Detecting Bicliques in GF[q]

Jan Ramon[1], Pauli Miettinen[2], and Jilles Vreeken[3]

[1] Department of Computer Science, KU Leuven, Belgium
Jan.Ramon@cs.kuleuven.be
[2] Max-Planck Institute for Informatics, Saarbrücken, Germany
pmiettin@mpi-inf.mpg.de
[3] Dept. of Mathematics and Computer Science, University of Antwerp, Belgium
Jilles.Vreeken@ua.ac.be

Abstract. We consider the problem of finding planted bicliques in random matrices over $GF[q]$. That is, our input matrix is a $GF[q]$-sum of an unknown biclique (rank-1 matrix) and a random matrix. We study different models for the random graphs and characterize the conditions when the planted biclique can be recovered. We also empirically show that a simple heuristic can reliably recover the planted bicliques when our theory predicts that they are recoverable.
Existing methods can detect bicliques of $O(\sqrt{N})$, while it is NP-hard to find the largest such clique. Real graphs, however, are typically extremely sparse and seldom contain such large bicliques. Further, the noise can destroy parts of the planted biclique. We investigate the practical problem of how small a biclique can be and how much noise there can be such that we can still approximately correctly identify the biclique. Our derivations show that with high probability planted bicliques of size logarithmic in the network size can be detected in data following the Erdős-Rényi model and two bipartite variants of the Barabási-Albert model.

1 Introduction

In this paper we study under what conditions we can recover a planted biclique from a graph that has been distorted with a noise. We consider the general setting of matrices under $GF[q]$, where the problem can be restated as finding the planted rank-1 matrix after noise has been applied. In addition to standard additive noise, we also allow destructive noise, that is, the noise can remove edges from the planted biclique. Therefore, we consider the planted biclique *recoverable* if it is the best rank-1 approximation of the noised matrix under $GF[q]$.

As tabular data essentially forms a bipartite graph, bicliques are meaningful for a wide variety of real data. Identifying bicliques, such as through factorization and bi-clustering, is an important topic in many fields, including machine learning, data mining, and social network analysis—each of these subfields naming bicliques differently, such as 'tiles', 'clusters', or 'communities'.

One of the main current challenges is the discovery of overlapping bicliques under noise. In particular, there is need for techniques that can model interactions where bicliques overlap. For example, say in our data we have records of male

conservatives, as well as of long-haired males, but none of long-haired male conservatives. Under $GF[2]$, where every subsequent factor can be seen to XOR the corresponding entries of a binary matrix, we only need two factors—one for conservatives, one for liberals, which corresponds to intuition. Methods unable to model interaction will need 3 factors, or have high errors. In bio-informatics, there are many examples of such complex interactions, such as inhibition and excitation in gene regulation as well as in protein-protein interaction [14, 13]. By factorizing matrices in $GF[q]$ we can model arbitrary levels of interaction.

An important step towards factorizing data under $GF[q]$—i.e. discovering *all* important cliques in the data—is the reliable detection of *individual* planted bicliques. To this end, in this paper we study bounds on the dimensions of planted bicliques such that we can still reliably approximately identify these in quasi-polynomial time under different noise models. As many adversarial attacks exist that render exact solutions exponential, we focus on approximations—also, in practice, data analysts often do not require optimal results, but rather obtain good approximations in much less time.

Existing approaches aim at finding *complete* bicliques of size $O(\sqrt{N})$ [2], and it has been shown to be NP-hard to find the largest such biclique [6, 11]. While most real-world graphs have very large number of vertices, they are, however, typically only very sparsely connected. Graphs that follow the popular Barabási-Albert model, for instance, only have a constant number of vertices with degrees of $O(\sqrt{N})$. Hence, finding a clique of size $O(\sqrt{N})$ can be trivially achieved by collecting those vertices with degree at least $O(\sqrt{N})$. As such, it is an interesting open question what the smallest size of a biclique and the largest amount of *destructive* noise are such that the biclique can still be approximately correctly discovered. In particular, we study Erdős-Rényi and Barabási-Albert background distributions for bipartite graphs.

Due to lack of space, we discuss the most fine-grained details of the proof of Lemma 6, i.e. Lemma 9, and Eq. 14, in the Appendix.[4]

2 Related Work

Finding large bicliques has many applications, and hence has received a lot of attention. Most research aims at finding *exact* bicliques, that is, complete bipartite subgraphs. One way of finding these is to remove edges from the graph until what is left is a complete bipartite subgraph. Hochbaum [6] showed that minimizing the number of edges to remove is NP-hard, though she also gave a 2-approximation algorithm for the problem. Later, Peeters [11] showed that finding the largest biclique is NP-hard in general.

Despite that finding the largest biclique is NP-hard, it is possible to recover a single planted biclique [2]. In particular, if the bipartite graph contains a biclique of $N + M$ nodes and an adversary adds up to $O(NM)$ edges, the planted biclique can still be recovered using nuclear norm minimization, provided that the added

[4] `http://www.mpi-inf.mpg.de/~pmiettin/gf2bmf/appendix.pdf`

edges do not have too many neighbors. Similarly, the random process can be characterized to add edges such that the biclique can still be found.

Our problem, however, is different as we do not aim to recover exact bicliques, but approximate *quasi-bicliques* (i.e. dense but not necessarily complete bipartite subgraphs). Compared to Ames et al. [2], we allow the noise to both *add* new edges as well as to *remove* edges from the planted bicliques.

An alternative approach is to consider the problem as rank-1 matrix factorization. If we work in $GF[2]$, any method discovering binary factor matrices works, including Boolean matrix factorization algorithms [10], binary matrix factorization algorithms [15], and PROXIMUS [7].

The problem of finding dense quasi-bicliques has been approached from different directions. In graph mining, a typical goal is to find all maximal quasi-bicliques satisfying a density condition. For example, Sim et al. [12] give an algorithm to mine all maximal quasi-bicliques where each vertex is connected to all but $\varepsilon \in \mathbb{N}$ vertices in the other side (for other algorithms, see [8]). Such algorithms can be used to find the quasi-biclique that best represents the data (in terms of error), but only by exhaustively iterating over density values.

There is existing research on finding rank-k approximations of given matrices under $GF[2]$. In fact, finding the rank of a matrix under $GF[2]$ is easy. This can be seen by noting that the problem is equivalent to the rank of the biadjacency matrix of the bipartite graph under the $GF[2]$, and therefore solvable in polynomial time using standard techniques. Finding the best $GF[2]$ rank-k factorization, however, is not so easy. In the Nearest Codeword Problem, we are given an N-by-M binary data matrix A and a binary N-by-k left factor matrix B, with the task to find the right factor matrix C such that we minimize $|A - B \oplus C|$. This problem is NP-hard to approximate within any constant factor, and there exists no polynomial-time algorithm for approximation within a factor of $2^{\log^{0.8-\varepsilon} N}$, unless NP $\subseteq$ DTIME($n^{poly(\log N)}$) [3]. There does exist, however, a polynomial-time randomized approximation algorithm with $O(k/\log N)$ approximation factor [4], and a deterministic approximation algorithm with the same factor and $N^{O(\log^* N)}$ running time [1].

3 Identifying Single Bicliques

We investigate bounds on discovering a single planted biclique under a given background distribution. As models for background noise we study resp. Erdős-Rényi graphs and the scale-free Barabási-Albert model.

3.1 A Generic Strategy

In the next sections, we will consider random graph models. In each of these cases, we assume that a 'planted biclique' is combined with 'noise' generated by the random graph model, and will consider the question of how easy it is to recognize the planted biclique. In the current section we present aspects common to the derivations for these random graph models.

In this section, we will denote the dimensions of the matrices by $N \times M$. We will use $\oplus$ and $\ominus$ to denote addition and subtraction in $GF[q]$ for vectors or matrices over $GF[q]$. Further, $x \equiv y$ denotes congruency in $GF[q]$, i.e. $x \equiv y \pmod{q}$. We also adopt the common notation $[n]$ for the set of all integers from 1 until n, i.e. $[n] = \{i \in \mathbb{N} \mid 1 \leq i \leq n\}$. We will use the indicator function $I(true) = 1$ and $I(false) = 0$.

We will often use vectors (Boolean or over $GF[q]$) to select a set of rows or columns, non-zero elements indicating a selected row or column, and similarly matrices to select a set of cells. Therefore, we define for two vectors a and b of the same dimensions that $a \setminus b$ is the vector for which $(a \setminus b)_i = I(a_i \neq 0 \wedge b_i = 0)$, and similarly $a \cap b$ and $a \cup b$ as the binary vectors for which $(a \cap b)_i = I(a_i \neq 0 \wedge b_i \neq 0)$ and $(a \cup b)_i = I(a_i \neq 0 \vee b_i \neq 0)$. We define the same operations for matrices, e.g. $(A \setminus B)_{i,j} = I(A_{i,j} \neq 0 \wedge B_{i,j} = 0)$. We will denoted with $|X|$ the number of non-zero elements of a vector or matrix X. We will denote the planted clique with uv where $u \in GF[q]^{N \times 1}$ and $v \in GF[q]^{1 \times M}$. We assume u and v fixed but unknown. We will denote approximations of u and v with x and y and express the quality of the approximation using a loss function

$$L(u, v, x, y) = \max(|u - x|, |v - y|) \ . \tag{1}$$

For sparse graphs, adding a planted clique to the graph usually increases the number of nonzero elements. We therefore adopt the following notations. Let A be a random graph according to some distribution $\mathcal{M}$. Let $B = A \oplus uv$ be the addition of the planted clique defined by u and v to this matrix. Let $x \in GF[q]^{m \times 1}$ and $y \in GF[q]^{1 \times n}$ be two vectors defining a biclique xy. We define the error of x and y wrt. identifying the biclique planted in B and characterized by u and v as

$$W'(x, y) = |B \ominus xy| = |\{(i, j) \mid B_{i,j} \not\equiv x_i y_j\}| = |\{(i, j) \mid A_{i,j} \oplus u_j v_j \not\equiv x_i y_j\}| \ ,$$

that is, $W'(x, y)$ counts the matrix cells which are nonzero after removing the hypothesized biclique xy (which we expect to be minimal if $xy = uv$). Furthermore, $W(x, y) = W'(x, y) - W'(u, v)$ characterizes whether xy yields better representation of B than uv ($W(x, y) < 0$) or vice versa ($W(x, y) > 0$). Clearly, $W(u, v) = 0$.

The set of elements where the approximated biclique xy differ from the planted biclique uv is denoted by $C_{\not\equiv}(x, y)$, i.e.

$$C_{\not\equiv}(x, y) = \{(i, j) \in [N] \times [M] \mid x_i y_j \neq u_i v_j\} \ .$$

If B is the matrix received on input, i.e. the matrix resulting from adding to a random graph a planted biclique, then we will denote with $\hat{u}$ and $\hat{v}$ the vectors minimizing $W'(\hat{u}, \hat{v}) = |B \ominus \hat{u}\hat{v}|$.

For each random graph model, our aim is to show that with reasonably high probability the planted biclique uv is well approximated by the biclique $\hat{u}\hat{v}$ minimizing $W'(\hat{u}, \hat{v})$. We will first show that the probability that $W(x, y) < 0$, with $xy \neq uv$, decreases exponentially with $|C_{\not\equiv}(x, y)|$. Then, by the following

lemma from such result we can derive that maximizing the objective function on the input will yield a good approximation of (u, v).

Lemma 1. *Let $\mathcal{M}$ be a distribution over $GF[q]^{N\times M}$, i.e. $N \times M$ matrices over $GF[q]$. Assume that there is an integer ζ and a constant c such that for any fixed $u \in GF[q]^{N\times 1}$ and $v \in GF[q]^{1\times M}$ with $|u| \geq \zeta$ and $|v| \geq \zeta$, with probability at least $1-\delta_1$ for a matrix A randomly drawn from $\mathcal{M}$, it holds for all $x \in GF[q]^{N\times 1}$ and $y \in GF[q]^{1\times M}$ that*

$$P(W(x, y) \leq 0) \leq \exp(-|C_{\not\equiv}(x, y)|c) \; . \tag{2}$$

Then, for all $\epsilon > 0$, u and v such that $|u| \geq \zeta$ and $|v| \geq \zeta$,

$$P_{A\sim\mathcal{M}}(L(u, v, \hat{u}, \hat{v}) \leq \epsilon) \geq 1 - \delta_1 - \delta_2$$

where $(\hat{u}, \hat{v}) = \arg\min_{(x,y)} |A \oplus uv \ominus xy|$ and

$$\delta_2 = T(\epsilon, |u|, |v|, |u|, |v|)T(\epsilon, N, M, |u|, |v|) \tag{3}$$

where

$$T(\epsilon, a, b, c, d) = \frac{\exp\left(\epsilon\left(\log\left(a+1\right) + \log\left(b+1\right) - \min\left(c, d\right)\right)c_{p,q}\right)}{1 - \exp\left(\left(\log(a+1) + \log(b+1) - \min(c, d)\right)c_{p,q}\right)} \; .$$

Proof. Equation (2) is a bound on the probability that for a given x and y, $W(x, y) < 0$. Several choices for x and y are possible. We will bound the probability that $\exists x, y : W(x, y) < 0$ by

$$P(\exists x, y : W(x, y) < 0) \leq \sum_{x,y} P(W(x, y) < 0)$$

For a given x and y we now bound $|C_{\not\equiv}(x, y)|$. First, we define

$$C_{uv\setminus xy} = \{(i, j) \mid (uv \setminus xy)_{i,j} = 1\} \tag{4}$$

and

$$C_{xy\setminus uv} = \{(i, j) \mid (xy \setminus uv)_{i,j} = 1\} \; , \tag{5}$$

such that $|C_{\not\equiv}(x, y)| = \left|C_{uv\setminus xy}\right| + \left|C_{xy\setminus uv}\right|$.

As we have both $\left|C_{uv\setminus xy}\right| \geq |u \setminus x|\,|v|$ and $\left|C_{uv\setminus xy}\right| \geq |v \setminus y|\,|u|$ it follows that

$$\left|C_{uv\setminus xy}(x, y)\right| \geq \max(|v \setminus y|\,, |u \setminus x|) \min(|u|\,, |v|) \; . \tag{6}$$

There are $\sum_{i=1}^{t} \binom{|u|}{i} \leq (|u| + 1)^t$ ways for choosing at most t rows out of the $|u|$ nonzero rows of u. Similarly, we have $\sum_{i=1}^{t} \binom{|v|}{i} \leq (|v| + 1)^t$ ways to choose at most t columns out of the $|v|$ nonzero columns of v. Hence, the number of ways to choose $u \setminus x$ and $v \setminus y$ such that both $|u \setminus x| \leq t$ and $|v \setminus y| \leq t$ hold is bounded by $(|u| + 1)^t(|v| + 1)^t$. Now, let us use $C^{(t)}_{uv\setminus.} = \{(x, y) \mid \max(|u \setminus x|\,, |v \setminus y|) = t\}$

for the set of (x, y)'s such that for each the largest intersection with the rows or columns of (u, v) is t elements. We can now write

$$\begin{aligned}
&\sum_{t=s}^{\max(|u|,|v|)} \sum_{(x,y)\in C^{(s)}_{uv\setminus\cdot}} \exp\left(-\left|C_{uv\setminus xy}\right| c\right) \\
&\leq \sum_{t=s}^{\max(|u|,|v|)} (|u|+1)^t (|v|+1)^t \exp\left(-t\min(|u|,|v|)c\right) \\
&\leq \sum_{t=s}^{\max(|u|,|v|)} \exp\left(t(\log(|u|+1)+\log(|v|+1)-\min(|u|,|v|)c)\right) \\
&= \frac{\exp\left(s(\log(|u|+1)+\log(|v|+1)-\min(|u|,|v|)c)\right)}{1-\exp\left((\log(|u|+1)+\log(|v|+1)-\min(|u|,|v|)c)\right)} \\
&= T(s,|u|,|v|,|u|,|v|)
\end{aligned}$$

Here, in the one-but last step we use the fact that $\sum_{i=0}^{\infty} x^i = 1/(1-x)$. Similarly, we define $C^{(t)}_{\cdot\setminus uv} = \{(x, y) \mid \max(|x \setminus u|, |y \setminus v|) = t\}$ by which we have

$$\sum_{t=s}^{\max(N,M)} \sum_{(x,y)\in C^{(s)}_{\cdot\setminus uv}} \exp\left(-\left|C_{xy\setminus uv}\right| c\right) \leq T(s,N,M,|x|,|y|)$$

Note that as u and v are fixed, we have only sets of size $|u|$ and $|v|$ to choose $u \setminus x$ and $v \setminus y$ from. Here however, x and y can be chosen from $N - |u|$ remaining rows and $M - |v|$ remaining columns, resp. Still, however, when $\log(N+1) + \log(M+1) < \min(|u|, |v|)$, $T(s, N, M, |u|, |v|) \leq 1$.

Finally, this allows us to combine these two inequalities into

$$P(W(x,y) < 0 \mid \max(|u-x|,|v-y|) \geq \epsilon) \leq T(\epsilon,|u|,|v|,|u|,|v|)T(\epsilon,N,M,|u|,|v|) \ .$$

This proves the lemma. □

According to the above, if we know the dimensions of a biclique, we have a clear bound on its detectability. Further, it follows that when $|u| \ll |v|$ or $|x| \ll |y|$ the problem becomes much harder. This follows intuition as under an independence assumption large square blocks are much less probable than thin bicliques—as these could just as well be the result of few very high degree nodes.

3.2 Erdős-Rényi

The Erdős-Rényi (ER) model is one of the most well-studied models for graph generation. The general idea is that every edge is equally probable, regardless of other edges in the graph. That is, graphs of the same number of nodes and same total number of edges are all equally likely. For the case of factorizing data under $GF[q]$ with noise distributed according to ER, we have the following definition.

Definition 1. *With $\mathcal{M}^{ER}(p, q, N \times M)$ we will denote the model of sparse random matrices in $GF[q]^{N\times M}$ according to the Erdős-Rényi model, in particular if $A \in \mathcal{M}^{ER}(p, q, N \times M)$, for each $(i, j) \in [N] \times [M]$, A_{ij} is zero with probability $1 - p$ and non-zero with probability p. Non-zero elements are chosen randomly from $GF[q]$, i.e. each non-zero element of $GF[q]$ has probability $1/(q - 1)$.*

We will now show that the probability that *some* biclique yields lower error (i.e. residual) than the planted biclique uv decays exponentially with the difference between that biclique and the planted one.

Lemma 2. *Let $p < 1/2$. Let N, M and q be integers, $u, x \in GF[q]^{N\times 1}$ and $v, y \in GF[q]^{1\times M}$. Then, there is a constant $c_{p,q}$ depending on p and q such that*

$$P_{A\sim\mathcal{M}^{ER}(p,q,N\times M)}(W(x, y) < 0) \leq \exp(c_{p,q}|C_{\not\equiv}|) \ .$$

Proof. Let A be randomly drawn from $\mathcal{M}^{ER}(p, q, N \times M)$. As above, let $B = A \oplus uv$ be the matrix obtained by adding to A the biclique uv.

Let $C_{i,j} = u_i v_j \ominus x_i y_j$ be the difference between uv and xy, and let

$$D_{i,j} = I(A_{i,j} \oplus C_{i,j} \not\equiv 0) - I(A_{i,j} \not\equiv 0) \ ,$$

where $I(\cdot)$ is the indicator function. Now, $W(x, y) = \sum_{i,j} D_{i,j}$. If $C_{ij} \equiv 0$, then $D_{ij} = 0$, so let $C_{\not\equiv}(x, y) = \{(i, j) \in [n] \times [m] \mid C_{ij} \not\equiv 0\}$ such that we have $W(x, y) = \sum_{(i,j)\in C_{\not\equiv}} D_{i,j}$.

Following Section 3.1, we bound the probability that xy gives better representation of B than uv. That is, we bound $P(W(x, y) < 0)$. To that end, we define, for $z \in \{-1, 0, +1\}$, $W_z(x, y) = \{(i, j) \in C_{\not\equiv} \mid D_{i,j} = z\}$, so we have

$$W(x, y) = |W_{+1}(x, y)| - |W_{-1}(x, y)| \ .$$

The three sets $W_z(x, y)$, $z \in \{-1, 0, 1\}$, partition the set $C_{\not\equiv}(x, y)$. Let $X_{i,j}$ be a random variable defined as $X_{i,j} = I\big((i, j) \in W_{-1}(x, y) \cup W_0(x, y)\big)$, so that $P(X_{i,j} = 1) = 1 - P\big((i, j) \in W_1(x, y)\big)$ for all $(i, j) \in C_{\not\equiv}(x, y)$, and $\sum_{i,j} X_{i,j} = |W_{-1}(x, y)| + |W_0(x, y)|$. We have for all $(i, j) \in C_{\not\equiv}(x, y)$ that

$$P\big((i, j) \in W_{-1}(x, y)\big) = \frac{p}{q - 1} \quad \text{and} \quad P\big((i, j) \in W_0(x, y)\big) = \frac{p(q - 2)}{q - 1} \ ,$$

where on the second equation we use the fact that $C_{i,j} \not\equiv 0$. Therefore

$$P(X_{i,j} = 1) = P\big((i, j) \in W_{-1}(x, y)\big) + P\big((i, j) \in W_0(x, y)\big) = \frac{p}{q - 1} + \frac{p(q - 2)}{q - 1} = p$$

for $(i, j) \in C_{\not\equiv}(x, y)$. Then, due to Chernoff's inequality, for any $\epsilon > 0$, we have

$$P\left(\frac{|W_{-1}(x, y)| + |W_0(x, y)|}{|C_{\not\equiv}(x, y)|} \geq p + \epsilon\right) \leq \exp\left[-|C_{\not\equiv}(x, y)| D_{KL}(p + \epsilon \parallel p)\right] \ ,$$

where

$$D_{KL}(p+\epsilon \parallel p) = (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (1-p-\epsilon)\log\left(\frac{1-p-\epsilon}{1-p}\right) . \qquad (7)$$

In order for $W(x,y) < 0$ to be possible, we need to have $|W_{-1}(x,y)| + |W_0(x,y)| \geq |C_{\neq}(x,y)| - |W_1(x,y)|$. Hence,

$$P\left(\frac{|W_{-1}(x,y)| + |W_0(x,y)|}{|C_{\not\equiv}(x,y)|} \geq 1 - \frac{|W_1(x,y)|}{|C_{\not\equiv}(x,y)|}\right)$$
$$\leq \exp\left(-|C_{\not\equiv}(x,y)| D_{KL}\left(1 - \frac{|W_1(x,y)|}{|C_{\not\equiv}(x,y)|} \,\middle\|\, p\right)\right) .$$

To have a chance to have $W(x,y) < 0$, we need at least $|W_{-1}(x,y)| + |W_0(x,y)| \geq |C_{\neq}(x,y)|/2$. Therefore, let $\epsilon = \frac{1}{2} - p$ to get

$$P\left(\frac{|W_{-1}(x,y)| + |W_0(x,y)|}{|C_{\not\equiv}(x,y)|} \geq \frac{1}{2}\right) \leq \exp\big(-|C_{\neq}(x,y)| D_{KL}(1/2 \parallel p)\big) , \qquad (8)$$

where

$$D_{KL}\left(\frac{1}{2} \,\middle\|\, p\right) = \frac{1}{2}\log\left(\frac{1}{2p}\right) + \frac{1}{2}\log\left(\frac{1}{2(1-p)}\right) .$$

This already gives us a bound on $P(W(x,y) < 0)$:

$$P(W(x,y) < 0) \leq P(|W_1(x,y)| < |C_{\not\equiv}(x,y)|/2) \leq \exp\big(-|C_{\not\equiv}(x,y)| D_{KL}(1/2 \parallel p)\big) .$$

In case $q > 2$, we can do better as we expect more (i,j)'s to land in $W_0(x,y)$ instead of $W_{-1}(x,y)$.

Suppose now $q > 2$. For a fixed value of $|W_{-1}(x,y)| + |W_0(x,y)|$, using $W_{-1}(x,y) \leq |C_{\neq}(x,y)|/2$ and Chernoff's inequality, we obtain

$$P\left(\frac{|W_{-1}(x,y)|}{|C_{\neq}(x,y)| - |W_1(x,y)|} > \frac{1}{q-1} + \left(\frac{|W_1(x,y)|}{|C_{\neq}(x,y)| - |W_1(x,y)|} - \frac{1}{q-1}\right)\right)$$
$$\leq \exp\left(-(|C_{\neq}(x,y)| - |W_1(x,y)|) D_{KL}\left(\frac{|W_1(x,y)|}{|C_{\neq}(x,y)| - |W_1(x,y)|} \,\middle\|\, \frac{1}{q-1}\right)\right)$$
$$\leq \exp\left(-\frac{|C_{\neq}(x,y)|}{2} D_{KL}\left(\frac{|W_1(x,y)|}{|C_{\neq}(x,y)| - |W_1(x,y)|} \,\middle\|\, \frac{1}{q-1}\right)\right) .$$

The above equations imply that there exists some constant $c_{p,q}$ depending on p and q such that

$$P(W(x,y) < 0) \leq \exp(-|C_{\neq}(x,y)| c_{p,q}) . \qquad (9)$$

This proves the lemma. □

The above lemma can be combined with Lemma 1 (substituting ζ with $\log(NM)$ and c with $c_{p,q}$) to show that one can retrieve a planted clique with

high confidence and small error (according to the trade-off given by Equation 3), and in time quasipolynomial in N and M.

It should be noted that for clarity of explanation and space limitations we keep our derivation simple, but a constant factor can be gained by calculating more precise expressions for $c_{p,q}$ and performing less rough estimations in Lemma 1. Moreover, Lemma 1 does not properly take the value of q into account and doing so would yield another q-dependent factor.

3.3 Graphs Constructed by the Barabási-Albert Process

Next, we consider the background noise distributed according to the well-known Barabási-Albert (BA) model, of which the main intuition is also known as 'preferential attachment'. Nodes are added one at a time, and while edges are still selected independently, their probability depends on the degree of the target node. Instead of the ER model's Gaussian degree distribution, for BA we see a powerlaw—as we see for many real-world graphs [5].

For simplicity of the derivations, below we will assume that $q = 2$. If $q > 2$, a similar but more involved derivation is possible.

Definition 2 (bipartite Barabási-Albert graph). *Let $G^{(0)}$ be a small graph on a vertex set $V^{(0)} = V^{(0)}_{row} \cup V^{(0)}_{col}$ consisting of row vertices $V^{(0)}_{row}$ and column vertices $V^{(0)}_{col}$. Let N and M be integers. A bipartite Barabási-Albert $N \times M$ graph is generated from seed $G^{(0)}$ with density parameter s by following Algorithm 1. We denote the obtained probability distribution over $N \times M$ adjacency matrices with $\mathcal{M}^{BA\text{-}gen}(s, N \times M)$.*

Lemma 3. *Let $G = (V, E)$ be a bipartite Barabási-Albert $N \times M$ graph generated according to Definition 2, let V_{row} be its row vertices and V_{col} its column vertices. Let $X_{row} \subseteq V_{row}$ and $X_{col} \subseteq V_{col}$. Then,*

$$|E \cap X_{row} \times X_{col}| \leq s(|X_{col}| + |X_{row}| - (s+1)/2) \ .$$

Proof. The proof is straightforward from Definition 2: each vertex connects with s vertices when added, but can only connect to vertices added before. □

Notice that analogue one can prove that a 'normal' (non-bipartite) Barabási-Albert graph can not contain an $(s+2)$-clique. The probability distribution over graphs induced by the Barabási-Albert generative process is rather hard to analyse in detail, but we can provide the following non-probabilistic result:

Lemma 4. *Let A be drawn from $\mathcal{M}^{BA\text{-}gen}(s, N \times M)$. Let $B = A \oplus uv$ for some fixed u and v with $|u| > 4s$ and $|v| > 4s$. Then, $|B \cap uv| > |u||v|/2$.*

Proof. We know from Lemma 3 that in the area covered by uv in B, uv made all cells except at most $s(|u| + |v| - (s+1)/2)$ nonzero: $|B \cap uv| \geq |u||v| - s|u| - s|v| + s(s+1)/2$. If $|u| \geq 4s$ and $|v| \geq 4s$ we have $|u||v|/2 - s|u| - s|v| \geq 0$ from which $|B \cap uv| > |u||v|/2$ follows. □

Algorithm 1 Generating a bipartite Barabási-Albert graph

Require: density parameter s; seed $G^{(0)}$; $M, N \in \mathbb{N}$
Ensure: an $N \times M$ bipartite Barabási-Albert graph G sampled from $\mathcal{M}^{BA\text{-}gen}$.

for $i = 1 \dots NM$ **do**
  $V_{row}^{(i)} \leftarrow V_{row}^{(i-1)}$; $V_{col}^{(i)} \leftarrow V_{col}^{(i-1)}$; $E^{(i)} \leftarrow E^{(i-1)}$
  if $|V_{row}^{(i)}|M < i$ **then**
    $v_{new} \leftarrow NewVertex()$
    $V_{row}^{(i)} \leftarrow V_{row}^{(i)} \cup \{v_{new}\}$
    Select a set A of s vertices from $V_{col}^{(i-1)}$ with probability proportional to their degree in $G^{(i-1)}$.
    $E^{(i)} \leftarrow E^{(i)} \cup (\{v_{new}\} \times A)$
  if $|V_{col}^{(i)}|N < i$ **then**
    $v_{new} \leftarrow NewVertex()$
    $V_{col}^{(i)} \leftarrow V_{col}^{(i)} \cup \{v_{new}\}$
    Select a set A of s vertices from $V_{row}^{(i-1)}$ with probability proportional to their degree in $G^{(i-1)}$.
    $E^{(i)} \leftarrow E^{(i)} \cup (A \times \{v_{new}\})$
  $V^{(i)} \leftarrow V_{row}^{(i)} \cup V_{col}^{(i)}$; $G^{(i)} \leftarrow (V^{(i)}, E^{(i)})$

Lemma 5. *Let A be drawn from $\mathcal{M}^{BA\text{-}gen}(s, N \times M)$. Let $B = A \oplus uv$ for some fixed u and v. Then, for any x and y such that $|C_{\not\equiv}(x, y)| > 2s|x| + 2s|y| - s(s+1) + |u||v|$, it holds that $|B \cap xy| \leq |x||y|/2$.*

Proof. We know from Lemma 3 that in the area covered by xy, at most $s(|x| + |y| - (s+1)/2)$ cells are nonzero in $A \cap xy$. For $B \cap xy$, at most the area $(|x||y| + |u||v| - |C_{\not\equiv}|)/2$ from overlap between xy and uv can be added. We get $|B \cap xy| \leq s|x| + s|y| - s(s+1)/2 + (|x||y| + |u||v| - |C_{\not\equiv}|)/2$. From $2s|x| + 2s|y| - s(s+1) + |u||v| < |C_{\not\equiv}|$ we can derive that $|B \cap xy| \leq |x||y|/2$. □

Hence, to detect a planted biclique in Barabási-Albert data, one only needs to search for bicliques of size $4s$ and expand greedily.

3.4 Graphs with Barabási-Albert Degree Distribution

Here we consider random graphs with the same degree distribution as Barabási-Albert model but without following the generative process, of which we showed in the previous section that it prohibits the creation of large bicliques.

For simplicity, w.l.o.g. we assume $N = M$.

Definition 3. *With $\mathcal{M}^{BA\text{-}deg}(s, q, N \times M)$ we will denote the model of sparse random matrices in $GF[q]^{N \times M}$ according to the Barabási-Albert degree distribution, in particular if $A \in \mathcal{M}^{BA\text{-}deg}(s, q, N \times M)$, it is the result of the following random construction procedure:*

- *Consider the probability distribution P_{deg} over the set $\{s, s+1, \dots, N\}$ such that $P_{deg}(i) = i^{-3}/Z$ with $Z = \sum_{j=s}^{N-1} j^{-3}$*

– *For all $i \in [N]$, choose d_i^{row} according to distribution P_{deg}. For all $j \in [N]$, choose d_j^{col} according to P_{deg}. Repeat this step until $\sum_{i=1}^{N} d_i^{row} = \sum_{j=1}^{M} d_j^{col}$.*
– *Draw X uniformly from the set of all matrices of $GF[q]^{N\times M}$ such that for all $i \in [N]$, the number of nonzero elements of row i equals d_i^{row} and for all $j \in [M]$, the number of nonzero elements of column i equals d_j^{row}.*

In order to say something on the discernibility of cliques we need access to the connectivity within rows and columns in the form of degree lists.

Definition 4. *For an adjacency matrix $A \in GF[q]^{N\times M}$, let $f^{row(A)} \in \mathbb{Z}^N$ and $f^{col(A)} \in \mathbb{Z}^M$ such that*

$$f_i^{row(A)} = \sum_{j=1}^{N} I(A_{i,j} \neq 0) \quad , \quad \text{and} \qquad f_j^{col(A)} = \sum_{i=1}^{M} I(A_{i,j} \neq 0) \; .$$

It is well-known that in Barabási-Albert graphs, the expected frequency of vertices with degree k is proportional to k^{-3}. Therefore, for sufficiently large N we can estimate the number of rows of with at least $c\sqrt{N}$ non-zero elements by

$$N \frac{\sum_{k=cN^{1/2}}^{N} k^{-3}}{\sum_{k=s}^{N}} \approx N \frac{\int_{k=cN^{1/2}}^{\infty} k^{-3}\mathrm{d}k}{\int_{k=s}^{\infty} k^{-3}\mathrm{d}k} = N \frac{\left(cN^{1/2}\right)^{-2}/2}{s^{-2}/2} = \frac{s^2}{c^2}$$

which is a constant, not depending on N. The same holds for columns.

Lemma 6. *Let s be an integer. Let N and q be integers, $u, x \in GF[q]^{N\times 1}$ and $v, y \in GF[q]^{1\times M}$ with $\log(N) \ll |u|$ and $\log(N) \ll |v|$. Then, there is a constant c_q^{BA} depending on q and δ_1 such that with probability at least $1 - \delta_1$*

$$P_{A\sim\mathcal{M}^{BA\text{-}deg}(s,q,N\times N)}(W(x,y) < 0) \leq \exp(c_q^{BA}|C_{\neq}|)$$

Proof. We will use notations similar to those used for the Erdős-Rényi case:

$$\begin{aligned} C_{i,j} &= u_i v_j \ominus x_i y_j \\ D_{i,j} &= I(A_{i,j} \oplus C_{i,j} \neq 0) - I(A_{i,j} \neq 0) \\ W'(x) &= |B \ominus x^\top x| = |A \oplus uv \ominus xy| \\ W(x) &= W'(x) - W'(u) = \sum_{i,j} D_{i,j} \end{aligned}$$

We have $W(u) = 0$. Given a fixed degree list pair (f^{row}, f^{col}), we have (approximately, for sufficiently large N)

$$\mu_{i,j} = \mathbb{E}[D_{i,j}] = 1 - p_{i,j} q/(q-1)$$

where $p_{i,j} = f_i^{row} f_j^{col} \left(\sum_{l=1}^{N} f_l^{row}\right)^{-1} \left(\sum_{l=1}^{N} f_l^{row}\right)^{-1}$, and

$$\begin{aligned} \sigma_{i,j}^2 &= \mathbb{E}[(I(x \oplus C_{i,j} \neq 0) - I(x \neq 0) - \mu_{i,j})^2] \\ &\leq p_{i,j}\,(1 - p_{i,j})\,\frac{q+2}{q-1} \\ &\leq p_{i,j}\frac{q+2}{q-1}. \end{aligned}$$

We again define $C_{\neq}(x) = \{(i,j) \in [n] \times [m] \mid u_i v_j \neq x_i y_j\}$ and for $v \in \{-1, 0, +1\}$ we have $W_v(x) = \{(i,j) \mid D_{i,j} = v\}$. Moreover, let

$$\mu = \mathbb{E}[W(x)] = \sum_{(i,j) \in C_{\neq}(x)} \mu_{i,j} \ ,$$

and

$$\sigma^2 = \sum_{(i,j) \in C_{\neq}(x)} \sigma_{i,j}^2 \ .$$

From (14) we can see[5] that

$$\mathbb{E}[\mu] \leq |C_{\neq}| 2(s - 1/2)^2 / N(s-1) \ .$$

By applying Chernoff's inequality, we then arrive at

$$P(W(x,y) < 0) = P\left(\mu - W(x,y) > \mu\right) \leq \exp\left(-\frac{\mu^2}{2(\sigma^2 + |C_{\neq}|)}\right) \ . \qquad (10)$$

Let $t = \lceil \Gamma^{-1}(2/\delta_1) \rceil - 1$ be such that $1/t! \leq \delta_1/2$. From Lemma 9 we know[5] that with probability $1 - \delta_1/2$ there is at most one i s.t. $f_i^{row} \geq s/\sqrt{N}$ and with probability $1 - \delta_1/2$ there is at most one j such that $f_j^{col} \geq s/\sqrt{N}$. Then, with probability at least $1 - \delta_1$,

$$\begin{aligned}
\sum_{(i,j) \in C_{\neq}} p_{i,j} &\leq \sum_{i \in u \setminus x} \sum_{j \in v} p_{i,j} + \sum_{i \in x \setminus u} \sum_{j \in y} p_{i,j} + \sum_{j \in v \setminus y} \sum_{i \in u} p_{i,j} + \sum_{j \in y \setminus v} \sum_{i \in x} p_{i,j} \\
&\leq R(|u \setminus x|, |v|, t) + R(|x \setminus u|, |y|, t) + R(|v \setminus y|, |u|, t) + R(|y \setminus v|, |x|, t)
\end{aligned}$$

with

$$R(a,b,t) = R'(\min(a,t), \min(b,t), \max(a-t,0), \max(b-t,0))$$

and

$$R'(a_H, b_H, a_L, b_L) = R_1(a_H, a_L) R_1(b_H, b_L)$$

with $R_1(H, L) = H + \frac{s}{\sqrt{N}} L$. For $b \geq \eta \geq t$, $R(a,b,t) = R^*(\min(a,t), t, \max(a - t, 0), b - t)$ with

$$\begin{aligned}
R^*(a_H, t, c_L, b-t) &= R_1(a_H, a_L) \left(t + \frac{s}{\sqrt{N}}(b-t)\right) \\
&\leq R_1(a_H, a_L) b \left(\frac{s}{\sqrt{N}} + (1 - \frac{s}{\sqrt{N}}) \frac{t}{b}\right) \\
&\leq R_1(a_H, a_L) b R_C
\end{aligned}$$

[5] See appendix: `http://www.mpi-inf.mpg.de/~pmiettin/gf2bmf/appendix.pdf`

with $R_C = \left(\frac{s}{\sqrt{N}} + \frac{t}{\eta}\right)$. Hence, as we assume $|u| \geq \eta$, $|v| \geq \eta$, $|x| \geq \eta$ and $|y| \geq \eta$, with probability at least $1 - \delta_1/2$ we have for every u, v, x and y

$$\sum_{(i,j)\in C_{\neq}} p_{i,j} \leq (R_1(|u \setminus x|)|v| + R_1(|v \setminus y|)|u| + R_1(|x \setminus u|)|y| + R_1(|y \setminus v|)|x|)R_C$$

As

$$|C_{\neq}| \leq |u \setminus x||v| + |x \setminus u||y| + |v \setminus y||u| + |y \setminus v||x| \leq 2|C_{\neq}|$$

and $R_1(a) \leq a$,

$$\sum_{(i,j)\in C_{\neq}} p_{i,j} \leq 2R_C|C_{\neq}| \tag{11}$$

and $\mu \geq |C_{\neq}|(1 - 2R_C)$. Combining Eq. (11) with the definition of σ^2 gives that $\sigma^2 \leq 2R_C|C_{\neq}|\frac{q+2}{q-1}$. Combining this with Eq. (10) results in

$$P(W(x,y) < 0) \leq \exp\left(-\frac{|C_{\neq}|(1-2R_C)^2}{2(1+2R_C(q+2)/(q-1))}\right) . \tag{12}$$

Setting $c_q^{BA} = (1-2R_C)^2/(2(1+2R_C(q+2)/(q-1)))$, we get

$$P(W(x,y) < 0) \leq \exp(-|C_{\neq}|c_q^{BA}) .$$

This proves the lemma. □

In practice, this means that as long as noise levels are not overly high, i.e. $s \ll N$, and the dimensions of the planted biclique are not overly small, i.e. $\log N \ll |u| \ll N^{-1/2}$, we can reliably identify the planted bicluster. We note that these assumptions are quite realistic under the *BA* model. More to the point, we find that a biclique uv is still discernible if $|u| > \log N$ and $|v| > \log N$.

4 Algorithm

In this section we describe a simple heuristic algorithm to recover the planted bicliques under $GF[2]$. We have already shown that the best biclique is the planted one (with high probability), and therefore we 'only' need to find the best biclique. Unfortunately, this problem is NP-hard (as finding the largest exact biclique is NP-hard [11]). Luckily, it seems that in practice a simple heuristic—which we present below—is able to recover the planted biclique very well.

Recall, that finding the best biclique in $GF[2]$ is equivalent to finding rank-1 binary matrix factorization that minimizes the Hamming distance. To compute find it, we used the `Asso` algorithm [10]. We note that our aim is not to perform a comparative study of different algorithms but to show that we can achieve the predicted performance using relatively simple, non-exhaustive method.

The crux of `Asso` is the use of pairwise *association confidences* for finding candidate column factors. Consider the N-by-M input matrix B. `Asso` will compute the association confidence between each row of B. The association confidence

from row b_i to row b_j is defined as $conf(b_i \to b_j) = (\sum_{k=1}^{m} b_{ik}b_{jk})/(\sum_{k=1}^{m} b_i k)$ and can be interpret as the (empirical) conditional probability that $b_{jk} = 1$ given that $b_{ik} = 1$. The intuition is that if rows i and j belong in the planted biclique, they should have relatively high confidence (each column of the biclique is 1 in both rows and each column not in the biclique is 0, save the effects of noise) and otherwise the confidence should be low. The `Asso` algorithm builds an N-by-N matrix D where $d_{ij} = conf(b_i \to b_j)$.

Matrix D is then round to binary matrix $\tilde{D}$ from some threshold τ. The *columns* of the binary matrix $\tilde{D}$ constitute the candidate columns of the biclique. The algorithm will then construct the optimum row for each of these columns, and select the best row-column pair (x, y) (measured by $A \ominus xy$). Computing the association accuracy takes $O(N^2M)$ time (where we assume $N \leq M$), there are N candidate vectors, and testing each of them takes $O(NM)$ time, giving the overall complexity for fixed τ as $O(N^2M)$.

The last detail is how to select the rounding threshold τ. We can try every value of D, but that adds N^2 factor to the second term in the time complexity. To avoid quadratic running times, we opt to evaluate a fixed set of thresholds.

5 Experiments

In this section we experimentally evaluate the above theory. As we need to measure against a ground truth, we will experiment on synthetic data. For practical reasons we focus on $GF[2]$: in order to evaluate our bounds we require an algorithm to extract candidate bicliques from the data. While no polynomial time (approximate) biclique discovery algorithm exists for $GF[q]$ in general, we have seen in Section 4 that `Asso` [10] is relatively easy to adapt to $GF[2]$.

We implemented the $GF[2]$ version of `Asso` in Matlab/C, and provide for research purposes the source code together with the generators for bipartite Erdős-Rényi and Barabási-Albert graphs.[6]

As synthetic data, we consider square matrices over $GF[q]$ of dimensions $N = M = 1000$, to which we add noise. We consider the *ER* model as discussed in Section 3.2, and the probabilistic *BA* model from Section 3.4. We focus on this *BA* variant, as by allowing larger bicliques to be generated it corresponds to the hardest problem setting. In this matrix A we plant a square biclique uv (i.e. $|u| = |v|$), such that we obtain $B = A \oplus uv$. We run `Asso` on B for all values of $\tau \in \{0, 0.01, \ldots, 1.0\}$, and select the best candidate biclique. We report the $L(u, v, x, y)$ error between this candidate and the planted biclique.

We evaluate performance for different noise ratios, defined as $|A|/(NM)$, and for different biclique sizes. Figure 1 shows the results averaged over five independent runs. For Erdős-Rényi, we see that bicliques of 10×10 are easily discerned even for high noise ratios, despite that `Asso` is uninformed of the shape or size of the biclique. In accordance with theory, bicliques of 5×5 can still be detected reliably for lower noise levels, while those of 3×3 only barely so.

[6] `http://www.mpi-inf.mpg.de/~pmiettin/gf2bmf/`

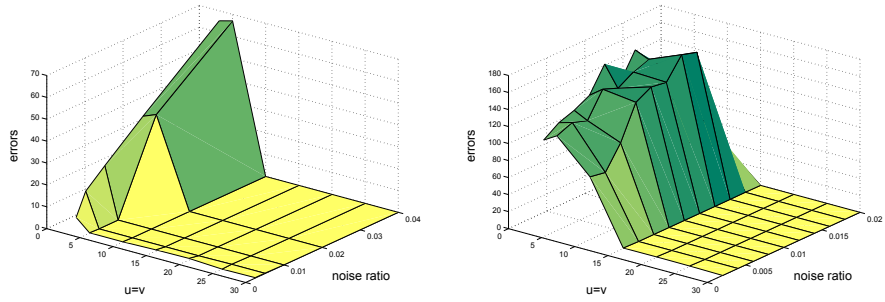

Fig. 1. Performance of `Asso` of finding a square planted biclique of dimensions ($|u| = |v|$), under different ratios of noise, resp. generated by the Erdős-Rényi model (left), and the Barabási-Albert model (right).

For Barabási-Albert (Figure 1, right), we also find that practice corresponds to theory. Per Section 3.4 bicliques need too have $|u| \gg \log N$ to be discernible; indeed, we here observe that for $|u| \geq 15$ the clique is discovered without error, while for smaller sized clusters error first rises and then stabilizes. By the scale-free property of the graph the noise ratio does not influence detection much.

6 Conclusion

We consider the problem of finding planted bicliques in random matrices over $GF[q]$. More in particular, we investigated the size of the smallest biclique such that we can still approximately correctly identify it as the best rank-1 approximation against a background of either Erdős-Rényi or Barabási-Albert distributed noise. Whereas existing methods can only detect bicliques of $O(\sqrt{N})$ under non-destructive noise, we show that bicliques of resp. $n \geq 3$, and $n \gg \log N$ are discernible even under destructive noise. Experiments with the `Asso` algorithm confirm that we can identify planted bicliques under $GF[2]$ with high precision.

While the ER and BA models capture important graphs properties, they are stark simplifications. Studying whether similar derivations are possible for more realistic models, such as Kronecker Delta [9] will make for engaging future work.

The key extension of this work will be the development of theory for matrix factorization in $GF[q]$, by which we will be able to identify and analyse interactions between bicliques such as found in proteomics data [13].

Acknowledgements

Jan Ramon is supported by ERC Starting Grant 240186 "MiGraNT". Jilles Vreeken is supported by a Post-doctoral Fellowship of the Research Foundation – Flanders (FWO).

References

1. N. Alon, R. Panigrahy, and S. Yekhanin. Deterministic Approximation Algorithms for the Nearest Codeword Problem. In *APPROX RANDOM '09*, pages 339–351, 2009.
2. B. P. W. Ames and S. A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Math. Program. B*, 129(1):69–89, May 2011.
3. S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The Hardness of Approximate Optima in Lattices, Codes, and Systems of Linear Equations. In *FOCS '93*, pages 724–733, 1993.
4. P. Berman and M. Karpinski. Approximating minimum unsatisfiability of linear equations. In *SODA '02*, January 2002.
5. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.
6. D. S. Hochbaum. Approximating clique and biclique problems. *J. Algorithm*, 29(1):174–200, October 1998.
7. M. Koyutürk and A. Grama. PROXIMUS: a framework for analyzing very high dimensional discrete-attributed datasets. In *KDD '03*, pages 147–156, 2003.
8. V. E. Lee, N. Ruan, R. Jin, and C. Aggarwal. A Survey of Algorithms for Dense Subgraph Discovery. In C. Aggarwal and H. Wang, editors, *Managing and Mining Graph Data*, pages 303–336. Springer, New York, January 2010.
9. J. Leskovec, D. Chakrabarti, J. M. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11:985–1042, 2010.
10. P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. *IEEE TKDE*, 20(10):1348–1362, 2008.
11. R. Peeters. The maximum edge biclique problem is NP-complete. *Discrete Appl. Math.*, 131(3):651–654, 2003.
12. K. Sim, J. Li, V. Gopalkrishnan, and G. Liu. Mining maximal quasi-bicliques: Novel algorithm and applications in the stock market and protein networks. *Statistical Analysis and Data Mining*, 2(4):255–273, November 2009.
13. M. E. Wall. Structure–function relations are subtle in genetic regulatory networks. *Math. Bioscience*, 231(1):61–68, 2011.
14. E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcriptionregulation and proteinprotein interaction. *PNAS*, 101(16):5934–5939, 2004.
15. Z.-Y. Zhang, T. Li, C. Ding, X.-W. Ren, and X.-S. Zhang. Binary matrix factorization for analyzing gene expression data. *Data Min. Knowl. Disc.*, 20(1):28–52, 2010.