permutations of the same matrix. To produce the picture on the right-hand side, we have used three singular vectors from the semi-scaled version of the left-hand matrix to reorder the rows and columns to reveal the block structure. Since it is not square, this matrix is not scalable to doubly stochastic form, but it can be semi-scaled with the use of the algorithm described in Section 3.

While the block structure of semi-scalable matrices is attractive there is no easy way to tell a priori whether a matrix is close to having this property or not. In practice, if we attempt to use current scaling algorithms on such matrices without pre-existing knowledge of the underlying block structure, then they will fail to converge to anything meaningful. To remedy this, we present a new iterative scaling algorithm, which simultaneously targets the row sums of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. We also prove that a matrix is semi-scalable if and only if our algorithm converges, providing in the limit a diagonal scaling so that $\mathbf{DAE}$ is semi-doubly stochastic. Additionally, we

[1] Only constant row and column sums scaling are addressed in [5] as stated in Section 2.1, but this generalises naturally to piecewise constant row and column sums scalings.

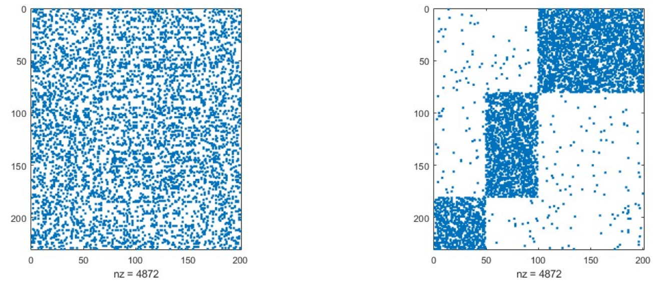

FIG. 1. *Approximate block structure revealed by scaling: raw matrix (left), and reordered matrix (right) after semi-scaling and block identification from the distribution of the entries in the singular vectors.*

illustrate the behaviour of the algorithm on matrices which are not semi-scalable. The algorithm still converges to a semi-doubly stochastic matrix but in this case it is one whose nonzero pattern is included in that of the original matrix, as certain entries are forced towards zero.

1.1. Notation. For a given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we will want to generate a number of associated quantities. The notation we use is detailed in Table 1. Note that if a bipartite graph has adjacency matrix $\begin{bmatrix} 0 & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix}$, we say the matrix $\mathbf{A}$ is the graph **bipartite matrix**.

Typeface	Definition
M, N	The sets $\{1,\dots,m\}$ and $\{1,\dots,n\}$, respectively.
$\mathbf{A}(R,C)$	The submatrix of $\mathbf{A}$ containing the intersection between rows in $R \subset M$ and columns in $C \subset N$.
$\mathcal{P}(\mathbf{A})$	The pattern of $\mathbf{A}$: $\mathcal{P}(\mathbf{A}) = \{(i,j) \in M \times N : \mathbf{A}(i,j) \neq 0\}$.
$\mathcal{B}(\mathbf{A})$	The bipartite graph for which $\mathbf{A}$ is the bipartite matrix.
$\mathcal{A}(\mathbf{A})$	The graph for which $\mathbf{A}$ is the adjacency matrix ($\mathbf{A}$ square).
$\mathbb{1}_p$	A column vector of 1s of dimension p.
$\mathcal{D}(\mathbf{r})$	The diagonal matrix given by some vector $\mathbf{r}$.
$\overline{T}$	Given $T \subset S$, then $\overline{T} = S \setminus T$.

TABLE 1
Notation.

2. The Class of Semi-Doubly Stochastic Matrices. In this section, we formally introduce the class of semi-doubly stochastic (SDS) matrices and detail some properties of this class. Our main result is a characterisation of SDS matrices, stated in Theorem 2.6.

DEFINITION 2.1. *A nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is said to be **semi-doubly***

stochastic *(SDS) if and only if its normal equations are both stochastic, that is*

$$\begin{cases} \mathbf{A}\mathbf{A}^T\mathbb{1}_m = \mathbb{1}_m \\ \mathbf{A}^T\mathbf{A}\mathbb{1}_n = \mathbb{1}_n \end{cases}.$$

Definition 2.1 is just a rewording of (1.1). It is clear that since $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are both symmetric, the fact that they are stochastic implies that they are doubly stochastic. However, the denomination *semi-doubly stochastic* means that both normal equation matrices are stochastic together, whereas $\mathbf{A}$ may not be.

We now analyse the structural properties of SDS matrices. We first state two general results for nonnegative sparse matrices that will be useful in defining the core blocks of SDS matrices.

Lemma 2.2. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with no zero row or column, then the following statements are equivalent:*

1. *$\mathbf{A}\mathbf{A}^T$ is bi-irreducible* [2]*.*
2. *$\mathbf{A}^T\mathbf{A}$ is bi-irreducible.*
3. *The bipartite graph $\mathcal{B}(\mathbf{A})$ is connected.*

Proof. (1) $\implies$ (3) : If $\mathcal{B}(\mathbf{A})$ is not connected, it means that $\exists U \subset M, V \subset N$ both nonempty, such that there is no edge between $\overline{U}$ and V and between $\overline{V}$ and U. Thus $\mathbf{A}$ can be simultaneously permuted to $\begin{bmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{bmatrix}$ with $\mathbf{A}_1 = \mathbf{A}(U,V)$, respectively $\mathbf{A}_2 = \mathbf{A}(\overline{U},\overline{V})$. This implies that $\mathbf{A}\mathbf{A}^T$ can be permuted as $\begin{bmatrix} \mathbf{A}_1\mathbf{A}_1^T & 0 \\ 0 & \mathbf{A}_2\mathbf{A}_2^T \end{bmatrix}$. Hence, $\mathbf{A}\mathbf{A}^T$ is not even irreducible, let alone bi-irreducible.

(3) $\implies$ (1) Given that $\mathbf{A}\mathbf{A}^T$ is symmetric and has a full diagonal, $\mathbf{A}\mathbf{A}^T$ is bi-irreducible iff it is irreducible, that is iff the graph $\mathcal{A}(\mathbf{A}\mathbf{A}^T)$ is connected.

An edge (u,v) in $\mathcal{A}(\mathbf{A}\mathbf{A}^T)$ coincides with a 2-path in $\mathcal{B}(\mathbf{A})$ whose external nodes are in M, that is a triplet $(u,y,v) \in M\times N\times M : \mathbf{A}(u,y) \neq 0$ and $\mathbf{A}(v,y) \neq 0$.

Since $\mathcal{B}(\mathbf{A})$ is connected, $\forall u,v \in M$, $\begin{cases} \exists\, y_1,\dots,y_k \in N, \\ \exists\, x_1,\dots,x_{k+1} \in M, \end{cases}$ such that $x_1 = u$, $x_{k+1} = v$, and $\forall i$, $\begin{cases} \mathbf{A}(x_i,y_i) \neq 0, \\ \mathbf{A}(x_{i+1},y_i) \neq 0. \end{cases}$ Since a triplet (x_i,y_i,x_{i+1}) is a 2-path in $\mathcal{B}(\mathbf{A})$, that is an edge in $\mathcal{A}(\mathbf{A}\mathbf{A}^T)$, this implies that, $\forall u,v \in M$ there is a path between u and v in $\mathcal{A}(\mathbf{A}\mathbf{A}^T)$. Thus, $\mathcal{A}(\mathbf{A}\mathbf{A}^T)$ is connected, which implies $\mathbf{A}\mathbf{A}^T$ is bi-irreducible.

(2) $\iff$ (3) is straightforward by considering $\mathbf{A}^T$ instead of $\mathbf{A}$ in the previous points. ∎

As matrices arising in Lemma 2.2 will be at the core of our study, we introduce the following useful definition.

Definition 2.3. *A rectangular matrix with no zero row or column that satisfies the conditions in Lemma 2.2 is called a **connected** matrix.*

The following corollary is a direct consequence of Lemma 2.2.

Corollary 2.4. *Any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with no zero row or column can be permuted into the direct sum of independent connected matrices. In other words,* $\mathbf{A}$

[2] That is there exist no row and column permutations that can rearrange $\mathbf{A}$ into a block triangular form.

can be permuted to

$$\begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_k \end{bmatrix},$$

where each $\mathbf{A}_i \in \mathbb{R}^{m_i \times n_i}$ *is a connected matrix.*

Proof. The blocks $\mathbf{A}_i$ are the bipartite matrices of the disjoint connected components of $\mathcal{B}(\mathbf{A})$. The rest follows from Lemma 2.2. ∎

The following theorem provides a characterisation of connected SDS matrices.

THEOREM 2.5. *A connected nonnegative matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *is SDS iff* $\mathbf{A}$ *has constant row sums equal to* $\sqrt{\dfrac{n}{m}}$*, respectively constant column sums equal to* $\sqrt{\dfrac{m}{n}}$*. In other words,* $\mathbf{A}$ *is SDS iff*

$$\begin{cases} \mathbf{A}\mathbb{1}_n = \sqrt{\dfrac{n}{m}}\mathbb{1}_m \\ \mathbf{A}^T\mathbb{1}_m = \sqrt{\dfrac{m}{n}}\mathbb{1}_n \end{cases}$$

Proof. ($\Longrightarrow$) Assume that $\mathbf{A}$ is SDS. Then

$$\begin{cases} \mathbf{f} = \mathbf{A}\mathbb{1}_n, \\ \mathbf{g} = \mathbf{A}^T\mathbb{1}_m, \end{cases} \quad \text{thus} \quad \begin{cases} \mathbf{A}^T\mathbf{f} = \mathbb{1}_n, \\ \mathbf{A}\mathbf{g} = \mathbb{1}_m, \end{cases} \quad \text{and finally} \quad \begin{cases} \mathbf{A}\mathbf{A}^T\mathbf{f} = \mathbf{f}, \\ \mathbf{A}^T\mathbf{A}\mathbf{g} = \mathbf{g}. \end{cases}$$

Therefore, $\mathbf{f}$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$ associated with an eigenvalue 1, and such that $\mathbf{f} > 0$. Since $\mathbf{A}\mathbf{A}^T$ is bi-irreducible and row-stochastic, we know from the Perron–Frobenius theorem that $\mathbf{f} = \alpha\mathbb{1}_m$, with $\alpha > 0$. With a similar argument, $\mathbf{g} = \beta\mathbb{1}_n$, with $\beta > 0$. This shows that $\mathbf{A}$ has constant row and column sums.

It is then necessary to have, $\alpha = \sqrt{n/m}$ and $\beta = \sqrt{m/n}$, as

$$\begin{cases} m &= \mathbb{1}_m^T \overbrace{\mathbb{1}_m}^{\mathbf{A}\mathbf{g}} &= \mathbf{g}^T \overbrace{\mathbf{A}^T\mathbf{A}\mathbf{g}}^{\mathbf{g}} &= \mathbf{g}^T\mathbf{g} &= n\beta^2, \\ n &= \mathbb{1}_n^T \underbrace{\mathbb{1}_n}_{\mathbf{A}^T\mathbf{f}} &= \mathbf{f}^T \underbrace{\mathbf{A}\mathbf{A}^T\mathbf{f}}_{\mathbf{f}} &= \mathbf{f}^T\mathbf{f} &= m\alpha^2, \end{cases}$$

together with $\alpha, \beta > 0$.

($\Longleftarrow$) The reciprocal is immediate. Since $\begin{cases} \mathbf{A}\mathbb{1}_n = \sqrt{n/m}\mathbb{1}_m, \\ \mathbf{A}^T\mathbb{1}_m = \sqrt{m/n}\mathbb{1}_n, \end{cases}$ we then get

$$\begin{cases} \mathbf{A}^T\mathbf{A}\mathbb{1}_n &= \sqrt{n/m}\mathbf{A}^T\mathbb{1}_m &= \sqrt{n/m}\sqrt{m/n}\mathbb{1}_n &= \mathbb{1}_n, \\ \mathbf{A}\mathbf{A}^T\mathbb{1}_m &= \sqrt{m/n}\mathbf{A}\mathbb{1}_n &= \sqrt{m/n}\sqrt{n/m}\mathbb{1}_m &= \mathbb{1}_m, \end{cases}$$

which shows that $\mathbf{A}$ is SDS. ∎

Combining Theorem 2.5 together with Corollary 2.4, we can then state the main result of the section, which characterises the class of SDS matrices.

THEOREM 2.6. *Any SDS matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *is the direct sum of connected matrices, each of size* $m_i \times n_i$ *and having constant row sums and constant column sums equal to* $\sqrt{n_i/m_i}$ *and* $\sqrt{m_i/n_i}$ *respectively.*

As seen in (1.2), for square matrices, being SDS is not equivalent to being doubly stochastic. The following corollary states the condition under which these two properties are equivalent.

COROLLARY 2.7. *A square nonnegative matrix is doubly stochastic if and only if it is semi-doubly stochastic with support.*

Proof. Any doubly stochastic matrix is SDS, and has support too, as it must have total support. At the same time, from Theorem 2.6 any SDS matrix whose connected subblocks are square is doubly stochastic. We need to show that the connected subblocks of any square matrix with support are square.

Assume that the nonnegative matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has support. From Corollary 2.4, we assume without loss of generality that $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & \\ & \ddots & \\ & & \mathbf{A}_k \end{bmatrix}$. A support of $\mathbf{A}$ is a set of n pairs $\{(i_t, j_t)\}_{t \in N}$, such that $\forall t \in N, \mathbf{A}(i_t, j_t) \neq 0$, and that covers all the rows and columns of $\mathbf{A}$. Thus, a support of $\mathbf{A}$ must contain a diagonal of maximum size for each block $\mathbf{A}_i$. If one block $\mathbf{A}_i$ is rectangular, say $m_i > n_i$, then the maximum diagonal for this block will cover at most n_i rows. Because of the independence of the $\mathbf{A}_i$s, it is not possible to cover the remaining $m_i - n_i$ rows from this block. Similarly, when $n_i > m_i$ it is impossible to cover $n_i - m_i$ columns from $\mathbf{A}_i$. Thus, if one block $\mathbf{A}_i$ is rectangular, then $\mathbf{A}$ has no support. ∎

While the purpose of our study is not to scale matrices to matrices with prescribed row and column sums, some results from this field provide interesting insights in our context when investigating connected matrices. In particular, combining the values of row and column sums from Theorem 2.5 together with the properties raised in Theorem 3.5 from [8], we obtain the following result.

LEMMA 2.8. *Given a connected nonnegative matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *with* $m \geq n$*:*

- *A necessary condition for* $\mathbf{A}$ *to be SDS is that*

$$\forall I \subset M, J \subset N, \left(\mathbf{A}(\bar{I}, J) = 0 \implies \frac{|J|}{|I|} < \frac{n}{m} \right). \tag{2.1}$$

- *If* $\mathbf{A}$ *satisfies (2.1), there exists an SDS matrix with the same pattern as* $\mathbf{A}$.

Remark 2.9. This lemma implies that every column of a tall SDS matrix must have strictly more than m/n nonzero entries. Or equivalently, that every column node in $\mathcal{B}(\mathbf{A})$ must be linked to strictly more than m/n row nodes. Moreover, any subset of $k \leq n$ columns in $\mathbf{A}$ must contain at least $k \times (m/n)$ non empty rows.

We can derive a corollary from the previous lemma that gives an interesting necessary condition about the pattern of an SDS matrix.

THEOREM 2.10. *Given a connected nonnegative matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$*, with* $m \geq n$*, the fact that* $\mathbf{A}$ *is SDS implies the two following statements.*

1. $\mathbf{A}$ *has the strong Hall property (Brualdi [2]).*
2. *Every nonzero entry of* $\mathbf{A}$ *lies on a column diagonal.*

Proof. $(SDS) \implies (1)$: For a connected matrix $\mathbf{A}$, the strong Hall property can be stated as

$$\forall R \subset M, \forall C \subset N, (\mathbf{A}(R,C) = 0 \implies |R| + |C| < m)\,,$$

(a slight adaptation from a statement in the Introduction in [2]).

Assume that we have $R \subset M$, $C \subset N$ such that $\mathbf{A}(R,C) = 0$. By replacing $\overline{I} = R$ and $J = C$ in Lemma 2.8, we have that $|C|/|\overline{R}| < n/m$. Since $m \geq n$ and $|\overline{R}| = m - |R|$, this leads to $|C| < m - |R| \iff |C| + |R| < m$, and thus $\mathbf{A}$ verifies the strong Hall property.

(1) $\iff$ (2) comes from Theorem 3.3 of [2]. □

Remark 2.11. Conversely, the fact that a matrix has the strong Hall property is not sufficient for ensuring that there exists an SDS matrix with the same pattern, as can be observed by considering a matrix with pattern

$$\begin{bmatrix} \times & \times \\ \times & \times \\ 0 & \times \\ 0 & \times \end{bmatrix}.$$

This is clearly a connected matrix, and it has the strong Hall property, since every nonzero entry lies on a column diagonal: for example consider the diagonals $\{(1,1),(2,2)\},\{(2,1),(1,2)\},\{(1,1),(3,2)\},\{(1,1),(4,2)\}$.

But with $I = \{1,2\}$, $J = \{1\}$, we have $\mathbf{A}(\overline{I}, J) = 0$, and yet $|J|/|I| = 1/2$, which is equal to n/m and the conditions of Lemma 2.8 do not hold.

3. Scaling Matrices to Semi-Doubly Stochastic Form.

If a nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ has no zero row and no zero column then both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have a total support, since they are both symmetric with a full diagonal, and so both normal equations can be independently scaled to doubly stochastic form. But, whether a given nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is diagonally equivalent to a semi-doubly stochastic matrix, or not, is not so obvious. The 2×2 counterexample given in the introduction clearly shows that is is not the case in general. We thus consider the class of nonnegative matrices that can actually be scaled to semi-doubly stochastic form.

Definition 3.1. *A nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is said to be* ***semi-scalable*** *if and only if it is diagonally equivalent to a semi-doubly stochastic matrix $\mathbf{B}$, i.e. there exist two positive diagonal matrices $\mathbf{D} \in \mathbb{R}^{m\times m}, \mathbf{E} \in \mathbb{R}^{n\times n}$ such that $\mathbf{B} = \mathbf{DAE}$ and*

$$\begin{cases} \mathbf{B}\mathbf{B}^T\mathbb{1}_m = \mathbb{1}_m, \\ \mathbf{B}^T\mathbf{B}\mathbb{1}_n = \mathbb{1}_n. \end{cases}$$

From Theorem 2.6, we know that scaling a nonnegative matrix to semi-doubly stochastic form is equivalent to a scaling to piecewise constant row and column sums, after some appropriate row and column permutations. Scaling a matrix to prescribed row and column sums is a well known problem, and many algorithms have been proposed to achieve such a target. The issue when trying to scale a matrix $\mathbf{A}$ to piecewise constant sums is to be able to determine in advance individual blocks within the pattern of $\mathbf{A}$, whose direct sum reproduces the matrix that we want to scale, and to verify also that each of these blocks corresponds to a matrix that can be scaled to constant row and column sums. Here, we describe an algorithm that will find

Algorithm 3.1 SDS-scaling algorithm

Input: A nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, with no zero row or column.

Output: Two diagonal matrices $\mathbf{D} \in \mathbb{R}^{m\times m}$ and $\mathbf{E} \in \mathbb{R}^{n\times n}$, and $\mathbf{S} \in \mathbb{R}^{m\times n}$ such that $\mathbf{S} = \mathbf{DAE}$.

1: $\mathbf{A}^{(0)} \leftarrow \mathbf{A}$
2: $\mathbf{d}^{(0)} \leftarrow \mathbb{1}_m$
3: $\mathbf{e}^{(0)} \leftarrow \mathbb{1}_n$
4: **for** $k = 0, 1, 2, \ldots$ **until** convergence **do**
5: $\quad \mathbf{r} \leftarrow \mathbf{A}^{(k)} {\mathbf{A}^{(k)}}^T \mathbb{1}_m$
6: $\quad \mathbf{c} \leftarrow {\mathbf{A}^{(k)}}^T \mathcal{D}(\mathbf{r})^{-1} \mathbf{A}^{(k)} \mathbb{1}_n$
7: $\quad \mathbf{A}^{(k+1)} \leftarrow \mathcal{D}\left(\sqrt{\mathbf{r}}\right)^{-1} \mathbf{A}^{(k)} \mathcal{D}\left(\sqrt{\mathbf{c}}\right)^{-1}$
8: $\quad \mathbf{d}^{(k+1)} \leftarrow \mathcal{D}\left(\sqrt{\mathbf{r}}\right)^{-1} \mathbf{d}^{(k)}$
9: $\quad \mathbf{e}^{(k+1)} \leftarrow \mathcal{D}\left(\sqrt{\mathbf{c}}\right)^{-1} \mathbf{e}^{(k)}$

Set $\mathbf{D} = \mathcal{D}\left(\mathbf{d}^{(k+1)}\right)$, $\mathbf{E} = \mathcal{D}\left(\mathbf{e}^{(k+1)}\right)$, and $\mathbf{S} = \mathbf{A}^{(k+1)}$

iteratively its way to scale a matrix to semi-doubly stochastic form, whenever this is possible, and we also analyse its convergence.

The SDS-scaling Algorithm 3.1 is rather simple, and does not require any prescribed row or column sums as input. The vectors $\mathbf{d}^{(k)}$ and $\mathbf{e}^{(k)}$ correspond to the scaling factors so that, at each iteration, $\mathbf{A}^{(k)}$ is diagonally equivalent to $\mathbf{A}$, with

$$\mathbf{A}^{(k)} = \mathcal{D}\left(\mathbf{d}^{(k)}\right) \mathbf{A}\, \mathcal{D}\left(\mathbf{e}^{(k)}\right) .$$

We will now analyse its convergence. We will use the fact that the algorithm is a diagonal product increasing algorithm (DPI) and exploit techniques introduced in [9].

Lemma 3.2. *The SDS-scaling algorithm produces a sequence of scaled matrices $\mathbf{A}^{(k)}$, diagonally equivalent to $\mathbf{A}$ for $k = 1, 2, \ldots$, which is bounded in $\mathbb{R}^{m\times n}$, and which contains convergent subsequences.*

Proof. In fact, for $k \geq 1$, we can verify that the spectral norm of $\mathbf{A}^{(k)}$ is equal to 1. Indeed, denoting the current iterate $\mathbf{A}^{(k)}$ as $\mathbf{A}$, and the next scaled iterate $\mathbf{A}^{(k+1)}$ as $\mathbf{S}$, the iteration in the SDS-scaling algorithm is essentially reduced to:

$$\begin{array}{ll} \text{form} & \mathbf{r} = \mathbf{A}\mathbf{A}^T \mathbb{1}_m, \\ \text{and set} & \widehat{\mathbf{A}} = \mathcal{D}\left(\sqrt{\mathbf{r}}\right)^{-1} \mathbf{A}, \\ & \mathbf{c} = \widehat{\mathbf{A}}^T \widehat{\mathbf{A}} \mathbb{1}_n, \\ \text{and finally} & \mathbf{S} = \mathcal{D}\left(\sqrt{\mathbf{r}}\right)^{-1} \mathbf{A}\, \mathcal{D}\left(\sqrt{\mathbf{c}}\right)^{-1}. \end{array}$$

Consequently, the non-zero eigenvalues of

$$\mathbf{S}\mathbf{S}^T = \widehat{\mathbf{A}}\, \mathcal{D}(\mathbf{c})^{-1} \widehat{\mathbf{A}}^T$$

are also the non-zero eigenvalues of

$$\mathbf{W} = \mathcal{D}(\mathbf{c})^{-1} \widehat{\mathbf{A}}^T \widehat{\mathbf{A}} ,$$

which is nonnegative and row-stochastic. The Perron–Frobenius theory enables us to conclude that the maximum eigenvalue of $\mathbf{W}$ is 1, and therefore that the largest singular value of $\mathbf{S}$ is equal to 1. Notice also that the same reasoning can be used to show that the largest eigenvalue of $\widehat{\mathbf{A}}^T \widehat{\mathbf{A}} = \mathbf{A}^T \mathcal{D}(\mathbf{r})^{-1} \mathbf{A}$ is also equal to 1.

Therefore, $\forall k \geq 1$, $\|\mathbf{A}^{(k)}\|_2 = 1$, and the sequence of scaled matrices is bounded in the finite dimensional space $\mathbb{R}^{m\times n}$, and there exist convergent subsequences. ∎

It follows that $\forall (i,j) \in \mathcal{P}(\mathbf{A})$, the pattern of $\mathbf{A}$, the sequences $\left(d_i^{(k)} e_j^{(k)}\right)_{k\geq 1}$ are bounded above, as

$$\forall k \geq 1,\ \ \forall (i,j) \in \mathcal{P}(\mathbf{A}),\ \ a_{ij}^{(k)} = a_{ij}\, d_i^{(k)} e_j^{(k)} \leq 1\,, \tag{3.1}$$

which implies that

$$d_i^{(k)} e_j^{(k)} \leq \frac{1}{\min\limits_{(i,j)\in\mathcal{P}(\mathbf{A})} a_{ij}} = L\,. \tag{3.2}$$

Lemma 3.3. *The SDS-scaling algorithm is diagonal product increasing (DPI) in the sense that*

$$\forall k \geq 1,\ \prod_{i=1}^{m} \frac{d_i^{(k+1)}}{d_i^{(k)}} \geq 1 \ \ \textit{and}\ \ \prod_{j=1}^{n} \frac{e_j^{(k+1)}}{e_j^{(k)}} \geq 1\,. \tag{3.3}$$

Proof. This is just a direct consequence of the fact that $\forall k \geq 1$, $\|\mathbf{A}^{(k)}\|_2 = 1$. From the arithmetic-geometric mean inequality, we get

$$\prod_{i=1}^{m} \frac{d_i^{(k)}}{d_i^{(k+1)}} = \prod_{i=1}^{m} \sqrt{r_i} \leq \left(\frac{1}{m}\sum_{i=1}^{m} \sqrt{r_i}\right)^m .$$

Now, by the Cauchy–Schwartz inequality, we also have $\sum_{i=1}^m \sqrt{r_i} \leq \sqrt{m}\sqrt{\sum_{i=1}^m r_i}$, and since

$$\sum_{i=1}^{m} r_i = \mathbb{1}_m^T \mathbf{r} = \mathbb{1}_m^T \mathbf{A}^{(k)} {\mathbf{A}^{(k)}}^T \mathbb{1}_m = \|{\mathbf{A}^{(k)}}^T \mathbb{1}_m\|_2^2 \leq m$$

(because $\|\mathbf{A}^{(k)}\|_2 = 1$), we can easily conclude that

$$\prod_{i=1}^{m} \frac{d_i^{(k)}}{d_i^{(k+1)}} \leq 1\,.$$

Similar considerations also imply that

$$\prod_{j=1}^{n} \frac{e_j^{(k)}}{e_j^{(k+1)}} = \prod_{j=1}^{n} \sqrt{c_j} \leq \left(\frac{1}{n}\sum_{j=1}^{n} \sqrt{c_i}\right)^n \leq 1\,,$$

as

$$\sum_{j=1}^{n} c_i = \mathbb{1}_n^T \mathbf{c} = \mathbb{1}_n^T {\mathbf{A}^{(k)}}^T \mathcal{D}(\mathbf{r})^{-1} \mathbf{A}^{(k)} \mathbb{1}_n,$$

and $\|{\mathbf{A}^{(k)}}^T \mathcal{D}(\mathbf{r})^{-1}\mathbf{A}^{(k)}\|_2 = 1$ as well. This establishes the DPI property (3.3) for the SDS-scaling algorithm. ∎

We now state our main convergence result, which shows that a rectangular matrix is semi-scalable if and only if the SDS-scaling algorithm converges. This is ensured whenever there exists a semi-doubly stochastic matrix $\mathbf{B}$, with the same pattern as that of the matrix $\mathbf{A}$ we wish to scale.

THEOREM 3.4. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a nonnegative matrix, and suppose that there exists a semi-doubly stochastic matrix $\mathbf{B}$ with $\mathcal{P}(\mathbf{A}) = \mathcal{P}(\mathbf{B})$. Then, the SDS-scaling algorithm produces a sequence of iterates $(\mathbf{A}^{(k)})_{k\geq 0}$ (starting from $\mathbf{A}$) that converges to a semi-doubly stochastic limit $\mathbf{Q}$. The scaling factors $(\mathbf{d}^{(k)}, \mathbf{e}^{(k)})_{k\geq 0}$ also have a limit, $(\mathbf{d}, \mathbf{e})$ say, and*

$$\mathbf{Q} = \mathcal{D}(\mathbf{d})\mathbf{A}\mathcal{D}(\mathbf{e})\,.$$

In the proof of this theorem we will make use of the following lemma which establishes two properties that are direct consequences of results in the literature.

LEMMA 3.5. *Suppose that $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a connected nonnegative matrix.*

(a) [Pretzel [11] – Proposition 1]
If $\mathbf{Q}$ and $\widehat{\mathbf{Q}}$ are two semi-doubly stochastic matrices diagonally equivalent to $\mathbf{A}$, e.g.

$$\mathbf{Q} = \mathcal{D}(\mathbf{x})\mathbf{A}\mathcal{D}(\mathbf{y}) \;\; \text{and} \;\; \widehat{\mathbf{Q}} = \mathcal{D}(\widehat{\mathbf{x}})\mathbf{A}\mathcal{D}(\widehat{\mathbf{y}})\,,$$

for some positive scaling vectors $\mathbf{x}$, $\mathbf{y}$, $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$, then $\mathbf{Q} = \widehat{\mathbf{Q}}$ and the scaling vectors $\mathbf{x}$ and $\widehat{\mathbf{x}}$ for the rows, and $\mathbf{y}$ and $\widehat{\mathbf{y}}$ for the columns, are unique up to some scaling factor.

(b) [Pretzel [11] – Lemma 2]
Consider a converging sequence of matrices diagonally equivalent to $\mathbf{A}$

$$\mathbf{Q} = \lim_{k\to+\infty} \mathcal{D}(\mathbf{x}^{(k)})\mathbf{A}\mathcal{D}(\mathbf{y}^{(k)})\,,$$

where $\left(\mathbf{x}^{(k)}\right)_k$ and $\left(\mathbf{y}^{(k)}\right)_k$ are two sequences of positive scaling vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. If $\mathbf{Q}$ has the same pattern as $\mathbf{A}$ (e.g. there are no vanishing elements in the limit), then for any given row index $i \in \{1, \ldots, m\}$ (or column index $j \in \{1, \ldots, n\}$), both sequences $(\mathbf{x}^{(k)}/x_i^{(k)})_k$ and $(\mathbf{y}^{(k)} \times x_i^{(k)})_k$ (or $(\mathbf{x}^{(k)} \times y_j^{(k)})_k$ and $(\mathbf{y}^{(k)}/y_j^{(k)})_k$, respectively) have a limit, which we denote as $\mathbf{d} \in \mathbb{R}^m$ and $\mathbf{e} \in \mathbb{R}^n$, and we have

$$\mathbf{Q} = \mathcal{D}(\mathbf{d})\mathbf{A}\mathcal{D}(\mathbf{e})\,.$$

We now briefly contextualise the two points in Lemma 3.5. Point (a) is a direct combination of Proposition 1 from [11], together with the characterisation for semi-doubly stochastic matrices given in Theorem 2.6. Indeed, from Theorem 2.6, we know that two semi-doubly stochastic matrices with the same pattern must share the same row sums and column sums, as their common pattern exhibits connected sub-components in the same place and with the same sizes. Since $\mathbf{Q}$ and $\widehat{\mathbf{Q}}$ are also diagonally equivalent, as

$$\widehat{\mathbf{Q}} = \mathcal{D}(\widehat{\mathbf{x}}/\mathbf{x})\mathbf{Q}\mathcal{D}(\widehat{\mathbf{y}}/\mathbf{y})\,,$$

Proposition 1 in [11] establishes the fact that $\mathbf{Q} = \widehat{\mathbf{Q}}$. Additionally, the demonstration of Proposition 1 in [11] shows that for a connected component we must have

$$\frac{\widehat{x}_i}{x_i} = \alpha, \; \forall i, \quad \text{and} \quad \frac{\widehat{y}_j}{y_j} = \frac{1}{\alpha}, \; \forall j.$$

Point (b) is actually included in the demonstration of Lemma 2 in [11], where the re-scaling of $\mathbf{x}^{(k)}$ by any of its entries, $x_1^{(k)}$ for instance, and the fact that $\mathbf{A}$ is connected, implies that all factors $x_i^{(k)}/x_1^{(k)}$ and $y_j^{(k)} \times x_1^{(k)}$ have a limit, d_i and e_j respectively, and as

$$\mathcal{D}\left(\mathbf{x}^{(k)}\right)\mathbf{A}\mathcal{D}\left(\mathbf{y}^{(k)}\right) = \mathcal{D}\left(\mathbf{x}^{(k)}/x_1^{(k)}\right)\mathbf{A}\mathcal{D}\left(\mathbf{y}^{(k)} \times x_1^{(k)}\right),\; \forall k,$$

in the limit we get $\mathbf{Q} = \mathcal{D}(\mathbf{d})\mathbf{A}\mathcal{D}(\mathbf{e})$.

We can now prove Theorem 3.4.

Proof. We can assume, without loss of generality, that $\mathbf{A}$ is connected. Indeed, if $\mathbf{A}$ is the direct sum of independent connected sub-components, then the SDS-scaling algorithm is reduced to scaling independently each sub-component. Additionally, Theorem 2.6 implies that $\mathbf{B}$ is also the direct sum of independent semi-doubly stochastic sub-components, each one of them associated to each sub-component in $\mathbf{A}$ since $\mathcal{P}(\mathbf{B}) = \mathcal{P}(\mathbf{A})$.

Now, for a connected semi-doubly stochastic matrix $\mathbf{B}$, we know from Theorem 2.5 that

$$\mathbf{B}\mathbb{1}_n = \sqrt{\frac{n}{m}}\mathbb{1}_m \text{ and } \mathbf{B}^T\mathbb{1}_m = \sqrt{\frac{m}{n}}\mathbb{1}_n\,.$$

From the DPI property (3.3), we know that

$$s^{(k)} = \left(\prod_{i=1}^{m} d_i^{(k)}\right)^{\sqrt{\frac{n}{m}}}\left(\prod_{j=1}^{n} e_j^{(k)}\right)^{\sqrt{\frac{m}{n}}}$$

is an increasing sequence in $\mathbb{R}^+$, and from (3.2) it is also bounded above as

$$s^{(k)} = \prod_{i=1}^{m}\left(\prod_{j=1}^{n}\left(d_i^{(k)}e_j^{(k)}\right)^{b_{ij}}\right) \leq \prod_{i=1}^{m}\left(\prod_{j=1}^{n} L^{b_{ij}}\right) \leq L^{\sqrt{mn}}\,,$$

in which the scalars b_{ij} are the elements of the SDS matrix $\mathbf{B}$. Therefore, the sequence $(s^{(k)})_{k\geq 1}$ must converge:

$$\lim_{k\to+\infty} s^{(k)} = \xi \geq s^{(1)} > 0\,.$$

Now, using the arithmetic-geometric mean inequality again, we can write that

$$\begin{aligned}
\frac{s^{(k)}}{s^{(k+1)}} &= \left(\prod_{i=1}^{m}\sqrt{r_i}\right)^{\sqrt{\frac{n}{m}}}\left(\prod_{j=1}^{n}\sqrt{c_j}\right)^{\sqrt{\frac{m}{n}}} \\
&\leq \left(\frac{1}{\sqrt{mn}}\left(\frac{1}{2}\sqrt{\frac{n}{m}}\sum_{i=1}^{m} r_i + \frac{1}{2}\sqrt{\frac{m}{n}}\sum_{j=1}^{n} c_j\right)\right)^{\sqrt{mn}} \\
&\leq 1\,.
\end{aligned}$$

As all the $a_{ij}^{(k)}$ are bounded by 1, we know that the values in vectors $\mathbf{r}$ and $\mathbf{c}$ stay bounded through every iterations so that if we consider any convergent subsequence, with limit given by vectors $\mathbf{x}$ and $\mathbf{y}$ respectively, we shall get in the limit

$$\begin{aligned}
1 &= \left(\prod_{i=1}^{m} x_i\right)^{\frac{1}{2}\sqrt{\frac{n}{m}}}\left(\prod_{j=1}^{n} y_j\right)^{\frac{1}{2}\sqrt{\frac{m}{n}}} \\
&= \left(\frac{1}{\sqrt{mn}}\left(\frac{1}{2}\sqrt{\frac{n}{m}}\sum_{i=1}^{m} x_i + \frac{1}{2}\sqrt{\frac{m}{n}}\sum_{j=1}^{n} y_j\right)\right)^{\sqrt{mn}}\,.
\end{aligned}$$

But the arithmetic-geometric mean inequality results in such an equality only when $x_i = 1, \forall i$, and $y_j = 1, \forall j$, showing that any convergent subsequence converges to 1 and therefore that the sequence of scalars in vectors $\mathbf{r}$ and $\mathbf{c}$ also all converge to one.

Next, for every non-zero element in $\mathbf{A}$, $0 < b_{ij} \leq min(\sqrt{\frac{n}{m}}, \sqrt{\frac{m}{n}}) \leq 1$, and we can write (using the fact that $\sum_{i,j} b_{ij} = \sqrt{mn}$)

$$\left(d_i^{(k)} e_j^{(k)}\right)^{b_{ij}} L^{\sqrt{mn}-b_{ij}} \geq s^{(k)} \geq s^{(1)} > 0\,,$$

and thus

$$d_i^{(k)} e_j^{(k)} \geq \left(\frac{s^{(1)}}{L^{\sqrt{mn}-b_{ij}}}\right)^{\frac{1}{b_{ij}}} \geq \frac{s^{(1)}}{L^{\sqrt{mn}}} = \alpha > 0$$

(as $0 < b_{ij} \leq \min(\sqrt{\frac{m}{n}}, \sqrt{\frac{n}{m}}) \leq 1$). Therefore

$$\forall k \geq 1,\ \ \forall (i,j) \in \mathcal{P}(\mathbf{A}),\ \ a_{ij}^{(k)} = a_{ij}\, d_i^{(k)} e_j^{(k)} \geq \alpha\, a_{ij} > 0\,,$$

thus, $\forall (i,j) \in \mathcal{P}(\mathbf{A})$ the sequence of iterates $(a_{ij}^{(k)})_{k\geq 1}$ is isolated from zero, and bounded above by 1 from (3.1). If we consider any convergent subsequence of $(\mathbf{A}^{(k)})_k$

$$\widehat{\mathbf{A}} = \lim_q \mathbf{A}^{(q)} = \lim_q \mathcal{D}(d^{(q)})\mathbf{A}\mathcal{D}(e^{(q)})\,,$$

it must have the same pattern as that of $\mathbf{A}$, and be semi-doubly stochastic too, that is

$$\lim_q \mathbf{A}^{(q)} {\mathbf{A}^{(q)}}^T \mathbb{1}_m = \widehat{\mathbf{A}}\widehat{\mathbf{A}}^T \mathbb{1}_m$$

and

$$\lim_q {\mathbf{A}^{(q)}}^T \mathcal{D}(\mathbf{r}^{(q)})^{-1} \mathbf{A}^{(q)} \mathbb{1}_n = \widehat{\mathbf{A}}^T \widehat{\mathbf{A}} \mathbb{1}_n$$

respectively.

From point (b) in Lemma 3.5, we know that $\widehat{\mathbf{A}}$ is diagonally equivalent to $\mathbf{A}$. Consequently, any limit of the bounded sequence $(\mathbf{A}^{(k)})_k$ is a semi-doubly stochastic matrix with the same pattern, and diagonally equivalent to $\mathbf{A}$, all with the same row sums equal to $\sqrt{\frac{n}{m}}$, and the same column sums equal to $\sqrt{\frac{m}{n}}$. These limits must then all be equal, from point (a) in Lemma 3.5, showing that

$$\mathbf{Q} = \lim_{k\to+\infty} \mathbf{A}^{(k)}$$

exists, is diagonally equivalent to $\mathbf{A}$, and semi-doubly stochastic.

To finish, we must now verify that the sequences of scaling factors $(\mathbf{d}^{(k)})_k$ and $(\mathbf{e}^{(k)})_k$ also converge. Both sequences are bounded, otherwise there would exist a subsequence of some scaling factor that diverges, say

$$d_i^{(q)} \underset{q\to+\infty}{\longrightarrow} +\infty$$

(the same reasoning can be made with the column scaling factors). Then, from point (b) in Lemma 3.5, we know that both $(\mathbf{d}^{(q)}/d_i^{(q)})_q$ and $(\mathbf{e}^{(q)} \times d_i^{(q)})_q$ have a limit, so that all $e_j^{(q)}$, $j = 1, \ldots, n$, must tend to zero when $q \to +\infty$. This is in contradiction with the DPI property (3.3) of the SDS-scaling algorithm.

Now, consider any two convergent subsequences of the bounded sequence $(\mathbf{d}^{(k)}, \mathbf{e}^{(k)})_k$. From point (a) in Lemma 3.5, their limits are essentially unique, in the sense that they can differ only by scaling factors α and $1/\alpha$. Finally, the DPI property (3.3) requires that $\alpha = 1$, so that all these limits are equal, which yields the required conclusion. ∎

Theorem 3.4 states that if $\mathbf{A}$ is semi-scalable, then the SDS-scaling algorithm converges to an SDS matrix with the same pattern as $\mathbf{A}$. We can also show that when the SDS-scaling algorithm converges, its limit is an SDS matrix with the same pattern as the input matrix $\mathbf{A}$ (which implies that $\mathbf{A}$ is semi-scalable).

THEOREM 3.6. *If a nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is connected and the SDS-scaling algorithm converges, in the sense that both row and column scaling factors have a limit (necessarily strictly positive), then the limit of the sequence of iterates $(\mathbf{A}^{(k)})_{k\geq 0}$ is SDS and $\mathbf{A}$ is thus semi-scalable.*

Proof. Since the algorithm is DPI, existence of a limit precludes any scaling factor converging to zero Again, without loss of generality, we assume that $\mathbf{A}$ is connected. The fact that the SDS-scaling algorithm converges, in the sense that both row and column scaling factors have a limit, means that

$$\begin{cases} \mathbf{d}^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbf{d} > 0 \\ \mathbf{e}^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbf{e} > 0 \\ \mathbf{A}^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbf{Q} = \mathcal{D}(\mathbf{d})\mathbf{A}\mathcal{D}(\mathbf{e}) \end{cases}.$$

Thus

$$\prod_{i=1}^{m} d_i^{(k)} \underset{k\to+\infty}{\longrightarrow} \prod_{i=1}^{m} d_i > 0$$

and

$$\prod_{j=1}^{n} e_j^{(k)} \underset{k\to+\infty}{\longrightarrow} \prod_{j=1}^{n} e_j > 0$$

both have a limit that is strictly positive, which implies

$$1 = \lim_{k\to+\infty} \left(\prod_{i=1}^{m} \frac{d_i^{(k)}}{d_i^{(k+1)}} \right) = \lim_{k\to+\infty} \left(\prod_{i=1}^{m} \sqrt{r_i^{(k+1)}} \right)$$

and

$$1 = \lim_{k\to+\infty} \left(\prod_{j=1}^{n} \frac{e_j^{(k)}}{e_j^{(k+1)}} \right) = \lim_{k\to+\infty} \left(\prod_{j=1}^{n} \sqrt{c_j^{(k+1)}} \right).$$

But $\forall k \geq 1$, we have

$$\prod_{i=1}^{m} \sqrt{r_i^{(k+1)}} \leq \left(\frac{1}{m} \sum_{i=1}^{m} \sqrt{r_i^{(k+1)}} \right)^m \leq 1$$

and

$$\prod_{j=1}^{n} \sqrt{c_j^{(k+1)}} \leq \left(\frac{1}{n} \sum_{j=1}^{n} \sqrt{c_j^{(k+1)}} \right)^n \leq 1,$$

as already observed in the proof of Lemma 3.3. Therefore

$$1 = \lim_{k\to+\infty} \left(\prod_{i=1}^{m} \sqrt{r_i^{(k+1)}} \right) = \lim_{k\to+\infty} \left(\frac{1}{m} \sum_{i=1}^{m} \sqrt{r_i^{(k+1)}} \right)^m$$

and

$$1 = \lim_{k\to+\infty} \left(\prod_{j=1}^{n} \sqrt{c_j^{(k+1)}}\right) = \lim_{k\to+\infty} \left(\frac{1}{n}\sum_{j=1}^{n} \sqrt{c_i^{(k+1)}}\right)^n .$$

Finally, since the two sequences $(\mathbf{r}^{(k)})_k$ and $(\mathbf{c}^{(k)})_k$ are bounded, by considering the limit $(\mathbf{x},\mathbf{y}) = \lim_{q\to+\infty}(\mathbf{r}^{(q)},\mathbf{c}^{(q)})$ of any converging subsequence, we get

$$1 = \prod_{i=1}^{m} \sqrt{x_i} = \left(\frac{1}{m}\sum_{i=1}^{m}\sqrt{x_i}\right)^m$$

and

$$1 = \prod_{j=1}^{n} \sqrt{y_j} = \left(\frac{1}{n}\sum_{j=1}^{n}\sqrt{y_j}\right)^n ,$$

which is feasible if and only if $x_i = 1, \forall i$ and $y_j = 1, \forall j$. This means that $\mathbf{r}^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbb{1}_m$ and $\mathbf{c}^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbb{1}_n$. We thus have that

$$\mathbb{1}_m = \lim_{k\to+\infty} \mathbf{r}^{(k)} = \lim_{k\to+\infty} \mathbf{A}^{(k)}\mathbf{A}^{(k)^T}\mathbb{1}_m = \mathbf{Q}\mathbf{Q}^T\mathbb{1}_m,$$

and

$$\mathbb{1}_n = \lim_{k\to+\infty} \mathbf{c}^{(k)} = \lim_{k\to+\infty} \mathbf{A}^{(k)^T}\mathcal{D}(\mathbf{r}^{(k)})^{-1}\mathbf{A}^{(k)}\mathbb{1}_n = \mathbf{Q}^T\mathcal{D}(\mathbb{1}_m)^{-1}\mathbf{Q}\mathbb{1}_n = \mathbf{Q}^T\mathbf{Q}\mathbb{1}_n.$$

Therefore, the matrix $\mathbf{Q}$ is SDS, and diagonally equivalent to $\mathbf{A}$. ∎

Theorem 3.4 along with Theorem 3.6 establish a characterisation for the class of semi-scalable matrices, in the sense that, a matrix is semi-scalable if and only if the SDS-scaling algorithm converges. This is analogous to the results that can be found in [11, 13], in the case of scaling to prescribed row and column sums through the Iterative Scaling Procedure.

4. Non Semi Scalable Matrices. In the previous section, we showed that, given a nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the algorithm converges towards an SDS matrix diagonally equivalent to $\mathbf{A}$ if and only if $\exists \mathbf{B} \in \mathbb{R}^{m\times n}$ an SDS matrix such that $\mathcal{P}(\mathbf{B}) = \mathcal{P}(\mathbf{A})$, that is, if and only if $\mathbf{A}$ is semi-scalable (SS). The natural question that arises is what happens when $\mathbf{A}$ is not SS?

Even in this case, extensive numerical experiments suggest that the algorithm seems to always produce a sequence of matrices $(\mathbf{A}^{(k)})_k$ that converges to an SDS matrix. However, the sequence of scaling factors $(\mathbf{d}^{(k)},\mathbf{e}^{(k)})_k$ diverges so that some nonzero elements from $\mathbf{A}$ vanish in $\lim_{k\to+\infty}\mathbf{A}^{(k)}$, as illustrated in Figure 2. On the left panel we display the output of the sequence provided by the algorithm applied on $\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$, where we consider that convergence is reached when

$$\max(\|\mathbf{A}^{(k)}\mathbf{A}^{(k)^T}\mathbb{1}_m - \mathbb{1}_m\|_\infty, \|\mathbf{A}^{(k)^T}\mathbf{A}^{(k)}\mathbb{1}_n - \mathbb{1}_n\|_\infty) \le 10^{-6},$$

with row and column sums highlighted on the y-and x-axes, respectively. Matrix $\mathbf{A}$ is not SS (it violates the conditions from Lemma 2.8), hence the nonzero in blue

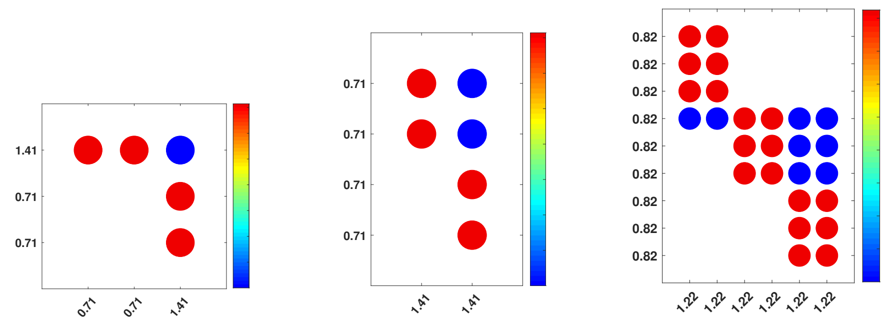

FIG. 2. *Limits of the converging sequence produced by the SDS-scaling algorithm on non-SS matrices, with row and column sums displayed on the y-and x-axes.*

vanishes, thus producing two disjoint connected blocks, as can be seen from the values of row and column sums. We remark that the vanishing element does not lie on a diagonal, thus by Theorem 2.10, it must not appear in an SS matrix whose pattern is included in $\mathcal{P}(\mathbf{A})$. On the other hand, all the nonzeros in the matrix from the middle panel lie on a diagonal, yet the matrix is not SS and the two top right elements vanish in the converging sequence, which again confirms that Theorem 2.10 provides a necessary **but not sufficient** condition for a matrix to be SS. Finally, the matrix in the right panel highlights that the algorithm is able to find its way towards an SDS limit, even when the SS submatrix within the initially connected matrix has more than two connected blocks.

This is in line with the Iterative Scaling Procedure (ISP) that scales matrices to prescribed row and column sums. Indeed, it is known that, when there exists a matrix $\mathbf{B} \in \mathbb{R}^{m\times n}$ whose row and column sums are as prescribed, and such that $\mathcal{P}(\mathbf{B}) \subset \mathcal{P}(\mathbf{A})$, then ISP produces a sequence of matrices whose limit has the prescribed row and column sums; and that is the largest submatrix from $\mathbf{A}$ whose pattern equals the one of a matrix having row and column sums as prescribed—see for instance Theorem 1 in [11].

Depending on the prescribed sums and the pattern of the input matrix, there is no guarantee that such a matrix exists, and thus that ISP will produce a converging sequence of matrices whose limit has row and column sums as prescribed. On the other hand, we have the guarantee that in any matrix with no zero row or column, there exists an SS submatrix, as proved below.

THEOREM 4.1. *From any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with no zero row or zero column, one can extract a semi-scalable matrix. That is, there exists a semi-scalable matrix $\mathbf{B}$ of dimension $m \times n$ such that $\mathcal{P}(\mathbf{B}) \subset \mathcal{P}(\mathbf{A})$.*

Proof. Algorithm 4.1 applied to the matrix $\mathbf{A}$ returns a SS matrix whose pattern is included in $\mathcal{P}(\mathbf{A})$. Two things must be ensured to guarantee the correctness of Algorithm 4.1: that the recursive calls can be done, and that the algorithm terminates.

For performing the recursive calls, we have to ensure that the subblocks $\mathbf{A}(I_0, J_0)$ and $\mathbf{A}(\overline{I_0}, \overline{J_0})$ have no zero row and no zero column. Assume that $\mathbf{A}(I_0, J_0)$ has a zero row, say i^*. Thus, $(I_0 \setminus \{i^*\}, J_0) \in Z$. But then, $|J_0|/|I_0 \setminus \{i^*\}| > |J_0|/|I_0|$, which contradicts the maximality of ρ. Similarly if $\mathbf{A}(\overline{I_0}, \overline{J_0})$ has an empty column j^*, by considering $(I_0, J_0 \cup \{j^*\}) \in Z$, the maximality of ρ is contradicted.

On the other hand, if $\mathbf{A}(I_0, J_0)$ has a zero column j^*, since $\mathbf{A}(\overline{I_0}, j^*)$ is zero, it

Algorithm 4.1 `SSExtract`

Input: A matrix $\mathbf{A} \in \mathbb{R}^{m \times n} : m \geq n$, with no zero row or column.

Output: A SS matrix whose pattern is included in $\mathbf{A}$.

1: $Z = \{(I, J) \subset M \times N, |I| < m : \mathbf{A}(\overline{I}, J) = 0\}$
2: **if** $Z = \emptyset$ **then**
3:     **return** $\mathbf{A}$ ▶ $\mathbf{A}$ is dense, thus it is SS.
4: $(I_0, J_0) \leftarrow$ an element from $Z : \dfrac{|J_0|}{|I_0|} = \max\limits_{(I,J)\in Z} \dfrac{|J|}{|I|} = \rho$
5: **if** $\rho < n/m$ **then**
6:     **return** $\mathbf{A}$ ▶ $\mathbf{A}$ satisfies Lemma 2.8, thus it is SS.
7: $\mathbf{A}(I_0, \overline{J_0}) \leftarrow 0$
8: **if** $|I_0| \geq |J_0|$ **then**
9:     $\mathbf{A}(I_0, J_0) \leftarrow$`SSExtract`$(\mathbf{A}(I_0, J_0))$
10: **else**
11:     $\mathbf{A}(I_0, J_0) \leftarrow$`SSExtract`$(\mathbf{A}(I_0, J_0)^T)^T$
12: **if** $|\overline{I_0}| \geq |\overline{J_0}|$ **then**
13:     $\mathbf{A}(\overline{I_0}, \overline{J_0}) \leftarrow$`SSExtract`$(\mathbf{A}(\overline{I_0}, \overline{J_0}))$
14: **else**
15:     $\mathbf{A}(\overline{I_0}, \overline{J_0}) \leftarrow$`SSExtract`$(\mathbf{A}(\overline{I_0}, \overline{J_0})^T)^T$

means that the column j^* is a zero column for the whole matrix $\mathbf{A}$, which contradicts the initial hypotesis. Similarly, if $\mathbf{A}(\overline{I_0}, \overline{J_0})$, has a zero row, it will be a zero row for the whole matrix $\mathbf{A}$.

Hence we have the guarantee that the subblocks on which the recursive calls are performed have no zero row and no zero column. This condition ensures that the algorithm terminates. Since each recursive call is run on a subblock which is strictly smaller than the previous block, if the end conditions provided by line 2 or 5 are not met before, the algorithm will eventually meet a subblock whose minimum dimension is $min(n, m) = 1$. Since such a block cannot have zero row or column, it is necessarily dense. Hence condition from line 2 holds and the algorithm terminates. ∎

Theorem 4.1 fits with our experimental observations that SDS-scaling algorithm always produces a converging sequence of matrices whose limit is SDS. However, we are not able to predict which element(s) will vanish in the produced sequence. While it is known that when the prescribed row and column sums can be achieved, ISP produces a sequence that converges to the largest possible submatrix (that is, making as few nonzeros as possible vanish), this is not the case for the SDS-scaling algorithm. This is illustrated in Figure 3: when applied to the matrix $\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$, the sequence of matrices produced by the algorithm converges to the identity matrix whilst $\mathbf{B} = \sqrt{1/2} \times \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ is an SDS matrix with $\mathcal{P}(\mathbf{B}) \subset \mathcal{P}(\mathbf{A})$, showing that the identity matrix is not the densest SS submatrix within $\mathbf{A}$.

Contrary to our convergence results from the previous section, a prediction cannot

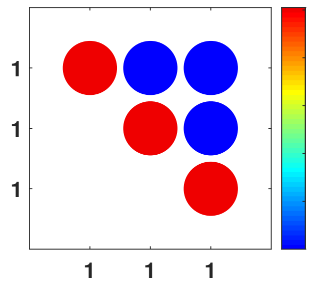

Fig. 3. *Limits of the converging sequence from the SDS-scaling algorithm–See Figure 2.*

be made from existing results on ISP, which rely on the knowledge of the target row and column sums. One purpose of the SDS-scaling algorithm—and more generally, of introducing the class of SDS matrices—is to avoid the need to prescribe row and column sums of the scaled matrix, which obviously implies that we do not know them a priori. Characterising the pattern of the limit of the converging sequence produced by the SDS-scaling algorithm will be the focus of further work.

5. Conclusion.

In this work, we have defined a new class of matrices, called *semi-doubly stochastic* (SDS), which are nonnegative $m \times n$ matrices whose normal equations are doubly stochastic. For matrices whose underlying bipartite graph is connected, we have shown that SDS matrices are exactly those having constant row and column sums equal to $\sqrt{n/m}$ and $\sqrt{m/n}$. In the general case, SDS matrices are exactly those having piecewise constant row and column sums, with pieces corresponding to the underlying connected components (of size $m_i \times n_i$) in their bipartite graph, such that row and column sums are equal to $\sqrt{n_i/m_i}$ and $\sqrt{m_i/n_i}$ in each component.

From this class of SDS matrices, we have derived a class of matrices that can be scaled to SDS, that is, that can be diagonally balanced to an SDS matrix. Such matrices are labelled *semi scalable* (SS). An algorithm to scale SS matrices to SDS has been derived, and its convergence demonstrated. Finally, some experimental observations have been made about the behaviour of the algorithm when the matrix is not SS. Of particular interest is the fact that the algorithm still produces a sequence of matrices whose limit is SDS, but contrary to classic ISP, this limit may not correspond to the densest SS submatrix with nonzeros in the input matrix's pattern. The next step to this work will then be to characterise the elements that vanish in the algorithm.

Matrices which are scalable but not semi-scalable have a block structure that should be exploitable in a manner used in [6] to uncover hidden structure in rectangular matrices. For large scale applications, it may be necessary to accelerate the algorithm presented in Section 3, and a Newton-based method akin to that in [3] may be possible.

REFERENCES

[1] R. A. Brualdi, *Convex sets of non-negative matrices*, Canadian Journal of Mathematics, 20 (1968), pp. 144–157.

[2] R. A. Brualdi and B. L. Shader, *Strong hall matrices*, SIAM Journal on Matrix Analysis and Applications, 15 (1994), pp. 359–365.

[3] P. A. Knight and D. Ruiz, *A fast algorithm for matrix balancing*, IMA Journal of Numerical

Analysis, 33 (2013), pp. 1029–1047.
[4] J. Kruithof, *Telefoonverkeersrekening*, De Ingenieur, 52 (1937), pp. 15–25.
[5] C. Laclau, I. Redko, B. Matei, Y. Bennani, and V. Brault, *Co-clustering through optimal transport*, in ICML, vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 1955–1964.
[6] L. Le Gorrec, S. Mouysset, I. S. Duff, P. A. Knight, and D. Ruiz, *Uncovering hidden block structure for clustering*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 140–155.
[7] M. Menon, *Matrix links, an extremization problem, and the reduction of a non-negative matrix to one with prescribed row and column sums*, Canadian Journal of Mathematics, 20 (1968), pp. 225–232.
[8] M. Menon and H. Schneider, *The spectrum of a nonlinear operator associated with a matrix*, Linear Algebra and its applications, 2 (1969), pp. 321–334.
[9] B. Parlett and T. Landis, *Methods for scaling to doubly stochastic form*, Linear Algebra and its Applications, 48 (1982), pp. 53–79.
[10] G. Peyré, M. Cuturi, et al., *Computational optimal transport: With applications to data science*, Foundations and Trends in Machine Learning, 11 (2019), pp. 355–607.
[11] O. Pretzel, *Convergence of the iterative scaling procedure for non-negative matrices*, Journal of the London Mathematical Society, 2 (1980), pp. 379–384.
[12] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al., *A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping*, Cell, 159 (2014), pp. 1665–1680.
[13] R. Sinkhorn and P. Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific Journal of Mathematics, 21 (1967), pp. 343–348.