Correlation clustering of graphs and integers

S. Akiyama, L. Aszalós, L. Hajdu, A. Pethő

***Abstract*—Correlation clustering can be modeled in the following way. Let A be a nonempty set, and $\sim$ be a symmetric binary relation on A. Consider a partition (clustering) $\mathcal{P}$ of A. We say that two distinct elements $a, b \in A$ are in conflict, if $a \sim b$, but a and b belong to different classes (clusters) of $\mathcal{P}$, or if $a \not\sim b$, however, these elements belong to the same class of $\mathcal{P}$. The main objective in correlation clustering is to find an optimal $\mathcal{P}$ with respect to $\sim$, i.e. a clustering yielding the minimal number of conflicts. We note that correlation clustering, among others, plays an important role in machine learning.**

In this paper we provide results in three different, but closely connected directions. First we prove general new results for correlation clustering, using an alternative graph model of the problem. Then we deal with the correlation clustering of positive integers, with respect to a relation $\sim$ based on coprimality. Note that this part is in fact a survey of our earlier results. Finally, we consider the set of so-called S-units, which are positive integers having all prime divisors in a fixed finite set. Here we prove new results, again with respect to a relation defined by the help of coprimality. We note that interestingly, the shape of the optimal clustering radically differs for integers and S-units.

***Index Terms*—correlation clustering, graphs, integers, S-units.**

I. INTRODUCTION

Correlation clustering was introduced in the field of machine learning. We refer to the paper of Bansal et al. [3], which also gives an excellent overview of the mathematical background. Let G be a complete graph on n vertices and label its edges with $+1$ or -1 depending on whether the endpoints have been deemed to be similar or different. Consider a partition of the vertices. Two edges are in conflict with respect to the partition if they belong to the same class, but are different, or they belong to different classes although they are similar. The ultimate goal of correlation clustering is to find a partition with minimal number of conflicts. The special feature of this clustering is that the number of clusters is not specified. In some applications G is not necessarily a complete graph like in [5] or the labels of the edges are real numbers like in [9].

Correlation clustering admits the following equivalent model too. Let A be a nonempty set, $\sim$ be a tolerance relation on A, i.e., a reflexive and symmetric binary relation. Consider a partition (clustering) $\mathcal{P}$ of A. We say that two elements $a, b \in A$ are in conflict, if $a \sim b$, but a and b belong to different classes (clusters) of $\mathcal{P}$, or if $a \not\sim b$, however, these elements belong to the same class of $\mathcal{P}$. The main objective is to find an optimal $\mathcal{P}$ with respect to $\sim$, i.e. a clustering yielding the minimal number of conflicts. It is worth to mention that if we also assume that $\sim$ is transitive, then it is an equivalence relation. In this case the optimal clustering is obviously provided by the equivalence classes of $\sim$. So this is the lack of the transitive property which makes the problem of correlation clustering interesting and important. Every clustering of A implies an equivalence relation on A. The number of conflicts in a clustering reflects a kind of distance of $\sim$ to this equivalence relation. An optimal correlation clustering causes the least number of conflicts among all clusterings, thus it induces a nearest equivalence relation to $\sim$.

A typical application of correlation clustering is the classification of unknown topics of (scientific) papers. In this case the papers represent the elements of A and two papers are considered to be similar (or being in relation $\sim$), if one of them refers to the other. The classes of an optimal clustering then can be interpreted as the topics of the papers. This kind of clustering has many applications: image segmentation [15], identifying biologically relevant groups of genes [4], examining social coalitions [16], reducing energy consumption in wireless sensor networks [6], modeling physical processes [12], etc.

The number of partitions of sets having n elements grows exponentially, so the exhaustive search is not available to find an optimal clustering. Bansal et al. [3] showed that to find an optimal clustering is NP-hard. Beside this, they also proposed and analyzed algorithms for approximate solutions of the problem. In fact the correlation clustering can be considered to be an optimization problem: one should find the clustering minimizing the number of conflicts. Thus it is possible to apply traditional and modern optimization algorithms to find almost optimal clusterings. Following this approach, Bakó and Aszalós [2] have implemented several traditional methods, and have also invented some new ones.

In this paper we consider infinite growing sequences of labeled graphs such that the labeling is hereditary (see Section II). Then we can define lower and upper densities of edges with label $+1$ as well as of the classes in an optimal correlation clustering. The aim of Section II is to show relations between these quantities. Our results show that the choice of the labeling heavily affects the structure of the optimal clustering. For example Theorem 1 implies that if the upper density of edges with $+1$ is less than $1/2$ then there are at least two classes in an optimal correlation clustering. The value $1/2$ is the best possible by Remark 1.

In Sections III and IV we investigate particular examples. To introduce them we switch to the relational model. In that case we may assume that $A_i, i = 1, 2, \ldots$ is a chain of subsets of $\mathbb{N}$ and $\sim_i$ is the restriction of $\sim$ to A_i. Here $\sim$ denotes a reflexive and symmetric relation on $\mathbb{N}$. After fixing the basic

Manuscript received September 11, 2014, revised December 8, 2014.

S. Akiyama is with the Institute of Mathematics, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan (zip:350-8571)

L. Aszalós is with the Faculty of Computer Science, University of Debrecen, H-4010 Debrecen, P.O. Box 12, Hungary

L. Hajdu is with the Institute of Mathematics, University of Debrecen, H-4010 Debrecen, P.O. Box 12, Hungary

A. Pethő is with the Faculty of Computer Science, University of Debrecen, H-4010 Debrecen, P.O. Box 12, Hungary

set to $\mathbb{N}$ it is natural to use the coprimality to define the relation $\sim$. More precisely, for positive integers $a, b \in A$ we set $a \sim b$ if $\gcd(a, b) > 1$ or $a = b = 1$. In Section III we consider this relation with sets A_n of positive integers not exceeding n ($n = 1, 2, \dots$). The results of this section are published in the paper [1], so we only outline the main results and methods here. We present a natural greedy algorithm, Algorithm 1, which computes locally optimal clustering and prove that it behaves regularly for $n < n_0 = 3{\cdot}5{\cdot}7{\cdot}11{\cdot}13{\cdot}17{\cdot}19{\cdot}23 = 111\,546\,435$, but from n_0 on this regularity disappears. Further we show that its optimal correlation clustering has at least two classes. In Section IV we give a similar analysis, but for the sets of S-units (or generalized Hamming numbers) not exceeding n ($n = 1, 2, \dots$). The results presented here are all new. We show that the optimal correlation clustering is in this case asymptotically trivial, i.e. has only one class. Although the asymptotic result is smooth, there are usually many growing classes in the early stages, but after a while the largest class starts to collect all elements like a black hole.

Finally, we give concluding remarks in Section V.

II. Correlation clustering of graphs

In this section we consider the problem of correlation clustering not for a single graph, but for an increasing family of graphs. For $n \geq 1$, let K_n be the complete graph of n vertices. Write $V(K_n)$ and $E(K_n)$ for the set of vertices and edges of K_n, respectively. Take an arbitrary labeling

$$c_n : \ E(K_n) \to \{-1, 1\}$$

of the edges of K_n, subject to the hereditary (consistency) condition that for some embedding

$$\sigma_{n-1} : K_{n-1} \to K_n$$

of K_n into K_{n-1} the coloring is invariant, that is

$$c_{n-1}(e) = c_n(\sigma_{n-1}(e)) \text{ for } e \in E(K_{n-1}).$$

Thus

$$K_1 \xrightarrow{\sigma_1} K_2 \xrightarrow{\sigma_2} \dots$$

can be considered as an increasing sequence of labeled graphs. We define the upper and lower densities of the edges having label 1 in the usual way:

$$\overline{g} = \limsup_{n\to\infty} \frac{|\{e \in E(K_n) \ : \ c_n(e) = 1\}|}{|E(K_n)|} =$$

$$= \limsup_{n\to\infty} \frac{|\{e \in E(K_n) \ : \ c_n(e) = 1\}|}{n(n-1)/2}$$

and

$$\underline{g} = \liminf_{n\to\infty} \frac{|\{e \in E(K_n) \ : \ c_n(e) = 1\}|}{|E(K_n)|} =$$

$$= \liminf_{n\to\infty} \frac{|\{e \in E(K_n) \ : \ c_n(e) = 1\}|}{n(n-1)/2}.$$

Here and later on, $|H|$ denotes the number of elements of the set H. Let $\mathcal{P}(n)$ be an optimal clustering of (K_n, c_n), with classes $\mathcal{P}_1(n), \mathcal{P}_2(n), \dots, \mathcal{P}_{m(n)}(n)$. Here without loss of generality we may assume that the classes are arranged in non-increasing order with respect to cardinality, that is

$$|\mathcal{P}_j(n)| \geq |\mathcal{P}_{j+1}(n)| \ (j = 1, \dots, m(n) - 1).$$

Define the upper and lower cluster densities of $\mathcal{P}_j(n)$ by

$$\overline{\rho}_j = \limsup_{n\to\infty} \frac{|\mathcal{P}_j(n)|}{n}, \qquad \underline{\rho}_j = \liminf_{n\to\infty} \frac{|\mathcal{P}_j(n)|}{n}$$

for $j = 1, \dots, m(n)$. If $j > m(n)$ then we set $\mathcal{P}_j(n) = \emptyset$. Clearly we have $\overline{\rho}_j \geq \overline{\rho}_{j+1}$ and $\underline{\rho}_j \geq \underline{\rho}_{j+1}$ for $j \geq 1$.

Theorem 1. *We have*

$$\sum_i \underline{\rho}_i^2 \leq 2\underline{g}, \qquad \overline{\rho_1}^2 \leq 2\overline{g}$$

and

$$\overline{g} - \sum_{i<j} \overline{\rho}_i \overline{\rho}_j \leq \sum_i \overline{\rho}_i^2, \qquad \underline{g} - \sum_{i<j} \underline{\rho}_i \underline{\rho}_j \leq \sum_j \underline{\rho}_i^2.$$

Proof. We claim that $\sum_i \underline{\rho}_i \leq 1$. In fact, for any $m \in \mathbb{N}$ and any $\varepsilon > 0$, there exists an $n_0 \in \mathbb{N}$ such that

$$\underline{\rho}_j - \varepsilon/m \leq |\mathcal{P}_j(n)|/n$$

for $j \leq m$ and $n \geq n_0$. Thus

$$\sum_{j=1}^{m} \underline{\rho}_j \leq \sum_{n=1}^{m} \frac{|\mathcal{P}_j(n)|}{n} + \varepsilon \leq 1 + \varepsilon.$$

As one can choose ε and m arbitrarily, the above inequality proves our claim. This fact is used in the last part of the proof of the first inequality.

As $\mathcal{P}_j(n)$ is an optimal cluster, in the induced graph to $\mathcal{P}_j(n)$ of (K_n, c_n), at least half of the edges belonging to each vertex of $\mathcal{P}_j(n)$ must have label 1. Indeed, if this does not hold for some vertex v of $\mathcal{P}_j(n)$, then the cluster $\mathcal{P}_j(n)$ can be divided into $\mathcal{P}_j(n) \setminus \{v\}$ and $\{v\}$, and in the new clustering the number of conflicts is less. This implies that among $|E(\mathcal{P}_j(n))|$ edges, there are at least $|E(\mathcal{P}_j(n))|/2$ edges with label 1. From the inequality

$$\frac{1}{2} \sum_j |E(\mathcal{P}_j(n))| \leq |\{e \in E(K_n) \ : \ c_n(e) = 1\}|, \qquad (1)$$

for any $m \in \mathbb{N}$ and $\varepsilon_1 > 0$, there exists an $n_1 \in \mathbb{N}$ such that

$$\frac{1}{2} \sum_{j=1}^{m} \frac{((\underline{\rho}_j - \varepsilon_1)n)((\underline{\rho}_j - \varepsilon_1)n - 1)}{2} \leq$$

$$\leq |\{e \in E(K_n) \ : \ c_n(e) = 1\}|$$

for $n \geq n_1$. Thus for any $\varepsilon_2 > 0$

$$\frac{1}{2} \sum_{j=1}^{m} \frac{((\underline{\rho}_j - \varepsilon_1)n)((\underline{\rho}_j - \varepsilon_1)n - 1)}{2} \leq (\underline{g} + \varepsilon_2)\frac{n(n-1)}{2}$$

holds for infinitely many n. Dividing by $n^2/2$ and letting n tend to ∞, we obtain the first inequality, since m, ε_1, and ε_2 are arbitrary.

It is also clear from (1) that for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, we have

$$\frac{((\overline{\rho}_1 - \varepsilon_1)n)((\overline{\rho}_1 - \varepsilon_1)n - 1)}{2} \leq$$

$$\leq |\{e \in E(K_n) \ : \ c_n(e) = 1\}| \leq (\overline{g} + \varepsilon_2)\frac{n(n-1)}{2}$$

for infinitely many n, giving the second inequality.

Consider now two clusters $\mathcal{P}_i(n)$ and $\mathcal{P}_j(n)$. Among the $|\mathcal{P}_i(n)|\cdot|\mathcal{P}_j(n)|$ edges joining the two clusters in K_n, the number of edges labeled by 1 can be at most $|\mathcal{P}_i(n)| \cdot |\mathcal{P}_j(n)|/2$. Indeed, otherwise we could decrease the number of conflicts by taking $\mathcal{P}_i(n) \cup \mathcal{P}_j(n)$ as a new cluster. So

$$\frac{1}{2}\sum_{i<j} |\mathcal{P}_i(n)| \cdot |\mathcal{P}_j(n)|$$

is an upper bound for the number of edges labeled by 1, connecting two distinct clusters $\mathcal{P}_i(n)$ and $\mathcal{P}_j(n)$. On the other hand, for any $\varepsilon_3 > 0$ and for infinitely many n, there exist at least

$$(\overline{g} - \varepsilon_3)|E(K_n)| - \sum |E(\mathcal{P}_j(n))|$$

edges labeled by 1, connecting two distinct clusters $\mathcal{P}_i(n)$ and $\mathcal{P}_j(n)$. Thus we get the inequality

$$\frac{1}{2}\sum_{i<j} |\mathcal{P}_i(n)| \cdot |\mathcal{P}_j(n)| \geq$$

$$\geq (\overline{g} - \varepsilon_3)|E(K_n)| - \sum |E(\mathcal{P}_j(n))|$$

for infinitely many n. Therefore for any $\varepsilon_2 > 0$, we have

$$\frac{1}{2}\sum_{i<j}((\overline{\rho}_i + \varepsilon_2)n)((\overline{\rho}_j + \varepsilon_2)n) \geq$$

$$\geq (\overline{g} - \varepsilon_3)\frac{n(n-1)}{2} - \sum_j \frac{((\overline{\rho}_j + \varepsilon_2)n)(((\overline{\rho}_j + \varepsilon_2)n) - 1)}{2}$$

for infinitely many n. This implies the third inequality. The proof of the last inequality is similar, and our statement follows. □

Corollary 1. *Using the previous notation, we have*

$$\sqrt{\overline{g}} \leq \sum_j \overline{\rho}_j \quad \textit{and} \quad \sqrt{\underline{g}} \leq \sum_j \underline{\rho}_j.$$

Proof. The first assertion follows from

$$\overline{g} \leq \overline{g} + \sum_{i<j} \overline{\rho}_i\overline{\rho}_j \leq \left(\sum_j \overline{\rho}_j\right)^2$$

using the third inequality of Theorem 1. The proof of the second inequality is similar. □

We say that the clusters are *full* if $\sum_{j=1}^{\infty} \underline{\rho}_j = 1$.

Not all clusterings are full. For example, we may introduce an ordering of vertices of K_n and let $\mathcal{P}_1(n)$ be the first half of the vertices and remaining $\mathcal{P}_j(n)$ be singletons for $j \geq 2$. Then we have $\underline{\rho}_1 = 1/2$ and $\underline{\rho}_j = 0$ for $j \geq 2$.

Corollary 2. *Assume that the optimal clusters are full. If $g = \overline{g} = \underline{g}$ then we have*

$$1/2 - \sum_{i<j} \underline{\rho}_i\underline{\rho}_j \leq g \leq 1 - \sum_{i<j} \underline{\rho}_i\underline{\rho}_j.$$

In particular, $\underline{\rho}_1 = 1$ and $\underline{\rho}_j = 0$ for $j > 1$ holds if and only if $g = 1$.

Proof. The statement follows from

$$g + \sum_{i<j} \underline{\rho}_i\underline{\rho}_j \leq \left(\sum_j \underline{\rho}_j\right)^2 \leq 2g + 2\sum_{i<j} \underline{\rho}_i\underline{\rho}_j$$

and $\sum_j \underline{\rho}_j = 1$. Since $\underline{\rho}_{j+1} \geq \underline{\rho}_j$ by definition, the inequality shows that $\underline{\rho}_1\underline{\rho}_2 > 0$ holds if and only if $g < 1$. □

The last statement shows that the edges labeled by 1 must be of density 1 in order to have only one non-empty class (namely, the whole K_n) in an optimal clustering. At this point we need to introduce some new notions.

A graph G is *locally stable* if the degree of each vertex is at least $\lceil(|G|-1)/2\rceil$. It is *globally stable* if for any partition $V(G) = A \cup B$ (disjoint), there are at least $\lceil|A| \cdot |B|/2\rceil$ edges connecting A and B. Denote by $G(\mathcal{P}_j(n))$ the graph obtained from the graph induced by the optimal cluster $\mathcal{P}_j(n)$ of (K_n, c_n), by removing all its edges labeled by -1. Then $G(\mathcal{P}_j(n))$ is globally stable, since otherwise a corresponding partition $A \cup B$ gives a lower number of conflicts. For brevity, we say that $\mathcal{P}_j(n)$ is globally stable if $G(\mathcal{P}_j(n))$ has this property.

Theorem 2. *If K_n is a single cluster, then its global stability implies that this is an optimal clustering (consisting of one cluster).*

Proof. Assume that $K_n = \mathcal{P}$ is globally stable and consider a different partition

$$K_n = \bigcup_{i=1}^{\ell} Q_i.$$

Let $c(Q_i, Q_j)$ be the total number of conflicts between Q_i and Q_j, that is

$$c(Q_i, Q_j) = \sum_e \frac{1 + c_n(e)}{2},$$

where the sum is taken over all edges between Q_i, Q_j. By the global stability of $\mathcal{P}$, the partition

$$\mathcal{P}_i = Q_i \cup \left(\bigcup_{j:j\neq i} Q_j\right)$$

gives not less conflicts. So we see that

$$c(Q_i, \bigcup_{j:j\neq i} Q_j) = \sum_{j:j\neq i} c(Q_i, Q_j) \geq 0$$

for all i. Summing these inequalities we obtain

$$(\ell - 1)\sum_{i<j} c(Q_i, Q_j) \geq 0.$$

Thus the partition $\bigcup_{i=1}^{\ell} Q_i$ gives not less conflicts than $\mathcal{P}$, and the statement follows. □

Lemma 1. *Complete bipartite graphs $K_{m,m}$ and $K_{m,m+1}$ are globally stable.*

Proof. Let U, V be the vertex sets of $K_{m,m}$. (That is, all edges of $K_{m,m}$ run between U and V.) Let $A \cup B$ be a partition of the vertex set $U \cup V$. Put $x = |A \cap U|$, $y = |A \cap V|$. Then

$$m - x = |B \cap U|, \quad m - y = |B \cap V|.$$

The number of edges between A and B is

$$x(m-y)+y(m-x)=m(x+y)-2xy.$$

This is not less than

$$|A|\cdot|B|/2=(x+y)(2m-x-y)/2.$$

For $K_{m,m+1}$, the computation is similar. With the same notation, the number of edges between A and B is

$$x(m+1-y)+y(m-x)=m(x+y)+x-2xy.$$

This is not less than

$$|A|\cdot|B|/2=(x+y)(2m+1-x-y)/2$$

since

$$x-2xy-(x+y)(1-x-y)/2=(x-y)(x-y+1)/2\geq 0$$

holds for $x,y\in\mathbb{Z}$. This implies the statement. □

Remark 1. *In the proof of Theorem 1 we only used the fact that $G(\mathcal{P}_j(n))$ is locally stable. However, the second inequality is the best possible in the sense that the constant 2 cannot be chosen smaller. Indeed, consider the natural embedding*

$$K_{m,m}\subset K_{m,m+1}\subset K_{m+1,m+1}\subset\dots,$$

and consider $K_{m,n}$ to be a subgraph of K_{m+n}. Define the labeling c_{m+n} by

$$c_{m+n}(e)=\begin{cases}1, & \text{if } e\in E(K_{m,n}),\\ -1, & \text{if } e\in E(K_{m+n})\setminus E(K_{m,n}).\end{cases}$$

Then this labeling c_k is consistent and asymptotically $k^2/4$ edges have label 1. This gives an example that $g=1/2$, $\rho_1=1$ and $\rho_j=0$ for $j>1$, which attains the equality for the second inequality of Theorem 1.

III. Correlation clustering of positive integers with respect to coprimality

As we have mentioned already in the introduction, it is obvious that the role of the relation $\sim$ (or, in the graph model, the definitions of the labels ±1) is crucial for the structure of the optimal clustering. This motivates the investigations in the present section. We mention that the results presented here were published in the paper [1].

Consider the sequence of complete graphs K_n together with a hereditary labeling $c_n : E(K_n)\mapsto\{-1,1\}$. Labeling the vertices of K_n by the integers $A_n=\{1,\dots,n\}$ the mapping c_n implies a reflexive and symmetric relation $\sim_n$ on A_n. By the hereditary property of c_n we may assume that $\sim_n$ admits this property, too, i.e., the restriction of $\sim_{n+1}$ to A_n is equal to $\sim_n$.

Let, more generally, $A_n\subset\mathbb{N}$ be finite and satisfying $A_n\subseteq A_{n+1}$ for $n=1,2,\dots$. Assume that there is a reflexive and symmetric relation $\sim_n$ on A_n. Further assume that the restriction of $\sim_{n+1}$ to A_n is equal to $\sim_n$. Setting $A=\cap_{n=1}^{\infty}A_n$ we have $A\subseteq\mathbb{N}$. Define the relation $\sim$ on A as follows: for $a,b\in A$ we set $a\sim b$ if there exist $n\geq1$ such that $a,b\in A_n$ and $a\sim_n b$. By the hereditary property of $\sim_n$ the relation $\sim$ is well defined on A, moreover it is reflexive and symmetric. Of course we can consider $\sim$ on $\mathbb{N}$, too. This justifies that in the sequel we consider a relation on $\mathbb{N}$. Divisibility is the best understood relation of integers. As it is not symmetric ($2|4$ but $4\nmid 2$) we cannot use it in our investigations. Fortunately the coprime relation, i.e. $a\sim b$ if $\gcd(a,b)>1$ or if $a=b=1$, is closely related to divisibility and is symmetric.

In this section we work with sets A_n of positive integers greater than 1, but not exceeding n (for $n=2,3,\dots$). Moreover we assume that A_n is equipped with the above defined coprime relation. Note that the behavior of the gcd among the first n positive integers has been investigated from many aspects; see e.g. a paper of Nymann, [13].

Bakó and Aszalós [2] have made several experiments concerning the optimal clustering of A_n with respect to $\sim$. They have discovered that the classes of a near optimal clustering have regular structure. In the sequel denote by p_i the i-th prime, i.e., $p_1=2,p_2=3,\dots$. Set

$$A_{i,n}=\{m\ :\ m\leq n,\ p_i|m,\ p_j\nmid n\ (j<i)\}.$$

In other words, $A_{i,n}$ is the set of integers at most n, which are divisible by p_i, but coprime to the smaller primes. Aszalós and Bakó found that

$$[2,n]\cap\mathbb{Z}=\bigcup_{j=1}^{\infty}A_{j,n}\tag{2}$$

is an optimal correlation clustering for $n\leq20$ and very probably for $n\leq500$, too. Notice that $A_{j,m}=\emptyset$ for all large enough j, i.e., the union on the right hand side is actually finite.

The main result of this section (and of [1]) is that for

$$n_0=3\cdot5\cdot7\cdot11\cdot13\cdot17\cdot19\cdot23=111\ 546\ 435$$

the decomposition (2) is not optimal. We prove that the number of conflicts in

$$[2,n_0]\cap\mathbb{Z}=(A_{1,n_0}\cup\{n_0\})\cup(A_{2,n_0}\setminus\{n_0\})\bigcup_{j=3}^{\infty}A_{j,n}\tag{3}$$

is less than in (2) with $n=n_0$. We cannot prove that (2) is optimal for $n<n_0$. However, we show that the natural greedy algorithm (Algorithm 1), presented below, produces the clustering (2) for all $n<n_0$, and it yields (3) for $n=n_0$.

As we have mentioned already in the introduction, by results from [3] we know that to find an optimal correlational clustering is an NP-hard problem. Hence to find an approximation of the optimal solution, it is natural to use some kind of greedy algorithm. For the sets A_n we use the following approach. The optimal clustering for $A_2=\{2\}$ is itself. Assume that we have a partition of A_{n-1} with $n>2$, and adjoin n to that class, which causes the less new conflicts. As a result we obtain a locally optimal clustering, which is not necessarily globally optimal on A_n.

Starting with a partition of A_{n-1} this algorithm returns a partition of A_n such that the conflicts caused by putting n into one of the classes is minimal. The output of Algorithm 1 on

Algorithm 1 Natural greedy algorithm

Require: an integer $n \geq 2$
Ensure: a partition $\mathcal{P}$ of N
1: $\mathcal{P} \leftarrow \{\{2\}\}$;
2: **if** $n = 2$ **then return** $\mathcal{P}$
3: **end if**
4: $m \leftarrow 3$
5: **while** $m \leq n$ **do**
6: $\mathcal{P}_M \leftarrow \mathcal{P} \cup \{\{m\}\}$
7: $M \leftarrow \text{CONFLICTS}(\mathcal{P}_M, m)$ ▷ the number of conflicts with respect to the partition $\mathcal{P}_M$ caused by the pairs $(m, a),\ a < m$
8: $C \leftarrow$ number of classes in $\mathcal{P}$
9: $j \leftarrow 1$
10: **while** $j \leq C$ **do**
11: $O \leftarrow \text{OP}(j, \mathcal{P})$ ▷ $OP(j, \mathcal{P})$ denotes the j-th class in the partition $\mathcal{P}$.
12: $\mathcal{P}_1 \leftarrow \mathcal{P} \setminus \{O\}$
13: $\mathcal{P}_1 \leftarrow \mathcal{P}_1 \cup \{O \cup \{m\}\}$
14: $M_1 \leftarrow \text{NUPAIR}(\mathcal{P}_1, m)$ ▷ the number of pairs (m, a) with $a < m$ causing a conflict in the partition $\mathcal{P}_1$
15: **if** $M_1 < M$ **then**
16: $M \leftarrow M_1$
17: $\mathcal{P}_M \leftarrow \mathcal{P}_1$
18: **end if**
19: **end while**
20: **end while**
21: **return** $\mathcal{P}_M$

the input n is denoted by $G(n)$. It is certainly a clustering of A_n. As one can easily check, we obtain

$$\begin{aligned} G(3) &= \{\{2\},\{3\}\} \\ G(4) &= \{\{2,4\},\{3\}\} \\ G(5) &= \{\{2,4\},\{3\},\{5\}\} \\ G(6) &= \{\{2,4,6\},\{3\},\{5\}\} \\ &\vdots \\ G(15) &= \{\{2,4,6,8,10,12,14\},\{3,9,15\}, \\ &\quad \{5\},\{7\},\{11\},\{13\}\}. \end{aligned}$$

For these values of n one can readily show that the above partitions provide (in fact the unique) optimal clusterings for A_n $(2 \leq n \leq 15)$, as well.

The main result of this section is the following

Theorem 3. *If* $m < n_0 = 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 = 111\ 546\ 435$ *then*

$$G(m) = \bigcup_{j=1}^{\infty} A_{j,m} \tag{4}$$

holds. However, we have

$$G(n_0) = (A_{1,n_0} \cup \{n_0\}) \cup (A_{2,n_0} \setminus \{n_0\}) \bigcup_{j=3}^{\infty} S_{j,n_0}.$$

Elements of the proof of Theorem 3. Since the complete proof of Theorem 3 is given in [1], here we only indicate the main ingredients of the proof. On this way, we recall several lemmas from [1], always without proofs.

The first important information we need is to characterize that class of $G(n-1)$ to which Algorithm 1 adjoins n. This is done with the following

Lemma 2. *Let* $n > 2$ *be an integer. Write* $G(n-1) = \{P_1, \ldots, P_M\}$ *and set* $P_0 = \emptyset$. *For* $1 \leq j \leq M$ *let*

$$E_{j,n} = \{m \ :\ m \in P_j, \gcd(m, n) = 1\}$$

and

$$B_{j,n} = \{m \ :\ m \in P_j, \gcd(m, n) > 1\}.$$

Define $E_{0,n} = B_{0,n} = \emptyset$. *Let* J *be the smallest index for which* $|B_{j,n}| - |E_{j,n}|$ $(j = 0, \ldots, M)$ *is maximal. Then* $G(n) = \{P'_0, \ldots, P'_M\}$ *such that*

$$P'_j = \begin{cases} P_j \cup \{n\}, & \text{if } j = J, \\ P_j, & \text{otherwise}. \end{cases}$$

This lemma has the following important consequence.

Corollary 3. *The following assertions are true.*
(1) *If* n *is even, then* $n \in A_{1,n}$.
(2) *If* n *is a prime, then* $\{n\} \in G(n)$.
(3) *If the smallest prime factor of* n *is* p_i *and* $n \in S_{j,n}$, *then* $j \leq i$.

The next result gives a useful bound for the sizes of the sets $A_{i,u}$.

Lemma 3. *Let* u *be an odd integer. Then we have* $|A_{1,u}| = \frac{u-1}{2}$. *Further, if* p_i *is an odd prime, then*

$$\left| |A_{i,u}| - \frac{u}{p_i} \prod_{\ell=1}^{i-1} \left(1 - \frac{1}{p_\ell}\right) \right| \leq 2^{i-2}.$$

The next lemma provides an estimation for $|B_{j,n}| - |E_{j,n}|$, where

$$B_{j,n} = \{m \ :\ m \in A_{j,n-1},\ \gcd(m, n) > 1\}$$

and

$$E_{j,n} = \{m \ :\ m \in A_{j,n-1},\ \gcd(m, n) = 1\}.$$

Note that the elements of $B_{j,n}$ and $E_{j,n}$ are those elements of $A_{j,n-1}$, which are, and which are not in the relation $\sim$ with n, respectively.

Lemma 4. *Let* $q_1 < \cdots < q_t$ *be odd primes,* $\alpha_1, \ldots, \alpha_t$ *positive integers and* $n = q_1^{\alpha_1} \cdots q_t^{\alpha_t}$. *Let* $j \geq 2$ *be such that* $p_j < q_1$. *Then*

$$\big| |B_{j,n}| - |E_{j,n}| - C_{n,j,t} \big| \leq 2^{t+j-2}$$

holds, where

$$C_{n,j,t} = \frac{n-1}{p_j} \prod_{\ell=1}^{j-1} \left(1 - \frac{1}{p_\ell}\right) \left(1 - 2 \prod_{k=1}^{t} \left(1 - \frac{1}{q_k}\right)\right).$$

The next lemma plays an important role in the proof of Theorem 3. We use the previous notation.

Lemma 5. *Let $n = q_1^{\alpha_1} \cdots q_t^{\alpha_t}$ with $q_1 < \cdots < q_t$ odd primes and $\alpha_1, \ldots, \alpha_t$ positive integers. Then*

$$|E_{1,n}| = \frac{\varphi(n)}{2} = \frac{n}{2}\left(1-\frac{1}{q_1}\right)\cdots\left(1-\frac{1}{q_t}\right),$$
$$|B_{1,n}| = \frac{n-1}{2} - |E_{1,n}|.$$

Now, in principle, we are ready to give the main steps of the proof of Theorem 3. However, the proof is rather detailed, tricky and complicated, so we restrict ourselves to indicate how the argument proceeds. We refer the interested reader again to [1] for details.

Step 1. We start with confirming the cases where n is odd and $3 \mid n$. The difficult part is to show that (4) holds for $n < n_0$. This assertion is verified by comparing the estimates of Lemmas 2, 3, 4, and 5, some computer search, and applying a tool from prime number theory namely, estimates for expressions of the form

$$\prod_{p<x}\left(1-\frac{1}{p}\right).$$

For the latter one can use e.g. formulas from [14].

Step 2. Next we check Theorem 3 for integers n with one or two prime factors. For this we need to combine Lemmas 2, 3, and 4, involved computer search, and the following two lemmas are also needed. The first one verifies Theorem 3 if n is a prime power.

Lemma 6. *Let $p = p_i \geq 3$ be a prime. If $p \leq 67$ and $p^\alpha < n_0$, $\alpha > 0$ then $p^\alpha \in A_{i,n}$. In general, $n = p^\alpha \in A_{i,n}$ holds for $\alpha \leq 4$.*

The second lemma proves our theorem for n with two distinct prime divisors, where the smaller one is at most 53.

Lemma 7. *Let $p = p_i \geq 3$ and $q > p$ be primes. If $p \leq 53$ and $p^\alpha q^\beta < n_0, \alpha, \beta > 0$ then $p^\alpha q^\beta \in A_{i,n}$. In general, $n = pq \in A_{i,n}$ is valid whenever $q < p^3$.*

Step 3. Consider now numbers n with three distinct prime factors. Unfortunately, for such values of n we could not find any general assertion or formula like in Lemma 6 or 7. However, our previous calculations yielded that here we may assume that the smallest prime factor of n is at least 19. For each prime $29 \leq p \leq 43$ we computed all integers, which are divisible by p, lie below a preliminary computed bound, and have three distinct prime divisors, which are at least p. Then we used a variant of the wheel algorithm, see e.g. [17], to handle these cases. Altogether, up to this point we could cover all values of n whose smallest prime factor is at most 47.

Step 4. To cover the remaining values of n (which have only "large" prime factors), we applied again formulas concerning the distribution of primes. Namely, we used estimates for $\pi(x)$, from [14]. This finishes the proof of Theorem 3. □

We proved that applying Algorithm 1 for A_n $(n \geq 1)$, the outputs (i.e. the clusterings of the A_n) have a regular shape until a certain large value of n (in fact up to n_0), but at that point the regularity vanishes. From the proof it is clear that n_0 is the first, but not at all the last integer, which behaves in this irregular way. For example, the number $3n_0$ is odd and is divisible by 3, however, adjoining it to $A_{1,n}$ causes less conflicts than adjoining it to $A_{2,n}$. Let $A^*_{i,n}$ denote the class containing p_i, produced by Algorithm 1. We can neither guess the structure of these sets, nor what is their asymptotic behavior. For example, it would be interesting to know whether the limit

$$\lim_{n\to\infty}\frac{|A^*_{1,n}|}{n}$$

exists or not, or is

$$\limsup_{n\to\infty}\frac{|A^*_{1,n}|}{n} = 1$$

or not.

On the other hand Theorem 1 shade some light to the asymptotic behavior of the optimal correlation clustering of $(A_n, \sim)$. Indeed denote by $\underline{g}, g, \overline{g}$, and $\overline{\rho_j}, j \geq 1$ the quantities defined in Section II in the case $(A_n, \sim)$. The next theorem is new.

Theorem 4. *With the above notation we have*

$$\overline{\rho_1} \leq \sqrt{2\left(1-\frac{6}{\pi^2}\right)} = 0.885520071...$$

In particular the optimal correlation clustering of $(A_n, \sim)$ has at least two classes.

Proof. First we prove that in the actual case g exists, i.e., $\underline{g} = g = \overline{g}$. Indeed we have

$$g = \lim_{n\to\infty}\frac{a_n}{b_n},$$

where

$$a_n = \left|\{(a,b) \,:\, 1 \leq a \leq b \leq n, \gcd(a,b) > 1\}\right|,$$

and

$$b_n = |\{(a,b) \,:\, 1 \leq a \leq b \leq n\}| = \frac{n(n-1)}{2}.$$

Obviously

$$a_n = b_n - |\{(a,b) \,:\, 1 \leq a \leq b \leq n, \gcd(a,b) = 1\}|$$
$$= b_n - \sum_{d=1}^{n}\varphi(d),$$

where $\varphi(x)$ denotes Euler's totient function. It is well known, see e.g. [10], that

$$\sum_{d=1}^{n}\varphi(d) = \frac{3}{\pi^2}n^2 + O(n\log n).$$

Combining everything together we get

$$g = 1 - \frac{6}{\pi^2}.$$

By the second assertion of Theorem 1 we get

$$\underline{\rho}_1^2 \leq 2g,$$

which together with the last inequality implies the statement of the theorem. □

IV. CORRELATION CLUSTERING OF S-UNITS WITH RESPECT TO COPRIMALITY

In this section we perform a similar analysis as in Section III, with the same relation, but the set of positive integers replaced by the set of positive integers having all prime divisors in a preliminary fixed finite set. As it will turn out, this modification changes the structure of the optimal clustering drastically. All the results presented in this section are new.

To formulate our results in this direction, we need to introduce some notions and notation. In what follows, let $S = \{p_1, \dots, p_k\}$ be a finite set of primes. Those positive integers which has no prime divisors outside S will be called S-units, and their set is denoted by $\mathbb{Z}_S$. This terminology is widely used in number theory, but in computer science the elements of the set $\mathbb{Z}_S$ for $S = \{2, 3, 5\}$ are also called Hamming numbers, see e.g. [8]. For a given positive $x \in \mathbb{R}$, let $\mathbb{Z}_S(x)$ denote the subset of $\mathbb{Z}_S$ consisting of S-units not greater than x. The sets $\mathbb{Z}_S(n), n = 1, 2, \dots$ play the same role as A_n in the last section. First we give a sharp upper bound for the number of elements of $\mathbb{Z}_S(x)$ and some of its subsets. For this we need some preparation. The following result due to Davenport [7] will be very useful.

Lemma 8 ([7, Theorem])**.** *Let $\mathcal{R}$ be a closed bounded region in the n dimensional space $\mathbb{R}^n$ and let $\mathrm{N}(\mathcal{R})$ and $\mathrm{V}(\mathcal{R})$ denote the number of points with integral coordinates in $\mathcal{R}$ and the volume of $\mathcal{R}$, respectively. Suppose that:*

- *Any line parallel to one of the n coordinate axes intersects $\mathcal{R}$ in a set of points which, if not empty, consists of at most h intervals.*
- *The same is true (with m in place of n) for any of the m dimensional regions obtained by projecting $\mathcal{R}$ on one of the coordinate spaces defined by equating a selection of $n - m$ of the coordinates to zero; and this condition is satisfied for all m from 1 to $n - 1$.*

Then

$$|\mathrm{N}(\mathcal{R}) - \mathrm{V}(\mathcal{R})| \leq \sum_{m=0}^{n-1} h^{n-m} V_m,$$

where V_m is the sum of the m dimensional volumes of the projections of $\mathcal{R}$ on the various coordinate spaces obtained by equating any $n - m$ coordinates to zero, and $V_0 = 1$ by convention.

The next lemma will also play an important role later on.

Lemma 9. *Let $y_1, \dots, y_r, x$ be positive real numbers, and let $N(\mathbf{y}, x)$ denote the number of non-negative integer solutions $n_1, \dots, n_r$ of the inequality*

$$0 \leq y_1 n_1 + \cdots + y_r n_r \leq x. \tag{5}$$

Then we have

$$N(\mathbf{y}, x) = c(\mathbf{y}) x^r + O(x^{r-1}),$$

where $c(\mathbf{y})$ is the volume of the r-dimensional polyhedron defined by the inequalities

$$\begin{aligned} x_i &\geq 0 \ (i = 1, \dots, r), \\ y_1 x_1 + \cdots + y_r x_r &\leq 1. \end{aligned}$$

Proof. It is clear that the non-negative integers $n_1, \dots, n_r$ satisfy (5) if and only if the lattice point $(n_1, \dots, n_r)$ belongs to the polyhedron $P(\mathbf{y}, x)$ defined by the inequalities

$$\begin{aligned} x_i &\geq 0 \ (i = 1, \dots, r), \\ y_1 x_1 + \cdots + y_r x_r &\leq x. \end{aligned}$$

Hence it is sufficient to bound the number of lattice points inside $P(\mathbf{y}, x)$. Obviously, $P(\mathbf{y}, x)$ satisfies the conditions of Lemma 8. Moreover, the volume of $P(\mathbf{y}, x)$ equals x^r times the volume of $P(\mathbf{y}, 1)$. The domains V_m occurring in Lemma 8 are polyhedra of dimensions at most $r - 1$, hence their total volume can be bounded by $O(x^{r-1})$, and the statement follows. □

Now using Lemma 9 we can easily bound the number of elements of $\mathbb{Z}_S(x)$.

Corollary 4. *Letting $c_S = c(\log p_1, \dots, \log p_k, 1)$, we have*

$$|\mathbb{Z}_S(x)| = c_S (\log x)^k + O((\log x)^{k-1}).$$

Proof. A positive integer m belongs to $\mathbb{Z}_S(x)$ if and only if $m \leq x$ and there exist non-negative integers $n_1, \dots, n_k$ such that

$$m = p_1^{n_1} \cdots p_k^{n_k}.$$

This implies that

$$0 \leq \log m = n_1 \log(p_1) + \cdots + n_k \log(p_k) \leq \log x.$$

Since $\log p_1, \dots, \log p_k > 0$ are real numbers, the conditions of Lemma 9 are satisfied, and the statement follows. □

Now we shall investigate correlation clustering on $\mathbb{Z}_S$ equipped with the same coprimality relation which we used in Section III. More precisely, for $a, b \in \mathbb{Z}_S$ let $a \sim b$, if and only if $\gcd(a, b) > 1$ or $a = b = 1$. For an integer $n \geq 1$ let $\mathcal{P}$ be a partition of $\mathbb{Z}_S(n)$. As before, we say that the S-units a and b are in conflict with respect to $\mathcal{P}$, if either they are in the same class but $a \not\sim b$, or they are in different classes and still $a \sim b$. As before, the purpose of correlation clustering is to find a partition with minimal number of conflicts.

The partition of $\mathbb{Z}_S(n)$ with only one class (i.e. when all elements of $\mathbb{Z}_S(n)$ belong to the same class) is called the trivial partition.

Lemma 10. *Let $C_S(n)$ denote the number of conflicts in the trivial partition of $\mathbb{Z}_S(n)$. Then we have*

$$C_S(n) \leq c(\log n)^k,$$

where c is a positive constant.

Proof. In case of the trivial partition, two different S-units are in conflict precisely when they are coprime. Thus we have to count those pairs of S-units (a, b), $a \neq b$ which are coprime. Let a, b be distinct S-units (not necessarily elements of $\mathbb{Z}_S(n)$), and assume that $\gcd(a, b) = 1$. Then there exists a $T \subseteq S$, such that

$$a = \prod_{p_j \in T} p_j^{\alpha_j}, \quad b = \prod_{p_j \in S \setminus T} p_j^{\alpha_j}.$$

If $T = \emptyset$ or $S \setminus T = \emptyset$ then the empty product is defined as 1, i.e. the pairs (a, b) with $a = 1$ or $b = 1$ are included. These equations mean that a is a T-unit, and b is an $(S \setminus T)$-unit. Hence for fixed $T \subseteq S$, the $(S \setminus T)$-units up to n are the positive integers which are coprime to the T-units in $\mathbb{Z}_S(n)$. Thus

$$C_S(n) \leq \sum_{T \subseteq S} |\mathbb{Z}_T(n)| \cdot |\mathbb{Z}_{S \setminus T}(n)|.$$

Using Corollary 4 and $|T| + |S \setminus T| = k$, we obtain that

$$\begin{aligned} C_S(n) &\leq \sum_{T \subseteq S} c_T c_{S \setminus T} (\log n)^{|T|} (\log n)^{|S \setminus T|} + \\ &+ O((\log n)^{|T|+|S \setminus T|-1} = \\ &= \left(\sum_{T \subseteq S} c_T c_{S \setminus T} \right) (\log n)^k + O((\log n)^{k-1}). \end{aligned}$$

Clearly, for fixed S the sum

$$\sum_{T \subseteq S} c_T c_{S \setminus T}$$

is independent of n, hence it is a constant. This proves the statement. □

Denoting by $g = g(\mathbb{Z}_S)$ and $\underline{\rho}_1 = \underline{\rho}_1(\mathbb{Z}_S)$ the densities defined in Section II, in the actual case we obtain

Corollary 5. *We have*

$$g = \underline{\rho}_1 = 1.$$

Proof. Corollary 4 and Lemma 10 imply $g = 1$ immediately. Thus by Corollary 1 we have $\sum_j \underline{\rho}_j \geq 1$, which together with the obvious inequality $\sum_j \underline{\rho}_j \leq 1$ implies that any optimal clustering of $\mathbb{Z}_S(n)$ is full. Hence we have $\underline{\rho}_1 = 1$ by Corollary 2. □

The equality $\underline{\rho}_1 = 1$ does not mean that the optimal correlation clustering is asymptotically trivial, i.e. the optimal correlation clustering of $\mathbb{Z}_S(n)$ is trivial for all large enough n. Thus the statement of the next theorem is much sharper as that of the last corollary.

Theorem 5. *Suppose that n is large enough. Then the optimal correlation clustering of $\mathbb{Z}_S(n)$ is given by the trivial partition.*

Proof. If $k = 1$, i.e. S contains only one element, then all members of $\mathbb{Z}_S$ are divisible by it. Thus $\sim$ is an equivalence relation on $\mathbb{Z}_S$, which corresponds to the trivial partition. Hence the proof is complete for $k = 1$ and we may assume $k \geq 2$ in the sequel.

Let $1 \leq i \leq k$ arbitrary, and let $a \in \mathbb{Z}_S(n)$ be such that $p_i \mid a$. Denote by Z_a the set of elements $b \in \mathbb{Z}_S(n)$ with $a \neq b$, $\gcd(a, b) > 1$. It is clear that Z_a contains all elements of $\mathbb{Z}_S(n) \setminus \{a\}$, which are multiples of p_i. Thus the number of these elements can be bounded by the number of elements of the set obtained from $\mathbb{Z}_S(n)$ by omitting its elements which are not divisible by p_i. Observe that the latter set is just $\mathbb{Z}_{S \setminus \{p_i\}}(n)$. Thus by Corollary 4 we obtain

$$\begin{aligned} |Z_a| &\geq c_S (\log n)^k + O((\log n)^{k-1}) - \\ &- c_{S \setminus \{p_i\}} (\log n)^{k-1} + O((\log n)^{k-2}) \qquad (6) \\ &= c_S (\log n)^k + O((\log n)^{k-1}). \end{aligned}$$

Consider now an optimal clustering of $\mathbb{Z}_S(n)$, denoted by $\mathcal{P}$. In view of Lemma 10, the number of conflicts with respect to $\mathcal{P}$ is at most $c(\log n)^k$, where c is a positive constant.

Assume that the elements of Z_{p_i} are distributed in r different classes, which contain $m_1, \ldots, m_r$ elements of Z_{p_i}, respectively. As the elements divisible by p_i corresponding to different classes are in conflict, thus the total number of conflicts with respect to $\mathcal{P}$ is at least

$$V = \sum_{1 \leq j_1 < j_2 \leq r} m_{j_1} m_{j_2}.$$

Assume that

$$\max_{1 \leq j \leq r} m_j \leq (\log n)^{1/2}.$$

Then by (6) we have

$$r \geq c_S (\log n)^{k-1/2},$$

whence

$$V \geq \frac{r(r-1)}{2} > \frac{r^2}{4} \geq \frac{c_S^2}{4} (\log n)^{2k-1},$$

whose magnitude is larger than that of the conflicts of the trivial partition. This implies that for an optimal clustering

$$\max_{1 \leq j \leq r} m_j > (\log n)^{1/2}$$

holds.

Let now $d > (c+1)/c_S$, where c is the constant given in Lemma 10. Suppose that

$$(\log n)^{1/2} < \max_{1 \leq j \leq r} m_j < |Z_{p_i}| - d.$$

Assume further that the m_j take their maximum for $j = 1$. Then

$$\begin{aligned} V \geq m_1(m_2 + \cdots + m_r) &= m_1(m_1 + \cdots + m_r - m_1) \\ &= m_1(|Z_{p_i}| - m_1). \end{aligned}$$

The function $m_1(|Z_{p_i}| - m_1)$ can attain its minimum in the endpoints of the given interval. If $m_1 = (\log n)^{1/2}$, then by (6) we have

$$V \geq c_S (\log n)^{k+1/2},$$

while $m_1 = |Z_{p_i}| - d$ implies

$$V \geq d(|Z_{p_i}| - d) > (c+1)(\log n)^k + O((\log n)^{k-1}).$$

Hence we get that the number of conflicts in both endpoints are larger than that of the trivial partition.

This implies that there exists a class, say $\mathcal{P}_1$, in which the number of multiples of p_i is at least $|Z_{p_i}| - d$. Suppose that $d > 0$ and let a be such that $p_i | a$, but $a \notin \mathcal{P}_1$. Then a is in conflict with all the elements of $\mathcal{P}_1 \cap Z_{p_i}$, and the number of such elements is

$$|Z_{p_i}| - d = c_S (\log n)^k + O((\log n)^{k-1}).$$

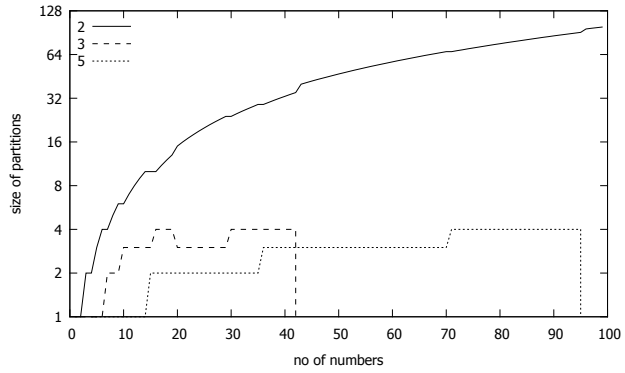

Fig. 1. Simulation result: from the 96^{th} Hamming numbers the correlation clustering gives only one cluster.

Moving a to $\mathcal{P}_1$, these conflicts will disappear. Of course, new conflicts can arise, on the one hand with those $d-1$ multiples of p_i which are not in $\mathcal{P}_1$, and on the other hand, with the elements not divisible by p_i. Altogether, the number of new conflicts can be at most

$$d + |\mathbb{Z}_{S\setminus\{p_i\}}(n)| = O((\log n)^{k-1}).$$

That is, if $\mathcal{P}$ is an optimal correlation clustering and if n is large enough, then all multiples of p_i must belong to the same class. Since p_i has been chosen arbitrarily, thus all elements must belong to the same class, and the theorem follows. □

Remark 2. *Theorem 5 describes completely the optimal correlation clustering of $\mathbb{Z}_S(n)$ only if n is large enough. Choosing S as the set of the first k primes, $\mathbb{Z}_S(n) = A_n$ if $n < p_{k+1}$, thus the optimal correlation clusterings of $\mathbb{Z}_S(n)$ and A_n are identical. By the results of Section III the number of clusters and their sizes of the optimal correlation clustering of $\mathbb{Z}_S(n)$ are growing for small n-s. After a certain point this tendency changes, all but one clusters become smaller until only one cluster survives, like a black hole. Our experiments show that this happens already for relatively small values of n, as you can see in Fig. 1 for the Hamming numbers.*

This means that locally optimal algorithms, like Algorithm 1, cannot give globally optimal solution for the correlation clustering problem.

V. CONCLUSION

In this paper we have considered the problem of correlation clustering, introduced in [3], from three different but closely related aspects. First we have derived new results for the graph model of the problem, considering an increasing family of graphs. We have obtained results concerning the optimal clustering in the general case. Then we have investigated particular sets with a specific relation. The reason for doing so is that clearly, the choice of the underlying relation strongly influences the structure of the optimal clustering. First we have considered positive integers and a relation based upon coprimality. Our main result here (recalled from [1]) has been that a natural greedy algorithm provides a "locally" optimal clustering up to a certain positive integer n_0, however, at some point the structure of such a clustering is deemed to change. Finally, we have considered the set of so-called S-units, under the same relation as for positive integers. Here we have proved that interestingly, in contrast with the case of positive integers, after some point the optimal clustering is always given by the trivial clustering (consisting of a single class).

ACKNOWLEDGMENT

We are grateful to the referees for their valuable remarks.

Our research was supported in part by the OTKA grants NK104208, NK101680, and K100339, by the Japanese Society for the Promotion of Science (JSPS), Grant in aid 24540012, and by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project is implemented through the New Hungary Development Plan, cofinanced by the European Social Fund and the European Regional Development Fund.

REFERENCES

[1] L. Aszalós, L. Hajdu, and A. Pethő, *On a correlational clustering of integers*, arXiv:1404.0904 [math.NT].
[2] M. Bakó, and L. Aszalós, *Combinatorial optimization methods for correlation clustering*, In: Coping with complexity/D. Dumitrescu, Rodica Ioana Lung, Ligia Cremene, Casa Cartii de Stiinta, Cluj-Napoca, 2–12, 2011.
[3] N. Bansal, A. Blum, and S. Chawla, *Correlational clustering*, Machine Learning, **56** (2004), 89–113.
[4] A. Bhattacharya, and R. K. De, *Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles.* Bioinformatics **24** 11. (2008): 1359–1366.
[5] Y. Chen, S. Sanghavi, and H. Xu, *Clustering sparse graphs.* Advances in neural information processing systems. (2012) 2204–2212.
[6] Z. Chen, S. Yang, L. Li, and Z. Xie, *A clustering approximation mechanism based on data spatial correlation in wireless sensor networks*, Wireless Telecommunications Symposium (WTS), 2010.
[7] H. Davenport, *On a principle of Lipschitz*, J. London Math. Soc. **26**, (1951). 179–183. *Corrigendum* ibid. **39** (1964), 580.
[8] E.W. Dijkstra, *A discipline of programming.* Vol. 4. Englewood Cliffs: Prentice-Hall, 1976. 129-133.
[9] T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan, *Improving recommendation accuracy by clustering social networks with trust.* Recommender Systems & the Social Web *532* (2009): 1-8.
[10] L.K. Hua, *Introduction to Number Theory*, Springer-Verlag, 1982.
[11] P. Kanani, and A. McCallum, *Resource-bounded information gathering for correlation clustering* In Computational Learning Theory 07, Open Problems Track, COLT 2007, 625-627, 2007.
[12] Z. Néda, F. Răzvan, M. Ravasz, A. Libál, and G. Györgyi, *Phase transition in an optimal clusterization model* Physica A: Statistical Mechanics and its Applications, **362** (2):357–368, 2006.
[13] J. E. Nymann, *On the probability that k positive integers are relatively prime*, J. Number Theory **4** (1972), 469–473.
[14] J. B. Rosser, and L. Schoenfeld, *Approximate formulas for some functions of prime numbers*, Illinois J. Math. **6** (1962), 64–94.
[15] K. Sungwoong, S. Nowozin, P. Kohli, and D. Y. Chang, *Higher-Order Correlation Clustering for Image Segmentation*, In: Advances in Neural Information Processing Systems 24. J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, Curran Associates, Inc. 1530–1538, 2011.
[16] B. Yang, W. K. Cheung, and J. Liu, *Community mining from signed social networks*, IEEE Transactions on Knowledge and Data Engineering **19** 10 (2007): 1333-1348.
[17] H. C. Williams, *Primality testing on a computer*, Ars Combin. **5** (1978), 127–185.

Shigeki Akiyama is a professor at Institute of Mathematics, University of Tsukuba, Japan. He got PhD degree in mathematics from Kobe University. His research interest is the interplay between number theory and ergodic theory. He thinks clustering looks quite similar to how we understand our nature by language. You can contact him: akiyama@math.tsukuba.ac.jp

László Aszalós is a senior lecturer and researcher at the University of Debrecen, member of the Computer Science Department since 1997. He received his PhD at 2002 at this university. His research interests belong to AI: automated theorem proving, multi-modal logics, optimization and rough set theory. Recently he examine the correlation clustering: its near-optimal solutions and applications in data mining.You can contact him: aszalos.laszlo@inf.unideb.hu.

Lajos Hajdu received his MSc in Mathematics from the Lajos Kossuth University, Hungary, in 1992. He obtained his PhD degree in Mathematics from the Lajos Kossuth University, Hungary, in 1998. He worked as a Post Doc researcher for the Mathematical Institute of Leiden University in 1999-2000. From 2000 he served as Assistant Lecturer, from 2003 as Assistant Professor and since 2012 he has been a Full Professor at the Institute of Mathematics, University of Debrecen. He is a member of the János Bolyai Mathematical Society, the Public Body of the Hungarian Academy of Sciences, and the Mathematical Committee of the Mathematical Division of the Hungarian Academy of Sciences. He has authored or co-authored 80 journal papers and 10 conference papers. His main interest lies in Diophantine number theory, in discrete tomography and in discrete mathematics with applications in digital image processing. His email address is hajdul@science.unideb.hu.

Attila Pethő is a professor of the Department of Computer Science, Faculty of Informatics, University of Debrecen, Hungary. He got PhD degree in mathematics from the Lajos Kossuth University. He is a corresponding member of the Hungarian Academy of Sciences. His research interest are number theory and cryptography. You can contact him: petho.attila@inf.unideb.hu.