Primes and the Internet

Attila Pethő and Norbert Bátfai

(University of Debrecen, Hungary)

Erdős 100, Budapest, July 1, 2013.

Research was partially supported by the OTKA grant K104208 and by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project.

Prologue

Since the Eighties of the last century Paul Erdős visited each year Debrecen. His talks were very popular. He started typically: "Let p be a prime."

Primes are important objects of the mathematic, but are widely considered not to be useful. Godfrey Hardy wrote in 1940: "I have never done anything useful. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world." He was one of greatest prime number theorists of the first half of the last century.

Primes and the Internet

Dramatic change:

• W. Diffie and M. Hellman, 1976, introduced the concept of public key cryptography including **digital signature**.

• R. Rivest, A. Shamir and L. Adleman, 1978, RSA algorithm.

• The Internet without secure communication and digital signature could not operate! These depend basically on properties of primes.

The RSA digital signature algorithm:

1. Setup

- The signer - S - chooses two large primes p, q.
- S computes $n = pq$ and $\varphi(n) = (p-1)(q-1)$.
- S chooses e, d such that $2 \leq e, d \leq \varphi(n) - 1$ and with $ed \equiv 1 \pmod{\varphi(n)}$.
- S publish e, n, but d, p, q keeps secret.

2. Signature If S will sign a message $1 \leq m < n$ then

- He computes $M \equiv m^d \pmod{n}$ and publish (m, M).

3. Verification If the verifier - V - will be sure that the message was signed by S then

- V computes $m' \equiv M^e \pmod{n}$.
- If $m' = m$ then V accept the signature, otherwise he reject it.

The RSA digital signature algorithm is practical because we can generate efficiently large primes.

- Primes in $\mathcal{P}$. M. Agrawal, N. Kayal and N. Saxena, 2002.
- Miller-Rabin test used in practice.

The RSA digital signature algorithm is considered to be secure.

- We are not able to factorize large composite numbers. RSA-768, 2009.
- P. Shor, 1994: Integers can factorize in polynomial time with quantum algorithm.

Internet and the Primes, PageRank

L. Page, S. Brin, R. Motwani, and T. Winograd, 1999, defined a ranking method of web pages. The success of Google is based on PageRank.

Naïve PageRank:

- $N(p)$: number of outgoing links from page p
- $B(p)$: set of pages that point to p
- $PageRank(p) = \sum_{q \in B(p)} PageRank(q)/N(q)$
- Intuition:

$--$ Each page q evenly distributes its importance to all pages that q points to

$--$ Each page p gets a boost of its importance from each page that points to p.

The PageRank can be computed easily:

• Let $p_1, \ldots, p_m$ be the pages of a network.

• Create the stochastic (normalized Google) matrix M for the link structure:

−− Each page i corresponds to row i and column i,

−− If page j has n outgoing links, then let

$$M(i,j) = \begin{cases} \frac{1}{n} & \text{if page } j \text{ points to page } i \\ 0 & \text{otherwise.} \end{cases}$$

• The vector $(PageRank(p_1), \ldots, PageRank(p_m))^t$ is the right eigenvector corresponding to the eigenvalue 1 of M.

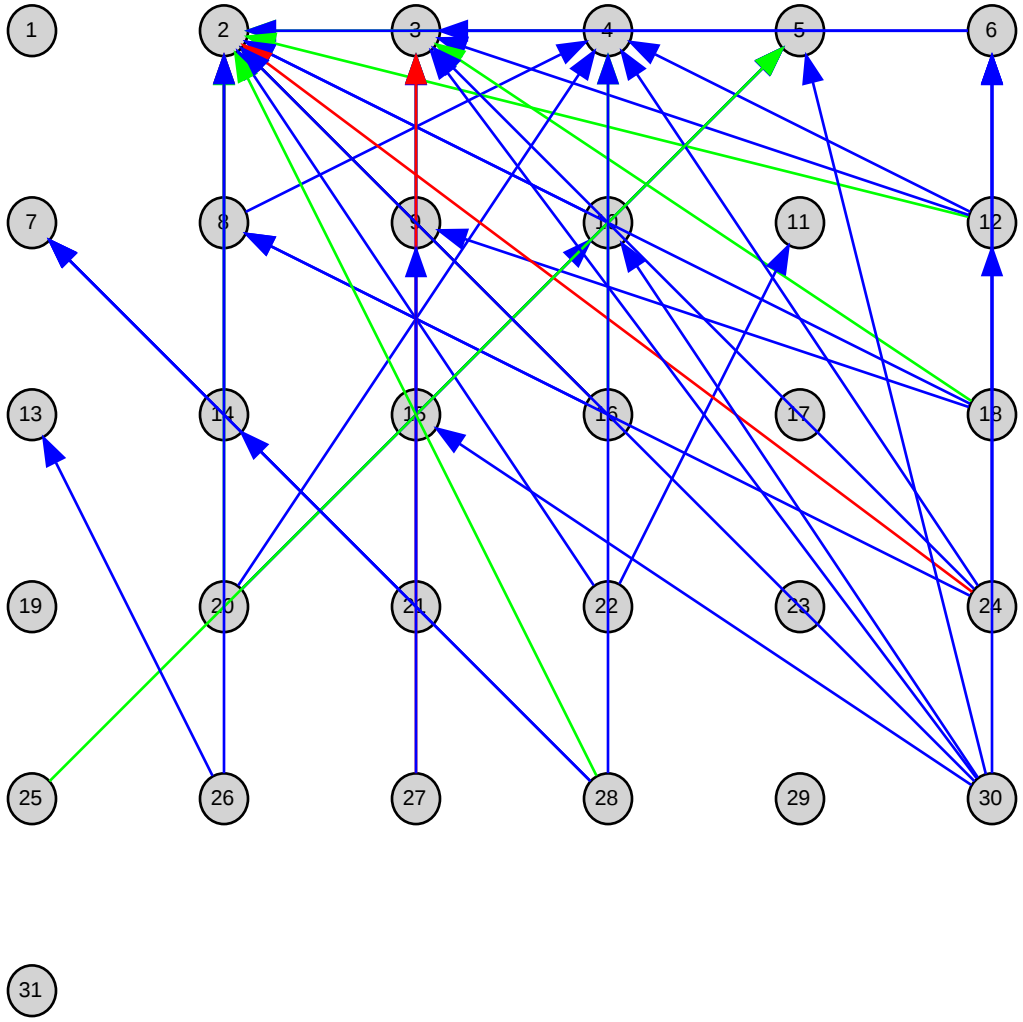
K.M. Frahm, A.D. Chepelianskii, and D.L. Shepelyansky, 2012, studied the PageRank of integers.

The pages (nodes) of the FCS-network are labeled by integers in $[1, N]$. From page a there is an edge of weight k to page b if $b^k|a$, but $b^{k+1} \nmid a$. Thus the entries of the Google matrix $A = (a_{mn}) \in \mathbb{N}^{N\times N}$ are

$$a_{mn} = \begin{cases} 0, & \text{if } m = 1 \text{ or } m = n, \\ k, & k = \max\{l : m^l|n\} \text{ otherwise.} \end{cases}$$

Its normalized form $S = (s_{mn}) \in \mathbb{R}^{N\times N}$ has entries

$$s_{mn} = \begin{cases} 1/N, & \text{if } \sum_{i=1}^{N} a_{in} = 0, \\ \frac{a_{mn}}{\sum_{i=1}^{N} a_{in}}, & \text{otherwise.} \end{cases}$$

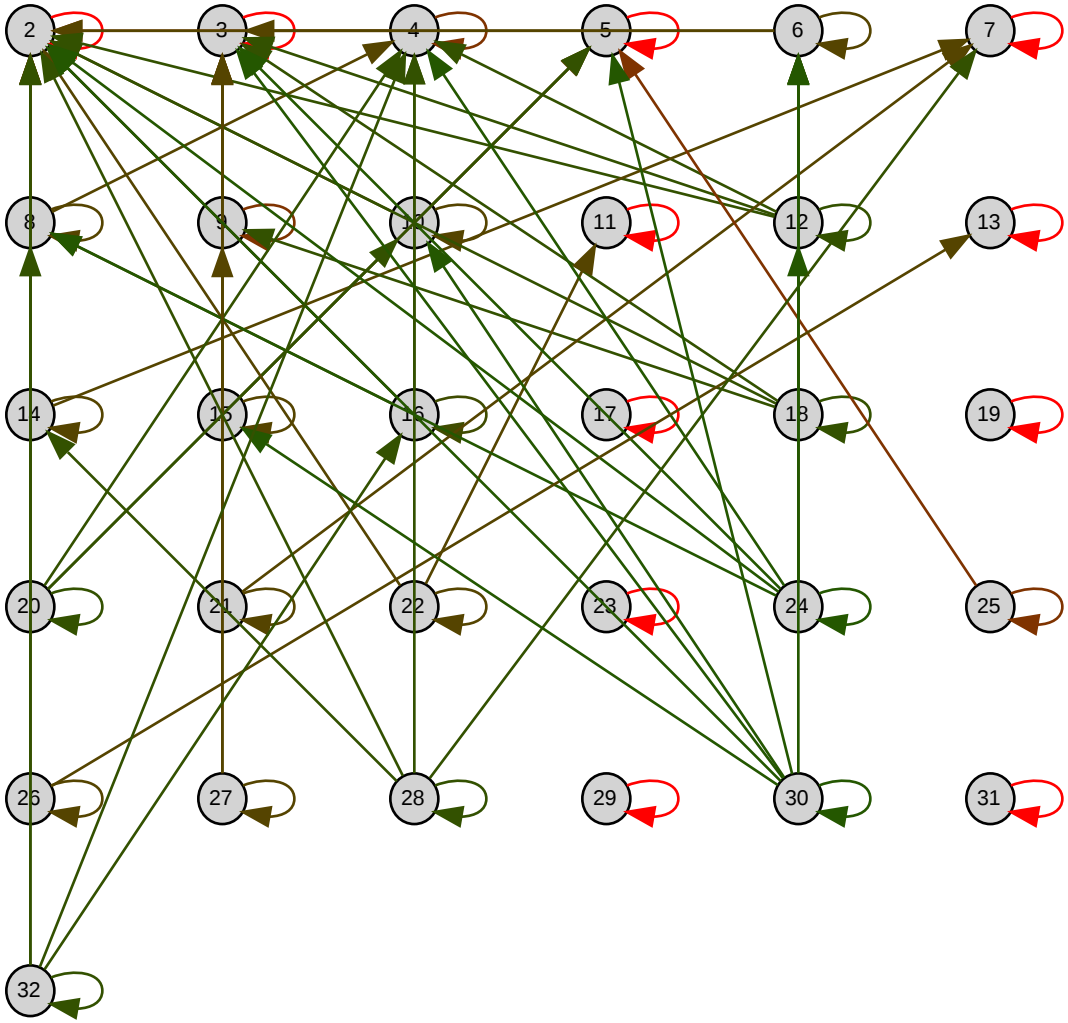


The FCS network before normalization, N=31.

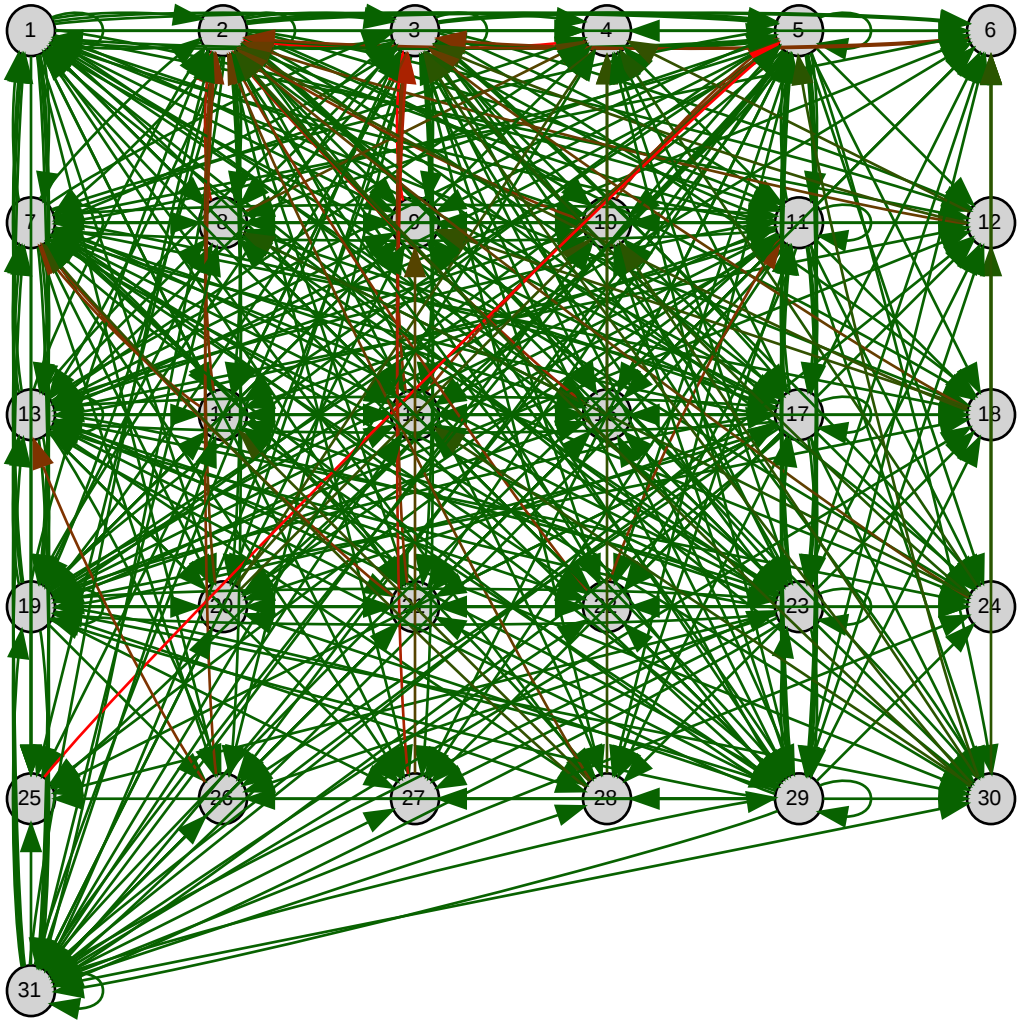
Independently from FCS we defined a different network on the integers $[2, N+1]$. It is called INN. In our case there is a path from page a to page b iff $b|a$. The entries of the corresponding Google matrix $A = (a_{mn}) \in \mathbb{N}^{N \times N}$ are

$$a_{mn} = \begin{cases} 0, & \text{if } (m+1) \not| (n+1), \\ 1, & \text{otherwise.} \end{cases}$$

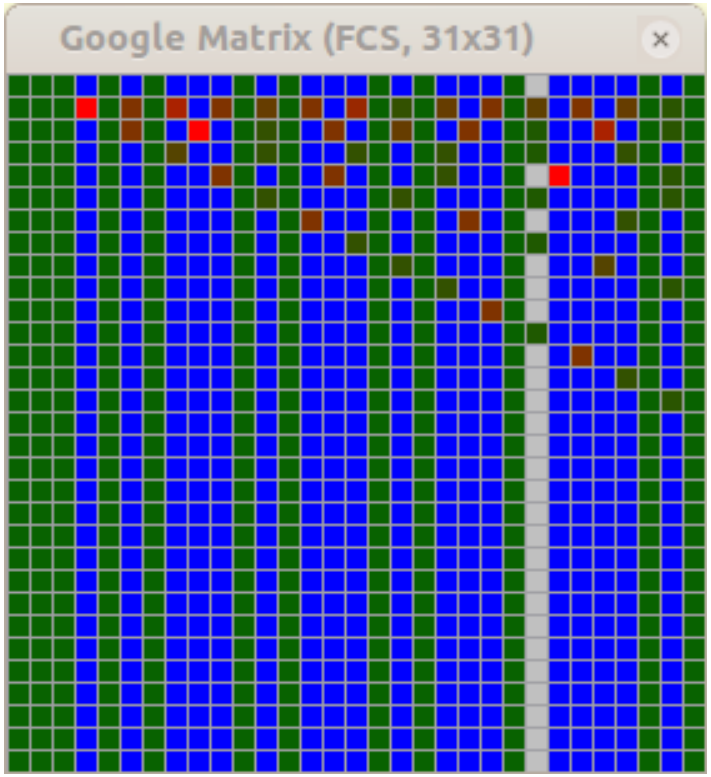
Its normalized form $S = (s_{mn}) \in \mathbb{R}^{N \times N}$ has entries $s_{mn} = \frac{a_{mn}}{\sum_{i=1}^{N} a_{in}}$.

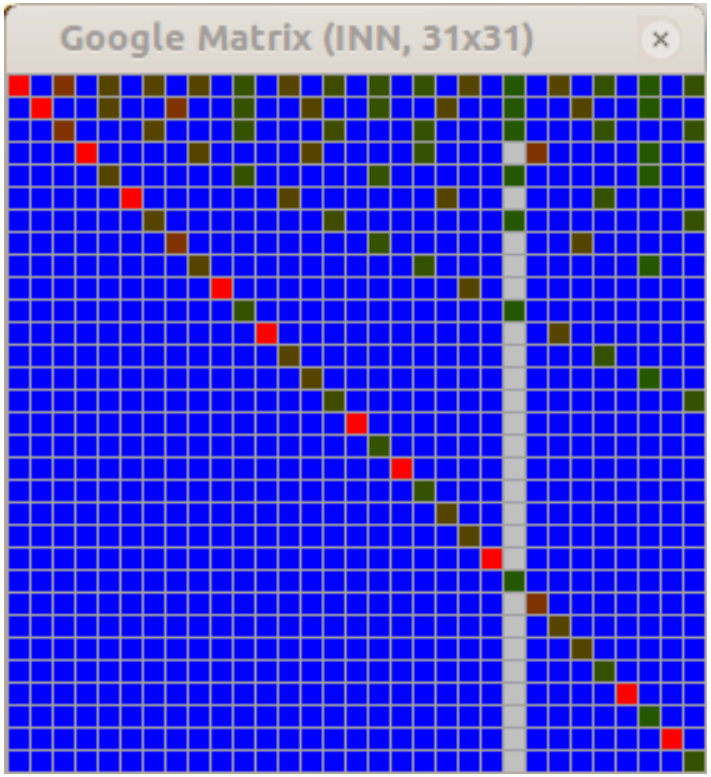


The INN network represented by a directed graph, N=31.



The FCS network after normalization, N=31.





The normalized Google matrices of FCS and INN respectively.

Let S be a normalized *Google matrix* and let $X \in \mathbb{R}^N$ denote the right eigenvector of S corresponding to the eigenvalue 1. Then the linear order $\prec$ on $\mathcal{N} \times \mathcal{N}$ where

$$a \prec b \text{ ,if} \begin{cases} x_b < x_a \\ a \leq b, \quad x_a = x_b \end{cases}$$

is called the Internet order of the first N natural numbers.

The PageRank vector of the normalized Google matrix was computed by Arnoldi's method in FCS and by the power iteration method for INN.

In the next table we show the comparison of FCS's original results and the results obtained with our algorithm applied to the FCS normalized Google matrix the difference is caused by rounding errors.

FCS N^9	FCS N^8	FCS N^7	INN^* N^5
2	2	2	2
3	3	3	3
5	5	5	5
7	7	7	7
4	4	4	4
11	11	11	11
13	13	13	13
17	17	17	17
6	6	6	6
19	19	19	19
9	9	9	9
23	23	23	23
29	29	29	8
8	8	8	29
31	31	31	10
10	10	10	31
37	37	37	37
41	41	41	41
43	43	43	14
14	14	14	43
47	47	47	15
15	15	15	47

Applying the iterative eigenvalue computation algorithm we obtained the next table. There

- $i = \max\{l | r_l = p_l\}$.
- The j denotes the index of number 4.
- The k denotes the maximum index such that for $k > i$ it holds that $r_k = p^2_{k-i}$, that is $k = \max\{h | h > i, r_h = p^2_{h-i}\}$.

N	i	$r_i = p_i$	$j \vert r_j = p_1^2$	$k \vert r_k = p_{k-i}^2$	p_{k-i}^2	ϵ
$\mathbf{10^2}$	26	101	**27**	30	49	10^{-7}
$\mathbf{10^3}$	168	997	**169**	179	961	10^{-7}
$\mathbf{10^4}$	1229	9973	**1230**	1254	9409	10^{-7}
$\mathbf{10^5}$	9592	99991	**9593**	9621	11881	10^{-7}
10^5	9592	99991	**9593**	9657	97969	$\mathbf{10^{-9}}$
10^5	9592	99991	**9593**	9657	97969	$\mathbf{10^{-11}}$
10^5	9592	99991	**9593**	9657	97969	$\mathbf{10^{-15}}$
$\mathbf{10^6}$	78498	999983	**78499**	78525	10609	10^{-7}
$\mathbf{10^6}$	78498	999983	**78499**	78666	994009	10^{-11}
$\mathbf{10^6}$	78498	999983	**78499**	78666	994009	10^{-15}

In the FCS case the highest ranked 32 numbers are:
$n = 2, 3, 5, 7, 4, 11, 13, 17, 6, 19, 9, 23, 29, 8, 31, 10, 37,$
$41, 43, 14, 47, 15, 53, 59, 61, 25, 67, 12, 71, 73, 22, 21.$

In contrast for the INN case the following seems to be true:

Conjecture 1 *The Internet order $\prec$ of the first N natural numbers satisfy:*
1. $n \prec 4$ *iff* n *is a prime.*
2. $n \prec 6$ *iff* n *is a prime power.*

By our computation the conjecture holds for $N \leq 10^6$.

Internet and the Primes, Correlation Clustering

N. Bansal, A. Blum and S. Chawla, 2003: "Given a fully-connected graph G with edges labeled + (similar) or − (different), find a partition of the vertices into clusters that agrees as much as possible with the edge labels." This means either

• maximizing agreements: the number of + edges inside clusters plus the number of − edges between clusters

or

• minimizing disagreements: the number of − edges inside clusters plus the number of + edges between clusters.

They proved that optimal clustering is NP-hard. They presented approximation algorithms for both minimizing disagreements and for maximizing agreements.

L. Aszalós and M. Bakó, 2013, compared several correlation clustering methods and applied they for graphs of integers too. Let $0 \leq d < D$ be integers. Label the vertices of a graph by the integers $1, \ldots, n$. If a and b has at least D proper common divisors then there is an edge between them with label $+$. If a and b has at most d proper common divisors then there is an edge between them with label $-$.

Consider the special case $d = 0$ and $D = 1$. Then there exist an edge between a and b
$--$ with label $+$, if $\gcd(a, b) > 1$,
$--$ with label $-$, if $\gcd(a, b) = 1$.

Denote this graph with n nodes by N_n.

Questions: What is the optimal correlation clustering of N_n? Doest the structure of the optimal correlation clustering of N_n depend on n?

For a partition $\mathcal{P}$ of N_n denote by $c_n(\mathcal{P})$ the number of conflicts. Let

$$R(a,b) = \begin{cases} 1 & \text{if } \gcd(a,b) > 1 \\ 0 & \text{if } \gcd(a,b) = 1. \end{cases}$$

$$\delta(\mathcal{P},a,b) = \begin{cases} 1 & \text{if } a,b \text{ belong to the same class} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$c_n(\mathcal{P}) = \sum_{1 \le a < b \le n} \left(R(a,b)(1-\delta(\mathcal{P},a,b)) + (1-R(a,b))\delta(\mathcal{P},a,b) \right).$$

Proposition 1 *Let $\mathcal{P}_0$ be the trivial partition of N_n, i.e, for which all nodes belong to one class. Then it is never an optimal clustering and*

$$c_n(\mathcal{P}_0) = \frac{3}{\pi^2} n^2 + O(n).$$

Proposition 2 *For $n = 2, 3$ the partitions $\{\{1\}, \{2\}\}$ and $\{\{1\}, \{2\}, \{3\}\}$ have 0 conflicts, but the trivial partitions $\{1, 2\}$ and $\{1, 2, 3\}$ have 1 and 3 respectively.*

If $n \geq 4$ then there exists at least one prime $n/2 < p \leq n$. Removing p from $\mathcal{P}_0$ and putting it in a new class, the number of conflicts decreases by $n - 1$.

$c_n(\mathcal{P}_0)$ *is exactly the number of pairs* (a, b) *with* $1 \leq a < b \leq n$ *and* $\gcd(a, b) = 1$. *Thus*

$$c_n(\mathcal{P}_0) = \sum_{b=2}^{n} \varphi(b) = \frac{3}{\pi^2} n^2 + O(n).$$

Let $2 = p_1 < p_2 < \ldots < p_k \leq n$ be all primes at most n. For a prime $p \leq n$ put

$$S_p = \{i \ : \ i \leq n, \text{if a prime } q \text{ divides } i \text{ then } q \geq p\}.$$

Based on computer experiments for $n \leq 2000$ we propose

Conjecture 2 *The optimal correlation clustering of N_n is*

$$\{1\} \cup_{j=1}^{k} S_{p_j}.$$