ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Indagationes Mathematicae ▮ (▮▮▮▮) ▮▮▮–▮▮▮

indagationes mathematicae

www.elsevier.com/locate/indag

# On a correlational clustering of integers☆

László Aszalós[b], Lajos Hajdu[a,*], Attila Pethő[b]

[a] *Institute of Mathematics, University of Debrecen, H-4010 Debrecen, P.O. Box 12, Hungary*
[b] *Department of Computer Science, University of Debrecen, H-4010 Debrecen, P.O. Box 12, Hungary*

Received 17 March 2015; received in revised form 17 September 2015; accepted 22 September 2015

Communicated by R. Tijdeman

**Abstract**

Let $A$ be a finite set, and let a symmetric binary relation be given on $A$. The goal of correlation clustering is to find a partition of $A$, with minimal conflicts with respect to the relation given. In this paper we investigate correlation clustering of subsets of the positive integers, based upon a relation defined by the help of the greatest common divisor.
© 2015 Published by Elsevier B.V. on behalf of Royal Dutch Mathematical Society (KWG).

*Keywords:* Greatest common divisor; Pairs of coprime integers; Correlation clustering; Greedy algorithm

## 1. Introduction and setting the problem

Correlation clustering is a concept originated in machine learning. It was introduced in Bansal et al. [2]; that paper gives a good overview of the mathematical background of the topic, as well. In [2] the problem was introduced through a graph model. Here we use an equivalent formulation, which is more appropriate for our purposes.

Let $A$ be a finite non-empty set, and let $\sim$ be a symmetric binary relation on $A$. Consider a partition $\mathbb{P}$ of $A$. Two distinct elements $a, b \in A$ are in conflict with respect to the partition $\mathbb{P}$ either if they belong to the same class of $\mathbb{P}$, but $a \nsim b$, or they belong to different classes of $\mathbb{P}$, although

☆ Research supported in part by the OTKA grants NK104208, NK101680, K100339 and K115479.
* Corresponding author.
*E-mail addresses:* Aszalos.Laszlo@inf.unideb.hu (L. Aszalós), hajdul@science.unideb.hu (L. Hajdu), Petho.Attila@inf.unideb.hu (A. Pethő).

http://dx.doi.org/10.1016/j.indag.2015.09.004
0019-3577/© 2015 Published by Elsevier B.V. on behalf of Royal Dutch Mathematical Society (KWG).

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

$a \sim b$. The goal of correlation clustering is to find a partition with minimal number of conflicts. Note the special feature of this clustering that the number of clusters is not specified in advance.

As one would expect, the structure of an optimal clustering should heavily depend on the relation $\sim$ defined on $A$. Note that in particular, if we assume that $\sim$ is transitive, then we may consider $\sim$ to be an equivalence relation on $A$. (Since the reflexive property does not have any effect on the number of conflicts.) Then the partition induced by $\sim$ is clearly an optimal clustering. The situation is much more interesting (and important) if $\sim$ is not transitive.

Being motivated by the above remarks, in this paper we work with subsets of the set of positive integers, and we choose $\sim$ to be a relation based upon the greatest common divisor. More precisely, for $n \geq 2$ let $A_n$ be the set of positive integers between 2 and $n$, and for $a, b \in A_n$ with $a \neq b$ let $a \sim b$ if and only if $\gcd(a, b) > 1$. (Obviously, it would have no point to include 1 into $A_n$.) Note that the behavior of the gcd among the first $n$ positive integers has been investigated from many aspects; see e.g. the paper of Nymann [3].

Bakó and Aszalós [1] have made several experiments on the set $A_n$ under $\sim$. They discovered that the classes of a certain "locally optimal" clustering have regular structures. In the sequel denote by $p_i$ the $i$th prime, i.e., $p_1 = 2, p_2 = 3, \ldots$. Set

$$S_{i,n} = \{m : 2 \leq m \leq n,\ p_i | m,\ p_j \nmid m\ (j < i)\}.$$

That is, $S_{i,n}$ is the set of integers between 2 and $n$, which are divisible by $p_i$, but coprime to the smaller primes.

**Remark 1.** For any indices $i$ and $i'$ with $1 \leq i < i'$ we have $|S_{i,n}| \geq |S_{i',n}|$, for every $n$ with $n \geq 2$. Indeed, if $m \in S_{i',n}$, then we can write $m = tp_{i'}^{\alpha}$ with some positive integers $t$ and $\alpha$ such that $p_j \nmid t$ for $1 \leq j \leq i'$. Since $p_i < p_{i'}$, by $m \leq n$ we have that $tp_i^{\alpha} \in S_{i,n}$. Thus the mapping $f : S_{i',n} \to S_{i,n}$ defined by $f(tp_{i'}^{\alpha}) = tp_i^{\alpha}$ is an injection. Hence the assertion follows.

Bakó and Aszalós found that

$$A_n = \bigcup_{j=1}^{\infty} S_{j,n} \tag{1}$$

is highly likely to be an optimal correlation clustering for $n \leq 500$. For brevity, here and later on we use the symbol $\cup$ for the standard union of sets, while $\bigcup$ is used for building clusterings from sets. That is, $A_n$ above is a set whose elements are the sets $S_{j,n}$. Notice that $S_{j,n} = \emptyset$ for all $j$ large enough, so $A_n$ is actually finite.

The aim of this paper is to show that for $n_0 = 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 = 111\,546\,435$ the decomposition (1) is not optimal. We prove that the number of conflicts in

$$A_n = (S_{1,n_0} \cup \{n_0\}) \bigcup (S_{2,n_0} \setminus \{n_0\}) \bigcup_{j=3}^{\infty} S_{j,n} \tag{2}$$

is less than in (1) with $n = n_0$. Unfortunately, we are not able to verify that (1) is optimal for $n < n_0$. However, we can prove that a natural greedy algorithm (Algorithm 1), presented in the next section, produces the clustering (1) for $n < n_0$, but (2) for $n = n_0$. Thus our results shed some light on the difficulty to find optimal clusterings of large sets.

Applying Algorithm 1 for $A_n$ and $\sim$, the results behave regularly until a certain large point, but then the regularity disappears. From our construction it will be clear that $n_0$ is the first, but not at all the last integer, which behaves irregularly in this sense. For example the numbers

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

$3n_0, 5n_0, 9n_0, \ldots$ are odd and are divisible by 3, but adjoining them to $S_{1,n}$ causes less conflicts than adjoining them to $S_{2,n}$, with $n = 3n_0 - 1, 5n_0 - 1, 9n_0 - 1$, respectively. Denote by $S^*_{i,n}$ the class, which contains $p_i$ and is produced by Algorithm 1. We have no idea whether these sets have some structure and what is their asymptotic behavior. For example, we do not know whether the limit $\lim_{n\to\infty} |S^*_{1,n}|/n$ exists, and if so, whether $\lim_{n\to\infty} |S^*_{1,n}|/n = 1$.

The paper is organized as follows. In Section 2 we present Algorithm 1 and the main theorem. Section 3 is devoted to the proof of combinatorial lemmata and in the last section we prove the theorem. In our proofs, besides certain combinatorial considerations, we apply arguments and estimates from prime number theory, e.g. we use bounds of Rosser and Schoenfeld [4].

## 2. Main result

Since the number of partitions of $n$ elements grows exponentially, it is not surprising that to find an optimal correlational clustering is an NP-hard problem (see [2]). To find an approximation of the optimal solution, it is natural to use some greedy algorithm. Working with $A_n = \{2, 3, \ldots, n\}$ and $\sim$, we use the following strategy. The optimal clustering for $\{2\}$ is itself. Assume that we have a partition of $A_{n-1}$ ($n > 2$), and adjoin $n$ to that class, which causes the least number of new conflicts. The result is a "locally optimal" clustering, which is not necessarily optimal globally. We formulate this method as Algorithm 1.

**Algorithm 1** Natural greedy algorithm

**Require:** an integer $n \geq 2$
**Ensure:** a partition $\mathcal{P}$ of $A_n$

$\mathcal{P} \leftarrow \{\{2\}\}$;
**if** $n = 2$ **then return** $\mathcal{P}$
**end if**
$m \leftarrow 3$
**while** $m \leq n$ **do**
  $\mathcal{P}_M \leftarrow \mathcal{P} \cup \{\{m\}\}$
  $M \leftarrow \text{CONFLICTS}(\mathcal{P}_M, m)$ ▷ the number of conflicts with respect to the partition $\mathcal{P}_M$ caused by the pairs $(m, a)$, $a < m$
  $C \leftarrow$ number of classes in $\mathcal{P}$
  $j \leftarrow 1$
  **while** $j \leq C$ **do**
    $O \leftarrow \text{OP}(j, \mathcal{P})$ ▷ $\text{OP}(j, \mathcal{P})$ denotes the $j$th class in the partition $\mathcal{P}$.
    $\mathcal{P}_1 \leftarrow \mathcal{P} \setminus \{O\}$
    $\mathcal{P}_1 \leftarrow \mathcal{P}_1 \cup \{O \cup \{m\}\}$
    $M_1 \leftarrow \text{NUPAIR}(\mathcal{P}_1, m)$ ▷ the number of pairs $(m, a)$ with $a < m$ causing a conflict in the partition $\mathcal{P}_1$
    **if** $M_1 < M$ **then**
      $M \leftarrow M_1$
      $\mathcal{P}_M \leftarrow \mathcal{P}_1$
    **end if**
  **end while**
**end while**
**return** $\mathcal{P}_M$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

Starting with a partition of $A_{n-1}$ this algorithm establishes a partition of $A_n$ such that the conflicts caused by $n$ is minimal. The output of Algorithm 1 on the input $n$ is denoted by $G(n)$. It is a partition of $A_n$. It is easy to see that

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

Moreover, one can also check that these partitions are optimal clusterings of $A_n$ for $n = 2, 3, \ldots, 15$.

Our main result is the following.

**Theorem 1.** *If* $n < n_0 = 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 = 111\,546\,435$ *then*

$$G(n) = \bigcup_{j=1}^{\infty} S_{j,n} \tag{3}$$

*holds. However, we have*

$$G(n_0) = (S_{1,n_0} \cup \{n_0\}) \bigcup (S_{2,n_0} \setminus \{n_0\}) \bigcup_{j=3}^{\infty} S_{j,n_0}. \tag{4}$$

We shall prove Theorem 1 inductively. We explain our method of proof in the beginning of Section 4.

## 3. Auxiliary results

To prove the main theorem we need some preparation. Throughout this paper the number of elements of a set $A$ will be denoted by $|A|$. In the first lemma we characterize that class of $G(n-1)$ to which Algorithm 1 adjoins $n$.

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

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

*and*

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

*Define* $U_{M+1} = V_{M+1} = \emptyset$. *Let* $J$ *be the smallest index for which the difference* $|U_j| - |V_j|$ $(j = 1, \ldots, M+1)$ *is maximal. Then* $G(n) = \{P'_1, \ldots, P'_{M+1}\}$ *such that*

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

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

**Proof.** Let $K_j$ denote the number of new conflicts, which arise adjoining $n$ to $P_j$ $(j = 1, \ldots, M+1)$. Then

$$K_j = |U_j| + \sum_{\substack{k=1 \\ k \neq j}}^{M+1} |V_k|.$$

Algorithm 1 adjoins $n$ to that $P_{\hat{J}}$ for which $K_{\hat{J}}$ is minimal and if there are more indices $j$ with minimal $K_j$ then $\hat{J}$ is minimal among them. This is equivalent to

$$|V_{\hat{J}}| - |U_{\hat{J}}| \geq |V_m| - |U_m| \quad (m = 1, \ldots, M+1).$$

Thus $|V_m| - |U_m|$ $(m = 1, \ldots, M+1)$ assumes its maximal value at $m = \hat{J}$ with $\hat{J} \leq m$ if the equality sign holds. Hence $J = \hat{J}$ and the lemma is proved. □

For positive integers $j$ and $n \geq 2$ put

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

and

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

Note that by the definition of $S_{j,n-1}$, the sets $B_{j,n}$ and $E_{j,n}$ are subsets of $\{1, \ldots, n-1\}$ for all $j$ and $n$. The elements of $B_{j,n}$ and $E_{j,n}$ are called the friends and enemies of $n$ in $S_{j,n-1}$, respectively.

**Corollary 1.** *Let $n \geq 2$. If $n > 2$, then suppose that for all $i = 1, \ldots, M$ we have $S^*_{i,n-1} = S_{i,n-1}$, that is*

$$G(n-1) = \{S_{1,n-1}, \ldots, S_{M,n-1}\}.$$

*Here $p_M$ is the largest prime $\leq n-1$. Then the following assertions are true.*

(i) *Algorithm 1 adjoins $n$ to that $S_J$ with minimal $J$ for which $|B_{J,n}| - |E_{J,n}| \geq |B_{j,n}| - |E_{j,n}|$ $(j = 1, \ldots, M+1)$.*
(ii) *If $n$ is even, then $S^*_{1,n} = S_{1,n}$.*
(iii) *If $n$ is a prime, then $\{n\} \in G(n)$.*
(iv) *Let $p_i$ be the smallest prime factor of $n$. Then $n \in S^*_{j,n}$ implies $j \leq i$.*

**Proof.** The assertions hold for $n = 2$. For $n > 2$ by our assumption we can write

$$G(n-1) = \{S_1, \ldots, S_M\},$$

where, for simplicity, we set $S_j = S_{j,n-1}$ $(j = 1, \ldots, M)$. Put $S_{M+1} = \emptyset$.

(i) Observe that now the sets $U_j$ and $V_j$ defined in the proof of Lemma 1 coincide with $E_{j,n}$ and $B_{j,n}$, respectively. Hence the statement immediately follows.

(ii) If $n$ is even with $n > 2$, then $B_{1,n} = S_1$, thus $|B_{1,n}| \geq n/2 - 1$. If $2 \leq j \leq M$ then $|B_{j,n}| \leq [(n-1)/p_j] < n/3$. As $n/2 - 1 > n/3$ for $n \geq 8$ we have

$$|B_{1,n}| - |E_{1,n}| > |B_{j,n}| - |E_{j,n}| \quad (j = 2, \ldots, M+1).$$

Hence Algorithm 1 adjoins $n$ to $S_1$, i.e., to the class of even numbers.

(iii) Let $n$ be an odd prime. By part (i) of Corollary 1 Algorithm 1 adjoins $n$ to that $S_J$ with minimal $J$ for which $|B_{j,n}| - |E_{j,n}|$ $(j = 1, \ldots, M+1)$ is maximal.

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

Since $n$ is a prime, $B_{j,n} = \emptyset$ for all $j = 1, \dots, M$. Thus $|B_{j,n}| - |E_{j,n}| < 0$ $(j = 1, \dots, M)$, but $|B_{M+1,n}| - |E_{M+1,n}| = 0$. Hence $n$ will be adjoined to the empty set, so it will form a class in $G(n)$ alone.

(iv) We may assume that $n$ is odd and composite. Let $n = q_1^{\alpha_1} \cdots q_t^{\alpha_t}$, where $q_1 < \cdots < q_t$ are odd primes and $\alpha_1, \dots, \alpha_t$ are positive integers. We obviously have $\{q_1, \dots, q_t\} \subseteq \{p_1, \dots, p_M\}$.

Suppose that $q_1 = p_i$. Then, since we assumed that $S^*_{i,n-1} = S_{i,n-1} = S_i$, we have $B_{i,n} = S_i$ and $E_{i,n} = \emptyset$. Hence $|B_{i,n}| - |E_{i,n}| = |S_i|$.

If $j > i$ then $|B_{j,n}| - |E_{j,n}| \le |S_j| \le |S_i|$. (The latter inequality follows from Remark 1 after the definition of the $S_{i,n}$.) Thus, by part (iv) of Corollary 1, if $n$ will be adjoined to $S_j$ then $j \le i$. □

The next lemma describes a simple, but useful property of the integer part function.

**Lemma 2.** *Let $q_1, \dots, q_t$ be pairwise different odd primes, $\alpha_1, \dots, \alpha_t$ positive integers. Let $u$ be a positive integer coprime to $q_i$ $(i = 1, \dots, t)$ and $n = q_1^{\alpha_1} \cdots q_t^{\alpha_t}$. If $\{i_1, \dots, i_k\} \subseteq \{1, \dots, t\}$ then*

$$\left[\frac{n-1}{u q_{i_1} \cdots q_{i_k}}\right] = \left[\frac{\frac{n}{q_{i_1} \cdots q_{i_k}} - 1}{u}\right]. \tag{5}$$

*In particular, if $u = 2$ then*

$$\left[\frac{n-1}{2 q_{i_1} \cdots q_{i_k}}\right] = \frac{n - q_{i_1} \cdots q_{i_k}}{2 q_{i_1} \cdots q_{i_k}} = \frac{n}{2 q_{i_1} \cdots q_{i_k}} - \frac{1}{2}. \tag{6}$$

**Proof.** We have

$$\left[\frac{n-1}{u q_{i_1} \cdots q_{i_k}}\right] = \frac{n-m}{u q_{i_1} \cdots q_{i_k}},$$

where $m$ is the smallest positive integer such that the fraction on the right hand side is an integer. As $q_{i_j} | n$, we must have $q_{i_j} | m$ $(j = 1, \dots, k)$ as well, which implies $q_{i_1} \cdots q_{i_k} | m$. This proves (5).

If $u = 2$ then $m = q_{i_1} \cdots q_{i_k}$ is the smallest positive integer with the required property, because $n - m$ is even. □

The next lemma gives a good approximation for the size of $S_{i,u}$.

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

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

**Proof.** The first statement is obvious. To prove the second one we start with the identity

$$\begin{aligned} S_{i,u} &= \{m : m \le u,\ p_i | m,\ p_\ell \nmid m \text{ for all } 1 \le \ell < i\} \\ &= \{m : m \le u, p_i | m\} \setminus \bigcup_{\ell=1}^{i-1} \{m : m \le u, p_i \cdot p_\ell | m\}. \end{aligned}$$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

In the rest of the proof we assume that the elements of the occurring sets are at most $u$. The law of inclusion and exclusion implies

$$|S_{i,u}| = \sum_{\ell=0}^{i-1}(-1)^{\ell} \sum_{1\le i_1<\cdots<i_\ell<i} |\{m : p_i \cdot p_{i_1}\cdots p_{i_\ell} | m\}|.$$

Thus

$$|S_{i,u}| = \sum_{\ell=0}^{i-1}(-1)^{\ell} \sum_{1\le i_1<\cdots<i_\ell<i} \left[\frac{u}{p_i \cdot p_{i_1}\cdots p_{i_\ell}}\right]. \tag{7}$$

Using $x - 1 < [x] \le x$ we obtain

$$-\sum_{\substack{\ell=0\\ \ell \text{ is even}}}^{i-1} \binom{i-1}{\ell} \le |S_{i,u}| - \frac{u}{p_i}\prod_{\ell=0}^{i-1}\left(1-\frac{1}{p_\ell}\right) < \sum_{\substack{\ell=0\\ \ell \text{ is odd}}}^{i-1} \binom{i-1}{\ell}.$$

As

$$\sum_{\substack{\ell=0\\ \ell \text{ is even}}}^{i-1} \binom{i-1}{\ell} = \sum_{\substack{\ell=0\\ \ell \text{ is odd}}}^{i-1} \binom{i-1}{\ell} = 2^{i-2},$$

the lemma is proved. □

In the next lemma we prove an estimate for $|B_{j,n}| - |E_{j,n}|$.

**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 $j \ge 2$ be such that $p_j < q_1$. Then*

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

**Proof.** As $|B_{j,n}| + |E_{j,n}| = |S_{j,n-1}|$ we have

$$|B_{j,n}| - |E_{j,n}| = 2|B_{j,n}| - |S_{j,n-1}| = |S_{j,n-1}| - 2(|S_{j,n-1}| - |B_{j,n}|). \tag{8}$$

For $|S_{j,n-1}|$ we can use the estimations of Lemma 3, thus we have to deal only with the second summand. As $p_j < q$ for all prime factors $q$ of $n$, we have

$$B_{j,n} = \bigcup_{\ell=1}^{t}\{m : m \in S_{j,n-1}, q_\ell | m\}.$$

Using again the law of inclusion and exclusion we get

$$|B_{j,n}| = \sum_{\ell=1}^{t}(-1)^{\ell-1} \sum_{1\le j_1<\cdots<j_\ell\le t} |\{m : m \in S_{j,n-1}, q_{j_1}\cdots q_{j_\ell} | m\}|.$$

Set $U_{j,\ell}(q_{j_1}, \ldots, q_{j_\ell}) = U_{j,\ell} = \{m : m \in S_{j,n-1}, q_{j_1}\cdots q_{j_\ell} | m\}$. Then

$$U_{j,\ell} = \{m : m \le n-1, p_j \cdot q_{j_1}\cdots q_{j_\ell} | m\} \setminus \bigcup_{i=1}^{j-1}\{m : p_i p_j q_{j_1}\cdots q_{j_\ell} | m\}.$$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

For the number of elements of $U_{j,\ell}$ by the law of inclusion and exclusion we obtain

$$|U_{j,\ell}| = \sum_{i=0}^{j-1}(-1)^i \sum_{1\le h_1<\cdots<h_i<j} \left[\frac{n-1}{p_j \cdot q_{j_1}\cdots q_{j_\ell} p_{h_1}\cdots p_{h_i}}\right].$$

Combining these formulae we get

$$|B_{j,n}| = \sum_{\ell=1}^{t}(-1)^{\ell-1} \sum_{1\le j_1<\cdots<j_\ell\le t} \sum_{i=0}^{j-1}(-1)^i \sum_{1\le h_1<\cdots<h_i<j} \left[\frac{n-1}{p_j \cdot q_{j_1}\cdots q_{j_\ell} p_{h_1}\cdots p_{h_i}}\right].$$

The last formula together with (7) implies

$$|S_{j,n-1}| - |B_{j,n}| = \sum_{\ell=0}^{t}(-1)^{\ell} \sum_{1\le j_1<\cdots<j_\ell\le t} \sum_{i=0}^{j-1}(-1)^i \times \sum_{1\le h_1<\cdots<h_i<j} \left[\frac{n-1}{p_j \cdot q_{j_1}\cdots q_{j_\ell} p_{h_1}\cdots p_{h_i}}\right].$$

Changing the order of the summation we get

$$|S_{j,n-1}| - |B_{j,n}| = \sum_{i=0}^{j-1}(-1)^i \sum_{1\le h_1<\cdots<h_i<j} \sum_{\ell=0}^{t}(-1)^{\ell} \times \sum_{1\le j_1<\cdots<j_\ell\le t} \left[\frac{n-1}{p_j \cdot q_{j_1}\cdots q_{j_\ell} p_{h_1}\cdots p_{h_i}}\right].$$

Put

$$C_1 := \sum_{\ell=0}^{t}(-1)^{\ell} \sum_{1\le j_1<\cdots<j_\ell\le t} \left[\frac{n-1}{p_j \cdot q_{j_1}\cdots q_{j_\ell} p_{h_1}\cdots p_{h_i}}\right] - \frac{n-1}{p_j \cdot p_{h_1}\cdots p_{h_i}} \prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)$$

and observe that $|C_1| \le 2^{t-1}$. We can write

$$|S_{j,n-1}| - |B_{j,n}| = \frac{n-1}{p_j}\prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)\prod_{\ell=1}^{j-1}\left(1-\frac{1}{p_\ell}\right) + C_2,$$

where

$$C_2 = \sum_{i=0}^{j-1}(-1)^i \sum_{1\le h_1<\cdots<h_i<j} C_1.$$

Hence $|C_2| \le 2^{t+j-2}$. This together with Lemma 3 and (8) gives

$$|B_{j,n}| - |E_{j,n}| = \frac{n-1}{p_j}\prod_{\ell=1}^{j-1}\left(1-\frac{1}{p_\ell}\right) - 2\frac{n-1}{p_j}\prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)\prod_{\ell=1}^{j-1}\left(1-\frac{1}{p_\ell}\right) + C_3,$$

where $|C_3| \le 2^{t+j-1} + 2^{j-2}$. Thus the statement follows. □

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

The next lemma plays a key role in the proof of Theorem 1. In contrast to the classes of odd numbers, it is possible to give the exact values of the difference of the number of friends and enemies of $n$ in $S_{1,n-1}$.

**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*

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

**Proof.** The statement could be proved by repeating the proof of Lemma 4 and using Lemma 2. However, there is a much more direct and simple way, which we present.

Let $H_1 = \{h : 1 \le h \le \frac{n-1}{2}, \gcd(h, n) = 1\}$ and $H_2 = \{h : \frac{n+1}{2} \le h < n, \gcd(h, n) = 1\}$. Then $H_1$ and $H_2$ are disjoint and their union is $H = \{h : 1 \le h < n, \gcd(h, n) = 1\}$. Plainly $|H| = \varphi(n)$. The mapping $\psi : h \mapsto n - h$ is bijective between $H_1$ and $H_2$. Moreover, $\psi(h)$ is odd if and only if $h$ is even. Thus the number of even positive integers, which are coprime to $n$ is $\varphi(n)/2$. As $E_{1,n}$ is exactly the set of even integers, less than and coprime to $n$, the proof is complete. □

## 4. Proof of Theorem 1

Despite of the lengthy preparation, the proof of Theorem 1 is complicated. The hard part is to prove that (3) is true for $n < n_0$. This is done by a combination of comparison of the estimates of Lemmata 3 and 4, some computer search and finally application of a tool from prime number theory.

To prove Theorem 1, we apply induction. We shall always assume, without any further mentioning, that $n$ is a positive integer with $n \le n_0$, and that Theorem 1 holds for all $m$ with $2 \le m < n$. (Note that the theorem is valid for $n = 2$.) That is, assuming that for all such $m$ we have

$$G(m) = \{S_{1,m}, \ldots, S_{u,m}\},$$

we prove that $S^*_{i,n} = S_{i,n}$ for all $i$ too, i.e.

$$G(n) = \{S_{1,n}, \ldots, S_{v,n}\}$$

is also valid (with the appropriate $u$ and $v$). In many cases it will be sufficient to use the induction hypothesis only for $m = n - 1$.

By part (ii) of Corollary 1, Theorem 1 is true for even $n$. In case of odd $n$, we start with the easy part, by showing that Theorem 1 holds if $3 \mid n$. Hence, in particular, we check (4) in the next subsection.

### 4.1. The case where $n$ is odd and $3 \mid n$

Suppose that the smallest prime factor of $n$ is $q_1 = 3$. Then by Lemma 5 we have $|B_{1,n}| - |E_{1,n}| = \frac{n-1}{2} - \varphi(n)$. A simple computation shows that $|S_2| = \frac{n-3}{6}$.

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

By part (i) of Corollary 1, $n$ is adjoined to $S_1$ precisely when $|S_2| \leq |B_{1,n}| - |E_{1,n}|$. This inequality implies $\frac{n-3}{6} \leq \frac{n-1}{2} - \varphi(n)$, which is equivalent to $\varphi(n) < \frac{n}{3}$. Using the explicit form of $\varphi(n)$ and dividing by $n$ we get $\left(1-\frac{1}{3}\right)\cdot\left(1-\frac{1}{q_2}\right)\cdots\left(1-\frac{1}{q_t}\right) \leq \frac{1}{3}$. Thus Algorithm 1 adjoins $n$ to $S_1$ if and only if

$$\left(1-\frac{1}{q_2}\right)\cdots\left(1-\frac{1}{q_t}\right) \leq \frac{1}{2}. \tag{9}$$

Inequality (9) is independent of the exponents $\alpha_1, \ldots, \alpha_t$, and for fixed $t$ the left hand side of (9) is minimal if $q_1, \ldots, q_t$ are consecutive odd primes starting with $q_1 = 3$. Using these observations, a simple computation shows that (9) does not hold for odd $n < n_0$ such that $3 \mid n$, however, $n_0$ satisfies (9).

Hence the induction step is proved for odd $n$ with $3 \mid n$.

**Remark 2.** An analogous computation shows that $n_1 = 5 \cdot p_4 \cdots p_{14} = 2\,180\,460\,221\,945\,005$ is a candidate to be the smallest odd integer, which is not divisible by 3 and is adjoined to $S_1$. However, as $n_1$ is much larger than $n_0$ and many odd integers between $n_0$ and $n_1$, e.g. $3n_0, 5n_0, 9n_0, \ldots$ are adjoined to $S_{1,3n_0-1}$, $S_{1,5n_0-1}$, $S_{1,9n_0-1}$, respectively, we are not sure whether for example $n_1' = 5 \cdot p_4 \cdots p_{13}$ will belong to $S^*_{1,n_1'}$, $S^*_{2,n_1'}$ or $S^*_{3,n_1'}$.

### 4.2. The case where $n$ has middle sized prime factors

Let $n < n_0$ and denote by $p$ and $t$ the smallest prime divisor of $n$ and the number of distinct prime divisors of $n$, respectively. Note that if $p \geq 37$ then since $37 \cdot 41 \cdot 43 \cdot 47 \cdot 53 > n_0$, the number of distinct prime factors of $n$ is at most four. In this subsection we prove the induction step for the cases where

- $p \leq 11$ and $t \geq 3$ or
- $p \leq 37$ and $t \geq 4$.

By our remark above, in the second case this establishes the validity of the induction step for all $n$ with $t \geq 5$.

Let $n = q_1^{\alpha_1} \cdots q_t^{\alpha_t} < n_0$ be such that $p_i = q_1 < q_2 \cdots < q_t$. For $i \leq 2$ the induction step is already proved, so we may assume that $i \geq 3$. By part (i) of Corollary 1, Algorithm 1 adjoins $n$ to $S_{i,n-1}$ if and only if

$$|B_{j,n}| - |E_{j,n}| < |S_{i,n-1}| \tag{10}$$

holds for all $1 \leq j < i$. (In view of part (iv) of Corollary 1, $n$ cannot be adjoined to $|S_{i',n-1}|$ with $i' > i$.) By Lemmata 3 and 4 we have

$$|S_{i,n-1}| \geq \frac{n-1}{p_i}\prod_{\ell=1}^{i-1}\left(1-\frac{1}{p_\ell}\right) - 2^{i-2}$$

and

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

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

Thus if

$$\frac{n-1}{p_i}\prod_{\ell=1}^{i-1}\left(1-\frac{1}{p_\ell}\right)+\frac{n-1}{p_j}\prod_{\ell=1}^{j-1}\left(1-\frac{1}{p_\ell}\right)\times\left(2\prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)-1\right)>2^{t+j-1}+2^{j-2}+2^{i-2}$$

then (10) holds. For fixed $t$ and $i$ the product $\prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)$ assumes its smallest value if the $q_k$-s are the $t$ consecutive primes starting with $p_i$. Thus if $n_1=n_1(i,j,t)$ denotes the smallest $n$ satisfying

$$n>1+(2^{t+j-1}+2^{j-2}+2^{i-2})/T$$

where

$$T=\frac{1}{p_i}\prod_{\ell=1}^{i-1}\left(1-\frac{1}{p_\ell}\right)+\frac{1}{p_j}\prod_{\ell=1}^{j-1}\left(1-\frac{1}{p_\ell}\right)\left(2\prod_{k=0}^{t-1}\left(1-\frac{1}{p_{i+k}}\right)-1\right),$$

then (10) holds for all $n\ge n_1$ having exactly $t$ distinct prime factors, of which the smallest is $p_i$.

We computed $n_1(i,j,t)$ for all triplets $(i,j,t)$ with $3\le i\le 19, 3\le j\le i-1, 1\le t\le t_i$, where $t_i$ is the largest $t$ such that $\prod_{k=0}^{t-1}p_{i+k}\le n_0$. Note that we have $n_1(i,j,t)>n_0$ for $i\ge 19$. According to our computation $n_1(i,j,t)$ is a monotone increasing function of $j$ for fixed $(i,t)$, thus we displayed in Table 1 only the values $n_1(i,i-1,t)$. The italic values in Table 1 indicate that the corresponding inequality $n\ge n_1$ is valid for all $n$ with smallest prime factor $p_i$, having $t$ distinct prime divisors. For example, the smallest $n$ with $i=5$ (with smallest prime factor $p_i=p_5=11$) and with $t=3$ distinct prime factors is $11\cdot 13\cdot 17=2431$. So the smaller number 1773 in row $(i,p_i)=(5,11)$ and column $t=3$ is in italics. Finally, if all integers $n$ with smallest prime factor $p_i$ and exactly $t$ distinct prime factors are $>n_0$, then the corresponding box in Table 1 is empty.

In particular, checking the boxes of Table 1 corresponding to the primes $p\le 11$ with $t\ge 3$, and $p\le 37$ with $t\ge 4$, we see that for these cases the induction step is established. Further, as we mentioned in the beginning of this subsection, the latter assertion implies that the induction step is also proved for all $n<n_0$ with $t\ge 5$.

### *4.3. The case where n has at most two distinct prime factors*

The following lemma verifies Theorem 1 if $n$ is a prime power.

**Lemma 6.** *Let $p=p_i$ be a prime and $n=p^\alpha$ $(\alpha>0)$. Then $S^*_{i,n}=S_{i,n}$.*

**Proof.** We already know that the assertion is valid for $p=2,3$. So we may assume that $p=p_i\ge 5$. If $\alpha=1$ then the statement follows from part (iii) of Corollary 1.

To treat the cases $\alpha>1$, we show that for any $j<i$

$$B_{j,p^\alpha}=\bigcup_{k=1}^{\alpha-1}p^kE_{j,p^{\alpha-k}}\tag{11}$$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

Table 1
Values of $n_1(i, i-1, t)$.

| $(i, p_i)$ | $t$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (3, 5) | 42 | 85 | *176* | *381* | *832* | *1 844* | *4073* |
| (4, 7) | 163 | 292 | *564* | *1 134* | *2 353* | *4 922* | |
| (5, 11) | 539 | 943 | *1 773* | *3 527* | *7 201* | *14 830* | |
| (6, 13) | 1 668 | 2 810 | 5 168 | *10 027* | *20 004* | *40 659* | |
| (7, 17) | 4 411 | 7 394 | 13 473 | *25 897* | *51 602* | | |
| (8, 19) | 11 430 | 18 856 | 33 849 | *64 646* | *127 553* | | |
| (9, 23) | 27 807 | 45 465 | 81 511 | *154 920* | *305 065* | | |
| (10, 29) | 71 314 | 116 355 | 207 415 | *393 138* | *773 398* | | |
| (11, 31) | 172 771 | 279 447 | 495 419 | *935 583* | *1 832 178* | | |
| (12, 37) | 400 411 | 646 688 | 1 146 625 | *2 163 465* | | | |
| (13, 41) | 948 529 | 1 528 461 | 2 701 477 | 5 078 938 | | | |
| (14, 43) | 2 098 084 | 3 371 377 | 5 939 601 | 11 137 433 | | | |
| (15, 47) | 4 590 463 | 7 360 443 | 12 945 781 | 24 269 102 | | | |
| (16, 53) | 10 391 079 | 16 631 063 | 29 240 976 | 54 743 149 | | | |
| (17, 59) | 23 720 841 | 37 937 473 | 66 587 159 | | | | |
| (18, 61) | 51 847 427 | 82 741 135 | | | | | |

holds. Indeed, if $m \in B_{j,p^\alpha}$, then either $m$ is divisible only by the first power of $p$, thus $m/p \in E_{j,p^{\alpha-1}}$ or $m$ is divisible by a higher power of $p$, in which case $m/p \in B_{j,p^{\alpha-1}}$, thus

$$B_{j,p^\alpha} = pE_{j,p^{\alpha-1}} \cup pB_{j,p^{\alpha-1}}.$$

Using this identity we get (11) by induction.

Now we split the proof of the lemma into four cases.

*Case* $\alpha = 2$. Then $|B_{j,p^2}| = |E_{j,p}|$ for $j < i$. By Bertrand's postulate there exists at least one prime $q$ with $p/p_j < q < p$. Hence $p_j q \in E_{j,p^2} \setminus E_{j,p}$, which implies $|E_{j,p^2}| > |B_{j,p^2}| = |E_{j,p}|$. Thus $S^*_{i,p^2} = S^*_{i,n} = S_{i,n}$.

*Case* $\alpha = 3$. Then for $j < i$, there exists a prime $q$ with $p^2/p_j < q < p^2$, hence $p_j q \notin E_{j,p^2}$. If $m \in E_{j,p}$ then $qm \le qp < p^3$, thus

$$|E_{j,p^3}| \ge |B_{j,p^3}| = |E_{j,p^2}| + |E_{j,p}|,$$

and this case is proved.

*Case* $\alpha = 4$. Identity (11) implies that for $j < i$,

$$|B_{j,p^4}| = |E_{j,p^3}| + |E_{j,p^2}| + |E_{j,p}|.$$

We plainly have $E_{j,p^2} = E_{j,p} \cup E_2 \cup E_3$, where the sets $E_2, E_3$ on the right-hand side include all elements of $E_{j,p^2}$ belonging to the intervals $(p, p^2/p_j]$ and $(p^2/p_j, p^2]$, respectively. Since $p \ge 5$, by a simple calculation based upon formulas of Rosser and Schoenfeld [4] concerning $\pi(x)$ (see also (12)), we get that there exist at least two different primes $q_1, q_2$ with $p^3/p_j < q_1, q_2 < p^3$. Hence

$$q_k E_{j,p} \cap E_{j,p^3} = q_1 E_{j,p} \cap q_2 E_{j,p} = \emptyset \quad (k = 1, 2).$$

Further, there exist primes $q_3, q_4$ with $p^2 < q_3 < 2p^2$ and $p^2/2 < q_4 < p^2$. By the construction we have

$$q_k E_{j,p} \cap q_3 E_2 = q_k E_{j,p} \cap q_4 E_3 = \emptyset \quad (k = 1, 2).$$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

Table 2
Values of $N(p)$.

| $p$ | 13 | 17 | 19 | 23 | 29 | 31 | 37 | 41 | 43 | 47 | 53 | 59 | 61 | 67 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N(p)$ | 2 | 7 | 29 | 77 | 203 | 566 | 1246 | 2964 | 6722 | 14 129 | 29 518 | 62 521 | 101 975 | 89 277 |

If $m \in q_3E_2 \cap q_4E_3$ then $m$ is divisible by the pairwise different primes $q_3, q_4, p_j$, thus $m \geq p_j q_3 q_4 > p_j p^4/2 \geq p^4$, which is a contradiction. Finally, it is clear that

$$E_{j,p^3} \cap q_3E_2 = E_{j,p^3} \cap q_4E_3 = \emptyset.$$

Summarizing the above facts, we obtain that the sets

$$q_1E_{j,p},\ q_2E_{j,p},\ q_3E_2,\ q_4E_3,\ E_{j,p^3}$$

are pairwise disjoint subsets of $E_{j,p^4}$. This implies that $|B_{j,p^4}| - |E_{j,p^4}| \leq 0$, and our claim is verified also in this case.

*Case* $\alpha \geq 5$. Since $n = p^\alpha < n_0$, this case may occur only for $p = p_i \leq 37$, i.e. $i \leq 12$. Now the assertion follows by the first column of Table 1. □

The next lemma verifies the induction step for integers, which have two different prime divisors and the smaller is at most 53.

**Lemma 7.** *Let* $p = p_i$ *and* $q > p$ *be primes. If* $p \leq 53$ *and* $n = p^\alpha q^\beta < n_0$ $(\alpha, \beta > 0)$, *then* $S^*_{i,n} = S_{i,n}$.

**Proof.** The idea of the proof is similar to the proof of Lemma 6. We omit the technical details. We remark that an alternative proof can also be given following the method of the next section. □

So altogether, in this subsection we have proved the induction step for prime powers and for integers with at most two distinct prime divisors, such that the smaller is at most 53.

### *4.4. The case where n has three distinct prime factors*

Unfortunately, we could not find any meaningful generalization of Lemmata 6 and 7 to integers with at least three prime divisors. By Table 1 the smallest prime factor of a candidate $n$ which could violate Theorem 1 is at least 13. For each prime $13 \leq p \leq 67$ we computed all integers, which are divisible by $p$, lie below the bound $\min\{n_0, n_1\}$, where $n_1$ is given in Table 1 and have three different prime divisors, which are at least $p$. Their number, $N(p)$ is given in Table 2.

Fix $p = p_i$. For each candidate $n$ we computed $|B_{i-1,n}| - |E_{i-1,n}|$. For this purpose we used a variant of the wheel algorithm, see e.g. [5]. In our case this listed efficiently the elements of $S_{i-1,n}$ because we know that they are divisible by $p_i$ and, on the other hand, relative prime to 2, 3, 5, 7. For each $m$ produced by the wheel algorithm we computed $\gcd(m, \prod_{j=5}^{i-2} p_j)$. If this is not 1, then $m$ does not belong to $S_{i-1,n}$, otherwise we added one to the counter of $|E_{i-1,n}|$ or $|B_{i-1,n}|$ according as $\gcd(m, n) = 1$ or not. We found in each case that $|B_{i-1,n}| - |E_{i-1,n}| < 0$, which means that $n$ cannot be adjoined to $S_{i-1,n}$. The total computational time on a notebook was about two days, from which one and a half was spent for 61 and 67 and the rest for the other ten primes.

After these calculations, the induction step is established for values of $n$ having three distinct prime divisors such that the smallest is at most 67.

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

### 4.5. The case where n has four distinct prime factors

For later use, we push forward the results obtained in Table 1 also in case of $t = 4$. Namely, in this subsection we establish the induction step for those $n$ having $t = 4$ distinct prime factors, from which the smallest is 41 or 43.

So let $n$ be of the form $n = pq_1q_2q_3$, where $p < q_1 < q_2 < q_3$ are distinct primes, and $p = 41$ or 43. Note that the exponents of these primes in $n$ are necessarily equal to one, otherwise $n > n_0$ would hold. In view of the values in the corresponding boxes of Table 1, we may assume that

$$n \le \begin{cases} 5\,078\,938, & \text{if } p = 41, \\ 11\,137\,433, & \text{if } p = 43, \end{cases}$$

since otherwise the induction step works for $n$. Thus by a simple calculation we obtain that $q_3 \le 61$ for $p = 41$, and $q_3 \le 103$ for $p = 43$. For the possible values of $n$, we check whether it is possible adjoin $n$ to $S_{j,n-1}$ for some $j$ with $2 \le p_j < p$ ($p_j$ prime). For this, first we apply the estimates for $S_{i,n-1}$ and $|B_{j,n}| - |E_{j,n}|$ given by Lemmata 3 and 4. After this check we are left only with 43 pairs $(n, j)$ such that it is still possible that $n$ gets adjoined to the class $S_{j,n-1}$. There are only 3 such pairs with $p = 41$, when we always have $j = 12$ (i.e. $p_j = 37$) and 40 such pairs with $p = 43$, when we have 13 (i.e. $p_j = 41$). Then, for each of the remaining cases, using Maple we count the number of friends and enemies in the class $S_{j,n-1}$ (with $j = 12$ and 13 for $p = 41$ and 43, respectively). In each case, we find $|B_{j,n}| - |E_{j,n}| < 0$. This shows that Algorithm 1 cannot adjoin $n$ to the class $S_{j,n-1}$. So the induction step works for the values of $n$ having $t = 4$ distinct prime factors, the smallest of which being 41 or 43.

### 4.6. Handling the remaining cases

So far we proved the induction step if $n$ has one or at least five different prime factors or

- $n$ has two different prime factors, from which the smaller is at most 53 or
- $n$ has three different prime factors, from which the smallest is at most 67 or
- $n$ has four different prime factors, from which the smallest is at most 43.

Now we shall consider the remaining values of $n < n_0$.

#### 4.6.1. Lower bounds for j with $n \in S^*_{j,n}$

Let $n = q_1^{\alpha_1} \cdots q_t^{\alpha_t} < n_0$ be such that $p_i = q_1 < q_2 \cdots < q_t$ and assume that Algorithm 1 adjoins $n$ to $S_{j,n-1}$. Then $j \le i$ by part (iv) of Corollary 1. We give here a lower bound for $j$ provided $t = 4$ and $q_1 > 43$, or $t = 3$ and $q_1 > 67$. As $|S_{i,n-1}| > 0$ we must have

$$|B_{j,n}| - |E_{j,n}| > 0$$

by part (i) of Corollary 1. By Lemma 4 we have

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

Thus if

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

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

then $|B_{j,n}| - |E_{j,n}| < 0$ and $n$ cannot be adjoined to $S_{j,n-1}$. The product $\prod_{\ell=1}^{j-1}\left(1-\frac{1}{p_\ell}\right)$ decreases with $j$ and for fixed $t$ the expression $2\prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)-1$ takes the smallest value if the distinct primes $q_1,\dots,q_t$ are as small as permitted. This means

$$2\prod_{k=1}^{t}\left(1-\frac{1}{q_k}\right)-1 \geq \begin{cases} 376\,783/409\,457, & \text{if } t=3,\\ 7\,683\,211/8\,965\,109, & \text{if } t=4.\end{cases}$$

We have $\prod_{\ell=1}^{11}\left(1-\frac{1}{p_\ell}\right)=\frac{13\,271\,040}{86\,822\,723}$. Thus if $n > 9\,544\,582$ and $t=4$ then $n$ cannot be adjoined to $S_{j,n-1}$ provided $j \leq 12$. As $47\cdot 53\cdot 59\cdot 67 = 9\,846\,923 > 9\,544\,581$ we proved that if $n \neq n' := 47\cdot 53\cdot 59\cdot 61$ has four prime factors then Algorithm 1 may adjoin $n$ to $S_{j,n-1}$ only if $j > 12$, i.e. $p_j \geq 41$. In the particular case of $n = n'$, the above considerations yield that

$$|B_{j,n'}| - |E_{j,n'}| < \begin{cases} 0, & \text{for } j=1,\dots,11,\\ 2052, & \text{for } j=12.\end{cases}$$

However, the class of 47, i.e. $S_{14,n-1}$ has more elements than 2052. Indeed, it contains all the numbers of the form $47q$, with $47 \leq q \leq 53\cdot 59\cdot 61$, $q$ prime. The number of such elements is already 17 209. This shows that for any $n$ having four distinct prime factors all greater than 43 can be adjoined to $S_{j,n-1}$ by Algorithm 1 only if $j > 12$, i.e. $p_j \geq 41$.

Similarly, we obtain that if $n > 318\,015$ and has three distinct prime factors, then Algorithm 1 may adjoin $n$ to $S_{j,n-1}$ only if $j > 9$, i.e. $p_j \geq 29$. As $71\cdot 73\cdot 79 > 318\,015$, we get that this assertion is valid whenever the smallest prime factor of $n$ is at least 71. However, we want to make one step further. Suppose that $n$ is as above, and it is attached to $S_{10,n-1}$. Then similarly as before, we get that $n > 838\,402$ implies that if $n$ has three prime factors, it cannot be adjoined to $S_{10,n-1}$ by Algorithm 1. Suppose that $n$ has three distinct prime factors from which the smallest is at least 71, and $n \leq 838\,402$. This implies that $n$ is of the form $n = q_1q_2q_3$ with $71 \leq q_1 < q_2 < q_3 \leq 157$ primes. There are 128 such values for $n$. A simple calculation with Maple yields that in case of each such $n$ we have

$$|B_{10,n}| - |E_{10,n}| < S_{i,n-1},$$

where $p_i$ is the smallest prime divisor of $n$. This, altogether with what we have proved previously, shows that if $n$ has three distinct prime factors from which the smallest is at least 71, then Algorithm 1 may adjoin $n$ to $S_{j,n-1}$ only if $j > 10$, i.e. $p_j \geq 31$.

### 4.6.2. *Completing the proof of Theorem 1*

**Proposition 1.** *Assume that one of the following properties is valid:*

- *$n$ has two different prime factors, from which the smallest is at least 59,*
- *$n$ has three different prime factors, from which the smallest is at least 71,*
- *$n$ has four different prime factors, from which the smallest is at least 47.*

*Then (3) is valid for $n$.*

**Proof.** Assume first that $n$ has precisely four prime divisors, i.e. $n$ is of the form $n = q_1^{\alpha_1}q_2^{\alpha_2}q_3^{\alpha_3}q_4^{\alpha_4}$ with $47 \leq q_1 < q_2 < q_3 < q_4$, and positive integers $\alpha_1,\dots,\alpha_4$. Then by what we have proved in Section 4.6.1, it is sufficient to show that $n \notin S^*_{\ell,n}$ with $41 \leq p_\ell < q_1$. A simple check shows that then $\alpha_1=\alpha_2=\alpha_3=\alpha_4=1$ must be valid, and any friend of $n$ in $S_{\ell,n-1}$ has

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

at most four prime factors, counted with multiplicity. Further, we must have $p_\ell \le 89$, otherwise $n > n_0$ would hold. We shall split the set of friends of $n$ in $S_{\ell,n-1}$ into two parts. Write $t_1$ and $t_2$ for the first two primes $> p_\ell$, distinct from the $q_i$. First consider the friends of $n$ of the form $p_\ell q_i r_1 r_2$ such that $r_1, r_2$ are primes with $p_\ell < r_1 \le r_2$, distinct from the $q_i$ and $t_1, t_2$. Further, we also require that

$$p_\ell p_{\ell+6} r_1 r_2 \le n.$$

To such friends of $n$ in $S_{\ell,n-1}$, we can adjoin the four enemies

$$p_\ell r_1 r_2,\ p_\ell^2 r_1 r_2,\ p_\ell t_1 r_1 r_2,\ p_\ell t_2 r_1 r_2$$

of $n$ in $S_{\ell,n-1}$. Indeed, observe that by our assumptions, all the four numbers above are distinct elements of $S_{\ell,n-1}$. Further, for distinct friends of $n$ of the shape $p_\ell q_i r_1 r_2$, these four numbers are also distinct.

Now we estimate the number of friends of $n$ in $S_{\ell,n-1}$ not of the above shape. For this, write $x_1$ for the number of friends of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q_i r_1 r_2$, where $r_1, r_2$ are primes $\ge p_\ell$, violating one of the above requirements. First observe that the number of such friends with one of $r_1, r_2$ in $\{p_\ell, t_1, t_2, q_1, q_2, q_3, q_4\}$ is at most $28(\pi(n/p_\ell^3) - \ell + 1)$. Further, for any $i = 1, 2, 3, 4$ the number of friends of $n$ of the form $p_\ell q_i r_1 r_2$ with $p_\ell p_{\ell+6} r_1 r_2 > n$, in view of

$$p_\ell^2 r_1 r_2 \le n,$$

is bounded by the number of integers in the interval $(n/p_\ell p_{\ell+6}, n/p_\ell^2]$, not divisible by 2 and 3. So we obtain that

$$x_1 \le 28(\pi(n/p_\ell^3) - \ell + 1) + 4(n/p_\ell^2 - n/p_\ell p_{\ell+6})/3 + 6.$$

Putting $x_2$ for the number of friends of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q_i r_1$, where $r_1$ is a prime with $r_1 \ge p_\ell$, in view of $r_1 \le n/p_\ell^2$ we get

$$x_2 \le 4(\pi(n/p_\ell^2) - \ell + 1).$$

Finally, $n$ also has the four friends $p_\ell q_i$ $(i = 1, 2, 3, 4)$ in $S_{\ell,n-1}$.

So to prove our claim in this case, it is sufficient to show that

$$x_1 + x_2 + 4$$

is less than the number of enemies of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q$, where $q$ is a prime distinct from the $q_i$. The number of such enemies of $n$ is clearly at least $\pi(n/p_\ell) - \ell - 3$. So in view of the inequalities

$$\frac{x}{\log(x)}\left(1 + \frac{1}{2\log(x)}\right) < \pi(x) < \frac{x}{\log(x)}\left(1 + \frac{3}{2\log(x)}\right) \tag{12}$$

holding for any $x \ge 59$ (see Rosser and Schoenfeld [4]), we only need to check that

$$\begin{aligned} f(n) := {} & \frac{4}{3}\left(\frac{n}{p_\ell^2} - \frac{n}{p_\ell p_{\ell+6}}\right) + 10 + 28\left(\frac{u_1}{\log(u_1)}\left(1 + \frac{3}{2\log(u_1)}\right) - \ell + 1\right) \\ & + 4\left(\frac{u_2}{\log(u_2)}\left(1 + \frac{3}{2\log(u_2)}\right) - \ell + 1\right) \\ & - \frac{v_1}{\log(v_1)}\left(1 + \frac{1}{2\log(v_1)}\right) + \ell + 3 < 0 \end{aligned}$$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

holds with $u_1 = n/p_\ell^3, u_2 = n/p_\ell^2, v_1 = n/p_\ell$ for $41 \le p_\ell \le 89$. (Note that we have $u_1, u_2, v_1 \ge 59$.) We used Maple to check the assertion. It turned out that for any possible value of $p_\ell$, the function $f(n)$ is monotone decreasing on the interval $[n_1, n_0]$ with $n_1 = p_{\ell+1}p_{\ell+2}p_{\ell+3}p_{\ell+4}$, and $f(n_1) < 0$. This proves our claim in this case.

Assume next that $n$ has precisely three prime divisors, i.e. $n$ is of the form $n = q_1^{\alpha_1}q_2^{\alpha_2}q_3^{\alpha_3}$ with distinct primes $q_1, q_2, q_3$, and positive integers $\alpha_1, \alpha_2, \alpha_3$. Then by what we have proved in Section 4.6.1, it is sufficient to show that $n \notin S^*_{\ell,n}$ with $31 \le p_\ell < q_1$. A simple check shows that then any friend of $n$ in $S_{\ell,n-1}$ has at most five prime factors, counted with multiplicity. Further, we must have $p_\ell \le 463$, otherwise $n > n_0$ would hold. We shall split the set of friends of $n$ in $S_{\ell,n-1}$ into two parts. Write $t_1$ for the first prime $> p_\ell$, distinct from the $q_i$. First consider the friends of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q_i r_1 r_2$ such that $r_1, r_2$ are primes with $p_\ell < r_1 \le r_2$, distinct from the $q_i$ and $t_1$. Further, we also require that

$$p_\ell p_{\ell+4} r_1 r_2 \le n.$$

To such friends of $n$ in $S_{\ell,n-1}$, we can adjoin the three enemies

$$p_\ell r_1 r_2,\ p_\ell^2 r_1 r_2,\ p_\ell t_1 r_1 r_2$$

of $n$ in $S_{\ell,n-1}$. Indeed, observe that by our assumptions, all the three numbers above are distinct elements of $S_{\ell,n-1}$. Further, for distinct friends of $n$ of the shape $p_\ell q_i r_1 r_2$, these three numbers are also distinct.

Now we estimate the number of friends of $n$ in $S_{\ell,n-1}$ not of the above shape. For this, write $x_1$ for the number of friends of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q_i r_1 r_2$, where $r_1, r_2$ are primes $\ge p_\ell$, violating one of the above requirements. First observe that the number of such friends with one of $r_1, r_2$ in $\{p_\ell, t_1, q_1, q_2, q_3\}$ is at most $15(\pi(n/p_\ell^3) - \ell + 1)$. Further, for each $i = 1, 2, 3, 4$ the number of friends of $n$ of the form $p_\ell q_i r_1 r_2$ with $p_\ell p_{\ell+4} r_1 r_2 > n$, in view of

$$p_\ell^2 r_1 r_2 \le n,$$

is bounded by the number of integers in the interval $(n/p_\ell p_{\ell+4}, n/p_\ell^2]$, not divisible by 2 and 3. So we obtain that

$$x_1 \le 15(\pi(n/p_\ell^3) - \ell + 1) + n/p_\ell^2 - n/p_\ell p_{\ell+4} + 4.$$

Write now $x_2$ for the number of friends of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q_i r_1 r_2 r_3$, where $r_1 \le r_2 \le r_3$ are primes $\ge p_\ell$. As one can easily check, such friends of $n$ may exist only if $p_\ell = 31, 37$ and $q_i \le n_0/p_\ell^4$ $(i = 1, 2, 3, 4)$. Then since $r_3 \le n/p_\ell^4$, we can easily bound $x_2$ in these cases. By checking the possibilities with Maple, we get

$$x_2 \le y_\ell := \begin{cases} 83, & \text{if } p_\ell = 31, \\ 11, & \text{if } p_\ell = 37, \\ 0, & \text{otherwise.} \end{cases}$$

Putting $x_3$ for the number of friends of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q_i r_1$, where $r_1$ is a prime with $r_1 \ge p_\ell$, in view of $r_1 \le n/p_\ell^2$ we get

$$x_3 \le 3(\pi(n/p_\ell^2) - \ell + 1).$$

Finally, $n$ also has the three friends $p_\ell q_i$ $(i = 1, 2, 3)$ in $S_{\ell,n-1}$.

So to prove our claim in this case, it is sufficient to show that

$$x_1 + x_2 + x_3 + 3$$

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

is less than the number of enemies of $n$ in $S_{\ell,n-1}$ of the form $p_\ell q$, where $q$ is a prime distinct from the $q_i$. The number of such enemies of $n$ is clearly at least $\pi(n/p_\ell) - \ell - 3$. So in view of the inequalities (12), we only need to check that

$$g(n) := \frac{n}{p_\ell^2} - \frac{n}{p_\ell p_{\ell+4}} + 7 + 15\left(\frac{u_1}{\log(u_1)}\left(1 + \frac{3}{2\log(u_1)}\right) - \ell + 1\right) + y_\ell + 3\left(\frac{u_2}{\log(u_2)}\left(1 + \frac{3}{2\log(u_2)}\right) - \ell + 1\right) - \frac{v_1}{\log(v_1)}\left(1 + \frac{1}{2\log(v_1)}\right) + \ell + 3 < 0$$

holds with $u_1 = n/p_\ell^3$, $u_2 = n/p_\ell^2$, $v_1 = n/p_\ell$ for $31 \le p_\ell \le 463$. (One can readily verify that $u_1, u_2, v_1 \ge 59$, so we can apply (12) without any problem.) This assertion can be checked by Maple, in a similar way as before. We obtain that the statement is valid also in this case.

Finally, assume that $n$ is of the form $q_1^{\alpha_1} q_2^{\alpha_2}$ with $q_2 > q_1 \ge 59$, and positive integers $\alpha_1, \alpha_2$. Let $p_\ell < q_1$. Then the number of friends of $n$ in $S_{\ell,n-1}$ is at most

$$\frac{n/p_\ell}{59} + \frac{n/p_\ell}{61}.$$

On the other hand, every number of the form $p_\ell q$ with $q \ge p_\ell$ prime, distinct from $q_1, q_2$, is an enemy of $n$ in $S_{\ell,n-1}$, provided that $p_\ell q < n$. By (12) and $p_\ell \nmid n$ we get that the number of such enemies of $n$ is at least

$$\pi(n/p_\ell) - \ell - 1 > \frac{n/p_\ell}{\log(n/p_\ell)}\left(1 + \frac{1}{2\log(n/p_\ell)}\right) - \ell - 1.$$

Put $x = n/p_\ell$. Observe that we have $59 \le x \le n_0$. Note that Theorem 3 of [4] implies $\ell \le 2\log(p_\ell)$, which by $n > p_\ell^2$ gives $\ell < 2\log(x)$. Consider the function

$$h(x) := \frac{x}{59} + \frac{x}{61} + 2\log(x) + 1 - \frac{x}{\log(x)}\left(1 + \frac{1}{2\log(x)}\right).$$

A simple calculation with Maple assures that $h(x)$ is negative on the interval $[59, n_0]$. This shows that $S_{\ell,n-1}$ contains more enemies than friends of $n$, implying $n \notin S^*_{\ell,n}$. Hence our claim follows also in this case. □

Since Proposition 1 covers all the cases not considered in the preceding subsections, the proof of Theorem 1 is now complete.

## Acknowledgment

The authors are very much grateful to the referee for the careful reading and for the many helpful suggestions, which helped to improve the quality of the presentation considerably.

## References

[1] M. Bakó, L. Aszalós, Combinatorial optimization methods for correlation clustering, in: D. Dumitrescu, Rodica Ioana Lung, Ligia Cremene (Eds.), Coping with Complexity, Casa Cartii de Stiinta, Cluj-Napoca, 2011, pp. 2–12.

[2] N. Bansal, A. Blum, S. Chawla, Correlational clustering, Mach. Learn. 56 (2004) 89–113.

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004

[3] J.E. Nymann, On the probability that $k$ positive integers are relatively prime, J. Number Theory 4 (1972) 469–473.
[4] J.B. Rosser, L. Schoenfeld, Approximate formulas for some functions of prime numbers, Illinois J. Math. 6 (1962) 64–94.
[5] H.C. Williams, Primality testing on a computer, Ars Combin. 5 (1978) 127–185.

Please cite this article in press as: L. Aszalós, et al., On a correlational clustering of integers, Indagationes Mathematicae (2015), http://dx.doi.org/10.1016/j.indag.2015.09.004