Algorithmica manuscript No.
(will be inserted by the editor)

Using Patterns to Form Homogeneous Teams

**Robert Bredereck*, Thomas Köhler*,
André Nichterlein, Rolf Niedermeier,
Geevarghese Philip****

Abstract Homogeneous team formation is the task of grouping individuals into teams, each of which consists of members who fulfill the same set of prespecified properties. In this theoretical work, we propose, motivate, and analyze a combinatorial model where, given a matrix over a finite alphabet whose rows correspond to individuals and columns correspond to attributes of individuals, the user specifies lower and upper bounds on team sizes as well as combinations of attributes that have to be homogeneous (that is, identical) for all members of the corresponding teams. Furthermore, the user can define a cost for assigning any individual to a certain team. We show that some special cases of our new model lead to NP-hard problems while others allow for (fixed-parameter) tractability results. For example, the problem is already

*Supported by the DFG, research project PAWS, NI 369/10.

**Major parts of this work were done while the author was with The Institute of Mathematical Sciences, Chennai, India, and visiting TU Berlin.

An extended abstract appeared under the title "Pattern-Guided Data Anonymization and Clustering" in *Proceedings of the 36th International Symposium on Mathematical Foundations of Computer Science (MFCS '11)*, volume 6907 of *LNCS*, pages 182-193, Springer 2011. That version concentrates on the anonymization aspects of the model. In our new version we slightly extend our model and show how it applies to (homogeneous) clustering of individuals, that is, to homogeneous team formation. Indeed, we now claim that the models and ideas better fit with these applications than with the previous data anonymization motivation. Apart from full proofs omitted in the extended abstract and also adapting our old ideas to the new extended model, the current article also contains a new and easier proof of NP-hardness, a new proof for showing that polynomial-time data reduction in term of so-called polynomial-size problem kernels is unlikely to exist with respect to certain parameterizations, and a new algorithm for the (still NP-hard) special case ignoring costs. Many of the new findings are part of the diploma thesis [18] of Thomas Köhler.

Robert Bredereck, André Nichterlein, Rolf Niedermeier
Institut für Softwaretechnik und Theoretische Informatik, TU Berlin, Germany
E-mail: {robert.bredereck, andre.nichterlein, rolf.niedermeier}@tu-berlin.de

Geevarghese Philip
Max-Planck-Institut für Informatik, Saarbrücken, Germany
E-mail: gphilip@mpi-inf.mpg.de

NP-hard even if (i) there are no lower and upper bounds on the team sizes, (ii) all costs are zero, and (iii) the matrix has only two columns. In contrast, the problem becomes fixed-parameter tractable for the combined parameter "number of possible teams" and "number of different individuals", the latter being upper-bounded by the number of rows.

Keywords Team selection, Team formation, k-Anonymity, Matrix modification problems, NP-hardness, Parameterized complexity, Fixed-parameter tractability, Kernelization

1 Introduction

The task of forming teams arises in different research areas with different models and optimization criteria. One line of approaches is to search for an allocation of individuals to teams such that the overall expertise per team is maximized [2, 6, 32, 33]. These approaches differ among themselves in their models of measuring and handling expertise and in the ways in which they find the solution. Another line of approaches [20, 22] is to form teams containing members that cover a prespecified set of skills while minimizing the communication costs indicated by the social network of the team members. Both models have in common that the resulting teams tend to be *heterogeneous*. To form *homogeneous* teams, different approaches have to be followed. Note that advantages and disadvantages of homogeneous versus heterogeneous teams are controversially discussed (e.g. [1, 3, 31]) but not in scope of this work.

Notably, the concept of homogeneity in teams is similar to the concept of k-anonymity in privacy-preserving data publishing [16]. An $n \times m$-matrix M over a fixed alphabet is said to be k-anonymous if for every row r in M there are at least $k-1$ further rows in M that are identical with r. The intuitive idea which motivates this notion for data privacy is as follows: Suppose each row in M contains data about a distinct person. Even if the table does not contain data—such as names, addresses, or dates of birth—which is usually slotted under "identifying information", it is possible—as has been remarkably illustrated using US Census data [29]—that rows can be associated with specific individuals by observing unique combinations of their attributes. If the matrix M is k-anonymous then, since there are no unique rows in M, one cannot associate a specific individual to one row of data M [27, 28, 30]. Complete homogeneity (when one cannot distinguish two rows) implies perfect anonymity of the individuals; in this case the corresponding matrix with n rows is n-anonymous. In this work, we will show how to use and extend concepts from data anonymization to compute homogeneous teams.

The following team formation task is central to our work. Given a set of individuals (e.g. employees, students, workshop participants) with several known attributes (e.g. abilities, interests, locations, qualifications, fitness level), the goal is to partition them into homogeneous groups (e.g. projects, exercise groups, social events). Being homogeneous means to agree on a certain subset of attributes which may differ from group to group depending on the respective

grouping purpose. For example, for one project it may be necessary that all employees work in the same city and have the same native language. Another project can only be realized if the employees use the same operating system and are experts for the same database management system. One social event (e.g. hiking) is only worth to be done when all participants have a comparable fitness level and agree on the destination. Another event (e.g. movie) requires that the participants agree on the type of movie and on the preferred language. Clearly, it could be the case that a group of certain type may have multiple instances (e.g. one has two hiking guides and three rooms with movie projectors).

The Basic Model The attributes of the individuals are stored row-wise in an $n \times m$-matrix M over a finite alphabet Σ. The homogeneity constraints are expressed by a $p \times m$-matrix P over a binary alphabet $\{\square, \star\}$, where p denotes the total number of allowed teams. That is, each team is represented by a *pattern vector* $\{\square, \star\}^m$, where $\square$ means that homogeneity is required for the corresponding attribute and $\star$ means that individuals in the group may have different values for the corresponding attribute. A mapping from input rows of M to pattern vectors of P is *consistent* if all rows that are mapped to the same pattern vector agree at the $\square$-positions. One arrives at the following basic decision problem.

Basic Homogeneous Team Formation

Input: A matrix $M \in \Sigma^{n \times m}$ and a homogeneity pattern $P \in \{\square, \star\}^{p \times m}$.

Question: Is there a consistent mapping φ from input rows of M to pattern vectors of P?

Example 1 Figure 1 depicts the assignment of students to project teams. Consider seven students who have to apply for implementation projects that are to be realized in teams. The corresponding professor provides two sorts of projects with at most two suitable supervisors each. Projects of the first sort comprise two implementations for which knowledge of some high-level programming language and an LP-solver is required. To work together on such a project the students must agree on the programming language as well as the LP-solver. Projects of the second sort consist of two different software implementations for a traffic monitoring system. The students are asked to test their implementations and to present their results in a collaborative talk. For testing in a real-world scenario the students should live in the same city. Clearly, for realizing the implementation and the talk they also have to agree on the programming language and the style of the slides. A solution respecting the given homogeneity pattern is given in the bottom table. Note that, for instance, there would be no solution if there was only one traffic monitoring project but three LP implementation projects.

Starting from this basic problem variant we also study more general versions. Particularly, we allow the user to specify a lower and an upper bound for the size of each team. Furthermore, we will also extend the model such that the

Attributes of the students:

prog. language	LP-solver	location	slides style
C++	CPLEX	Berlin	LibreOffice
Java	CPLEX	Saarbrücken	LibreOffice
Haskell	Gurobi	Berlin	Latex Beamer
C++	CPLEX	Jena	Latex Beamer
C++	CPLEX	Saarbrücken	LibreOffice
Java	Gurobi	Saarbrücken	LibreOffice
Haskell	CPLEX	Berlin	Latex Beamer

Homogeneity pattern of the projects:

2× LP implementation	□	□	⋆	⋆
2× Traffic monitoring	□	⋆	□	□

Homogeneous teams respecting the pattern matrix:

Team 1	C++	CPLEX	⋆	⋆
	C++	CPLEX	⋆	⋆
	C++	CPLEX	⋆	⋆
Team 2	Java	⋆	Saarbrücken	LibreOffice
	Java	⋆	Saarbrücken	LibreOffice
Team 3	Haskell	⋆	Berlin	Latex Beamer
	Haskell	⋆	Berlin	Latex Beamer

Fig. 1 Example assignment of students to project teams.

user may fix some costs for assigning an individual to a team and ask for solutions not exceeding some prespecified cost bound. A formal definition of the extended model follows in Section 2.

Relation to k-Anonymity and Related Work The well-studied problem of making a matrix k-anonymous by suppressing a minimum number of entries, that is, by replacing a minimum number of matrix entries with the $\star$-symbol, is closely related to homogeneous team formation. Each group of at least k identical rows can be seen as homogeneous team. Our full model can be seen as extension of this concept (see Section 2). We also provide a cost measure similarly to counting the number of suppressions and allow for specifying bounds on the team sizes similar to the degree k of anonymity. Additionally, we allow for specifying homogeneity patterns expressing which combination of attributes have to be identical, thus incorporating user guidance.

For $k \geq 3$ it is NP-hard to make a given matrix k-anonymous by suppressing a minimum number of entries [7, 23]. However, it was shown that homogeneity in the input as well as in the solution has a (positive) effect on the computational complexity of the problem [10, 11]. For example, the problem becomes fixed-parameter tractable for the parameter "number of different input rows" or for the parameter combination "number of different output row types" and "number of suppressions" [10].

Our research is also related to the work of Aggarwal et al. [4] who proposed a new model of data anonymization based on clustering. While they

develop several polynomial-time approximation algorithms, their basic modeling idea is—roughly—to cluster the rows of the input matrix and then to publish the "cluster centers"; importantly, it is required that each cluster contains at least k rows, and this corresponds to the k-anonymity concept.

In companion work [11], the pattern concept also has been studied for anonymization purposes, including positive experimental findings.

We are not aware of any combinatorial models for homogeneous team formation in the literature.

Our Contributions We formally define a simple combinatorial model of user-specified homogeneous team formation using some concepts of k-anonymity. We show that the central problem Basic Homogeneous Team Formation is NP-complete even when there are no constraints on the team sizes and the matrix alphabet Σ is binary. We also show that the problem is NP-complete for matrices containing just two columns. On the positive side, we show that Homogeneous Team Formation is fixed-parameter tractable (FPT)[1] for the parameter t (the number of different kinds of matrix rows). Since t is a lower bound for n (the number of matrix rows) this implies fixed-parameter tractability for the parameter n. Moreover, it can be solved in polynomial time for a constant number p of given pattern vectors—in other words, Homogeneous Team Formation is contained in the parameterized complexity class XP for the parameter p. Membership in XP also holds for the parameter s, the cost bound. Since several of our findings indicate computational hardness, this motivates investigations in the spirit of multivariate algorithmics [14, 25], that is, the study of combined parameters. Here, we have the following: Homogeneous Team Formation is fixed-parameter tractable for the combined parameters $(m, |\Sigma|)$ and (s, p) (due to upper bound arguments using t) whereas the parameterized complexity status (fixed-parameter tractability vs W[1]-hardness) is open for the combined parameter (m, p). We also show that the problem is unlikely to have polynomial-size problem kernels for the combined parameter $(m, p, |\Sigma|)$, excluding hope for efficient data reduction in terms of this parameter combination. See Table 1 for a list of our results with respect to several parameterizations.

Organization of the Paper In the next section we formally introduce the new model and the notation which we use in the paper. In Section 3 we show that even very restricted cases of the problem are NP-hard. Furthermore, we show an "impossibility result" concerning efficient and effective data reduction. In Section 4 we show fixed-parameter tractability for several parameterizations of the problem. We conclude in Section 5 with a discussion of directions for future research.

[1] Informally speaking, a problem with input size x and parameter p is called fixed-parameter tractable if it can be solved in $f(p) \cdot x^{O(1)}$ time, where f may be an arbitrarily computable function solely depending on p.

Table 1 Results for the computational complexity of Basic Homogeneous Team Formation with respect to various parameters. FPT stands for "fixed-parameter tractability" and XP stands for polynomial-time solvability when the parameter is a constant; see Section 2 for definitions.

Results for Basic Homogeneous Team Formation	
alphabet size $\|\Sigma\|$	NP-complete for $\|\Sigma\| = 2$
number m of columns	NP-complete for $m = 2$
number n of rows	FPT
number t of different input rows	FPT
number p of pattern vectors	XP
combined parameter $(\|\Sigma\|, m)$	FPT*

*Does not admit polynomial-size problem kernels unless coNP⊆NP/poly.

2 Preliminaries and the Full Model

As mentioned in the introduction, the Basic Homogeneous Team Formation problem allows the user to specify, for each possible team, the combination of attributes which have to be homogeneous for that team. For most of our results we consider an extended model where the user is not only allowed to specify the homogeneity pattern for each possible team but also to specify lower and upper bounds on the team sizes. Furthermore, assigning individuals to teams may cause some costs; for instance, each team member may require a workstation. To model such constraints we allow the user to specify a cost for each possible team. Recall that each team is represented by a pattern vector p which is a row of the homogeneity pattern P, and that the team consists of all the individuals (rows from matrix M) which are mapped to p by a consistent mapping. So we take the cost of assigning an individual to a team, and associate this cost with the pattern vector which represents the team.

We now formally define a model that captures all this. To this end, it is helpful to interpret a matrix simply as a multiset of rows:

Definition 1 Let $M \in \Sigma^{n\times m}$ be a matrix over a finite alphabet Σ. Then $\mathrm{R}(M)$ is the multiset of all the rows in M.

The heart of our homogeneous team formation model lies in a function that "consistently" maps input matrix rows to some given pattern vectors.[2] This is described in the following definition, where we use $v[i]$ and $x[i]$ to refer to the i^{th} entry in the vector v and the row x, respectively.

Definition 2 Let Σ be a finite alphabet, let $M \in \Sigma^{n\times m}$, and $P \in \{\square, \star\}^{p\times m}$ be two matrices. A function $\varphi : \mathrm{R}(M) \to \mathrm{R}(P)$ is *consistent* if $\forall x, y \in \mathrm{R}(M)$ with $v := \varphi(x) = \varphi(y)$ and for all $1 \le i \le m$ it holds that

$$(v[i] = \square) \Rightarrow (x[i] = y[i]).$$

[2] Although the input table as well as the given patterns formally are matrices, we use different terms to distinguish between them: The "input matrix" consisting of "rows" and the "pattern mask" consisting of "pattern vectors".

As mentioned above, we let the user specify the cost of each pattern vector. Formally, this is expressed by a *cost function* $c : \mathrm{R}(P) \to \mathbb{N}$. Then the cost of a mapping is defined as follows.

Definition 3 Let $M \in \Sigma^{n\times m}$ and $P \in \{\square, \star\}^{p\times m}$ be two matrices and let $\varphi : \mathrm{R}(M) \to \mathrm{R}(P)$ be a mapping from the rows of M to the pattern vectors of P. For $v \in \mathrm{R}(P)$, let $\#(v) := |\{x \in \mathrm{R}(M) \mid \varphi(x) = v\}|$. Then, the *cost* of φ is $\sum_{v\in\mathrm{R}(P)} c(v) \cdot \#(v)$.

The bounds of the team sizes are expressed by functions $l, u : \mathrm{R}(P) \to \mathbb{N}$.

Definition 4 Let Σ be a finite alphabet, let $M \in \Sigma^{n\times m}$, and $P \in \{\square, \star\}^{p\times m}$ be two matrices and let $l, u : \mathrm{R}(P) \to \mathbb{N}$ be two functions. A function $\varphi : \mathrm{R}(M) \to \mathrm{R}(P)$ *fulfills the size constraints* l *and* u if

$$\forall x \in \mathrm{R}(M) : l(\varphi(x)) \leq \#(\varphi(x)) \leq u(\varphi(x)).$$

Finally, we are ready to define the central computational problem (formulated in its decision version) of this work.

Homogeneous Team Formation

Input: A matrix $M \in \Sigma^{n\times m}$, a homogeneity pattern $P \in \{\square, \star\}^{p\times m}$, three functions $l, u, c : \mathrm{R}(P) \to \mathbb{N}$, and a cost bound $s \in \mathbb{N}$.

Question: Is there a consistent mapping $\varphi : \mathrm{R}(M) \to \mathrm{R}(P)$ that fulfills the size constraints l and u and has cost at most s?

Note that Basic Homogeneous Team Formation is a special case of Homogeneous Team Formation where $l = \langle 1\rangle$, $u = \langle n\rangle$, $c = \langle 0\rangle$, and $s = 0$, where $\langle a\rangle$ denotes the constant function that maps all rows to the value a.

We use the following notation in the rest of the paper. A consistent mapping φ (see Definition 2) plays a central role in the definition of Homogeneous Team Formation. We often talk about it implicitly when we say that a row is mapped to a pattern vector. Moreover, we speak about assigning a $\square$-symbol of a pattern vector v to a symbol $a \in \Sigma$; this means that every row mapped to v has an a at the position of this $\square$-symbol.

Parameterized Complexity Our algorithmic results mostly rely on concepts of parameterized complexity analysis [13, 15, 24]. The fundamental idea herein is, given a computationally hard problem L, to identify a parameter k (typically a positive integer or a tuple of positive integers) for L and to determine whether a size-x input instance of L can be solved in $f(k) \cdot x^{O(1)}$ time, where f is an arbitrary computable function. If this is the case, then one says that L is *fixed-parameter tractable* for the parameter k. The corresponding complexity class is called FPT. If L could only be solved in polynomial running time where the degree of the polynomial depends on k (such as $x^{f(k)}$), then, for parameter k, the problem L is said to lie in the—strictly larger [13]—parameterized complexity class XP. Finally, we also consider the parameterized complexity class W[1] with FPT $\subseteq$ W[1] $\subseteq$ XP. It is widely believed that a parameterized problem

which is W[1]-hard—based on so-called parameterized reductions [13]—does not have FPT algorithms.

A core tool in the development of fixed-parameter algorithms is polynomial-time preprocessing by *data reduction* [8, 17, 21]. Here, the goal is to transform a given problem instance I with parameter k in polynomial time into an equivalent instance I' with parameter $k' \leq k$ such that the size of I' is upper-bounded by some function g only depending on k. If this is the case, then we call I' a (problem) *kernel* of size $g(k)$. It is well-known that every fixed-parameter tractable problem has a problem kernel; however, in general the corresponding function $g(k)$ is only exponentially bounded. Thus, it is a central question to decide whether such a problem has a problem kernel of size $g(k)$ polynomial in k.

3 Intractability Results

In this section, we show that Basic Homogeneous Team Formation is NP-complete even in very restricted cases. Note that all intractability results for Basic Homogeneous Team Formation imply intractability results for the more general Homogeneous Team Formation. Membership in NP is easy to see: Guessing a mapping φ of the rows from M to pattern vectors from P, it is easy to verify in polynomial time that φ is consistent, fulfills the size constraints, and has cost at most s. In the following, we provide a polynomial-time many-one reduction from the NP-complete Constrained Bipartite Vertex Cover problem [19] to show NP-hardness for Basic Homogeneous Team Formation with $m = 2$.

Before doing the reduction we show how to get rid of non-binary alphabets. The structural properties of Homogeneous Team Formation (and its special case Basic Homogeneous Team Formation) allow us to replace any alphabet with a binary alphabet.

Lemma 1 *Let $I = (M, P, l, u, c, s)$ be an instance of* Homogeneous Team Formation *with $M \in \Sigma^{n \times m}$. Then there is a polynomial-time algorithm that computes an equivalent instance $I' = (M', P', l, u, c, s)$ such that $M' \in \{0,1\}^{n \times m'}$ and $m' = \lceil \log |\Sigma| \rceil \cdot m$.*

Proof For $I = (M, P, l, u, c, s)$, construct $I' = (M', P', l, u, c, s)$ as follows. Assign to each symbol in Σ a unique integer from $\{0, 1, \ldots, |\Sigma| - 1\}$. Each column of M will be replaced with $\lceil \log |\Sigma| \rceil$ columns. The corresponding columns are used to binary encode (filling up with zeros on the left) the identifier of the original symbol. The pattern vectors from P are extended analogously: Each $\star$- (respectively $\square$-) symbol is replaced by $\lceil \log |\Sigma| \rceil$ many consecutive $\star$- (respectively $\square$-) symbols. The size constraint functions u and l and the cost function c remain unchanged for the extended pattern vectors. Observe that the new instance is equivalent to the original one: Let φ map the rows $r_1, \ldots, r_n$ of M to the pattern vectors $v_1, \ldots, v_p$ of P and, correspondingly, let φ' map the rows $r'_1, \ldots, r'_n$ of M to the pattern vectors $v'_1, \ldots, v'_p$

of P such that $\varphi(r_i) = v_j \iff \varphi'(r'_i) = v'_j$. Then it is easy to see that φ is a solution for I if and only if φ' is a solution for I'. □

Now we present our NP-completeness result.

Theorem 1 Basic Homogeneous Team Formation *is NP-complete, even if the number m of columns is two.*

Proof We provide a polynomial-time many-to-one reduction from the NP-complete Constrained Bipartite Vertex Cover [19] problem. A *vertex cover* of a graph $G = (V, E)$ is a set $S \subseteq V$ of vertices such that for every $\{u, v\} \in E$ it holds that $u \in S$ or $v \in S$.

Constrained Bipartite Vertex Cover

Input: A bipartite graph $G = (L \uplus R, E)$ and two positive integers k_ℓ and k_r.

Question: Is there a vertex cover $S \subseteq L \uplus R$ with $|S \cap L| \leq k_\ell$ and $|S \cap R| \leq k_r$?

For an instance (G, k_ℓ, k_r) of Constrained Bipartite Vertex Cover we construct an equivalent instance (M, P) of Basic Homogeneous Team Formation as follows. First, for each edge $\{u, v\} \in E$ add a two-column row $\boxed{u}\boxed{v}$ to the input matrix M. Second, add k_ℓ pattern vectors $P_\ell = \{p^\ell_1, \ldots, p^\ell_{k_\ell}\}$ of type $(\square, \star)$ and k_r pattern vectors $P_r = \{p^r_1, \ldots, p^r_{k_r}\}$ of type $(\star, \square)$ to the pattern mask P.

We next prove that (G, k_ℓ, k_r) is a yes-instance of Constrained Bipartite Vertex Cover if and only if (M, P) is a yes-instance of Basic Homogeneous Team Formation.

"⇒": Let S be a vertex cover with $|S \cap L| \leq k_\ell$ and $|S \cap R| \leq k_r$. Thus, there exist two total injective functions $f^\ell : |S \cap L| \to P_\ell$ and $f^r : |S \cap R| \to P_r$. Now, we construct a solution mapping φ. For each edge $\{u, v\} \in E$ with $u \in S \cap L$ set $\varphi(\boxed{u}\boxed{v}) = f^\ell(u)$ and for each edge $\{u, v\} \in E$ with $v \in S \cap R$ and $u \notin S \cap L$ set $\varphi(\boxed{u}\boxed{v}) = f^r(v)$. Since S is a vertex cover, by construction, each row is consistently assigned to exactly one pattern vector.

"⇐": Let φ be a solution for the instance (M, P). For a row $\boxed{u}\boxed{v}$ with $\varphi(\boxed{u}\boxed{v}) \in P_r$, we call v not suppressed and, correspondingly, we call u not suppressed if $\varphi(\boxed{u}\boxed{v}) \in P_\ell$. Let $S = \{v_1, \ldots, v_q\}$ be the elements that are not suppressed in the mapped rows. By construction it holds that $q \leq k_\ell + k_r$. Observe that S is a vertex cover for G: Every row is assigned to a pattern vector and, hence, every edge in G is covered. Since there are k_ℓ pattern vectors of type $(\square, \star)$, it holds that $|S \cap L| \leq k_\ell$. Analogously, it holds that $|S \cap R| \leq k_r$. □

Combining Lemma 1 with Theorem 1 gives NP-completeness for the binary case.

Corollary 1 *Even if the alphabet Σ is binary,* Basic Homogeneous Team Formation *is NP-complete.*

As we will show later, Basic Homogeneous Team Formation is fixed-parameter tractable with respect to the combined parameter $(m, |\Sigma|)$; however, now we show that it is unlikely that Basic Homogeneous Team Formation admits a polynomial-size problem kernel with respect to the combined parameter (m, p) and binary alphabet, directly implying the same non-existence result with respect to $(m, p, |\Sigma|)$ and $(m, |\Sigma|)$.

Bodlaender et al. [9] introduced a refined concept of parameterized reduction that allows to transfer "non-existence results" for polynomial-size problem kernels to other problems. It is defined as follows.

Definition 5 [9] Let P and Q be two parameterized problems over $\Sigma^* \times \mathbb{N}$. We say that P is *polynomial time and parameter reducible* to Q, written $P \leq_{Ptp} Q$, if there exists a polynomial-time computable function $f : \Sigma^* \times \mathbb{N} \to \Sigma^* \times \mathbb{N}$ and a polynomial p, such that for all $(x, k) \in \Sigma^* \times \mathbb{N}$:

1. $(x, k) \in P \Leftrightarrow (x', k') := f(x, k) \in Q$, and
2. $k' \leq p(k)$.

The function f is called *polynomial time and parameter transformation.*

Bodlaender et al. [9] showed that, for two parameterized problems P and Q whose unparameterized versions are NP-complete, if $P \leq_{Ptp} Q$, then a polynomial problem kernel for Q implies a polynomial problem kernel for P.

Using this type of parameterized reduction, we show the following:

Theorem 2 Basic Homogeneous Team Formation *parameterized by the combined parameter* (m, p) *has no problem kernel of polynomial size unless coNP* $\subseteq$ *NP/poly, even if the alphabet* Σ *is binary.*

Proof We give a polynomial-time and parameter transformation from the Set Splitting problem.

Set Splitting

Input: A set family $\mathcal{F} = \{F_1, \ldots, F_{|\mathcal{F}|}\}$ over a universe $U = \{u_1, \ldots, u_{|U|}\}$.

Question: Does there exists a subset $X \subseteq U$ such that each set in $\mathcal{F}$ contains both an element from X and from $U \setminus X$?

Cygan et al. [12] showed that Set Splitting parameterized by $|U|$ does not admit a problem kernel of size $|U|^{O(1)}$ unless an unexpected complexity-theoretic collapse occurs, namely coNP $\subseteq$ NP/poly.

Given an instance $(\mathcal{F}, U)$ of Set Splitting, construct an instance (M, P) of Basic Homogeneous Team Formation as follows. We create a matrix M with $m = |U|$ columns, each column corresponding to one element in U. For each set $F \in \mathcal{F}$ we add $2m + 2$ rows to the initially empty input matrix. In what follows, let d_i be a dummy symbol:

- For each $1 \leq i \leq m + 1$, we add a row r_i^F where
  - $r_i^F[j] = 0$ if $u_j \in F$;

  - $r_i^F[j] = d_i$ otherwise.
- For each $1 \leq i \leq m+1$, we add a row $r_i^{F'}$ where
  - $r_i^{F'}[j] = 1$ if $u_j \in F$;
  - $r_i^F[j] = d_i$ otherwise.

The pattern mask P is an $m \times m$ vector. All the non-diagonal entries of P are $\star$, and its diagonal entries are all $\square$.

To prove the correctness of the construction we show that $(\mathcal{F}, U)$ is a yes-instance of Set Splitting if and only if (M, P) is a yes-instance of Basic Homogeneous Team Formation.

"$\Rightarrow$:" Let $X \subseteq U$ be a solution to the Set Splitting-instance. Let v be the characteristic vector of X, that is, $v[i] = 1$ if $u_i \in X$ and $v[i] = 0$ otherwise. Since X is a solution for the Set Splitting-instance, each row r in M coincides in at least one entry with v. For row r of M, let i_r be an index such that $r[i_r] = v[i_r]$. Then mapping each row r of M to pattern vector $P[i_r]$ yields a solution to the Basic Homogeneous Team Formation-instance.

"$\Leftarrow$:" Let φ be a solution for the Basic Homogeneous Team Formation instance (M, P). Observe that by construction if rows r_a, r_b of M are such that $\varphi(r_a) = \varphi(r_b) = p = P[i]$, then $r_a[i] = r_b[i]$. Let v be a $1 \times m$ vector constructed as follows. For $1 \leq i \leq m$:

- If there exists a row r of M such that $\varphi(r) = p = P[i]$, then $v[i] = r[i]$;
- Otherwise, $v[i] = \perp$, a dummy symbol denoting "undefined".

We claim that $X := \{u_j : v[j] = 1\}$ is a solution for $(\mathcal{F}, U)$. To see this, consider an arbitrary set $F \in \mathcal{F}$. Since there are $m+1$ dummy symbols in M and only m pattern vectors in P, we can assume without loss of generality that the dummy symbol d_1 is not contained in vector v. Hence, the row $r_1^{F'}$ is mapped to a pattern vector such that its $\square$-symbol is assigned to 1. Thus, by construction of X, it follows that $F \cap X \neq \emptyset$. Analogously, the row r_1^F is mapped to a pattern vector such that its $\square$-symbol is assigned to 0. Thus, $F \cap (U \setminus X) \neq \emptyset$.

This proves the correctness of our reduction. Note that in this reduction the alphabet size is $|U| + 3$. By applying Lemma 1, we get an equivalent instance with $|\Sigma| = 2$ and the number of columns is $|U| \log |U|$. Thus, the statement of the theorem follows. $\square$

4 Tractable Cases

In the previous section, we showed computational intractability results for various special cases of Basic Homogeneous Team Formation. Now, we complement these hardness results by presenting some relevant tractable cases. To this end, we consider several parameterizations of Homogeneous Team Formation. Since Homogeneous Team Formation allows the user to specify pattern vectors to influence the homogeneity structure of the solution, the number of pattern vectors p appears to be one of the most natural

problem-specific parameters. There are instances with a small amount of pattern vectors, for instance, when the user wants to form a small number of teams.

We start with a general observation on the solution structure of Homogeneous Team Formation instances. To this end, we introduce the concept of row types. A *row type* is a string from Σ^m. We say that a set of rows in the matrix has a certain row type if the rows in the set are all are identical.

A General Scheme in our Algorithms Most tractability results in this section are based on a general algorithmic scheme which we will introduce first. The central point is that Homogeneous Team Formation is polynomial-time solvable if some additional information about its solution, called *hint* in the following, is known. Thus, our algorithms consist of two phases.

1. Hint computation (using fixed-parameter algorithms).
2. Polynomial-time hint checking.

The intuition behind our algorithmic approach is the following. In the first phase, one determines a hint for the solution and calls the second phase. The second phase checks whether there is a solution which respects the given hint. If no such solution exists, then the first phase will generate another hint. The decisive point is to find a realization of the first phase which generates, for all yes-instances, at least one "correct" hint, that is, a hint which leads to a solution.

More precisely, we have the following.

Hint computation

Input: A matrix $M \in \Sigma^{n\times m}$ and a pattern mask $P \in \{\square, \star\}^{p\times m}$.

Task: Compute a "hint" function $h : \mathrm{R}(P) \to \mathrm{R}(M) \cup \{\varnothing\}$ which maps each pattern vector either to a row of the input matrix M or to $\varnothing$.

The hint gives information about the solution φ: For each pattern vector, one either fixes that it is not used in the solution, that is, it is mapped to $\varnothing$, or one fixes one row from the input matrix which is mapped to the pattern vector in the solution. A hint function h is *correct* if there is a solution φ of the Basic Homogeneous Team Formation instance (M, P) such that $\forall x \in \mathrm{R}(P)$:

$$\big(h(x) \neq \varnothing \to \varphi(h(x)) = x\big) \wedge \big(h(x) = \varnothing \to \nexists y \in \mathrm{R}(M) : \varphi(y) = x\big).$$

The second phase efficiently computes a solution that respects the hint whenever there is such a solution. We use the term *preimage type* to denote those row types which can safely be mapped to a specific pattern vector in the context of a hint.

Definition 6 Let $(M_{n\times m}, P)$ be an instance of Basic Homogeneous Team Formation, and let $h : R(P) \to R(M)$ be a hint function. We say that a row type $\mathbf{r}$ of M is a *preimage type* of a pattern vector v of P if rows from $\mathbf{r}$ can

potentially be mapped to v while respecting the hint h. More precisely, row type $\mathbf{r}$ is a *preimage type* of pattern vector v if there is a row x of M such that if $h(v) = x$, then for all $1 \leq i \leq m$ it holds that

$$(v[i] = \square) \implies (\mathbf{r}[i] = x[i]).$$

Polynomial-time hint checking

Input: A matrix $M \in \Sigma^{n \times m}$, a pattern mask $P \in \{\square, \star\}^{p \times m}$, a hint function $h : \mathrm{R}(P) \to \mathrm{R}(M) \cup \{\varnothing\}$, three functions $l, u, c : \mathrm{R}(P) \to \mathbb{N}$, and a cost bound $s \in \mathbb{N}$.

Task: Compute a consistent function φ mapping the rows of M to the pattern vectors of P with cost at most s, fulfilling the size constraints l and u, and respecting the hint h, or answer "no" if there is no such mapping.

In Phase 2, we have the following situation. Suppose that there are t input row types, and a solution uses $p' \leq p$ pattern vectors; these are the vectors which the hint function h does *not* map to $\varnothing$. In the following we represent the set of input row types by $T_{\text{in}} := \{1, \ldots, t\}$ and the set of pattern vectors used in the solution by $T_{\text{out}} := \{1, \ldots, p'\}$. Let $\kappa : T_{\text{in}} \times T_{\text{out}} \to \{0, 1\}$ be the function expressing whether an input row type is a preimage type of a pattern vector. The size constraints are expressed by the integers α_i and β_i with $i \in T_{\text{out}}$, where $\alpha_i = l(\rho_i)$ and $\beta_i = u(\rho_i)$ with ρ_x denoting the x^{th} pattern vector used in the solution. Furthermore, let ω_i with $i \in T_{\text{out}}$ denote the cost $c(i)$ of the i^{th} pattern vector and let n_j with $j \in T_{\text{in}}$ denote the number of rows in the j^{th} input row type. A consistent mapping g that fulfills the size constraints l and u, has cost at most s, and respects the preimage types corresponds to a solution of a slight modification[3] of the Row Assignment [10] problem. It is defined as follows.

Row Assignment*

Input: Nonnegative integers s, $\alpha_1, \ldots, \alpha_{p'}$, $\beta_1, \ldots, \beta_{p'}$, $\omega_1, \ldots, \omega_{p'}$, and $n_1, \ldots, n_t$ with $\sum_{i=1}^{t} n_i = n$ and a function $\kappa : T_{\text{in}} \times T_{\text{out}} \to \{0, 1\}$.

Question: Is there a function $g : T_{\text{in}} \times T_{\text{out}} \to \{0, \ldots, n\}$ such that

$$\kappa(i, j) \cdot n \geq g(i, j) \qquad \forall i \in T_{\text{in}}, \forall j \in T_{\text{out}} \tag{1}$$

$$\alpha_j \leq \sum_{i=1}^{t} g(i, j) \leq \beta_j \qquad \forall j \in T_{\text{out}} \tag{2}$$

$$\sum_{j=1}^{p'} g(i, j) = n_i \qquad \forall i \in T_{\text{in}} \tag{3}$$

$$\sum_{i=1}^{t} \sum_{j=1}^{p'} g(i, j) \cdot \omega_j \leq s \tag{4}$$

[3] In Inequality 2 the modified Row Assignment* has a specific lower bound α_j and a specific upper bound β_j for each $j \in T_{\text{out}}$ instead of a uniform upper bound k.

Let us see why Row Assignment* correctly captures the *Polynomial-time hint checking* problem. We interpret $g(i,j) = \ell$ in the former problem to mean that the function φ of the latter problem maps exactly ℓ rows of input type i to pattern vector j. Inequality 1 ensures that for each pattern vector $v \in P$, only rows from its preimage types are mapped to v. Inequality 2 ensures that the mapping fulfills the size constraints l and u of *Polynomial-time hint checking*. Equation 3 states that all rows of each input type are mapped to *some* pattern vector; this ensures that each input row is mapped to a pattern vector. Inequality 4 ensures that the costs of the mapping are at most s.

The following lemma shows that Row Assignment* can be be solved in polynomial time. The proof is similar to to the original proof showing that Row Assignment is polynomial-time solvable [10, Lemma 2].

Lemma 2 *There is an algorithm that solves* Row Assignment* *in time* $O(tp \cdot \log(t+p)(tp + (t+p)\log(t+p)))$.

Proof We reduce Row Assignment* to the Capacitated Minimum Cost Flow problem, which is defined as follows [26]:

Capacitated Minimum Cost Flow

Input: A network (directed graph) $D = (V, A)$ with demands $d : V \to \mathbb{Z}$ on the nodes, costs $c : V \times V \to \mathbb{N}$, and capacities $\delta : V \times V \to \mathbb{N}$.

Task: Find a function f which minimizes $\sum_{(u,v)\in A} c(u,v) \cdot f(u,v)$ and satisfies:

$$\sum_{\{v|(u,v)\in A\}} f(u,v) - \sum_{\{v|(v,u)\in A\}} f(v,u) = d(u) \qquad \forall u \in V$$

$$0 \leq f(u,v) \leq \delta(u,v) \qquad \forall (u,v) \in A$$

We first describe the construction of the network with demands, costs, and capacities. For each n_i, $1 \leq i \leq t$, add a node v_i with demand $-n_i$ (that is, a supply of n_i) and for each $1 \leq j \leq p'$ add a node u_j with demand α_j. If $\kappa(i,j) = 1$, then add an arc (v_i, u_j) with cost ω_j and capacity ∞. Finally, add a sink z with demand $(\sum n_i - \sum \alpha_i)$ and the arcs (u_j, z) with cost zero and capacity $\beta_j - \alpha_j$. See Figure 2 for an example of the construction.

The Capacitated Minimum Cost Flow problem is solvable in $O(|A| \cdot \log(|V|)(|A|+|V|\cdot\log(|V|)))$ time in a network (directed graph) $D = (V, A)$ [26]. Since our constructed network has $O(t+p)$ nodes and $O(t \cdot p)$ arcs, we can solve our Capacitated Minimum Cost Flow-instance in $O(tp \cdot \log(t+p)(tp + (t+p)\log(t+p)))$ time.

It remains to prove that the Row Assignment*-instance is a yes-instance if and only if the constructed network has a minimum cost flow of cost at most s.

"⇒": Assume that g is a function fulfilling constraints 1 to 4. Then define a flow f as follows: For each $1 \leq i \leq t, 1 \leq j \leq p$, set $f(v_i, u_j) = g(i,j)$ and $f(u_j, z) = \sum_{i=1}^{t} g(i,j) - \alpha_j$. Since g satisfies Equation 3 and Inequality 2,

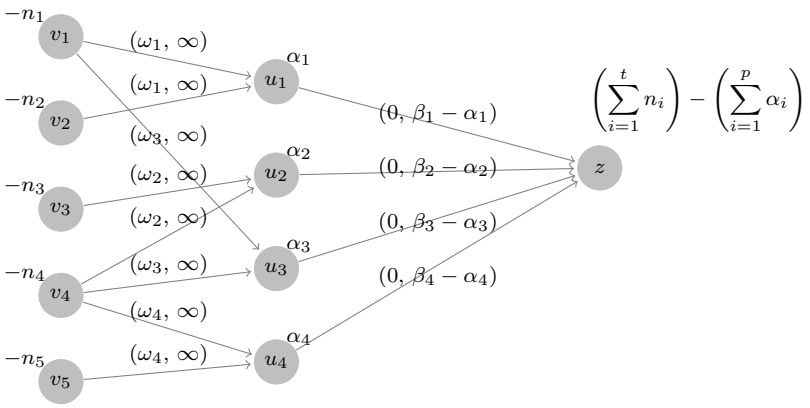

Fig. 2 Example of the constructed network with $t = 5$ and $p = 4$. The pair (x, y) on each arc denotes costs x and capacity y. The number next to each node denotes its demand.

we get that the flow f fulfills the demands on the nodes. Since g fulfills Inequality 4 and the cost of each arc (u_j, z), $1 \leq j \leq p$, is zero, flow f has cost of at most s.

"$\Leftarrow$": Assume that f is a flow with cost of at most s. All costs, constraints, and demands are integer-valued, and hence, due to the Integrality Property [5] of network flow problems, there exists an optimal flow with integer values. Then set $g(i, j) = f(v_i, u_j)$ for each $1 \leq i \leq t, 1 \leq j \leq p$. Note that g fulfills Equation 3 and Inequality 2 due to the demands on the nodes of the network and the capacities of the ingoing arcs of z. Since $n_i \leq n$ for all $1 \leq i \leq t$, also Inequality 1 is fulfilled. Note that f has cost at most s and, hence, g fulfills Inequality 4. □

Computing function κ as needed in Row Assignment* takes $O(p \cdot t \cdot m)$ time and as preprocessing we have to compute the input row types in $O(n \cdot m)$ time (by constructing a trie on the rows [10]). We obtain the following lemma.

Lemma 3 *Phase* 2 *can be solved in* $O(t^2 \cdot p^2 \cdot \log(tp) + t \cdot p \cdot m + n \cdot m)$ *time.*

Next, we describe several fixed-parameter algorithms for Phase 1 of the above described algorithmic scheme. The respective algorithms differ in the varying parameters that are used.

Parameters p and t We first study whether Homogeneous Team Formation is still intractable (that is, NP-hard) when the number p of pattern vectors, that is, the number of possible teams, is a constant. Combining with p the parameter t denoting the number of input row types, we show that Homogeneous Team Formation is fixed-parameter tractable with respect to the combined parameter (p, t). To this end, we use a brute-force realization of the hint computation in Phase 1. The corresponding algorithm (consisting of both phases) can also be interpreted as an XP-algorithm for Homogeneous Team

Formation parameterized by p, that is, Homogeneous Team Formation is polynomial-time solvable for constantly many pattern vectors.

Theorem 3 *There is an algorithm solving* Homogeneous Team Formation *in time* $O(t^p \cdot 2^p \cdot (t^2 \cdot p^2 \cdot \log(tp) + t \cdot p \cdot m) + n \cdot m)$.

Proof The parameterized hint computation works in two steps as follows.

1. For each pattern vector v, determine whether it is *used* in the solution, that is, determine whether v occurs in the image of the mapping.
2. For each pattern vector v that is used in the solution, guess one of the rows which are mapped to v in the solution.

We realize both steps by branching over all possibilities. Step 1 can be realized by branching on 2^p possibilities. In Step 2, we have to consider up to t^p possibilities. Since we consider all possibilities, one clearly finds a correct hint function for every yes-instance. □

Theorem 3 shows fixed-parameter tractability for Homogeneous Team Formation with respect to the combined parameter (t, p). Next, we develop a fixed-parameter algorithm for the individual parameter t when there are no upper bounds on the team sizes. This is mainly a classification result because its current running time is impractical. Furthermore, it seems reasonable to assume that the number of possible teams can be bounded by a function in the number of different individuals in most realistic instances. Then, however, one would always prefer the algorithm from Theorem 3.

We begin with an important observation on the solution mappings which holds when there are no upper bounds on the team sizes. The following lemma says that without loss of generality one may assume that there is an optimal solution that uses at most t pattern vectors.

Lemma 4 *Let* (M, P, l, u, c, s) *be a yes-instance of* Homogeneous Team Formation *with* $u = \langle n \rangle$*, that is, there are no upper bounds on the team sizes. If* M *has* t *row types, then there exists a solution mapping* φ *for* (M, P, l, u, c, s) *whose image contains at most* t *elements.*

Proof Let φ be a consistent mapping fulfilling the size constraint l and having cost at most s. If the image of φ has at most t elements, then there is nothing to prove. So let the image of φ contain more than t elements, that is, φ uses more than t pattern vectors. We now describe an operation that reduces the number of pattern vectors used by φ without increasing its cost.

We call a pattern vector v used by φ *redistributable* if, for each row r mapped to v, there is another pattern vector v' used by φ such that $c(v') \leq c(v)$ and mapping r to v' instead of v does not violate the consistency of the mapping. Observe that if a pattern vector used by φ is redistributable, then we can eliminate this row type from the image of φ by "moving" each of its rows to a different, at most as expensive pattern vector, while preserving the lower bound condition on pattern vectors from the image of φ. This operation reduces the number of pattern vectors used by φ without increasing the cost. As

long as there are redistributable pattern vectors left, we repeatedly eliminate pattern vectors from the image of the mapping in this manner. Let φ' denote the mapping which results from exhaustive application of this procedure to φ.

Now, we analyze the properties of the modified mapping φ'. Clearly, its image contains only pattern vectors that are not redistributable. Consider any pattern vector v' in the image of φ. Since v' is not redistributable, there exists a row r' mapped to v' such that no row that has the same row type as r' can be consistently mapped to another pattern vector from the image of φ with at most the same cost. In this sense, v' is the "cheapest possible" pattern vector for at least one row type. Hence, every pattern vector which is used by φ' is the cheapest possible pattern vector for some row. Since there are only t row types, at most t pattern vectors can be the "cheapest possible" pattern vector for any row. Hence, the image of φ' contains at most t pattern vectors. □

Theorem 4 *If there are no upper bounds on the team sizes, then* Homogeneous Team Formation *can be solved in* $O(2^{t^2}t^{2t+2} \cdot (m + t^2 \log t) + n \cdot m)$ *time.*

Proof To show fixed-parameter tractability for the single parameter t, we need a more refined realization of the hint computation phase. Clearly, whenever $p \leq t$, we use the brute-force realization from Theorem 3 without any modification. The corresponding running time is $O(t^t \cdot 2^t \cdot (t^4 \cdot \log t + t^2 \cdot m) + n \cdot m)$. For $p > t$, we slightly modify the Step 1 in the algorithm behind Theorem 3.

Recall that in Step 1 one determines a set $P' \subseteq P$ of pattern vectors that are used in the solution. Due to Lemma 4 we know that without loss of generality $|P'| \leq t$. In Theorem 3 we simply try all size-at-most-t subsets of P. Here, we show that for guessing we only have to take into account a relatively small subset $P^* \subseteq P$ with $|P^*| \leq g(t)$ and g being a function which only depends on t.

Consider a pattern vector v of the unknown P'. In Phase 2 of the algorithm (polynomial-time solving by the help of the hint), we determine the preimage types, that is, the set of input row types that may contain rows that are mapped to v in the solution. Assume that the preimage types for all pattern vectors from P' are fixed. To determine which concrete pattern vector corresponds to a set of preimage types, we only have to take into account the t cheapest compatible pattern vectors, where compatible means that all rows of these preimage types coincide at the □-symbol positions. By definition, there exist at most 2^t different sets of preimage types. Thus, keeping for each set of preimage types the t cheapest pattern vectors and removing the rest results in a set P^* of size $2^t \cdot t$.

Summarizing, when $p > t$, we realize Step 1 by computing P^* as described above and branch on all subsets $P' \subseteq P^*$ of size at most t. This can be done in $O\binom{2^t \cdot t}{t} \leq O(2^{t^2}t^t)$ time. Step 2 in the algorithm behind Theorem 3 remains unchanged, that is, for pattern vector $v \in P'$ we guess one row from M which is mapped to v in the solution. Altogether, we solve Homogeneous Team Formation in $O(t^t \cdot (2^{t^2}t^t) \cdot (t^2 \cdot m + t^4 \log t) + n \cdot m)$ time. Clearly, since $t \leq n$, our result also holds for the parameter n. □

For Theorem 3 we described an XP-algorithm with respect to the parameter p, that is, an algorithm with polynomial running time for constant values of p. We leave it open whether there also exists an algorithm where the degree of the polynomial is independent of p, that is, whether Homogeneous Team Formation is fixed-parameter tractable for parameter p.[4] However, at least for the special case with $s \geq n \cdot \max_{v \in \mathrm{R}(P)} c(v)$, that is, effectively the costs are unbounded, we can show fixed-parameter tractability.

Theorem 5 *If there are no upper bounds on the team sizes and no cost bound, then* Homogeneous Team Formation *can be solved in* $O(p! \cdot m \cdot n^2 \cdot t^4 \cdot \log t)$ *time.*

Proof To prove this statement, we describe a greedy algorithm for computing a hint function h.

We will show that there is a *correct permutation* of pattern vectors such that applying the following greedy procedure leads to a correct hint function h. For now, assume that a specific permutation of pattern vectors is given.

Greedy hint construction

1. Start with $h(x) := \varnothing$ for all pattern vectors $x \in P$.
2. Take the first pattern vector z with $h(z) = \varnothing$.
3. Set $h(z)$ to be the first row in M.
4. Remove from M the row $h(z)$ and remove all rows which coincide with $h(z)$ at the $\square$-positions of z. If M is nonempty, then go to Step 2.

The hint function constructed by this greedy procedure highly depends on the ordering of the pattern vectors. Thus, we simply try all possible orderings to realize Phase 1 of the algorithmic scheme. It remains to show the following:

Claim. *Let φ be a solution for* Homogeneous Team Formation *which uses a minimal number of pattern vectors and let $P' \subseteq P$ be the set of pattern vectors used by φ. Then, there is a permutation of pattern vectors such that applying the "Greedy hint construction" procedure to this permutation of pattern vectors produces a correct hint, that is, it assigns one row r_i to each pattern vector $p_i \in P'$ such that $\varphi(r_i) = p_i$, and no row to any pattern vector in $P \setminus P'$.*

Proof (of Claim) Given φ and P' we show the existence of the correct ordering of pattern vectors by construction. To this end, let $\mathrm{first}(M)$ denote the first row in the matrix M and let π be an (initially empty) list of pattern vectors.

(a) Insert $p^* := \varphi(\mathrm{first}(M))$ at the end of π.
(b) Remove all rows x from M which coincide with $\mathrm{first}(M)$ at the $\square$-positions of p^*.
(c) If M is nonempty, then go to Step (a).

[4] As a consequence of the "binarization" in Lemma 1, the question whether Homogeneous Team Formation is fixed-parameter tractable with respect to the combined parameter $(p, |\Sigma|)$ is equivalent to the question whether Homogeneous Team Formation is fixed-parameter tractable with respect to p alone.

(d) Finally insert all pattern vectors from $P \setminus P'$ at the end of π.

The ordering of the pattern vectors in π is correct: Observe that since φ uses a minimal number of pattern vectors, it holds that for each pattern vector $p_i \in P'$ there is at least one row which cannot be mapped to any other pattern vector from P'. Hence, every pattern vector from P' has some position in π. Now, apply "Greedy hint construction" with the ordering of the pattern vectors given by π to obtain the hint function h. Consider p_i, the i^{th} pattern vector in π. By construction of π, row $h(p_i)$ was the first row in the matrix in the i^{th} iteration of the construction procedure for π. Thus, by Step (a), $\varphi(h(p_i)) = p_i$. Furthermore, "Greedy hint construction" terminates after $|P'|$ iterations which means that $\forall y \in P \setminus P' : h(y) = \varnothing$ and, hence, the hint is correct. □

This completes the proof of the theorem. □

Corollaries for Further (Combined) Parameters As corollaries of Theorem 3 and Theorem 4 we show fixed-parameter tractability for some natural parameter combinations. All results rely on the fact that one can bound the number t of input row types from above by a function only depending on the respective combined parameter. In particular, $|\Sigma|^m$ and n are both upper bounds for the number t of input row types. This yields the following corollary. Observe that n is *not* the input size in this problem, and can indeed be much smaller than the total input size: hence it is not trivially the case that the problem with parameter n is fixed-parameter tractable.

Corollary 2 Homogeneous Team Formation *is fixed-parameter tractable with respect to the combined parameter* $(|\Sigma|, m)$*. If there are no upper bounds on the team sizes, then* Homogeneous Team Formation *is fixed-parameter tractable with respect to the parameter* n*.*

For the next two corollaries we require the following rather technical restriction on the cost function. The pattern vectors without $\star$-symbols are the only pattern vectors that have cost zero, or, equivalently, all pattern vectors containing at least one $\star$-symbol have cost at least one.

Corollary 3 *If the cost function* $c : \mathrm{R}(P) \to \mathbb{N}$ *fulfills the requirement that* $(c(v) = 0) \Rightarrow (v = \square^m)$*, then* Homogeneous Team Formation *is fixed-parameter tractable with respect to the combined parameter* (p, s)*.*

Proof Subsequently, we call rows that are mapped to pattern vectors v with cost at least one, that is, $c(v) \geq 1$, *costly rows* and their corresponding row types *costly row types*. Analogously, rows that are mapped to pattern vectors with cost zero are called *costless rows* and row types that only contain costless rows are called *costless row types*. Clearly, every input row type is costly or costless. Note that costly row types may also contain some costless rows. There are at most s costly rows and, hence, at most s costly row types. Since the pattern vectors with cost zero contain no $\star$-symbol, two costless

rows from different input row types cannot be mapped to the same pattern vector. Furthermore, the number of pattern vectors without $\star$-symbols is at most p. Hence, the number of costless row types is also at most p. Thus, in a yes-instance the number t of input row types is at most $s + p$. Applying Theorem 3 yields fixed-parameter tractability. □

Corollary 3 shows fixed-parameter tractability with respect to the combined parameter (p, s). Fixed-parameter tractability for the single parameters p as well as s remains open. However, in parallel to the case for p (Theorem 3) we show that there is a polynomial-time algorithm for constant values of s.

Corollary 4 *If the cost function c fulfills the requirement that $(c(v) = 0) \Rightarrow (v = \square^m)$, then* Homogeneous Team Formation *is in XP with respect to the parameter s.*

Proof Using the definitions of costly and costless from Corollary 3, we give a simple algorithm that shows membership in XP. The first step is to guess, from $\sum_{i=0}^{s} \binom{n}{i}$ possibilities, the rows which are costly. The second step is to guess, from $\sum_{i=0}^{s} \binom{p}{i}$ possibilities, the pattern vectors that contain $\star$-symbols which are used in the solution. Then, guess the mapping between at most s rows and at most s pattern vectors and check whether it is consistent and k-anonymous. In the last step, the costless rows are greedily mapped to pattern vectors without $\star$-symbols. □

5 Conclusion

We introduced a natural and simple combinatorial model for homogeneous team formation and provided a first theoretical analysis. Our model allows to specify the homogeneity structure as well as lower and upper bounds on the team sizes. The corresponding combinatorial problem Homogeneous Team Formation is NP-hard and has a number of tractable and intractable special cases; most of our results are listed in Table 1 in the introductory section. Besides the general quest for improving our worst-case upper bounds, several concrete questions remain open. For example, the parameterized complexity (fixed-parameter tractability vs W[1]-hardness) of Homogeneous Team Formation for the parameters cost bound s and the combined parameter (m, p), where m is the number of columns and p is the number of pattern vectors, remains open. A particularly interesting open question is whether Homogeneous Team Formation for parameter p is fixed-parameter tractable or W[1]-hard when s is bounded. It also seems worth investigating Homogeneous Team Formation from the viewpoint of polynomial-time approximation. Finally, it remains to complement our purely theoretical investigations with empirical studies concerning the practical usefulness of our newly proposed model. In particular, how practical are our algorithms, and to what extent can or should these be replaced by efficient and effective heuristics?

Are there further practically relevant parameterizations to be exploited in a practical application?

Acknowledgements We are grateful to the anonymous referees of the *MFCS '11* conference for helping to improve this work by spotting some flaws and providing the idea behind Corollary 4. Furthermore, we thank an anonymous referee for providing the idea of the proof of Theorem 2 which is significantly simpler than the one in the conference version of this paper. We are also grateful to two anonymous *Algorithmica* reviewers for their constructive feedback.

References

1. M. G. Aamodt and W. W. Kimbrough. Effect of group heterogeneity on quality of task solutions. *Psychological Reports*, 50(1):171–174, 1982.
2. H. Abdelsalam. Multi-objective team forming optimization for integrated product development projects. In *Foundations of Computational Intelligence Volume 3*, volume 203 of *Studies in Computational Intelligence*, pages 461–478. Springer, 2009.
3. S. O. Adodo and J. O. Agbayewa. Effect of homogenous and heterogeneous ability grouping class teaching on student's interest, attitude and achievement in integrated science. *International Journal of Psychology and Counselling*, 3(3):48–54, 2011.
4. G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. *ACM Transactions on Algorithms*, 6(3):1–19, 2010.
5. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.
6. A. Baykasoglu, T. Dereli, and S. Das. Project team selection using fuzzy optimization approach. *Cybernetics and Systems*, 38(2):155–185, 2007.
7. J. Blocki and R. Williams. Resolving the complexity of some data privacy problems. In *Proceedings of the 37th International Colloquium on Automata, Languages and Programming (ICALP '10)*, volume 6199 of *LNCS*, pages 393–404. Springer, 2010.
8. H. L. Bodlaender. Kernelization: New upper and lower bound techniques. In *Proceedings of the 4th International Workshop on Parameterized and Exact Computation (IWPEC '09)*, volume 5917 of *LNCS*, pages 17–37. Springer, 2009.
9. H. L. Bodlaender, S. Thomassé, and A. Yeo. Kernel bounds for disjoint cycles and disjoint paths. *Theoretical Computer Science*, 412(35):4570–4578, 2011.
10. R. Bredereck, A. Nichterlein, R. Niedermeier, and G. Philip. The effect of homogeneity on the computational complexity of combinatorial data anonymization. *Data Mining and Knowledge Discovery*, 2012. Online available.

11. R. Bredereck, A. Nichterlein, and R. Niedermeier. Pattern-guided k-anonymity. In *Proceedings of the Joint Conference of the 7th International Frontiers of Algorithmics Workshop and the 9th International Conference on Algorithmic Aspects of Information and Management (FAW-AAIM '13)*, volume 7924 of *LNCS*, pages 350–361. Springer, 2013.
12. M. Cygan, M. Pilipczuk, M. Pilipczuk, and J. Wojtaszczyk. Solving the 2-disjoint connected subgraphs problem faster than 2^n. In *Proceedings of the 10th Latin American Symposium on Theoretical Informatics (LATIN' 12)*, volume 7256 of *LNCS*, pages 195–206. Springer, 2012.
13. R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.
14. M. R. Fellows, B. M. Jansen, and F. Rosamond. Towards fully multivariate algorithmics: Parameter ecology and the deconstruction of computational complexity. *European Journal of Combinatorics*, 34:541–566, 2013.
15. J. Flum and M. Grohe. *Parameterized Complexity Theory*. Springer, 2006.
16. B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53, 2010.
17. J. Guo and R. Niedermeier. Invitation to data reduction and problem kernelization. *ACM SIGACT News*, 38(1):31–45, 2007.
18. T. Köhler. Benutzergeführtes Anonymisieren von Daten mit Pattern Clustering: Algorithmen und Komplexität (in German, English title: User-guided data anonymization with pattern clustering: Algorithms and complexity). Diploma thesis, Friedrich-Schiller-Universität Jena, 2011. Available at `http://fpt.akt.tu-berlin.de/publications/pattern_D.pdf`.
19. S. Kuo and W. Fuchs. Efficient spare allocation for reconfigurable arrays. *IEEE Design & Test of Computers*, 4(1):24–31, 1987.
20. T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, pages 467–476. ACM, 2009.
21. D. Lokshtanov, N. Misra, and S. Saurabh. Kernelization – preprocessing with a guarantee. In *The Multivariate Algorithmic Revolution and Beyond*, volume 7370 of *LNCS*, pages 129–161. Springer, 2012.
22. A. Majumder, S. Datta, and K. Naidu. Capacitated team formation problem on social networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, pages 1005–1013. ACM, 2012.
23. A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '04)*, pages 223–228. ACM, 2004.
24. R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.
25. R. Niedermeier. Reflections on multivariate algorithmics and problem parameterization. In *Proceedings of the 27th International Symposium*

on Theoretical Aspects of Computer Science (STACS '10), volume 5 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 17–32. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2010.
26. J. Orlin. A faster strongly polynomial minimum cost flow algorithm. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC '88)*, pages 377–387. ACM, 1988.
27. P. Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
28. P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '98)*, pages 188–188. ACM, 1998.
29. L. Sweeney. Uniqueness of simple demographics in the U.S. population. Technical report, Carnegie Mellon University, School of Computer Science, Laboratory for International Data Privacy, 2000.
30. L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 557–570, 2002.
31. K. B. White. A preliminary investigation of information systems team structures. *Information & Management*, 7(6):331–335, 1984.
32. H. Wi, S. Oh, J. Mun, and M. Jung. A team formation model based on knowledge and collaboration. *Expert Systems with Applications*, 36(5): 9121–9134, 2009.
33. A. Zzkarian and A. Kusiak. Forming teams: An analytical approach. *IIE Transactions*, 31(1):85–97, 1999.