J Glob Optim (2014) 58:439–464
DOI 10.1007/s10898-013-0053-2

A continuous characterization of the maximum-edge biclique problem

Nicolas Gillis · François Glineur

Received: 11 May 2012 / Accepted: 27 February 2013 / Published online: 20 March 2013
© Springer Science+Business Media New York 2013

Abstract The problem of finding large complete subgraphs in bipartite graphs (that is, bicliques) is a well-known combinatorial optimization problem referred to as the maximum-edge biclique problem (MBP), and has many applications, e.g., in web community discovery, biological data analysis and text mining. In this paper, we present a new continuous characterization for MBP. Given a bipartite graph G, we are able to formulate a continuous optimization problem (namely, an approximate rank-one matrix factorization problem with nonnegativity constraints, R1N for short), and show that there is a one-to-one correspondence between (1) the maximum (i.e., the largest) bicliques of G and the global minima of R1N, and (2) the maximal bicliques of G (i.e., bicliques not contained in any larger biclique) and the local minima of R1N. We also show that any stationary points of R1N must be close to a biclique of G. This allows us to design a new type of biclique finding algorithm based on the application of a block-coordinate descent scheme to R1N. We show that this algorithm, whose algorithmic complexity per iteration is proportional to the number of edges in the graph, is guaranteed to converge to a biclique and that it performs competitively with existing methods on random graphs and text mining datasets. Finally, we show how R1N is closely related to the Motzkin–Strauss formalism for cliques.

Keywords Maximum-edge biclique problem · Biclique finding algorithm · Algorithmic complexity · Nonnegative rank-one approximation

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. The first author is a postdoctoral researcher with the Fonds de la Recherche Scientifique-FNRS (F.R.S.-FNRS).

N. Gillis (✉) · F. Glineur
ICTEAM Institute, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium
e-mail: nicolas.gillis@uclouvain.be

F. Glineur
Center for Operations Research and Econometrics, Université catholique de Louvain,
Voie du Roman Pays, 34, 1348 Louvain-la-Neuve, Belgium
e-mail: francois.glineur@uclouvain.be

1 Introduction

Many real-world applications rely on the discovery of complete bipartite subgraphs, i.e., *bicliques*; for example in web community discovery, biological data analysis and text mining, see [12,13,16] and the references therein. In fact, in many practical situations, two distinct groups of objects interact (e.g., Internet users vs. web sites, genes vs. experimental conditions, and texts vs. words) and one would like to find highly connected pairs of subgroups in these datasets.

Some algorithms aim at detecting all bicliques, which is computationally challenging. In fact, there might be an exponential number of such subgraphs [1] and, for large datasets (e.g., in web community discovery or in text mining), it might therefore be hopeless to write down all of them.

Finding instead only the largest biclique(s) is comparatively easier. However, the corresponding optimization problem, called the maximum-edge biclique problem, is NP-hard [15]. In practice, one then often tries to find good but not necessarily optimal solutions, i.e., to find large or maximal bicliques.

1.1 Outline of the paper

After introducing a formulation for the maximum-edge biclique problem in Sect. 2, we propose in Sect. 3 a new continuous characterization based on a rank-one matrix approximation problem with nonnegativity constraints, herein referred to as approximate rank-one nonnegative factorization (R1N). Hence, given a bipartite graph G, we are able to construct an instance of R1N, namely $\mathrm{R1N}_d(G)$, with the following properties:

- The set of global minima of $\mathrm{R1N}_d(G)$ coincides with the set of largest bicliques of G (i.e., the maximum bicliques).
- The set of local minima of $\mathrm{R1N}_d(G)$ coincides with the set of bicliques of G not contained in any larger biclique (i.e., the maximal bicliques).
- Any stationary point of $\mathrm{R1N}_d(G)$ is close to a biclique of G.

Building on these facts, Sect. 4 introduces a new type of biclique finding algorithm that relies on the application of a simple nonlinear optimization scheme (block-coordinate descent) to $\mathrm{R1N}_d(G)$, whose iterations only require a number of operations proportional to the number of edges of the graph. This method is then compared to a greedy heuristic, to the existing algorithm of Ding et al. [4] and to the root node level heuristics of the commercial mixed-integer programming solver GUROBI [11] on some synthetic and text datasets, and is shown to perform competitively. Finally, we show how our formulation is closely related to the Motzkin–Strauss formalism for the maximum clique problem [14].

1.2 Notation

The set of m-by-n real matrices is denoted $\mathbb{R}^{m\times n}$; for $A \in \mathbb{R}^{m\times n}$, we denote the ith column of A by $A_{:i}$ or $A(:, i)$, the jth row of A by $A_{j:}$ or $A(j, :)$, and the entry at position (i, j) by A_{ij} or $A(i, j)$; for $b \in \mathbb{R}^{m\times 1} = \mathbb{R}^m$, we denote the ith entry of b by b_i. Notation $A(I, J)$ refers to the submatrix of A with row and column indices, respectively in I and J. The matrix A^T is the transpose of A. The ℓ_2-norm $\|.\|_2$ is defined as $\|b\|_2^2 = b^T b$; $\|.\|_F$ is the related matrix norm called Frobenius norm with $\|A\|_F^2 = \sum_{i,j}(A_{ij})^2$. The ℓ_1-norm $\|.\|_1$ is defined as $\|b\|_1 = \sum_i |b_i|$. The support of x is denoted $\mathrm{supp}(x)$, it is the set of nonzero entries of x. The cardinality of the set S is denoted $|S|$. For $M \in \mathbb{R}^{m\times n}$, we also let $M_+ = \max(0, M)$, $M_- = \max(0, -M)$, $\min(M) = \min_{i,j}(M_{ij})$ and $\|M\|_2$ be the standard matrix 2-norm of M, i.e.,

$\|M\|_2 = \max_{x\in\mathbb{R}^n, \|x\|_2=1} \|Mx\|_2 = \sigma_{max}(M)$ where $\sigma_{max}(M)$ is the largest singular value of M. We note $A \circ B$ the component-wise multiplication of matrices A and B with $(A \circ B)_{ij} = A_{ij} B_{ij}$. The inequality $M \geq 0$ means that M is component-wise greater or equal to zero, and for $N \in \mathbb{R}^{m\times n}$ inequality $M \geq N$ means that M is component-wise greater or equal than N. The m-by-n matrix of all ones (resp. zeros) is denoted $\mathbf{1}_{m\times n}$ (resp. $\mathbf{0}_{m\times n}$).

2 Maximum-edge biclique problem

A bipartite graph $G = (V, E)$ is a graph whose vertices can be divided into two disjoint sets V_1 and V_2 such that there is no edge between two vertices in the same set, with $V = V_1 \cup V_2$ and $E \subseteq (V_1 \times V_2)$. A *biclique* is a subset of vertices that induce a complete bipartite subgraph, i.e., a bipartite subgraph where all the vertices are connected by an edge. The so-called maximum-edge biclique problem in a bipartite graph G is the problem of finding a biclique in G with maximum number of edges. The corresponding decision problem: *Given B, does G contain a biclique with at least B edges?* has been shown to be NP-complete [15]. Therefore, the problem of finding the biclique with maximum number of edges in G is at least NP-hard.

Let $A \in \{0, 1\}^{m\times n}$ be the biadjacency matrix of the bipartite graph $G = (V_1 \cup V_2, E)$ with $V_1 = \{s_1, \ldots s_m\}$ and $V_2 = \{t_1, \ldots t_n\}$, i.e., $A(i, j) = 1$ if and only if $(s_i, t_j) \in E$. With this notation, the maximum-edge biclique problem in G can be formulated as follows

$$\begin{array}{ll} \max_{u,v} & \sum_{ij} u_i v_j \\ & u_i + v_j \leq 1 + A_{ij}, \quad \forall i, j, \\ & u \in \{0, 1\}^m, v \in \{0, 1\}^n, \end{array} \tag{1}$$

where $u_i = 1$ (resp. $v_j = 1$) means that node s_i (resp. t_j) belongs to the solution, $u_i = 0$ (resp. $v_j = 0$) otherwise. The first constraints ensure that if $A_{ij} = 0$ then either u_i or v_j is equal to zero, i.e., if there is no edge between s_i and t_j then they cannot both belong to a feasible solution. They are equivalent to the more natural constraints $u_i v_j \leq A_{ij} \forall i, j$, but present the advantage of being linear. Hence, there is one-to-one correspondence between the bicliques of G and the feasible solutions of (1).

In the rest of this paper, we will use a more convenient formulation: because A, u and v are binary, and $u_i v_j \leq A_{ij} \forall i, j$, one can check that

$$\sum_{ij} u_i v_j = \sum_{ij} (u_i v_j)^2 = \sum_{ij} A_{ij} u_i v_j = ||A||_F^2 - ||A - uv^T||_F^2 = |E| - ||A - uv^T||_F^2.$$

Hence (1) can be equivalently reformulated as follows

$$\begin{array}{ll} \min_{u,v} & ||A - uv^T||_F^2 \\ & u_i + v_j \leq 1 + A_{ij}, \quad \forall i, j, \\ & u \in \{0, 1\}^m, v \in \{0, 1\}^n, \end{array} \tag{MB(G)}$$

where the objective function counts the number of edges outside the biclique, and its minimization is therefore equivalent to maximizing the edges contained in the biclique. Notice that the optimal objective function value of MB(G) is equal to $|E| - |E^*|$, where $|E^*|$ is the size of the largest biclique(s) of G (i.e., the optimal value of (1)). We will be particularly interested in the

- Maximum bicliques, which are the largest bicliques in G (i.e., of size $|E^*|$), corresponding to the optimal solutions of MB(G), and the

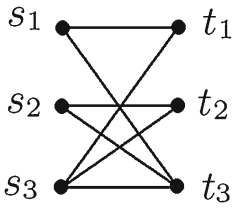

Fig. 1 Graph corresponding to the biadjacency matrix A from Eq. (3)

- Maximal bicliques, which are bicliques not contained in any larger biclique.

3 Continuous characterization of the maximum-edge biclique problem

First, let us define the following problem: given an m-by-n real matrix $R \in \mathbb{R}^{m\times n}$, find its best nonnegative rank-one approximation, i.e., solve

$$\min_{u\in\mathbb{R}^m, v\in\mathbb{R}^n} ||R - uv^T||_F^2 \quad \text{such that} \quad u \geq 0,\ v \geq 0. \tag{R1N}$$

From now on, we say that a pair of vectors (u, v) *coincides* with another pair (u', v') if and only if they correspond to the same rank-one matrix, i.e., if and only if $uv^T = u'v'^T$.

Then, given a parameter $d \geq 0$, a bipartite graph G and its biadjacency matrix $A \in \{0, 1\}^{m\times n}$, we define the following instance of R1N:

$$\min_{u\in\mathbb{R}^m, v\in\mathbb{R}^n} ||M - uv^T||_F^2 \quad \text{such that} \quad u \geq 0,\ v \geq 0, \tag{R1N$_d(G)$}$$

where M is the matrix A where the zero values have been replaced by $-d$, i.e.,

$$M_{ij} = \begin{cases} 1 & \text{if } A_{ij} = 1 \\ -d & \text{if } A_{ij} = 0 \end{cases},\ 1 \leq i \leq m, 1 \leq j \leq n. \tag{2}$$

Although R1N$_d(G)$ is a continuous optimization problem, we are going to show that for any d sufficiently large

- Any of its global minimum coincides with a binary optimal solution of the corresponding (discrete) biclique problem MB(G), and vice versa (Theorem 2).
- Any local minima of R1N$_d(G)$ coincides with a maximal biclique of G, and vice versa (Theorem 1).
- Any stationary point of R1N$_d(G)$ is close to a biclique of G (Sect. 3.4).

Intuitively, the reason is the following. If a $-d$ entry of M is approximated by a positive value in R1N$_d(G)$, say p, the corresponding term in the objective function will be equal to $(-d-p)^2 = d^2+\mathbf{2pd}+p^2$. As d increases, it becomes more and more costly to approximate $-d$ by a positive number (because of the $2pd$ term) and we will show that, for d sufficiently large, negative values of M have to be approximated by zeros. Since the remaining values of M (not approximated by zeros) are all ones, the optimal rank-one solutions will be binary, as in MB(G).

To illustrate this, let us consider the bipartite graph G displayed on Fig. 1, its biadjacency matrix A, and the corresponding matrix M as defined in Eq. (2) with $d = \max(m, n) = 3$, i.e.,

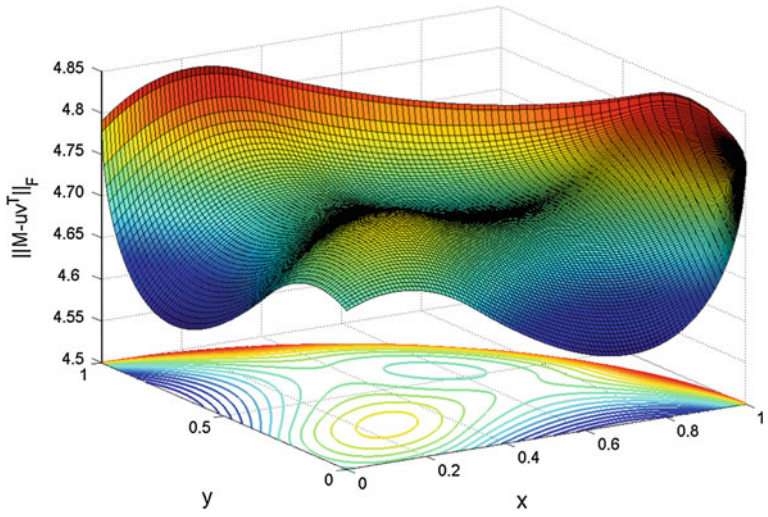

Fig. 2 Objective function $||M - u(x, y)v(x, y)^T||_F$

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \text{and} \quad M = \begin{pmatrix} 1 & -3 & 1 \\ -3 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}. \tag{3}$$

The bipartite graph G contains

- Two maximum bicliques (s_2, s_3, t_2, t_3) and (s_1, s_3, t_1, t_3), corresponding to the two optimal solutions of MB(G), $\left[u^* = (0, 1, 1), v^* = (0, 1, 1)\right]$ and $\left[u' = (1, 0, 1), v' = (1, 0, 1)\right]$, respectively.
- Two maximal (but not maximum) bicliques (s_3, t_1, t_2, t_3) and (s_1, s_2, s_3, t_3) corresponding to the two feasible solutions of MB(G),

  $$\left[u^\dagger = (0, 0, 1), v^\dagger = (1, 1, 1)\right] \text{ and } [u^\ddagger = (1, 1, 1), v^\ddagger = (0, 0, 1)]$$

  respectively.

Let us now consider R1N$_d(G)$. Without loss of generality, one can impose the norm of u to be equal to one, with

$$u(x, y) = \begin{pmatrix} x \\ y \\ \sqrt{1 - x^2 - y^2} \end{pmatrix}, \text{ where } \begin{cases} x \geq 0, y \geq 0 \\ x^2 + y^2 \leq 1 \end{cases}.$$

For u fixed, the optimal solution in v is given[1] by

$$v(x, y) = \text{argmin}_{w \geq 0} ||M - u(x, y)w^T||_F = \max\left(0, M^T u(x, y)\right).$$

Figure 2 displays the surface of the objective function $||M - u(x, y)v(x, y)^T||_F$ with respect to parameters x and y. We distinguish two global minima:

[1] The first-order stationarity condition of R1N$_d(G)$ for variables v is given by $v = \max\left(0, M^T u/||u||_2^2\right)$, see Sect. 3.3. Therefore, local and global minimizers of R1N$_d(G)$ must satisfy this condition, hence they exactly correspond to the local and global minimizers of the problem in the new variables (x, y).

1. $\left[x^* = 0, y^* = \frac{\sqrt{2}}{2}\right]$ with $u(x^*, y^*) = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ and $v(x^*, y^*) = \left(0, \sqrt{2}, \sqrt{2}\right)$, coinciding with the maximum biclique (u^*, v^*).
2. $\left[x' = \frac{\sqrt{2}}{2}, y' = 0\right]$ with $u(x', y') = \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$ and $v(x', y') = \left(\sqrt{2}, 0, \sqrt{2}\right)$, coinciding with the maximum biclique (u', v').

We also distinguish two local minima $\left[x^\dagger = 0, y^\dagger = 0\right]$ and $\left[x^\ddagger = \frac{\sqrt{3}}{3}, y^\ddagger = \frac{\sqrt{3}}{3}\right]$: one can check that they coincide with $(u^\dagger, v^\dagger)$ and $(u^\ddagger, v^\ddagger)$, respectively.

In conclusion, we have then observed a one-to-one correspondence between the global (resp. local) minimizers of $\mathrm{R1N}_d(G)$ and the maximum (resp. maximal) bicliques of G.

3.1 Additional definitions and notations

Let G be a bipartite graph, A its biadjacency matrix, and M the matrix defined in Eq. (2) depending on the parameter d. The pair (u, v) is a *stationary point* of $\mathrm{R1N}_d(G)$ if and only if it satisfies its first-order optimality conditions, i.e., if and only if

$$u \geq 0, \quad \mu = (uv^T - M)v \geq 0 \quad \text{and} \quad u_i\, \mu_i = 0 \quad \forall i, \tag{4}$$

$$v \geq 0, \quad \lambda = u^T(uv^T - M) \geq 0 \quad \text{and} \quad v_i\, \lambda_i = 0 \quad \forall i. \tag{5}$$

Of course, we are only interested in nontrivial solutions and, assuming that $u \neq 0$ and $v \neq 0$, one can check that conditions (4)–(5) are equivalent to

$$u = \max\left(0, \frac{Mv}{||v||_2^2}\right) \qquad \text{and} \qquad v = \max\left(0, \frac{M^T u}{||u||_2^2}\right). \tag{6}$$

For $x \in \mathbb{R}^n$, let us define

$$\mathcal{B}_+(x, r) = \{\, y \in \mathbb{R}^n_+ \mid ||y - x||_2 \leq r \,\},$$

the ball centered at x of radius r intersected with the nonnegative orthant. The pair (u, v) is a *local minimum* of $\mathrm{R1N}_d(G)$ if and only if there exists $\epsilon > 0$ such that for all $u' \in \mathcal{B}_+(u, \epsilon)$ and $v' \in \mathcal{B}_+(v, \epsilon)$, we have $||M - uv^T||_F^2 \leq ||M - u'v'^T||_F^2$. The pair (u, v) is a *global minimum* of $\mathrm{R1N}_d(G)$ if and only if $||M - uv^T||_F^2 \leq ||M - u'v'^T||_F^2$ for all $u' \in \mathbb{R}^m_+$ and $v' \in \mathbb{R}^n_+$.

Given a positive real number d, we define the following three sets of rank-one matrices:

- $\mathcal{S}_d(G)$, corresponding to the set of nontrivial stationary points of $\mathrm{R1N}_d(G)$, i.e.,

$$\mathcal{S}_d(G) = \{uv^T \in \mathbb{R}^{m \times n} \mid (u, v) \text{ satisfies (6), } u \neq 0 \text{ and } v \neq 0 \,\}.$$

- $\mathcal{L}_d(G)$, corresponding to the set of nontrivial local minima of $\mathrm{R1N}_d(G)$.
- $\mathcal{G}_d(G)$, corresponding to the set of nontrivial global minima of $\mathrm{R1N}_d(G)$.

By definition, $\mathcal{G}_d(G) \subseteq \mathcal{L}_d(G) \subseteq \mathcal{S}_d(G)$.

Let us also define the following three sets of binary rank-one matrices:

- $F(G)$, corresponding to the set of feasible solutions of $\mathrm{MB}(G)$, i.e.,

$$F(G) = \{uv^T \in \mathbb{R}^{m \times n} \mid (u, v) \text{ is a feasible for } \mathrm{MB}(G)\}.$$

- $B(G)$, corresponding to the set of maximal bicliques of $\mathrm{MB}(G)$, i.e., $uv^T \in B(G)$ if and only if $uv^T \in F(G)$ and uv^T corresponds to a maximal biclique of G.
- $H(G)$, corresponding to the set of maximum bicliques of $\mathrm{MB}(G)$, i.e., $uv^T \in H(G)$ if and only if $uv^T \in F(G)$ and uv^T corresponds to a maximum biclique of G.

By definition, $H(G) \subseteq B(G) \subseteq F(G)$.

In the rest of this section, we show that if the graph G contains at least one edge (i.e., if $A \neq 0$), then

- For any $d \geq \max(m, n)$, $\mathcal{G}_d(G) = H(G)$, see Theorem 2.
- For any $d \geq \max(m, n)$, $\mathcal{L}_d(G) = B(G) = \mathcal{S}_d(G) \cap F(G)$, see Theorems 1 and 3.
- For any $d \geq 2\max(m, n)\sqrt{|E|}$, there is a simple rounding operator Φ such that $\Phi(\mathcal{S}_d(G)) \subseteq F(G)$, see Sect. 3.4.

3.2 Key lemmas

Throughout this section, we will need several results concerning the following (unconstrained) rank-one approximation problem: given a m-by-n real matrix $M \in \mathbb{R}^{m\times n}$, find its best rank-one approximation, i.e., solve

$$\min_{u\in\mathbb{R}^m, v\in\mathbb{R}^n} ||M - uv^T||_F^2. \tag{R1U(M)}$$

The following lemma is a well-known result concerning R1U(M) see, e.g., [7, Ch. 2].

Lemma 1 *The local minima of R1U(M) are global minima. All other stationary points are saddle points.*

We also have that (u, v) is a *pair of singular vectors* of matrix M with singular value $\sigma = u^T Mv$ if and only if

$$\sigma u = Mv, \quad \sigma v = M^T u \quad \text{and} \quad ||u||_2 = ||v||_2 = 1,$$

or, equivalently, if and only if

$$u = \frac{Mv}{||Mv||_2} \quad \text{and} \quad v = \frac{M^T u}{||M^T u||_2}.$$

The pair (u, v) is a stationary point of R1U(M) if and only if $\left(\frac{u}{||u||_2}, \frac{v}{||v||_2}\right)$ is a pair of singular vectors of M, and it is an optimal solution if it is associated with the maximum singular value of M, i.e., $||u||_2||v||_2 = \sigma_{\max}(M)$ [9].

We will also need the following Lemma which shows that if the minimum entry $\min(M)$ of matrix M is smaller than the Frobenius norm of the nonnegative part of M, then the best rank-one approximation of M must contain at least one nonpositive entry.

Lemma 2 *For any matrix M such that* $\min(M) \leq -||M_+||_F$*, any optimal solution* (u, v) *of R1U(M) is such that* uv^T *contains at least one nonpositive entry.*

Proof If $M = 0$, the result is trivial. Otherwise we have $\min(M) < 0$ since $\min(M) \leq -||M_+||_F$. Let (u, v) be an optimal solution of R1U(M) and assume uv^T does not contain any nonpositive entry, i.e., $uv^T > 0$. Since the negative values of M are approximated by positive ones and since M has at least one negative entry, we have

$$||M - uv^T||_F^2 > ||M_-||_F^2. \tag{7}$$

By the Eckart–Young theorem (see, e.g., [9]), the optimal rank-one approximation uv^T must satisfy

$$||M - uv^T||_F^2 = ||M||_F^2 - \sigma_{max}(M)^2 = ||M||_F^2 - ||M||_2^2 .$$

Moreover,

$$||M||_F^2 = ||M_+||_F^2 + ||M_-||_F^2 \quad \text{and} \quad ||M||_2^2 \geq \min(M)^2,$$

hence we have

$$||M - uv^T||_F^2 \leq ||M_+||_F^2 + ||M_-||_F^2 - \min(M)^2 \leq ||M_-||_F$$

which is in contradiction with Eq. (7) hence we cannot have $uv^T > 0$. □

3.3 Local and global optima of $\mathrm{R1N}_d(G)$

In this section, we show that, for any $d \geq \max(m, n)$, $\mathcal{L}_d(G) = B(G)$ and $\mathcal{G}_d(G) = H(G)$.

Lemma 3 *If $G = (V, E)$ is a bipartite graph with at least one edge and $d \geq \sqrt{|E|}$, then $\mathcal{L}_d(G) \subseteq B(G)$.*

Proof Let $A \in \{0, 1\}^{m \times n}$ be the biadjacency matrix of G with $A \neq 0$, and $M \in \{-d, 1\}^{m \times n}$ be defined as in Eq. (2). Let (u, v) be a nontrivial local minimum of $\mathrm{R1N}_d(G)$, i.e., $uv^T \in \mathcal{L}_d(G)$. Let us denote the (non-empty) support of u as $K = \mathrm{supp}(u)$ and the (non-empty) support of v as $L = \mathrm{supp}(v)$, and define $u' = u(K)$, $v' = v(L)$ and $M' = M(K, L)$ to be the subvectors and submatrix with indexes in K, L and $K \times L$, respectively. Let us also define G' as the bipartite graph whose biadjacency matrix is given by $A(K, L)$. Observe that (u', v') must be a local minimum of $\mathrm{R1N}(G')$ otherwise (u, v) would not be a local minimum of $\mathrm{R1N}_d(G)$. In fact, the objective functions of these two problems differ only by a constant factor: we have $||M - uv||_F^2 = ||M' - u'v'^T||_F^2 + ||M||_F^2 - ||M'||_F^2$. Suppose now there is a $-d$ entry in M', we have

$$\min(M') = -d \leq -\sqrt{|E|} = -||M_+||_F \leq -||M'_+||_F.$$

Moreover, (u', v') is located in the interior of the feasible domain $\mathbb{R}_+^{|K|} \times \mathbb{R}_+^{|L|}$ of $\mathrm{R1N}(G')$ since it is positive. Therefore, it is also a local minimum of the unconstrained problem, i.e., it is a local minimum of $\mathrm{R1U}(M')$. By Lemma 1, this must be a global minimum. This is a contradiction with Lemma 2: (u', v') should contain at least one nonpositive entry since $\min(M') \leq -||M'_+||_F$. Therefore M' does not contain any $-d$ entry, and we have $M' = \mathbf{1}_{|K| \times |L|}$.

Since (u', v') is a global minimum of $\mathrm{R1U}(M')$ and $M' = \mathbf{1}_{|K| \times |L|}$, we must have $u'v'^T = M' = \mathbf{1}_{|K| \times |L|}$. Therefore uv^T is binary and coincides with a feasible solution (u_b, v_b) of $\mathrm{MB}(G)$, implying that $uv^T \in F(G)$.

It remains to show that $uv^T \in B(G)$. Assume the pair (u, v) corresponds to a biclique of G which is not maximal, i.e., without loss of generality $\exists i \notin \mathrm{supp}(u)$ such that $(u_b + e_i, v_b)$ corresponds to a larger biclique of G where e_i is the ith column of the identity matrix. Then for any $0 < \epsilon \leq \max(v')^{-1}$, one can check that the solution $(u + \epsilon e_i, v)$ is strictly better than (u, v) for $\mathrm{R1N}_d(G)$: in fact, entries of M corresponding to edges contained only in the larger biclique $\{i\} \times L$ are now approximated by values between 0 and 1 (instead of 0); a contradiction which implies $uv^T \in B(G)$. □

It is interesting to notice that the converse of Lemma 3 above is not true, i.e., $B(G) \nsubseteq \mathcal{L}_d(G)$ for any $d \geq \sqrt{|E|}$. For example, with

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}, \tag{8}$$

the maximal biclique corresponding to the first row of M, i.e., $u = (1, 0)$ and $v = (1, 1, 1, 1)$ does not correspond to a local minima of $\mathrm{R1N}_d(G)$. In fact, it is not even a stationary point since, for $d = \sqrt{|E|} = \sqrt{7}$, we have $\frac{Mv}{||v||_2^2} = (1\ 0.88)^T \neq u$.

However, this holds for any $d \geq \max(m, n)$:

Theorem 1 *If G is a bipartite graph with at least one edge and $d \geq \max(m, n)$, then $B(G) = \mathcal{L}_d(G)$.*

Proof Because $\sqrt{|E|} \leq \sqrt{mn} \leq \max(m, n)$, by Lemma 3, we have $\mathcal{L}_d(G) \subseteq B(G)$ for any $d \geq \max(m, n)$. It remains to show that $B(G) \subseteq \mathcal{L}_d(G)$, which is done in Appendix A. □

Remark 1 (*Tightness of the bound*) The smallest lower bound on d which guarantees that $B(G) \subseteq \mathcal{L}_d(G)$ is given by $\max(m, n) - 1$. In fact, it is shown in Appendix 6 that for any $d > \max(m, n) - 1$, $B(G) \subseteq \mathcal{L}_d(G)$. Moreover, using the biadjacency matrix from Eq. (8), one can check that the maximal biclique $u = (1, 0)$ and $v = (1, 1, 1, 1)$ is not a local minimum of $\mathrm{R1N}_d(G)$ for any $d < \max(m, n) - 1$ because $\frac{Mv}{||v||_2^2} \neq u$ (since $M(2, :)v > 0$).

We can now prove that $\mathcal{G}_d(G) = H(G)$ for any $d \geq \max(m, n)$, which is straightforward:

Theorem 2 *If G is a bipartite graph with at least one edge and $d \geq \max(m, n)$, then $\mathcal{G}_d(G) = H(G)$.*

Proof By Theorem 1, any local minimum (u, v) of $\mathrm{R1N}_d(G)$ coincides with a feasible solution (u_b, v_b) of $\mathrm{MB}(G)$ corresponding to a maximal biclique of G, i.e., $uv^T = u_b v_b^T$. In that case, the objective functions of $\mathrm{R1N}_d(G)$ and $\mathrm{MB}(G)$ only differ by a constant factor, with $||M - uv^T||_F^2 = ||A - u_b v_b^T||_F^2 + (mn - |E|)d^2$. Hence, (u, v) is globally optimal if and only if (u_b, v_b) corresponds to a maximum biclique of G. □

Corollary 1 *R1N is NP-hard.*

Proof This is a consequence of Theorem 2 and NP-hardness of $\mathrm{MB}(G)$ [15]. □

In other words, it is NP-hard to find the best possible rank-one nonnegative approximation of a matrix which contains negative entries. Note that if the matrix to be approximated is nonnegative, then an optimal solution can be computed in polynomial time: this is a well-known result combining Eckart–Young and Perron–Frobenius theorems.

Corollary 1 is closely related to the complexity of nonnegative matrix factorization (NMF), defined as follows: given a nonnegative matrix $M \in \mathbb{R}_+^{m \times n}$ and a factorization rank r, solve

$$\min_{u_i \in \mathbb{R}_+^m, v_i \in \mathbb{R}_+^m, 1 \leq i \leq r} \left\| M - \sum_{i=1}^{r} u_i v_i^T \right\|_F^2. \qquad \text{(NMF)}$$

In fact, each rank-one subproblem in NMF (i.e., finding the best $u_i v_i^T$ with respect to the corresponding residual $M - \sum_{k \neq i} u_k v_k^T \not\geq 0$) is a R1N problem. We refer the reader to [7, Ch. 5] and [8] for more information on the link between R1N and NMF.

3.4 Stationary points of $\mathrm{R1N}_d(G)$

In this section, we focus on stationary points of $\mathrm{R1N}_d(G)$: we show how they are related to the feasible solutions of $\mathrm{MB}(G)$. These results, combined with the ones above, will be used in Sect. 4 to design a new type of biclique finding algorithm.

3.4.1 Stationarity of maximal bicliques

The next theorem states that, for $d \geq \max(m, n)$, the only nontrivial feasible solutions of MB(G) that are stationary points of R1N$_d(G)$ are the maximal bicliques, i.e., $B(G) = \mathcal{S}_d(G) \cap F(G)$.

Theorem 3 *If G is a bipartite graph with at least one edge and $d \geq \max(m, n)$, then $B(G) = F(G) \cap \mathcal{S}_d(G)$.*

Proof Let us show that $uv^T \in B(G)$ if and only if $uv^T \in F(G)$ and $uv^T \in \mathcal{S}_d(G)$. Let then $uv^T \in B(G)$ and let us assume without loss of generality that u and v are binary. By definition, uv^T belongs to $B(G)$ if and only if uv^T belongs to $F(G)$ and is maximal, i.e.,

(*) $\nexists i$ such that $u_i = 0$ and $M(i, j) = 1, \forall j$ s.t. $v_j = 1$,
(**) $\nexists j$ such that $v_j = 0$ and $M(i, j) = 1, \forall i$ s.t. $u_i = 1$.

Noting $L = \text{supp}(v)$, we have

$$v_j = \frac{||v||_1}{|L|} = 1, \quad \forall\, j \in L.$$

Moreover we have $d \geq \max(m, n)$ so that (*) is equivalent to

$$\nexists\, i \quad \text{such that} \quad u_i = 0 \quad \text{and} \quad M(i, :)v > 0.$$

Therefore, either $u_i = 0$ and $M(i, :)v \leq 0$, or $u_i = 1 = \frac{||v||_1}{|L|} = \frac{M(i,:)v}{||v||_2^2}$. These are exactly the stationarity conditions for u, cf. Eq. (6). By symmetry, (**) is equivalent to the stationarity conditions for v, so that we can conclude that $uv^T \in B(G)$ if and only if $uv^T \in F(G)$ and $uv^T \in \mathcal{S}_d(G)$. □

3.4.2 Limit points of $\mathcal{S}_d(G)$

It would be interesting to have the opposite affirmation: for d sufficiently large, does any stationary point of R1N$_d(G)$ correspond to a maximal biclique of MB(G)? Unfortunately, we will see later that this property does not hold. However, as d goes to infinity, we now show that the points in $\mathcal{S}_d(G)$ get closer to feasible solutions of MB(G), see Theorem 4 below.

Lemma 4 *For any bipartite graph G and $d \geq 0$, the set $\mathcal{S}_d(G)$ is bounded; in fact, $\forall uv^T \in \mathcal{S}_d(G)$:*

$$||uv^T||_2 = ||u||_2||v||_2 \leq \sqrt{|E|}.$$

Proof For any $uv^T \in \mathcal{S}_d(G)$, we have by (6)

$$||u||_2 = \left\| \max\left(0, \frac{M^T v}{||v||_2^2}\right) \right\|_2 \leq \frac{||\max(0, M^T)v||_2}{||v||_2^2} \leq \frac{||\max(0, M^T)||_F}{||v||_2} = \frac{\sqrt{|E|}}{||v||_2}.$$

□

Lemma 5 *For any bipartite graph G and $uv^T \in \mathcal{S}_d(G)$, if $M_{ij} = -d$ and $(uv)_{ij} > 0$, we have*

$$0 < u_i < \frac{||u||_1}{d+1} \quad \text{and} \quad 0 < v_j < \frac{||v||_1}{d+1}.$$

Proof By optimality condition (6), we have

$$0 < v_j||u||_2^2 = M(:, j)^T u \leq ||u||_1 - (d+1)u_i \;\Rightarrow\; 0 < u_i < \frac{||u||_1}{d+1}.$$

The corresponding result for v is obtained similarly. □

Theorem 4 *For any bipartite graph G, as d goes to infinity, every stationary point of $R1N_d(G)$ gets arbitrarily close to some feasible solutions of MB(G), i.e., $\forall \epsilon > 0, \exists D$ s.t. $\forall d > D$:*

$$\min_{u_b v_b^T \in F(G)} ||uv^T - u_b v_b^T||_F < \epsilon, \quad \forall uv^T \in \mathcal{S}_d(G). \tag{9}$$

Proof Let G be a bipartite graph, and A be its biadjacency matrix. Let $uv^T \in \mathcal{S}_d(G)$. Without loss of generality, we assume that $uv^T > 0$; otherwise, we consider the subproblem with the vectors $u(K)$ and $v(L)$ where K (resp. L) is the support of u (resp. v) and the graph G' corresponding to the biadjacency matrix $A(K, L)$. In fact, it is clear that if $(u(K), v(L))$ is close to a feasible solution of MB(G'), then (u, v) is for MB(G). We also assume without loss of generality that $||v||_2 = 1$ (this is because $uv^T \in S_d \Rightarrow \left(\lambda u \frac{1}{\lambda} v^T\right) \in S_d, \forall \lambda > 0$). Lemma 4 implies that $||u||_2 \leq \sqrt{|E|}$. By optimality condition (6),

$$u = Mv \quad \text{and} \quad v = \frac{M^T u}{||u||_2^2}. \tag{10}$$

Therefore, $(u/||u||_2, v) > 0$ is a pair of singular vectors of M associated with the singular value $||u||_2 > 0$. If $M = \mathbf{1}_{m\times n}$, the only pair of positive singular vectors of M is $\left(\frac{1}{\sqrt{m}}\mathbf{1}_m, \frac{1}{\sqrt{n}}\mathbf{1}_n\right)$ so that $uv^T = M$ coincides with a feasible solution of MB(G).

Otherwise, when $M \neq \mathbf{1}_{m\times n}$, we define

$$I = \left\{ i \;\middle|\; M_{ij} = 1, \quad \forall j \right\} \quad \text{and} \quad J = \left\{ j \;\middle|\; M_{ij} = 1, \quad \forall i \right\}, \tag{11}$$

and their complements $\bar{I} = \{1, 2, \ldots, m\}\backslash I$, $\bar{J} = \{1, 2, \ldots, n\}\backslash J$; with

$$M(I, :) = \mathbf{1}_{|I|\times n} \quad \text{and} \quad M(:, J) = \mathbf{1}_{m\times|J|}.$$

These two sets clearly correspond to a biclique $I \times J$ in G (since $A(I, J) = \mathbf{1}_{|I|\times|J|}$) or, equivalently, to a (binary) feasible solution $(\bar{u}_I, \bar{v}_J)$ for problem MB(G), where $\bar{u}_I$ is equal to one for indices in I and to zero otherwise (similarly for $\bar{v}_J$ and J). We are now going to show that uv^T gets arbitrarily close to $\bar{u}_I \bar{v}_J^T$ as d increases, which will prove our claim.

Using Lemma 5 and the fact that $||x||_1 \leq \sqrt{n}||x||_2 \forall x \in \mathbb{R}^n$, we get

$$0 < u(\bar{I}) < \frac{\sqrt{m|E|}}{d+1}\mathbf{1}_{|\bar{A}|} \quad \text{and} \quad 0 < v(\bar{J}) < \frac{\sqrt{n}}{d+1}\mathbf{1}_{|\bar{B}|}. \tag{12}$$

Since $||v||_2 = 1$ and $||u||_2 \leq \sqrt{|E|}$, we obtain

$$||u(\bar{I})v^T - 0||_F = ||u(\bar{I})||_2||v||_2 < \frac{1}{d+1}\left(m\sqrt{|E|}\right), \quad \text{and} \tag{13}$$

$$||uv(\bar{J})^T - 0||_F = ||u||_2||v(\bar{J})||_2 < \frac{1}{d+1}\left(n\sqrt{|E|}\right). \tag{14}$$

It remains to show that $u(I)v(J)^T$ coincides with a biclique of the (complete) graph generated by $A(I, J) = \mathbf{1}_{|I|\times|J|}$ since $u(\bar{I})v^T$ and $uv(\bar{J})^T$ tend to zero as d goes to infinity.

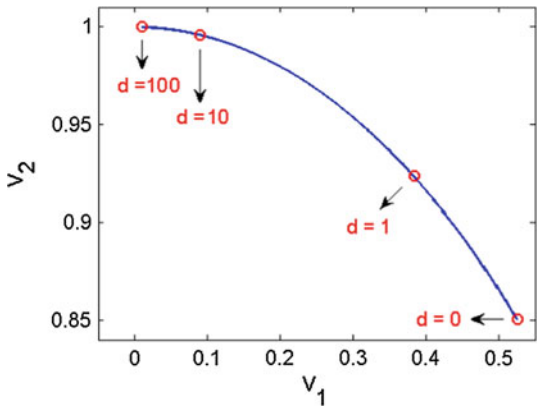

Fig. 3 Evolution of (v_1, v_2)

Noting $k_v = \frac{||u||_1}{||u||_2^2}$ and using Eq. (10), we get $v(J) = \frac{\mathbf{1}_{|J|\times m} u}{||u||_2^2} = k_v \mathbf{1}_{|J|}$. Combining this with Eq. (12) gives

$$1 - |\bar{J}|\frac{\sqrt{n}}{d+1} < ||v||_2^2 - ||v(\bar{J})||_2^2 = ||v(J)||_2^2 = |J|k_v^2 \leq ||v||_2^2 = 1. \tag{15}$$

Moreover, Eq. (10) also gives $u(I) = \mathbf{1}_{|I|\times n} v = ||v||_1 \mathbf{1}_{|I|}$ so that

$$|J|k_v \mathbf{1}_{|I|} \leq u(I) = (||v(J)||_1 + ||v(\bar{J})||_1)\mathbf{1}_{|I|} < \left(|J|k_v + |\bar{J}|\frac{\sqrt{n}}{d+1}\right)\mathbf{1}_{|I|}. \tag{16}$$

Finally, multiplying equation (16) by $k_v \mathbf{1}_{|J|}^T$, combining it with (15) and noting that we have $k_v \leq 1$ since $||v||_2 = 1$, we obtain

$$\left(1 - \frac{|\bar{J}|\sqrt{n}}{d+1}\right)\mathbf{1}_{|I|\times|J|} < u(I)v(J)^T < \left(1 + \frac{|\bar{J}|\sqrt{n}}{d+1}\right)\mathbf{1}_{|I|\times|J|}. \tag{17}$$

We can conclude that uv^T gets arbitrarily close to a feasible solution $\bar{u}_I \bar{v}_J^T$ of MB(G) as d increases; more precisely, $||uv^T - \bar{u}_I \bar{v}_J^T||_F \leq \mathcal{O}(\frac{1}{d})$. □

Example 1 Let

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} -d & 1 \\ 1 & 1 \end{pmatrix},$$

and G be the graph corresponding to the biadjacency matrix A. Clearly, $\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ belongs to the set $H(G)$, i.e., it corresponds to a maximum biclique of G. By Theorem 2, for $d \geq 2$, it belongs to $\mathcal{G}_d(G)$, i.e., $[u = (1, 1), v = (0, 1)]$ is a global minimum of R1N$_d(G)$.

For any $d > 1$, one can check that the singular values of M are different and that the outer product of the singular vectors associated with the second singular value is positive. Since it is a positive stationary point of the unconstrained problem, it is also a stationary point of R1N$_d(G)$. As d goes to infinity, it must get closer to a biclique of MB(G) (Theorem 4). Moreover, M is symmetric, so that the right and left singular vectors are equal to each other. Figure 3 shows the evolution[2] with respect to d of this positive singular vector (v_1, v_2),

[2] By Wedin's theorem (cf. matrix perturbation theory [17]), singular subspaces of M associated with a positive singular value depend continuously on d.

which is such that $(v_1\ v_2)^T(v_1\ v_2) \in \mathcal{S}_d(G)$. It converges to $(0\ 1)$, which means that the outer product of the left and right singular vectors converges to $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, a biclique, i.e., a member of $F(G)$ (not in $B(G)$).

Let us define the following rounding operator:

$$\Phi : \mathbb{R}_+^{m\times n} \to \{0,1\}^{m\times n} : X \to \left\{ \Phi(X)_{ij} = \left\{ \begin{array}{l} 0 \text{ if } X_{ij} \leq 0.5 \\ 1 \text{ if } X_{ij} > 0.5 \end{array} \right\}_{1\leq i\leq m, 1\leq j\leq n} \right.$$

Corollary 2 *For any bipartite graph G,*

$$d \geq 2\max(m,n)\sqrt{|E|}, \tag{18}$$

and $uv^T \in \mathcal{S}_d(G)$, we have $\Phi(uv^T) \in F(G)$.

Proof Let G be any bipartite graph and its biadjacency matrix A. The condition

$$\max_{uv^T\in\mathcal{S}_d(G)} \min_{u_b v_b^T \in F(G)} \max_{ij} \left(uv^T - u_b v_b^T\right)_{ij} < \frac{1}{2},$$

is clearly sufficient to guarantee that rounding any stationary point of $\mathrm{R1N}_d(G)$ will generate a biclique of G. Looking back at Theorem 4, one can check that this is satisfied, cf. Eqs. (13), (14) and (17), for d given by (18). We use the fact that $|E| \geq \max(m,n)$ can be assumed without loss of generality, i.e., that each row and each column of A has at least one nonzero entry. In fact, if A contains a column (resp. row) of all zeros then it can be discarded because, for any stationary point (u, v), the corresponding entry of u (resp. v) must be equal to zero, cf. optimality conditions (6). □

4 Biclique finding algorithm

In this section, we present a heuristic scheme designed to find large bicliques in a given graph, whose main iteration requires a number of operations proportional to the number of edges $|E|$ in the graph. It is based on the previously established links between the maximum-edge biclique problem $\mathrm{MB}(G)$ and the approximate rank-one nonnegative factorization problem $\mathrm{R1N}_d(G)$, see Theorems 1 and 2, and Corollary 2. We compare its performance on random graphs and text mining datasets with that of three other algorithms requiring $\mathcal{O}(|E|)$ operations per iteration, and to the heuristics applied at the root node level by GUROBI, a commercial mixed-integer programming solver [11].

4.1 A new biclique finding algorithm

For d sufficiently large, stationary points of $\mathrm{R1N}_d(G)$ are close to bicliques of $\mathrm{MB}(G)$ (Corollary 2). Since $\mathrm{R1N}_d(G)$ is a continuous optimization problem, *any* standard nonlinear optimization technique can in principle be used to compute such a stationary point. One can therefore think of applying an algorithm that finds a (good) stationary point of $\mathrm{R1N}_d(G)$ in order to localize a (large) biclique of the graph generated by A (the better the stationary point, the larger the biclique).

Of course, solving $\mathrm{R1N}_d(G)$ up to global optimality, i.e., finding the best stationary point, is as hard as solving $\mathrm{MB}(G)$. However, one can hope that the nonlinear optimization scheme used will converge to a relatively large biclique of G (i.e., with an objective function close to the global optimum); this hope will be confirmed empirically later in this section.

We choose to use a block-coordinate descent method, i.e., solve alternatively the problem in the variable u for v fixed, then in the variable v for u fixed, since the optimal solutions for each of these steps can be written in closed form, cf. Eq. (6). We also propose, instead of fixing the value of parameter d to the value recommended by Corollary 2, to start with a lower initial value d_0 and gradually increase it (with a multiplicative factor $\gamma > 1$) until it reaches an upper bound D equal to the recommended value. Convergence of the resulting scheme, Algorithm 1, is proved in the next theorem.

Theorem 5 *The Φ-rounding of every limit point of Algorithm 1 generates a biclique of G, the bipartite graph generated by A.*

Proof When an exact two-block coordinate descent is applied to an optimization problem with a continuously differentiable objective function and a feasible domain equal to the Cartesian product of two closed convex sets (the two blocks correspond to $\mathbb{R}^m_+$ and $\mathbb{R}^n_+$ in this case), every limit point of the iterates is a stationary point [10].

After a finite number of steps of Algorithm 1, parameter d attains the upper bound $D = 2\max(m, n)\sqrt{|E|}$ and no longer changes, so that we can invoke this result and, using Corollary 2, guarantee that the resulting limit points can be rounded to generate a feasible solution of MB(G), i.e., a biclique of G. □

Note that the normalization of u ($u \leftarrow u/\max(u)$) performed by Algorithm 1 at each iteration only changes the scaling of the solution uv^T and allows (u, v) to converge to binary vectors. Also note that the stationary points of $\mathrm{R1N}_d(G)$ which do not correspond to maximal bicliques are either saddle points or local maxima. In fact, Theorems 1 and 3 state that, for $d \geq \max(m, n)$, $\mathcal{L}_d(G) = B(G) = \mathcal{S}_d(G) \cap F(G)$. We can actually prove the following.

Theorem 6 *Let (u, v) be a nontrivial saddle point of $R1N_d(G)$ and let us note K and L the supports of u and v, respectively. Then $M(K, L)$ contains at least one $-d$ entry and $(u(K), v(L))$ is a saddle point of $R1U(M(K, L))$.*

Proof The proof is similar to the one of Theorem 1, see Appendix B. □

Theorem 6 suggests that it is very unlikely for Algorithm 1 to converge to a biclique of G which is not maximal: in fact, when restricted to positive entries of u and v, that is, $u(K)$ and $v(L)$, updates (19) and (20) correspond to the power method [9]. It is well-known that the power method applied to matrix M, when initialized with a vector which is not orthogonal to

Algorithm 1 Biclique Finding Algorithm based on R1N

Require: Bipartite graph $G = (V, E)$ described by biadjacency matrix $A \in \{0, 1\}^{m \times n}$, initial values $v_0 \in \mathbb{R}^n_{++}$, $d_0 > 0$, and parameter $\gamma > 1$.

1: **Set** parameter $D = 2\max(m, n)\sqrt{|E|}$ and **initialize** variables $d \leftarrow d_0$, $v \leftarrow v_0$;

2: **for** $k = 1, 2, \ldots$ **do**

3:

$$u \leftarrow \max\left(0, (1+d)Av - d||v||_1\right) \left(= \max(0, Mv)\right) ; \tag{19}$$

$$u \leftarrow u/\max(u) ;$$

$$v \leftarrow \max\left(0, \frac{(1+d)A^T u - d||u||_1}{||u||_2^2}\right)\left(= \max\left(0, \frac{M^T u}{||u||_2^2}\right)\right) ; \tag{20}$$

$$d \leftarrow \min(\gamma d, D) ;$$

4: **end for**

the singular subspace corresponding to the largest singular value of M, necessarily converges to a singular vector associated with the largest singular value (corresponding to a global minimum of R1U(M)). When initialized with a randomly generated vector, the probability for the power method to converge to a saddle point is therefore equal to zero. On all the numerical experiments we performed, Algorithm 1 never converged to a biclique which was not maximal.

Finally, one can easily check that Algorithm 1 requires only $\mathcal{O}(|E|)$ operations per iteration, the main cost being the computation of the matrix-vector products Av and $A^T u$ (the rest of an iteration requiring only $\mathcal{O}(\max(m, n))$ operations).

4.1.1 Parameters

It is not clear a priori how the initial value d_0 should be selected. We observed that it should not be chosen too large: otherwise, the algorithm often converges to the trivial solution: the empty biclique. In fact, in that case, the negative terms ($d||v||_1$ and $d||u||_1$) in (19) and (20) will dominate, even during the initial steps of the algorithm, and the solution will be set to zero.[3]

On the other hand, the algorithm with $d = 0$ is equivalent to the power method applied to $A \geq 0$, and then converges (under the condition stated above) to the best rank-one approximation of A. We observed that when d_0 is chosen small, the iterates will in general converge to the same solution (the one obtained when initializing the algorithm with the best rank-one approximation of A).

In order to balance positive and negative entries in M, we found appropriate to choose an initial value of d such that $||M_+||_F \approx ||M_-||_F$, i.e.,

$$d_0 \approx \frac{||A||_F}{\sqrt{|Z|}} = \sqrt{\frac{|E|}{|Z|}}, \tag{21}$$

where $|Z|$ is the number of zero entries in A, with $|E| + |Z| = mn$. We chose $d_0 = \sqrt{\frac{|E|}{|Z|}}$ for our tests, which appears to work well in practice.

The algorithm does not seem to be very sensitive to the multiplicative factor γ, and selecting values around 1.1 gives good results; this value will be used for the computational test below.

We use an a priori limit on the number of iterations as main stopping criterion. Moreover, if the solution becomes nearly integer in the sense that

$$v_i \leq 0.01 \text{ or } v_i \geq 0.99 \text{ for all } i, \quad \text{and} \quad w_j \leq 0.01 \text{ or } w_j \geq 0.99 \text{ for all } j,$$

the algorithm is terminated prematurely, as it is unlikely to further modify the rounded solution. It turns out that the algorithm converges rather fast in practice, as fewer than 50 iterations are usually required. In particular, for all runs on the synthetic datasets presented in the next section (i.e., a total of 1,700 graphs, each initialized with 100 different randomly generated vectors), the algorithm converged in at most 44 iterations (the average being 22.5). The code is available at https://sites.google.com/site/nicolasgillis/code.

[3] In practice, we used a safety procedure which reduces the value of d whenever u or v is set to zero and reinitializes u and v to their previous value.

4.2 Other algorithms in $\mathcal{O}(|E|)$ operations

We briefly present here two other algorithms designed to find large bicliques using $\mathcal{O}(|E|)$ operations per iteration.

4.2.1 Greedy heuristic

The simplest heuristic one can imagine is to add, at each step, the vertex which is connected to the most vertices in the other side of the bipartite graph. Once a vertex is selected, the vertices which are not connected to the chosen vertex are deleted. The procedure is repeated on the remaining graph until one obtains a biclique, which is necessarily maximal.

4.2.2 Motzkin–Strauss formalism

In Ding et al. [4], extend the generalized Motzkin–Strauss formalism, defined for cliques (see Sect. 5 for more details), to bicliques. They define the optimization problem

$$\max_{x \in F_x^\alpha, y \in F_y^\beta} x^T A\, y, \tag{22}$$

where A is the biadjacency matrix of G, $F_x^\alpha = \{x \in \mathbb{R}_+^n | \sum_{i=1}^n x_i^\alpha = 1\}$, $F_y^\beta = \{y \in \mathbb{R}_+^n | \sum_{i=1}^n y_i^\beta = 1\}$ and $1 < \alpha, \beta \ll 2$. Multiplicative updates for this problem are then provided:

$$x \leftarrow \left(x \circ \frac{A\, y}{x^T A\, y} \right)^{\frac{1}{\alpha}}, \quad y \leftarrow \left(y \circ \frac{A^T x}{x^T A\, y} \right)^{\frac{1}{\beta}}. \tag{MS}$$

This algorithm does not necessarily converge to a biclique: if α and β are not sufficiently small, it may only converge to a dense bipartite subgraph (a bicluster). In particular, for $\alpha = \beta = 2$, it converges to an optimal rank-one solution of R1U(A), as Algorithm 1 does for $d = 0$. For our tests, we choose $\alpha = \beta = 1.05$ as recommended in [4]. The updates MS have a computational cost comparable to that of Algorithm 1 ($\mathcal{O}(|E|)$ operations) since their main cost is the computation of the matrix-vector products Ay and $A^T x$.

In order to evaluate the quality of the solutions provided by this algorithm when it did not converge to a biclique, we considered the following two different post-processing procedures to convert a bicluster into a biclique:

1. Extract from the generated bicluster a biclique using the greedy heuristic presented above. We will refer to this variant of the algorithm based on the MS updates post-processed with the Greedy algorithm as *Greedy MS*.
2. Use the updates MS recursively on the extracted bicluster, i.e., rerun it on the positive submatrix while decreasing the values of parameters α and β with $\alpha \leftarrow 1 + \frac{\alpha - 1}{2}$ and $\beta \leftarrow 1 + \frac{\beta - 1}{2}$. We will refer to this variant of the algorithm based on the MS updates used recursively as *Recursive MS*.

Both variants will be tested in Sect. 4.4.

4.3 MIP root node heuristics

Besides the three $\mathcal{O}(|E|)$ algorithms described above, we also compare our algorithm against the sophisticated procedures implemented at the root node level in GUROBI, a commercial mixed-integer programming (MIP) solver [11]. Three different formulations were considered:

1. The original formulation (1), which is a (nonconvex) binary integer program with a quadratic objective function.
2. A convex mixed integer reformulation of (1) where a new continuous nonnegative variable t appears as the objective function and an additional convex constraint $t^2 \leq (\sum_i u_i)(\sum_j v_j)$ is introduced. As the latter can be rewritten as $4t^2 + (\sum_i u_i - \sum_j v_j)^2 \leq (\sum_i u_i + \sum_j v_j)^2$, the resulting problem is a mixed integer binary second-order cone problem.
3. A linear binary reformulation of (1) where each quadratic term $u_i v_j$ in the objective function is linearized. More precisely, the objective function is replaced by a sum of binary variables s_{ij} for all pairs (i, j) such that $A_{ij} > 0$, and the corresponding linking constraints $2s_{ij} \leq u_i + v_j$ are introduced.

Parameters `TimeLimit=1` and `NodeLimit=1` were provided to the solver[4] in order to use only root node heuristics and limit the total CPU time spent to one second (which is already more than an order of magnitude larger than what the competing $\mathcal{O}(|E|)$ algorithms require, see Table 1).

Preliminary testing revealed that the convex second-order cone formulation is not competitive at all, and that the results of the linear reformulation are inferior to those of the original quadratic formulation, probably because of the large number of additional variables introduced in the linear reformulation. Hence, in the next section, we only report results for the original quadratic formulation.

4.4 Results

In this section, we present some numerical results for synthetic and text mining datasets. All the experiments were performed on a desktop computer running MATLAB R2012b (64 bits) on an Intel® CORE i5-2320 CPU @ 3GHz processor equipped with 6 Go of RAM.

4.4.1 Synthetic data

For each density (0.1, 0.3, 0.5, 0.7 and 0.9), 100 bipartite graphs with 200 vertices (100 on each side, i.e., $m = n = 100$) are randomly generated (the probability that an edge belongs to the graph is equal to the density). We then perform, for each graph, 100 runs with the same random initializations and each algorithm is allotted a maximum of 100 iterations, except for the greedy heuristic, which is always run until completion and only once for each graph (since it does not require a random initialization), and the MIP heuristic, also run once with a one-second time limit.

Table 1 gives the average computational times measured by MATLAB for the different algorithms when tested on graphs with different densities (note that, on a multi-processor machine, MATLAB reports the sum of the CPU times used on each processor).

We observe that

- Greedy and Algorithm 1 are the fastest algorithms. On dense graphs, for which the greedy heuristic requires more iterations to identify a biclique, Algorithm 1 is slightly faster.
- Greedy MS and Recursive MS require roughly the same computational time on sparse graphs and actually return the same solutions (see Fig. 5): the reason is that the MS updates are able to identify a biclique by themselves, and no post-processing is required.

[4] Additional tweaking of parameters `MIPFocus`, `Heuristics`, `PreQLinearize`, `MIQCPMethod` and `RINS` did not lead to better results.

Table 1 Average computational time for solving 100 biclique problems on 100-by-100 randomly generated bipartite graphs, for various densities of the biadjacency matrix

Density	Greedy	Algorithm 1	Greedy MS	Recursive MS	MIP
0.05	0.19	0.19	1.06	1.07	150.5
0.1	0.22	0.16	1.16	1.16	145.8
0.15	0.21	0.19	1.18	1.20	141.4
0.2	0.19	0.22	1.18	1.18	137.1
0.3	0.23	0.20	1.19	1.22	129.4
0.4	0.17	0.21	1.21	1.25	122.5
0.5	0.25	0.25	1.24	1.37	116.6
0.6	0.28	0.29	1.24	1.52	157.7
0.7	0.23	0.30	1.31	1.79	175.4
0.8	0.42	0.33	1.43	2.23	108.2
0.85	0.51	0.33	1.57	2.67	102.8
0.9	0.62	0.34	1.74	3.82	102.7
0.95	0.82	0.36	2.02	5.26	76.0

For dense graphs, Recursive MS is slower because it recursively calls the MS updates until it identifies a biclique (while Greedy MS only calls the greedy heuristic once).

- MIP is the slowest, as we allowed it to run for one second on each graph (also note that the one-second time constraint provided to GUROBI applies to wall time and not total CPU time).

Figure 4 displays the performance profile for these experiments [5], where the performance function at $\rho \leq 1$ is defined as the percentage, among all graphs and all runs, of bicliques whose sizes (i.e., number of edges) is larger than ρ times the size the largest biclique found by any algorithm in the corresponding graph, i.e.,

$$\text{performance}(\rho) = \frac{\#\{\text{bicliques} \mid \text{size} \geq \rho \times \text{size of best biclique found}\}}{\#\text{runs}}.$$

On such a performance profile, the higher the curve, the better; more specifically, the left part of the graph measures efficiency, i.e., how often a given algorithm produces the best biclique among its peers, while the right part estimates robustness, i.e., how far from the best non-optimal solutions are. These two aspects are also reported more quantitatively in Table 2, which displays the value of the performance function at $\rho = 1$ (Efficiency, i.e., how often a given algorithm finds a biclique with largest size) and the smallest value of ρ such that the performance function is equal to 100 % (Robustness, i.e., the relative size of the worst biclique found).

We observe on the performance profile that both Algorithm 1 and MS perform better than the greedy heuristic. The variant of MS using recursive post-processing performs slightly better than the one based on the use of the greedy heuristic. Nevertheless, Algorithm 1 generates in general better solutions: it is more efficient (16 % of its solutions are 'optimal', the second best being the MS algorithms with 6 %) and more robust (all solutions are at most a factor 0.31 away from the best solution, the second best being the greedy heuristic with 0.29). Despite the larger amount of CPU time spent, GUROBI performs on average rather poorly on these instances.

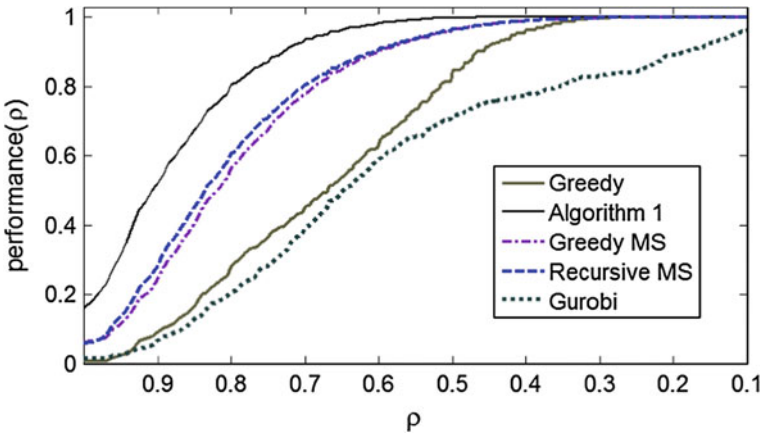

Fig. 4 Performance profile for random graphs (densities from 0.1 to 0.9)

Table 2 Efficiency and Robustness of the different algorithms on 100 randomly generated graphs

	Greedy	Algo. 1	Greedy MS	Rec. MS	GUROBI
All	1 %—0.29	16 %—0.31	6 %—0.17	6 %—0.17	1 %—0
Sparse	0 %—0.27	33 %—0.32	18 %—0.19	18 %—0.19	9 %—0.38
Dense	7 %—0.65	26 %—0.79	8 %—0.62	2 %—0.63	0 %—0.07

All corresponds to Fig. 4, Sparse to Fig. 5 left, and Dense Fig. 5 right

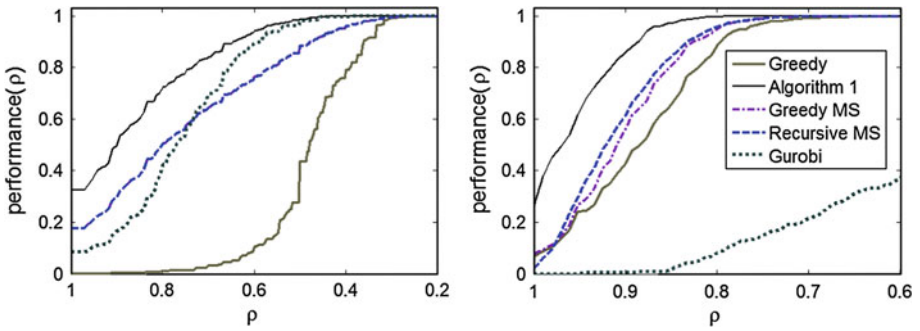

Fig. 5 Performance profiles for random graphs: sparse (*left*, from 0.05 to 0.2) and dense (*right*, from 0.8 to 0.95)

It is worth noting that the algorithms behave quite differently on sparse and dense graphs. Using the same setting as before, Fig. 5 displays performance profiles for sparse graphs (on the left, with densities 0.05, 0.1, 0.15 and 0.2) and dense graphs (on the right, with densities 0.8, 0.85, 0.9 and 0.95).

For sparse graphs, Algorithm 1 is the best overall performer. Both versions of MS coincide, and the greedy heuristic performs significantly worse. The GUROBI heuristics now perform better, especially for smaller values of ρ: they are slightly more robust than Algorithm 1 (0.38 vs. 0.32), but are still significantly less efficient (9 vs. 33 %).

Table 3 Text mining datasets [18] (sparsity is given in %: $100 \times |Z|/(mn)$)

Data	m	n	$\|E\|$	Sparsity
Classic	7,094	41,681	223,839	99.92
Sports	8,580	14,870	1,091,723	99.14
Reviews	4,069	18,483	758,635	98.99
hitech	2,301	10,080	331,373	98.57
ohscal	11,162	11,465	674,365	99.47
la1	3,204	31,472	484,024	99.52

For dense graphs, the GUROBI heuristics seem quite ineffective (in fact, for 0.95 density, they terminate before the one-second time limit and return a very poor solution, see Fig. 5). Recursive MS performs slightly better than the Greedy MS (although it is slightly less efficient) which performs slightly better than the greedy heuristic. Algorithm 1 performs the best: it is more efficient as it finds the best solution in 26 % of the runs (the second best being Greedy MS with 8 %), and it is more robust as all solutions are at most a factor 0.79 away from the best solution (the second best being the greedy heuristic with 0.65).

4.4.2 Text datasets

If parameter D in Algorithm 1 is chosen smaller than the value recommended by Corollary 2, the algorithm is no longer guaranteed to converge to a biclique. However, the negative entries in M will force the corresponding entries of the solutions of $\mathrm{R1N}_d(G)$ to be small (cf. Theorem 4). Therefore, instead of a biclique, one gets a dense submatrix of A, i.e., a *bicluster*. Algorithm 1 can then be used as a *biclustering algorithm* and the density of the corresponding submatrix will depend on the choice of parameter D between 0 and $2\max(m,n)\sqrt{|E|}$. We test this approach on the six text mining datasets (with sparse matrices) described in Table 3.

Figure 6 compares Algorithms 1 and MS for varying values of their parameters: in the Motzkin–Strauss formalism, we tested each value for $\alpha = \beta$ in the interval $[1.3,\ 1.9]$ with step size 0.025 and, for Algorithm 1, we tried $D = d_0 10^x$ for each value of x in the interval $[3,\ 9]$ with step size 0.25 (d_0 given by Eq. 21). For each value, we performed 10 runs (same initializations for both algorithms and 500 iterations) and plotted all the non-dominated solutions (i.e., for which no other solution has both larger size and higher density) for each dataset.

We observe that our approach consistently generates better results since its curves dominate the ones of the Motzkin–Strauss formalism, i.e., the biclusters it finds are denser for the same size or larger for the same density.

Table 4 gives the average computational time for the different algorithms to computing one bicluster for the different datasets. We observe that both algorithms spent roughly the same computational time: the main effort per iteration of both algorithms is essentially the same (two matrix-vector products) while a fixed number of 500 iterations is performed (no early termination or post-processing were performed, as opposed to the experiments performed in the previous section).

Finally, we mention that Algorithm 1 can be further enhanced in the following ways:

- It is applicable to weighted graphs, i.e., non-binary biadjacency matrices; and Theorems 1, 2 and 4 can be adapted: Lemma 2 still holds for $d \geq ||M_+||_F$, and the lower bound

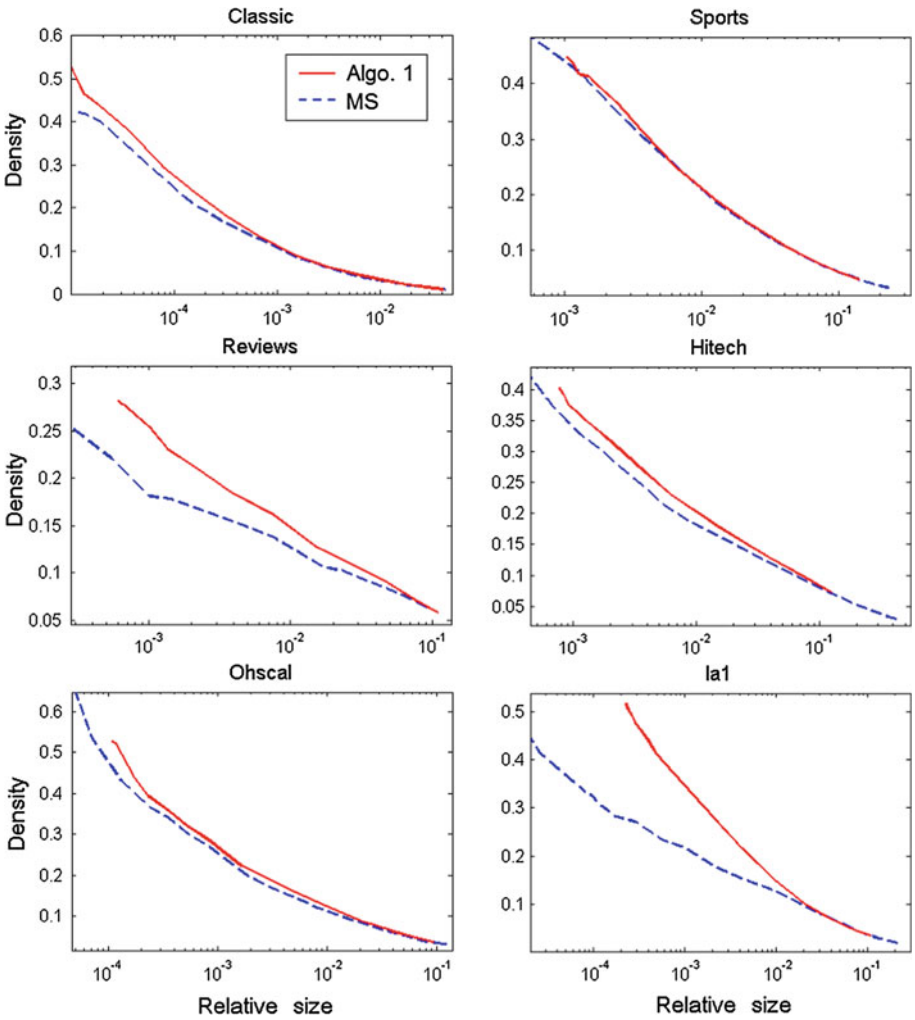

Fig. 6 Relative size versus density for the Motzkin–Strauss formalism for biclique (MS, *dashed line*) and Algorithm 1 based on $\mathrm{R1N}_d(G)$ (*solid line*). The x-axis indicates the relative sizes of the extracted clusters (i.e., number of entries in the extracted submatrix divided by the number of entries in the original matrix) while the y-axis indicates the density of these clusters (number of nonzero entries divided by the total number of entries) for the text datasets of Table 3

$\max(m, n)$ on d in Theorem 1 can be replaced by $\max(m, n) \max_{ij}(A)$, where A is the weighted biadjacency matrix.

- It is possible to give more weight to a given side of the biclique by adding regularization terms to the cost functions. For example, one can consider the following objective function

$$\min_{u \geq 0, v \geq 0} \quad ||M - uv^T||_F^2 + \alpha ||u||_2^2 + \beta ||v||_2^2$$

which our algorithm can handle after some straightforward modifications (namely, the optimal solution for u when v is fixed can still be written in closed-form, and vice versa).

Table 4 Average computational time in seconds spent by the different algorithms for computing one bicluster on each text dataset

	Classic	Sports	Reviews	Hitech	Ohscal	La1
MS	3.97	6.23	4.81	2.27	4.81	5.24
Algorithm 1	2.39	6.81	4.89	2.13	4.57	3.74

- If $A \in \{0,1\}^{n\times n}$ is the adjacency matrix of a (non bipartite) graph $G = (V, E)$ with no self loop where $V = \{v_1, \ldots, v_n\}$, i.e., $A(i, j) = 1 \Leftrightarrow (v_i, v_j) \in E\, i \neq j$ and $A(i, i) = 0$ for all i, then MB(G) corresponds to the problem of identifying two disjoint sets of nodes where all the nodes of one set are connected by an edge to all the nodes of the other set. In fact, MB(G) reduces to identifying the largest block of ones in any binary matrix. Therefore, all the results of this paper apply to this particular problem.

5 Maximum clique and Motzkin–Strauss formalism

In this last section, we show how our formulation is related to the Motzkin–Strauss formalism for maximum clique finding.

Given a graph G, the maximum clique problem looks for a complete subgraph (i.e., a clique) with maximum number of vertices (or equivalently with maximum number of edges, because a clique with n vertices has $\binom{n}{2}$ edges). The optimal solution is denoted $\omega(G)$ and called the clique number of G.

Given the adjacency matrix B of G, it was shown by Motzkin and Strauss [14] that the following quadratic program, called Motzkin–Strauss QP,

$$c^* = \max_{x\in\mathbb{R}^n} x^T Bx \quad \text{such that } x \geq 0 \text{ and } e^T x = ||x||_1 = 1, \tag{23}$$

satisfies $c^* = (1 - 1/\omega(G))$. Moreover, there is a close link between local maxima of (23) and maximal cliques of G [6]. In particular, slightly modifying the problem using a quadratic penalty (namely, setting $B_{ii} = \frac{1}{2}$ for all i) leads to a one-to-one correspondence between these two sets [2].

If G is bipartite, Ding et al. extended this formalism to bicliques [3,4], see Eq. (22). They proved that optimal solutions (x^*, y^*) of (22) such that the nonzero elements of x^* (resp. y^*) are equal to each other define maximal bicliques in G. However, no other theoretical guarantee is provided, and it is not clear for example whether there is a one-to-one correspondence between global (resp. local) minima of (22) and the maximum (resp. maximal) bicliques of G.

We now briefly explain how our continuous formulation R1N$_d(G)$ of MB(G) is actually closely related to the formalism introduced by Motzkin and Strauss for the clique problem. First, observe that R1N$_d(G)$ can be equivalently reformulated as

$$q^* = \min_{\sigma\geq 0, u\geq 0, v\geq 0} ||M - \sigma uv^T||_F^2 \quad \text{such that } ||u||_2 = 1, ||v||_2 = 1. \tag{24}$$

Since $||M - \sigma uv^T||_F^2 = ||M||_F^2 - 2\sigma u^T Mv + \sigma^2$, the function $||M - \sigma uv^T||_F^2$ is (convex) quadratic in σ and, for any (u, v), the optimal value for σ (corresponding to the stationarity conditions) is given by $\sigma^* = \max(0, u^T Mv)$. In fact, either $u^T Mv$ is nonnegative and

$\sigma^* = u^T M v$, or it is negative and $\sigma^* = 0$; the corresponding objective value is $||M||_F^2 - \max(0, u^T M v)^2$.

If the graph G has at least one edge (otherwise the problem is trivial since $M < 0$), then M has at least one positive entry (i.e., $M_{ij} = 1$ for some i, j), implying $q^* < ||M||_F^2$ (take for example $u = e_i$ and $v = e_j$, where e_k is the kth column of the identity matrix, as a feasible solution). In that case, we must then have $\sigma^* = u^{*T} M v^* > 0$ for any nontrivial stationary point (σ^*, u^*, v^*) of (24). Since $||M - \sigma^* u^* v^{*T}||_F^2 = ||M||_F^2 - (u^{*T} M v^*)^2$, the triplet (σ^*, u^*, v^*) minimizes $||M - \sigma u^* v^{*T}||_F^2$ if and only if it maximizes $(u^{*T} M v^*)^2$ or equivalently $u^{*T} M v^*$ since it is positive. Finally, if M has at least one positive entry, the problem

$$p^* = \max_{u \geq 0, v \geq 0} u^T M v \quad \text{such that } ||u||_2^2 = 1 \text{ and } ||v||_2^2 = 1, \tag{25}$$

satisfies $q^* = ||M||_F^2 - (p^*)^2$. This problem is very similar to the Motzkin–Strauss formalism (23), except that we now have a constraint on the ℓ_2-norm of the variables (instead of ℓ_1), and that matrix M is not the biadjacency matrix of G (but is closely related to it, similarly as in [2] for the clique problem).

Therefore, all results of this paper actually apply to formulation (25) above. In fact, one can check that the first-order stationarity conditions for (25) are, up to a constant factor, the same as for $\mathrm{R1N}_d(G)$, so that there is a one-to-one correspondence between global minima, local minima and stationary points of (25) and $\mathrm{R1N}_d(G)$.

6 Conclusion

Given a graph G, we have proposed a new continuous characterization for the maximum-edge biclique problem based on an approximate rank-one matrix factorization problem, namely $\mathrm{R1N}_d(G)$. We proved that there is a one-to-one correspondence between the maximal (resp. maximum) bicliques of G and the local (resp. global) minima of $\mathrm{R1N}_d(G)$. We also showed that the stationary points of $\mathrm{R1N}_d(G)$ are close to bicliques of G. Based on these results, we presented a heuristic biclique-finding algorithm whose iterations require $\mathcal{O}(|E|)$ operations per iteration. We experimentally demonstrated its efficiency on random graphs and text mining datasets. Finally, we showed how $\mathrm{R1N}_d(G)$ is closely related to the Motzkin–Strauss formalism for cliques.

Appendix A: Proof of Theorem 1

Let us show that $B(G) \subseteq \mathcal{L}_d(G)$ for any $d \geq \max(m, n)$. Let $uv^T \in B(G)$, with u and v binary without loss of generality. The binary rank-one matrix uv^T belongs to $\mathcal{L}_d(G)$ if and only if there exists $\epsilon > 0$ such that for all $x \in \mathcal{B}_+(u, \epsilon)$ and $y \in \mathcal{B}_+(v, \epsilon)$, we have $||M - uv^T||_F^2 \leq ||M - xy^T||_F^2$.

Let then $x \in \mathcal{B}_+(u, \epsilon)$ and $y \in \mathcal{B}_+(v, \epsilon)$, and let us note S_u, S_v, S_x and S_y the supports of u, v, x and y, respectively. For $\epsilon < 1$, since u and v are binary, we have $S_u \subseteq S_x$ and $S_v \subseteq S_y$ (i.e., $u_i = 1 \Rightarrow x_i > 0$ and $v_j = 1 \Rightarrow y_j > 0$). This implies that for $\epsilon < 1$, $||M - uv^T||_F^2 \leq ||M - xy^T||_F^2$ if and only if

$$||M(S_x, S_y) - u(S_x)v(S_y)^T||_F^2 \leq ||M(S_x, S_y) - x(S_x)y(S_y)^T||_F^2.$$

Let us note $\bar{S_u} = S_x \backslash S_u$ and $\bar{S_v} = S_y \backslash S_v$. Since $x \in \mathcal{B}_+(x, \epsilon)$, there exists δu such that $x = u + \epsilon\delta u$ with $||\delta u||_2 \leq 1$ and $\delta u(\bar{S_u}) \geq 0$ since $u(\bar{S_u}) = 0$; symmetrically there exists δv such that $y = v + \epsilon\delta v$ with $||\delta v||_2 \leq 1$ and $\delta v(\bar{S_v}) \geq 0$.

Let us analyze the four submatrices of $M(S_x, S_y)$ corresponding to the decomposition $S_x = S_u \cup \bar{S_u}$ and $S_x = S_u \cup \bar{S_u}$.

1. Submatrix (S_u, S_v). Since $M(S_u, S_v) = \mathbf{1}_{|S_u|\times|S_v|}$, $u(S_u) = \mathbf{1}_{|S_u|}$ and $v(S_v) = \mathbf{1}_{|S_v|}$,

$$e_1 = ||M(S_u, S_v) - x(S_u)y(S_v)^T||_F^2 \geq ||M(S_u, S_v) - u(S_u)v(S_v)^T||_F^2 = 0.$$

2. Submatrix $(\bar{S_u}, \bar{S_v})$. Since $u(\bar{S_u}) = 0$, $v(\bar{S_v}) = 0$ and $||M(\bar{S_u}, \bar{S_v})||_F^2 \leq |\bar{S_u}||\bar{S_v}|d^2 \leq mnd^2$ for $d \geq 1$,

$$e_2 = ||M(\bar{S_u}, \bar{S_v}) - x(\bar{S_u})y(\bar{S_v})^T||_F^2 = ||M(\bar{S_u}, \bar{S_v}) - \epsilon^2\delta u(\bar{S_u})\delta v(\bar{S_v})^T||_F^2$$
$$||\delta u(\bar{S_u})\delta v(\bar{S_v})^T||_F^2.$$

   In fact, recall that $||A-B||_F^2 = ||A||_F^2 - 2\sum_{ij} A_{ij}B_{ij} + ||B||_F^2 \geq ||A||_F^2 - 2||A||_F||B||_F$.

3. Submatrix $(S_u, \bar{S_v})$. Since $u(S_u) = \mathbf{1}_{|S_u|}$, $v(\bar{S_v}) = \mathbf{0}_{|\bar{S_v}|}$, $d \geq 1$ and $\epsilon < 1$,

$$\begin{aligned} e_3 &= ||M(S_u, \bar{S_v}) - x(S_u)y(\bar{S_v})^T||_F^2 \\ &= ||M(S_u, \bar{S_v}) - \epsilon(\mathbf{1}_{|S_u|} + \epsilon\delta u(S_u))\delta v(\bar{S_v})^T||_F^2 \\ &= ||M(S_u, \bar{S_v}) - \epsilon\mathbf{1}_{|S_u|}\delta v(\bar{S_v})^T - \epsilon^2\delta u(S_u)\delta v(\bar{S_v})^T||_F^2 \\ &\geq ||M(S_u, \bar{S_v}) - \epsilon\mathbf{1}_{|S_u|}\delta v(\bar{S_v})^T||_F^2 - 2\sqrt{mn}(d+1)\epsilon^2||\delta u(S_u)\delta v(\bar{S_v})^T||_F. \end{aligned}$$

   In fact, one can check that $|M(S_u, \bar{S_v}) - \epsilon\mathbf{1}_{|S_u|}\delta v(\bar{S_v})^T| \leq d + 1$ for $\epsilon < 1$ since $|\delta v(\bar{S_v})| \leq 1$ implying that $||M(S_u, \bar{S_v}) - \epsilon\mathbf{1}_{|S_u|}\delta v(\bar{S_v})^T||_F^2 \leq mn(d+1)^2$.
   Because (u, v) corresponds to a maximal biclique, there must be at least one $-d$ entry in each column of $M(S_u, \bar{S_v})$. Let us analyze each column separately. For any $i \in \bar{S_v}$, let us note $n_i \geq 1$ the number of $-d$ entry in the column $M(S_u, i)$. We have

$$\begin{aligned} ||M(S_u, i) - \epsilon\mathbf{1}_{|S_u|}\delta v(i)||_F^2 &= n_i(-d - \epsilon\delta v(i))^2 + (|S_u| - n_i)(1 - \epsilon\delta v(i))^2 \\ &\geq n_i d^2 + (|S_u| - n_i) + 2\epsilon\delta v(i)(n_i d - |S_u| + n_i) \\ &= ||M(S_u, i)||_F^2 + 2\epsilon\delta v(i)(n_i d + n_i - |S_u|) \\ &\geq ||M(S_u, i)||_F^2 + 2\epsilon\delta v(i). \end{aligned}$$

   In fact, $n_i d \geq d \geq \max(m, n) \geq |S_u|$ (it is then actually sufficient to take $d > \max(m, n) - 1$). Finally, recalling that $\delta v(\bar{S_v}) \geq 0$ and summing on index $i \in \bar{S_v}$, we obtain

$$\begin{aligned} e_3 \geq\ & ||M(S_u, \bar{S_v}) - u(S_u)v(\bar{S_v})^T||_F^2 + 2\epsilon||\delta v(\bar{S_v})||_1 \\ & -2\sqrt{mn}(d+1)\epsilon^2||\delta u(S_u)\delta v(\bar{S_v})^T||_F. \end{aligned}$$

4. Submatrix $(\bar{S_u}, S_v)$. By symmetry, the same can be done as for the submatrix $(S_u, \bar{S_v})$, and we have

$$\begin{aligned} e_4 &= ||M(\bar{S_u}, S_v) - x(\bar{S_u})y(S_v)^T||_F^2 \\ &\geq ||M(\bar{S_u}, S_v) - u(\bar{S_u})v(S_v)^T||_F^2 + 2\epsilon||\delta u(\bar{S_u})||_1 \\ &\quad -2\sqrt{mn}(d+1)\epsilon^2||\delta u(S_u)\delta v(\bar{S_v})^T||_F. \end{aligned}$$

Combining the above results and noting $C = 2\sqrt{mn}(d+1)$, we have

$$\begin{aligned} e_T &= e_1 + e_2 + e_3 + e_4 \\ &= ||M(S_x, S_y) - x(S_x)y(S_y)^T||_F^2 \\ &\geq ||M(S_x, S_y) - u(S_x)u(S_y)^T||_F^2 + 2\epsilon||\delta u(\bar{S}_u)||_1 + 2\epsilon||\delta v(\bar{S}_v)||_1 \\ &\quad -C\epsilon^2||\delta u(\bar{S}_u)\delta v(\bar{S}_v)^T||_F^2 - C\epsilon^2(||\delta u(\bar{S}_u)\delta v(S_v)^T||_F^2 + ||\delta u(S_u)\delta v(\bar{S}_v)^T||_F^2). \end{aligned}$$

Recalling that $||x||_1 \geq ||x||_2$ for any $x \in \mathbb{R}^n$, $||xy^T||_F = ||x||_2||y||_2$ for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, and that $||\delta u||_2 \leq 1$ and $||\delta v||_2 \leq 1$, we have that for any $0 < \epsilon < \frac{1}{C}$

$$\begin{aligned} e_T &\geq ||M(S_x, S_y) - u(S_x)u(S_y)^T||_F^2 \\ &\quad +\epsilon||\delta u(\bar{S}_u)||_2^{\frac{1}{2}}\left(2 - C\epsilon||\delta u(\bar{S}_u)||_2^{\frac{1}{2}}||\delta v(\bar{S}_v)^T||_2 - C\epsilon||\delta u(\bar{S}_u)||_2^{\frac{1}{2}}||\delta v(S_v)^T||_2\right) \\ &\quad +\epsilon||\delta v(\bar{S}_v)||_2^{\frac{1}{2}}\left(2 - C\epsilon||\delta v(\bar{S}_v)||_2^{\frac{1}{2}}||\delta u(\bar{S}_u)^T||_2 - C\epsilon||\delta v(\bar{S}_v)||_2^{\frac{1}{2}}||\delta u(S_u)^T||_2\right) \\ &\geq ||M(S_x, S_y) - u(S_x)u(S_y)^T||_F^2 + 2\epsilon(1 - C\epsilon)(||\delta u(\bar{S}_u)||_2^{\frac{1}{2}} + ||\delta v(\bar{S}_v)||_2^{\frac{1}{2}}) \\ &\geq ||M(S_x, S_y) - u(S_x)v(S_y)^T||_F^2. \end{aligned}$$

Finally, for any $d \geq \max(m, n)$, $uv^T \in B(G)$, $0 < \epsilon < \frac{1}{2mn(d+1)^2}$, $x \in \mathcal{B}_+(u, \epsilon)$ and $y \in \mathcal{B}_+(v, \epsilon)$, we have $||M - uv^T||_F^2 \leq ||M - xy^T||_F^2$.

Appendix B: Proof of Theorem 6

Let (u, v) be a nontrivial saddle point of $\mathrm{R1N}_d(G)$ (hence $uv^T \in \mathcal{S}_d(G)$). Let us denote the (non-empty) support of u as $K = \mathrm{supp}(u)$ and the (non-empty) support of v as $L = \mathrm{supp}(v)$, and define $u' = u(K)$, $v' = v(L)$ and $M' = M(K, L)$ to be the subvectors and submatrix with indexes in K, L and $K \times L$, respectively. Let us also define G' as the bipartite graph whose biadjacency matrix is given by $A(K, L)$.

Observe that (u', v') must be a saddle point of $\mathrm{R1N}(G')$ otherwise (u, v) would not be a saddle point of $\mathrm{R1N}_d(G)$. In fact, the objective functions of these two problems differ only by a constant factor: we have $||M - uv||_F^2 = ||M' - u'v'^T||_F^2 + ||M||_F^2 - ||M'||_F^2$. By stationarity of (u, v), Eq. (6) gives

$$u' = \frac{M'v'}{||v'||_2^2} \quad \text{and} \quad v' = \frac{M'^T u'}{||u'||_2^2}.$$

Therefore, $(u'/||u'||_2, v'/||v'||_2) > 0$ defines a pair of singular vectors of M' associated with the singular value $||u'||_2||v'||_2 > 0$.

If M' does not contain any $-d$ entries, then $(u', v') = (\mathbf{1}_{|K|}, \mathbf{1}_{|L|})$ is the unique pair of positive singular vectors (up to a constant factor). We then have that $uv^T \in F(G)$. By Theorem 3, $uv^T \in B(G) = \mathcal{L}_d(G)$ is then a local minima since $F(G) \cap \mathcal{S}_d(G) = B(G) = \mathcal{L}_d(G)$ for any $d \geq \max(m, n)$, a contradiction.

Therefore M' contains at least one $-d$ entry. By Lemma 2, any pair of singular vectors of M' associated with the largest singular value of M' must contain a least one non-positive entry. Therefore, (u', v') is a pair of positive singular vectors of M' not associated with the largest singular value of M', i.e., it is a saddle point of $\mathrm{R1U}(M')$.

An example of such a saddle point is given in Example 1.

References

1. Alexe, G., Alexe, S., Crama, Y., Foldes, S., Hammer, P., Simeone, B.: Consensus algorithms for the generation of all maximal bicliques. Discret. Appl. Math. **145**(1), 11–21 (2004)
2. Bomze, I.: Evolution towards the maximum clique. J. Glob. Opt. **10**, 143–164 (1997)
3. Ding, C., Li, T., Jordan, M.: Nonnegative matrix factorization for combinatorial optimization: spectral clustering, graph matching, and clique finding. In: IEEE International Conference on Data Mining, pp. 183–192 (2008)
4. Ding, C., Zhang, Y., Li, T., Holbrook, S.: Biclustering protein complex interactions with a biclique finding algorithm. In: IEEE International Conference on Data Mining, pp. 178–187 (2006)
5. Dolan, E., Moré, J.: Benchmarking optimization software with performance profiles. Math. Prog. Ser. A **91**, 201–213 (2002)
6. Gibbons, L., Hearn, D., Pardalos, P., Ramana, M.: Continuous characterizations of the maximum clique problem. Math. Oper. Res. **22**(3), 754–768 (1997)
7. Gillis, N.: Nonnegative Matrix Factorization: Complexity, Algorithms and Applications. Ph.D. Thesis, Université catholique de Louvain (2011)
8. Gillis, N., Glineur, F.: Nonnegative Factorization and The Maximum Edge Biclique Problem (2008). CORE Discuss. pap. 2008/64
9. Golub, G., Van Loan, C.: Matrix Computation, 3rd edn. The Johns Hopkins University Press, Baltimore (1996)
10. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. Oper. Res. Lett. **26**, 127–136 (2000)
11. Gurobi Optimization, I.: Gurobi Optimizer Reference Manual (2012). http://www.gurobi.com
12. Lehmann, S., Schwartz, M., Hansen, L.: Biclique communities. Phys. Rev. E **78**(1), 016108 (2008)
13. Liu, G., Sim, K., Li, J.: Efficient Mining of Large Maximal Bicliques, Lect. Notes in Comput. Sci. pp. 437–448. Springer, Berlin (2006)
14. Motzkin, T., Strauss, E.: Maxima for graphs and a new proof of a theorem of Turan. Can. J. Math. **17**, 533–540 (1965)
15. Peeters, R.: The maximum edge biclique problem is NP-complete. Discret. Appl. Math. **131**(3), 651–654 (2003)
16. Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissemb, W., Hennig, L., Thiele, L., Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22**(9), 1122–1129 (2008)
17. Stewart, G., Sun, J.G.: Matrix Perturbation Theory. Academic Press, San Diego (1990)
18. Zhong, S., Ghosh, J.: Generative model-based document clustering: a comparative study. Knowl. Inf. Syst. **8**(3), 374–384 (2005)