COMBINATORIAL STRUCTURES FOR ANONYMOUS DATABASE SEARCH

Klara Stokes

Dipòsit Legal: T-1799-2011

Combinatorial Structures For Anonymous Database Search

Dissertation submitted to the Department of Computer Engineering and Mathematics of Universitat Rovira i Virgili, Tarragona, in the fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Author:
Klara Stokes

Director:
Dr Maria Bras Amorós

Klara Stokes

Combinatorial Structures For Anonymous Database Search

PH.D. DISSERTATION

Directed by Dr Maria Bras Amorós

Department of Computer Engineering and Mathematics

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2011

UNIVERSITAT
ROVIRA I VIRGILI

DEPARTAMENT D'ENGINYERIA INFORMÀTICA
I MATEMÀTIQUES

I STATE that the present study, entitled "Combinatorial Structures For Anonymous Database Search", presented by Klara Stokes for the award of the degree of Doctor, has been carried out under my supervision at the Department of Computer Engineering and Mathematics of this university, and that it fulfils all the requirements to be eligible for the European Doctorate Award.

Tarragona, October 25, 2011

Dr Maria Bras Amorós, Doctoral Thesis Supervisor

Acknowledgements

Many thanks go to my director for her great patience and support. I am also very grateful to my father, my mother, Nora, Vicenç and the numerous people who have helped me in the elaboration of this thesis. I also want to thank the Department of Mathematics of the Stockholm University for their hospitality during my stay there during the winter 2010-2011.

Contents

Chapter 1

Introduction

The society of today is sometimes classified as a society of information. Indeed, many aspects of modern life are based on the easy access to information. The quantity of information that is available from any computer connected to the internet is increasing rapidly and without much control. Although the quality of information does not seem to increase with the same velocity, resources are for example invested in order to make it possible to access the humanities whole collection of information resources, “with only a click”. These projects include the digitalization of vast quantities of printed material that form part of the cultural heritage of humanity.

In the old days navigation was an art. In our days the navigation on the internet is challenging because of the large quantities of available information. The web based search engines were invented as an answer to this challenge. These search engines use advanced sorting algorithms and crawling techniques combined with user profiling and interest analysis in order to find the information that the user wants. Originally, the contact between the search engine and the user was based only on the queries posted by the user and the query answers given by the search engine. However, nowadays most search engines also provide other services to their users, like for example web-based email and social network services. Of course, the use of all these services reveals information about the user to the server. This information is collected by the server in a user profile database that presumedly is used mainly for commercial purposes, for example to send directed advertisements to the users on the webpages provided by the server. But the same information can also be sold to other companies. Indeed the user has no control at all of the information about himself that he gives to the server. He can not know what the server will do with the information and he is usually not even aware of the fact that he is revealing sensitive information about himself to a powerful company.

1.1 Privacy protection in communications

There are several disciplines that study different aspects of privacy in communications. Here we will describe some of these shortly with the purpose to contrast them with the protocols that will be treated in this thesis.

1.1.1 Private information retrieval

The discipline that studies how a user should retrieve an element from a database or a search engine, without the system or the server being able to deduce which element is the object of the user's interest, is called Private Information Retrieval (PIR). The name and the discipline were introduced in the works of Chor, Goldreich, Kushilevitz and Sudan [20, 22].

A first result, found in these works, says that the only way to guarantee complete privacy, when using one single database, is by making the user access all the information in the database. Because of this, the first PIR protocols were initially designed for situations where there exist several copies of the same database, without these copies being intercommunicated. In this case, privacy refers to each of the servers individually [20, 22].

Later, computational PIR (cPIR) was introduced, dealing with privacy against one single database [21, 51]. In this case, there is a unique server with limited computational capacity and the privacy is relaxed to computational privacy. This means that the computations the server has to perform in order to gather enough information on the queries of a user to vulnerate her privacy, exceeds the capacity of the server. To distinguish the original PIR from the computational PIR, the former is called information theoretic PIR.

A major issue with cPIR schemes is that they are computationally expensive. The database needs to process all its entries for every query sent by the users, since otherwise it would be able to deduce in what entries the user is not interested. New protocols have been presented lately, based on noise over lattices instead of on number theoretical problems. The computational cost is then lowered, but communication performance obtained is worse [3]. One interesting example of this is [38] where a higher efficiency is obtained accepting a minor probability of error in the answer of the query.

Apart from cPIR, there are PIR protocols based on the assumption that a trusted hardware is installed in the database, so-called trusted-hardware based PIR (thPIR). This is a rather prosperous assumption, and several ideas for thPIR protocols have been presented, see for example [86]. However, the assumption of the existence of a trusted hardware, restricts the applications of these protocols to particular situations.

The drawbacks we observe with existing PIR protocols are the following:

- PIR protocols usually model the database as a vector in which the user knows the physical address of the item she is interested in. This is a very unreal assumption, e.g. think of a user querying a search engine. One exception is the protocol described in [23], that provides PIR for keyword queries;
- Theoretical PIR protocols have complexity that is linear in the size of the database. To avoid giving the server any clues of the interests of the user, the protocol must be such that the server processes all entries in the database for every query;
- It is assumed that the database server cooperates in the PIR protocol. But it is the user who is interested in her own privacy, whereas the motivation for the database server is dubious; actually, PIR is likely to be unattractive to most companies running queryable databases, as it limits their profiling ability.

1.1.2 Mixing

Digital mixes were invented by David Chaum in 1981, in order to provide anonymous email. A mix works at the network layer, and tries to hide the meta-data associated to a communication, i.e. avoid traffic analysis. One common approach for the construction of mixes is to use a chain of multiple untrusted relays, e.g. MorphMix [63] and Tarzan [36], but it is also possible to use a single trusted relay, as is the case with the example of PIR applied to mixing that can be found in [4].

A well-known software, providing traffic analysis resistance for interactive communication, is Tor [81]. It is an example of mixing using repeated public key cryptography through a chain of untrusted relays, called *onion-routing*. However, these are not intended to offer private information retrieval. They protect the transport of data, but give no end-to-end protection (at the application level). A server may link the successive queries submitted by the same user (e.g by using cookies), and in that way be able to profile and re-identify the user.

This last observation is generally true for all systems working on the network layer, hence for all mixers. Although anonymity on network level is achieved, so that the user's IP is maintained in secret, the collection of network traffic originating from her (secret) IP will reveal her by its content, e.g. through user names, query contents, etc.

There exist other systems that provide privacy protection for communications that do not classify under the disciplines described above. Here we mention some of these.

1.1.3 Anonymous database search

Anonymous database search was introduced in order to provide a system that makes it possible for a server (i.e. data owner) to publish data in a controlled way that allows for an authorized client to anonymously and securely query a server for documents containing a desired keyword [62]. We choose to interpret the concept in a more general way, and define anonymous database search as the discipline that studies how a client can retrieve an element from a database server without the server being able to tell who posted the query among all clients using the server.

The protocol described in [62] claims to protect not only the identity of the client from the server, but also the content of his query, and ensure that the client does not learn more about the database than he asked for. As pointed out in the same article, such high ambitions are impossible to realize without involving more parties in the protocol than the pair client-server only. The solution that they provide is based on the introduction of two trusted third parties, one index server and one query router. The introduction of trusted third parties in a protocol usually implies a security risk.

1.1.4 Privacy Preserving Keyword Search

In a paper by Chang et al. a protocol that provides Privacy Preserving Keyword Search is described [19]. The protocol belongs to a class of protocols that are thought to protect the privacy of the client of a distributed file system. It is assumed that the server does not own the data it holds. Instead it is the client who uploads his encrypted data to the server and afterwards wants to consult the data without revealing his interests to the server. This scenario adapts well for example to clients of a server for remotedly stored email, but can in general not be applied in the case of a web-based search engine and its clients.

1.1.5 Goopir

In [31] a system named Goopir is proposed in which a user masks her target query by ORing it with $k-1$ fake queries and then submits the resulting masked query to a search engine or large database which does not need to cooperate (in fact, it does not even need to know that the user is trying to protect her privacy). Strictly speaking, Goopir does not achieve PIR as defined above; rather, it provides $h(k)-$private information retrieval, in that it cloaks the target query within a set of k queries of entropy at least $h(k)$. This system works fine but it assumes that the frequencies of keywords and phrases that can appear in a query are known and available: for maximum privacy, the frequencies of the target and the fake queries should be similar, so that the uncertainty $h(k)$ of the search engine about the real target query is maximum.

1.1.6 TrackMeNot

TrackMeNot [47] is a software available as a plugin for Firefox. It periodically issues randomized search-queries to popular search engines, e.g., AOL, Yahoo!, Google, and MSN. In this way it hides the users actual search trails in a cloud of ’ghost’ queries, significantly increasing the difficulty of aggregating such data into accurate or identifying user profiles. While practical at a small scale, if the use of TrackMeNot became generalized, the overhead introduced by ghost queries would significantly degrade the performance of search engines and communications networks. Also, the way the automatic ghost queries are submitted may be distinguishable from the way real queries are submitted, which could provide clues on how to identify the latter type of queries.

1.1.7 User-private information retrieval or user-controlled anonymous database search

In [28, 29] a protocol was presented for the protection of the query profile of the web-based search engine user. This protocol was based on the collaboration between several users who post each others queries in order to cause confusion on the origin of the query. The users of this protocol upload and download their queries to so-called "communication spaces". A communication space is a memory sector together with a cryptographic key that is used to encrypt and decrypt the content on the memory sector. The distribution of these communication spaces among the users is defined by a mapping of the users to the points of an incidence structure. An incidence structure is a set of so-called points and a family of subsets of the point set called blocks, or sometimes lines. The communication spaces are then represented by the blocks so that two users share a communication space if and only if their points are both incident with the same block.

The incidence structures used by the protocol are the combinatorial configurations. For references on combinatorial configurations, see for example [39, 42], or below. A combinatorial (r, k)-configuration is an incidence structure such that all blocks have the same number k of points, all points are on the same number r of lines, and every pair of points is contained in at most one block. These properties are useful, the regularity ensures that the same privacy is given to everyone and the last property ensure that the users only share the number of queries that we have assigned them to share. Indeed if a communication space was shared twice by the same pair of users, then these two users would share twice the number of queries compared to what was planned when the protocol was designed.

The protocol protecting the query profile of the users is called peer-to-peer user-private information retrieval, or shorter, P2P UPIR [28, 29]. User-private information retrieval is different from private information retrieval in that it does not protect the content of user queries from the server, but the identity of the user, so that the server can not know who posted the query. Regarding the P2P UPIR protocol, we consider that the following should be stressed:

- P2P UPIR has none of the disadvantages of cPIR, e.g. it does not need the cooperation of the server, it has sublinear complexity, and the database does not have to be modeled as a vector. Of course it is not a fair comparison, since the P2P UPIR protocol does something quite different from what usually is meant by PIR;

- Unlike mixers, P2P UPIR hides the profile of the user in front of the database/server. The users send queries on behalf of others, i.e. it is something like a mixer on application level;

- Unlike the Anonymous Database Search protocol in [62], the P2P UPIR protocol does not assume the collaboration of the server. Also, P2P UPIR does not use trusted third parties in order to introduce anonymity, but it distributes the trust between the other users (peers) of the protocol. The privacy risk in front of these peers is managed;

- Unlike the Privacy Preserving Keyword Search from [19] and other privacy preserving protocols for distributed file systems, P2P UPIR assumes that the data is owned by the server and that the server has no interest in preserving the privacy of the user;

- Unlike Goopir, no knowledge of the frequencies of all possible keywords and phrases that can be queried is required;

- Unlike TrackMeNot, the overhead of ghost query submission is avoided.

As we understand it, user-private information retrieval and anonymous database search denominate the same thing, except for the important difference that UPIR does not assume any collaboration from the server, so that UPIR can be used in situations when the server is not interested in preserving the privacy of the user. Some protocols for anonymous database search, like the one described in [62], require the collaboration of the server, and the user has to be authorized in order to follow the protocol. We consider that this difference is important, but we also think that there is no reason why the collaboration of the server should be assumed for anonymous database search. Therefore we use the concepts of anonymous database search and user-private information retrieval without distinction, but we also recommend, for clarifying purposes, the name user-controlled anonymous database search.

Using this notation, the peer-to-peer user-private information retrieval protocol is a peer-to-peer user-controlled anonymous database search protocol.

This thesis is about this protocol and also about combinatorial configurations, their construction and existence.

1.2 Configurations: a short background

In a combinatorial configuration no geometrical meaning is given to the terms point and line. A geometrical configuration, on the other hand, is a combinatorial configuration that can be embedded into the real euclidean or the real projective plane.

Combinatorial and geometric configurations appear in many areas of mathematics, classical geometry, combinatorics, topology, algebraic geometry, etc. When configurations are studied for their own sake, usually the interest is focused on problems of existence. Given a set of parameters, can we tell if a configuration with these parameters exists? And if there exists at least one, how many different configurations exist with these parameters?

The history of the study of configurations is long. One early example of a geometric configuration is the $(3,3)$-configuration of 9 points and 9 lines defined by the Pappus' (hexagon) theorem. Other early results in the subject were provided by Desargues, Steiner, Möbius and Cayley.

The name configuration was coined in Reyes' book from 1876 [64]. In the subsequent 35 years many basic results on configurations were published. Some authors from this era are Reye, Kantor, Martinetti, Schröter, Schönflies, Brunel, Burnside, Daublebsky and Steinitz. For example, Kantor counted the number of combinatorial and geometric $(3,3)$-configurations with 8,9 and 10 points. Kantor was however wrong about the number of geometric $(3,3)$-configurations on 10 points, as was discovered shortly afterwards. Also other results from this era were erroneous, but the errors remained undiscovered for almost a century, perhaps because nobody actually read their works during this time. Indeed, no major publications on configurations were made between 1910 and 1990, with some important exceptions, like the book by Levi [55], the book by Hilbert and Cohn-Vossen [46], and some publications by Coxeter.

After 1990, many results on combinatorial configurations were published by Gropp, see for example [39, 41, 40], and also Grünbaum has important contributions on geometric configurations. The new progresses induced a new interest for configurations, and recent publications on configurations have been made by Kaski, Östergård, Betten, Brinkman, Pisanski and Boben, see for example [49, 8, 59, 13]. For further background, the book [42] by Grünbaum, and the book chapter [39] by Gropp serve as recent general references on configurations, and the former has a very useful bibliography.

As commented before, this thesis treats on one hand some questions on the existence and the construction of combinatorial configurations, and on the other hand, an application of combinatorial configurations to user-private information retrieval. Regarding other examples of applications of combinatorial configurations in computer science, we have for example the application to key distribution for distributed sensor networks that can be found in [53, 54]. There are also applications for combinatorial configurations in coding theory, for example in the construction of LDPC codes, see [34, 35, 57, 82]. More applications in cryptography and coding theory can be found in a book about finite projective geometry by Beutelspacher and Rosenbaum [9]. The applications described there include message authentication codes (MAC) or in particular Cartesian authentication schemes (see [27]), secret sharing schemes (see for example [50]), and Reed-Muller codes.

1.3 Contents and contributions

The second chapter of this thesis contains the preliminaries. It is divided into two sections; the first treats the mathematical background and the second contains relevant notions from computer science about privacy and anonymity.

The use of a combinatorial configuration for the design of the protocol implies that the geometric properties of these combinatorial objects may have influences on the performance of the protocol. The third chapter is dedicated to the analysis of this phenomenon. The most important results of this analysis are the following.

1. The $(v, k, 1)$-BIBD, or with a different name, the $S(2, k, v)$ Steiner systems, are identified as optimal combinatorial configurations for P2P UPIR with respect to the diffusion of the real profile of the protocol user;

2. The finite projective planes are identified as the optimal combinatorial configurations for P2P UPIR, with respect to criteria as privacy in front of the server and storage efficiency;

3. The neighborhood of a point in a combinatorial configuration is recognized as a quasi-identifier of that point.

4. A modification of the P2P UPIR protocol is proposed in order to avoid the neighborhood problem;

5. The $(v, k, 1)$-BIBD are identified as optimal combinatorial configurations for the modified P2P UPIR protocol;

6. The theory of n-anonymity is applied to the neighborhood problem;

7. The use of transversal designs for n-anonymous P2P UPIR is proposed with respect to the neighbors of the points;

8. The use of $(v, k, 1)$-BIBD for v-anonymous modified P2P UPIR is proposed with respect to the neighbors of the points;

9. Collusions of adversary protocol users communicating only over the channels provided by the protocol are recognized as a privacy risk;

10. Triangle-free combinatorial configurations are proposed in order to
    avoid the privacy risk caused by collusions of users communicating over channels provided by the protocol;

11. For collusions of users communicating also over external channels the magnitude of this privacy risk is calculated.

The fourth chapter is dedicated to the existence and construction of combinatorial configurations. The following results are presented there.

1. An algorithm for the construction of projective planes is provided;

2. A subset of the natural numbers $D_{(r,k)}$ is associated to the parameter tuples of combinatorial (r, k)-configurations and then it is proved that this subset is a numerical semigroup. This result implies, for example, the following:

    - For any pair of integers $r, k \geq 2$ there exist infinitely many combinatorial (r, k)-configurations;
    - Given a pair of integers $r, k \geq 2$ there exists a positive number N such that for all integers $n \geq N$ there exists at least one combinatorial configuration with parameters

$$\left(n\frac{k}{\gcd(r,k)}, n\frac{r}{\gcd(r,k)}, r, k \right),$$

that is, when the number of points (and lines) is big enough, there is at least one configuration for any admissible parameter set;

- When $\gcd(r,k) = 1$, then every prime power $q \geq \max(r,k)$ belongs to $D_{(r,k)}$;
- Two different constructions of combinatorial configurations from other combinatorial configurations are provided.

3. A subset of the natural numbers $D^{\bigtriangledown}_{(r,k)}$ is associated to the triangle-free (r,k)-configurations and it is proved that for every pair of integers $r,k \geq 2$, the set $D^{\bigtriangledown}_{(r,k)}$ is a numerical semigroup. This implies for example the following:

   - For any pair of integers $r,k \geq 2$ and prime power
   $$q \geq (r-1)(k-1)$$
   there exists a triangle-free (r,k)-configuration with
   $$2(r-1)(k-1)kq^2$$
   points and
   $$2(r-1)(k-1)rq^2$$
   lines. This is to be compared to a previous bound given by the generalized Gray/$LC(r)$ configuration with r^r points and r^r lines. Our result is more general, since we also treat non-balanced configurations.
   - It is proved that there exist infinite families of triangle-free (r,k)-configurations. These families are different from the families which can be constructed from the results presented in [42], and also in this case it should be noticed that we treat both balanced and unbalanced configurations;
   - Given a pair of integers $r,k \geq 2$ it is proved that there exists a positive number N such that for all integers $n \geq N$ there exists at least one configuration with parameters
   $$\left(n\frac{k}{\gcd(r,k)}, n\frac{r}{\gcd(r,k)}, r, k\right),$$
   that is, when the number of points (and lines) is big enough, there is at least one configuration for any admissible parameter set;

- The proofs are constructive and can be used to define an algorithm for the construction of triangle-free configurations.

In the fifth and last chapter another view of the existence and construction problem of combinatorial configurations is presented. The chapter contains an analysis of the numerical semigroups $D_{(r,k)}$ with small multiplicity.

Some results in this thesis can also be found in the following articles:

1. M. Bras-Amorós, J. Domingo-Ferrer and K. Stokes (2009) *Configuraciones combinatóricas y recuperación privada de información por pares.* In Congreso de la Real Sociedad Matemática Española-RSME 2009, Oviedo, Spain.

2. M. Bras-Amorós and K. Stokes, *The semigroup of combinatorial configurations*. Semigroup Forum, accepted (also available as preprint arXiv: 0907.4230v3).

3. M. Bras-Amorós, K. Stokes and M. Greferath (2010) *Using (0,1)-geometries for collusion-free P2P user private information retrieval.* In Proceedings of The 19th International Symposium on Mathematical Theory of Networks and Systems, Budapest.

4. K. Stokes and M. Bras-Amorós (2010) *Optimal configurations for peer-to-peer user-private information retrieval.* Computers & Mathematics with Applications, 59:4, pp. 1568 – 1577.

5. M. Bras-Amorós and K. Stokes (2010) *On the existence of combinatorial configurations.* In Proceedings of the 3rd International Workshop on Optimal Networks Topologies (IWONT), 9-11 June 2010, Barcelona, pp. 145–168.

6. K. Stokes and M. Bras-Amorós (2011) *Associating a numerical semigroup to the triangle-free configurations.* Advances in Mathematics of Communications, 5:2, pp. 351 – 371.

7. K. Stokes and M. Bras-Amorós (2011) *On query self-submission in peer-to-peer user-private information retrieval.* In Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society (PAIS '11), Uppsala, Sweden.

8. K. Stokes, Oriol Farràs and Maria Bras-Amorós (2011) *Transversal designs: n-anonymous combinatorial configurations for anonymous database search.* Manuscript.

9. K. Stokes and V. Torra (2011) *Reidentification and k-anonymity: a model for disclosure risk in graphs.* Submitted.

The content in the thesis is distributed in these articles as follows. Section 3.1 can be found in articles 1, 4, 7, 8, Section 3.2 in articles 7, 8 and 9, Section 3.3 is unpublished, Section 4.1 in article 4, Section 4.2 in articles 2 and 5, Section 4.3 in articles 3 and 6 and Section 5 is unpublished. Section 2.2 and 3.2 has inspired article 9, although article 9 is not contained in this thesis.

Chapter 2

Preliminaries

2.1 Mathematical background

2.1.1 Graphs

Graphs are combinatorial objects with many applications. A graph can represent something which we see in our daily life, like the urban transport network in our city, or a social network, but it can also represent an abstract mathematical object, like the isogeny classes of elliptic curves, or as in this thesis, the incidences of an incidence structure. A graph is also in itself an incidence structure.

There are directed and undirected graphs.

Definition 2.1.1. *An undirected graph (V, E) is a set V and a set of 2-element subsets $E \subseteq 2^V$. The elements in V are called vertices and the elements in E are called edges.*

We write $e = v_1v_2$ when we mean $e = \{v_1, v_2\}$. Therefore we have that $v_1v_2 = v_2v_1$.

Definition 2.1.2. *A directed graph is a set V together with a set of ordered pairs of vertices $A \subseteq V \times V$. The elements in V are called vertices and the elements in A are called arcs or directed edges. The left vertex and the right vertex in an arc a are called the origin and the end of a and are denoted by $o(a)$ and $f(a)$, respectively.*

We write $a = v_1v_2$ when we mean that $o(a) = v_1$ and $f(a) = v_2$. We say that $o(a)$ and $f(a)$ are the extremities of a.

An undirected graph (V, E) is isomorphic to a directed graph (V, A) such that

$$a = o(a)f(a) \in A$$

implies that

$$\bar{a} = f(a)o(a) \in A.$$

The edge set of the undirected graph is then the set of pairs of arcs

$$E = \{e = (a, \bar{a}), a \in A\},$$

or if one prefers, the set of arcs A modulo the inversion $\overline{v_1v_2} = v_2v_1$. An undirected edge has two extremities $o(a) = f(\bar{a})$ and $f(a) = o(\bar{a})$. Two vertices v_1 and v_2 in a graph are said to be adjacent or neighbors if there is an arc a such that v_1 and v_2 are the extremities of a. The outgoing degree of a vertex $v \in V$ is the number of arcs $a \in A$ with $o(a) = v$. The ingoing degree of a vertex $v \in V$ is the number of arcs $a \in A$ with $f(a) = v$. The ingoing degree and the outgoing degree of a vertex in an undirected graph coincide and define the degree of that vertex. Hence the degree of a vertex v is the number of edges $e \in E$ such that $v \in e$. An undirected graph, such that all the vertices in V have the same degree r, is said to be r-regular. If all vertices in a subset $U \subseteq V$ of an undirected graph have the same degree r, then we say that the graph is r-regular over U.

Definition 2.1.3. *A bipartite undirected graph is an undirected graph in which the vertex set can be partitioned into two disjoint sets (or parts) $V = V_1 \cup V_2$, such that no two vertices in the same part are adjacent.*

Hence, in a bipartite undirected graph $(V_1 \cup V_2, E)$, all edges in E contain one vertex from V_1 and one vertex from V_2. We say that a bipartite graph is (n, m)-biregular if it is n-regular over V_1 and m-regular over V_2.

The path graph on n vertices is the directed graph with vertex set

$$V = \{1, \dots, n\}$$

and arc set

$$A = \{i(i + 1) : i, i + 1 \in V\}.$$

For an example of a path graph, see Figure 2.1. A graph morphism between two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is a map $\phi : V_1 \to V_2$, such that if two vertices $v_1, v_2 \in V_1$ are connected by an edge

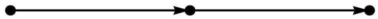

Figure 2.1: The path graph on 3 vertices

$v_1v_2 \in E_1$, then the two vertices $\phi(v_1), \phi(v_2)$ are either equal or there is an edge $\phi(v_1)\phi(v_2) \in E_2$. A path of length n in a graph G is an injective morphism from the path graph on n vertices into G. A graph G is connected if there is at least one path between any two vertices in G.

The cycle graph on n vertices is the directed graph with vertex set

$$V = \mathbb{Z}/n\mathbb{Z}$$

and arc set

$$A = \{i(i+1) : i \in V\}.$$

For an example of a cycle graph, see Figure 2.2. A cycle of length n in a graph G is an injective morphism from the cycle graph on n vertices into G.

A cycle of length 1 is called a loop. We defined graphs so that they have no loops. Also, we only consider graphs that have at most one edge with the same origin and end. A more general definition of graph would permit for several edges with the same origin and end. A family of at least two edges with the same origin and end is called a multiedge. A graph without loops and multiedges is called a simple graph. For an example of an undirected graph with a loop and a multiedge, see Figure 2.3.

In this document most graphs will be undirected and simple. If not otherwise indicated, the word graph will always refer to an undirected, simple graph.

Definition 2.1.4. *A tree is a connected, non-empty graph without cycles.*

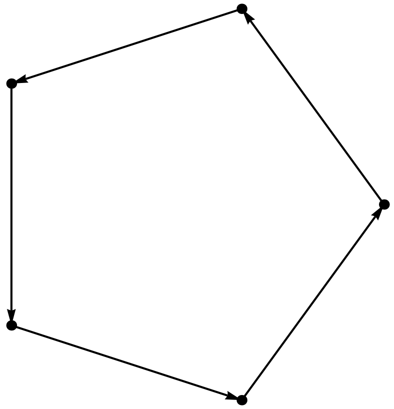

Figure 2.2: The cycle graph on 5 vertices

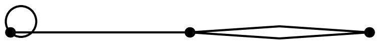

Figure 2.3: A graph with a loop and a multiedge

It can be proved that any connected graph G contains a tree which is maximal for the relation of inclusion. Such a tree is called a spanning tree of G.

The manner in which an object is represented determines in what ways we can analyze it. By representing a graph by a matrix, we can for example apply linear algebra and operator theory to the theory of graphs.

Definition 2.1.5. *Let $G = (V, A)$ be a directed graph and index the vertices by I so that $V = (v_i)_I$. The matrix $M = (x_{i,j})_{I \times I}$ where*

$$x_{i,j} = \sharp\{a \in A : o(a) = v_i, f(a) = v_j\}$$

is called the adjacency matrix of G.

If the directed graph is isomorphic to an undirected graph (V, E), then its matrix is symmetric. Since we have chosen to concentrate on undirected graphs we also give the definition of adjacency matrix of an undirected graph in terms of Definition 2.1.1.

Definition 2.1.6. *Let $G = (V, A)$ be an undirected graph and index the vertices by I so that $V = (v_i)_I$. The matrix $M = (x_{i,j})_{I \times I}$ where*

$$x_{i,j} = \sharp\{e \in E : v_i, v_j \in e\}$$

is called the adjacency matrix of G.

The adjacency matrix of an undirected graph is symmetric. The coefficients on the diagonal represent loops. The adjacency matrix of a graph without loops has only zeros in the diagonal. The adjacency matrix of a graph without multiple edges has its coefficients in $\{0, 1\}$. The adjacency matrix of a bipartite undirected graph $(V_1 \cup V_2, E)$ is of the form

$$\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$$

where B is a $|V_2| \times |V_1|$ matrix with coefficients in $\{0, 1\}$ and B^T is the transpose of B.

2.1.2 Elements of finite geometry

Definition 2.1.7. *An incidence structure $S = (\mathcal{P}, \mathcal{L}, I)$ is*

- *a set of "points" $\mathcal{P}$ and*

- *a set of "blocks" $\mathcal{L}$ together with*
- *a symmetric incidence relation $I \subseteq (\mathcal{P} \times \mathcal{L}) \cup (\mathcal{L} \times \mathcal{P})$.*

Equivalently, an incidence structure is a set $\mathcal{P}$ and a family of subsets $\mathcal{L} \subseteq 2^{\mathcal{P}}$. This latter definition permits us to use the shorter notation $(\mathcal{P}, \mathcal{L})$, so that the incidence relation becomes implicit.

We say that a point p and a block l are incident if $(p, l) \in I$ and we denote this relation by $p\ I\ l$ or $l\ I\ p$. If a point p and a block l are incident, then, following the latter definition of incidence structure, we say that p is in l or that l contains p. If two blocks l_1 and l_2 contain the same point p then we say that l_1 and l_2 meet in p.

Sometimes the terminology block design is used instead of incidence structure. However, the word block design is also used as a synonym for 2-design (see Section 2.1.2). We will prefer the notation incidence structure, but will use the words block design or design as synonyms for incidence structure in contexts which by tradition belong to design theory.

Initially, in an incidence structure there is no restriction on the cardinality of the blocks, so that they do not need to contain the same number of points. Also, a more general definition would allow for a block in $\mathcal{L}$ to appear several times, in which case $\mathcal{L}$ would be a multiset. However, we will always consider incidence structures in which $\mathcal{P}$ and $\mathcal{L}$ are disjoint, nonempty sets. We will also suppose that the incidence structures are connected, so that for any two points $p \neq q$ of an incidence structure, there are points $p_1, \ldots, p_{n-1}$ and blocks $l_1, \ldots, l_n$ and a chain of incidences

$$p\ I\ l_1\ I\ p_1\ I \cdots I\ p_{n-1}\ I\ l_n\ I\ q.$$

The rank of a block l is the number of points in l. If all the blocks in S have the same rank k then we say that S is uniform of rank k. The degree of a point p is the number of blocks incident with p. If all points in S have the same degree, then we say that S is regular of degree r.

There exists a notion of duality. By interchanging the roles of the points and the blocks in an indidence structure $(\mathcal{P}, \mathcal{L}, I)$ we obtain another incidence structure $(\mathcal{L}, \mathcal{P}, I)$.

To an incidence structure we associate a $|\mathcal{P}| \times |\mathcal{L}|$ matrix $M = (a_{ij})$ with coefficients

$$a_{ij} = \begin{cases} 1 \text{ if the point } p_i \text{ is on the block } l_j; \\ \\ 0 \text{ otherwise.} \end{cases}$$

We call this matrix the incidence matrix of the incidence structure.

We will also use a shorter representation of the incidence structure in which the points are represented by integers $\mathcal{P} = \{1, \ldots, n\}$ and the blocks by either the rows or the rows of a table with entries in $\mathcal{P}$. We call this table the incidence table of the incidence structure.

To an incidence structure we also associate a graph $G = (V, E)$ with vertex set $V := \mathcal{P} \cup \mathcal{L}$ and edge set the incidence relation $E := I$. We call this graph the incidence graph of the incidence structure. The incidence graph is then a bipartite graph with the points $\mathcal{P}$ in one set and the blocks $\mathcal{L}$ in the other. The edges between the two sets are defined by the incidence relation, so that two vertices p and l are connected if and only if the point p is incident with the block l.

The incidence graph has associated an adjacency matrix $A(G)$. The adjacency matrix of the incidence graph is a $(|\mathcal{P}| + |\mathcal{L}|) \times (|\mathcal{P}| + |\mathcal{L}|)$ matrix. The properties of the adjacency matrix from Section 2.1.1 together with the fact that the incidence graph is bipartite implies that the adjacency matrix of the incidence graph is symmetric and that only the two $|\mathcal{P}| \times |\mathcal{L}|$ minors which represent the adjacencies between the partition sets are different from 0. We also know that the symmetry implies that these two minors are the transpose of each other, so that only one of these two minors is necessary in order to represent all available data without redundancy. The non-redundant representation of the adjacency matrix A of the incidence graph is nothing but the incidence matrix M of the incidence structure. The relation between both is indeed

$$A = \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}.$$

By adding different restrictions to the definition of incidence structure, different combinatorial objects are obtained.

Graphs as incidence structures

Graphs are incidence structures. Indeed we can define an undirected graph as a set of vertices V and a set of edges E, together with a symmetric incidence relation $I \subseteq (V \times E) \cup (E \times V)$ such that for every element $e \in E$ there are exactly two elements $v_1, v_2 \in V$ such that $(e, v_1), (e, v_2) \in I$. Then the symmetry of I implies that also (v_1, e) and (v_2, e) is in I. Hence, comparing with Definition 2.1.7, we see that an incidence structure in which there are two points on every block is exactly a graph. A graph has an associated adjacency matrix as defined in

Definition 2.1.6. Since a graph is also an incidence structure, it also has associated an incidence matrix and an incidence graph. The incidence matrix of a graph is the $|V| \times |E|$ matrix $M = (a_{ij})$ with coefficients

$$a_{ij} = \begin{cases} 1 \text{ if the vertex } v_i \text{ is on the edge } e_j; \\ \\ 0 \text{ otherwise.} \end{cases}$$

The incidence graph of a graph is the bipartite graph with vertex set $V \cup E$ and edge set I. This graph is bipartite with V in one set and E in the other. The edges between the two sets are defined by the incidence relation, so that two vertices v and e are connected if and only if v is incident with e. In the original graph every edge is incident with exactly two vertices, implying that the incidence graph is 2-regular over the edge set E. The degrees of the vertex set V are the same as the degrees of the vertex set V in the original graph.

In this thesis the role of the graphs is therefore twofold;

- as incidence structures, graphs will be analyzed for their own sake, and
- as incidence graphs of incidence structures, graphs will be used as a tool for the analysis of incidence structures.

Balanced incomplete block designs

A $t-(v,k,\lambda)$ design, or simply a t-design, is a uniform incidence structure (or design) $S = (\mathcal{P}, \mathcal{L})$ of rank k with $|\mathcal{P}| = v$ and such that every t points occur in exactly λ blocks.

Definition 2.1.8. *A (v,k,λ)- balanced incomplete block design, or, in short notation form, a (v,k,λ)-BIBD is a $2-(v,k,\lambda)$ design.*

A $(v,k,1)$-BIBD is a $2-$design with $\lambda = 1$. That is, for every pair of points p_i, p_j there is exactly one block l with $p_i, p_j \in l$. The BIBD with $\lambda = 1$ will play an important role in this thesis. More generally, t-designs with $\lambda = 1$ are called Steiner systems.

Definition 2.1.9. *A Steiner system $S(t,k,v)$ is defined as a $t-(v,k,1)$ design.*

Hence a Steiner system is a uniform incidence structure of rank k, such that every tuple of t blocks meet in exactly one point. The most well-known Steiner systems are the Steiner triple systems $S(2,3,v)$. The $(v,k,1)$-BIBD are Steiner systems $S(2,k,v)$.

Necessary and sufficent conditions for the existence of a $(v,k,1)$-BIBD, or if the notation of Steiner systems is preferred, for the existence of a Steiner system $S(2,k,v)$, can be found for example in [2, 25].

Combinatorial configurations and partial linear spaces

A line in an incidence structure is a non-empty intersection of all blocks that are incident with a fixed pair of points. In particular, there are always at least two points on a line. Consider an incidence structure such that every pair of points are incident with at most one block. In this incidence structure there is a bijection between the blocks and the lines and we say that the incidence structure has points and lines. Many of the incidence structures that we will consider will be of this type.

Consider two points p and q in a line l. We say that p and q are on l, that they span l, that they are collinear by l, or that l joins p and q. We also use the notation $l = pq$. Consider two lines l and m that have a point p in common. We say that l and m meet in p or that l and m intersect in p and we use the notation $l \cap m = p$.

Definition 2.1.10. *A partial linear space is an incidence structure in which*

- *each point is on at least two lines,*
- *each line has at least two points and*
- *any two different points are incident with at most one line, or equivalently, any two different lines are incident with at most one point.*

If there are natural numbers s and t such that the partial linear space has $r = t + 1$ lines through every point and $k = s + 1$ points on every line, then we say that it has order (s,t).

Definition 2.1.11. *A combinatorial configuration is an incidence structure in which*

- *there are r lines through every point,*
- *there are k points on every line and*
- *through any pair of points there is at most one line, or equivalently, any pair of lines meet in at most one point.*

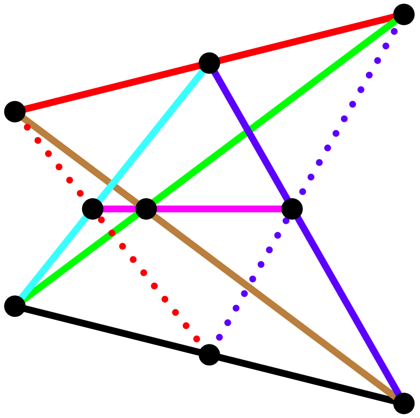

Figure 2.4: The Pappus configuration

We use the notation combinatorial (v, b, r, k)-configuration to say that the combinatorial configuration has

- v points,
- b lines,
- r lines through every point and
- k points on every line.

When v and b is not known or is not important,then we use the shorter notation (r, k)-*configuration.* We say that a combinatorial configuration is balanced if $r = k$. This implies that $v = b$. Note that a combinatorial (r, k)-configuration is a partial linear space of order (k, r). A combinatorial (v, b, r, k)-configuration is also a $1-(v, k, r)$ design such that for every pair of points p_i, p_j there is at most one block l such that $p_i, p_j \in l$. The parameter $|\mathcal{L}| = b$ of a configuration can be calculated from the three parameters v, k and r of the design, see Theorem 2.1.55.

Example 2.1.12. *The sets of points and lines in the classical Theorem of Pappus, result in the balanced combinatorial* $(9, 9, 3, 3)$*-configuration in Figure 2.4.*

Example 2.1.13. *A* $(v, k, 1)$*-BIBD is a* $2-$*design with* $\lambda = 1$*, that is, for every pair of points* p_i, p_j *there is exactly one block* l *with* $p_i, p_j \in l$*. Therefore any* $(v, k, 1)$*-BIBD is a combinatorial* (v, b, r, k)*-configuration with* $r(k-1) = v$*.*

Example 2.1.14. *An r-regular undirected graph is a combinatorial $(r, 2)$-configuration.*

A geometric configuration is a combinatorial configuration which can be embedded into the real euclidean or the real projective plane. In this text, when we use the word configuration we mean combinatorial configuration, and we do not attach any geometric significance to the terms point and line. Also, we consider the empty configuration, that is, a (v, b, r, k)-configuration with $v = |\mathcal{P}| = 0$ and $b = |\mathcal{L}| = 0$, to be a $(0, 0, r, k)$-configuration for every $r, k \in \mathbb{N}$, $r, k \geq 2$.

2.1.3 Polygons in combinatorial configurations

Definition 2.1.15. *In a configuration, by a triangle we mean a triplet of points, pairwise connected by lines, such that there is no line incident with all the three points.*

More generally we have the following definition of n-gon in a combinatorial configuration.

Definition 2.1.16. *In a combinatorial configuration, by an n-gon we mean a set of n distinct points $\{p_1, \dots, p_n\}$ and a set of n distinct lines $\{l_1, \dots, l_n\}$ such that in the incidence relation of the configuration there is the incidence chain*

$$p_1 \; I \; l_1 \; I \; p_2 \; \dots \; p_n \; I \; l_n \; I \; p_1.$$

In a combinatorial configuration, a triangle is an n-gon with $n = 3$. We say that an n-gon with $n = 4$ is a quadrangle.

Definition 2.1.17. *A triangle-free configuration is a configuration without triangles.*

Definition 2.1.18. *An (α, β)-geometry is a connected partial linear space such that if p is a point and l is a line not incident with p, then there are exactly m points $p_1, \dots p_m$ and m lines $l_1, \dots, l_m$ such that there is a chain of incidences $p \; I \; l_i \; I \; p_i \; I \; l$, for $i = 1, 2, \dots, m$ and m can take the values α or β.*

The triangle-free configurations are (α, β)-geometries with $(\alpha, \beta) = (0, 1)$. An equivalent definition of triangle-free configuration is therefore that, given a line l and a point p not on l, there is at most one line through p intersecting l, or equivalently, at most one point on l collinear with p.

Definition 2.1.19. *A partial geometry is an (α, β)-geometry with $\alpha = \beta$.*

A partial geometry with $\alpha = 1$ is a generalized quadrangle.

Definition 2.1.20. *A (finite) generalized quadrangle is a combinatorial configuration such that if p is a point and l is a line not incident with p, then there is a unique pair $(q, m) \in \mathcal{P} \times \mathcal{L}$ such that there is the chain of incidences $p \; I \; m \; I \; q \; I \; l$.*

Consider a generalized quadrangle C and let p_1 be a point and l_1 a line not incident with p_1 in C. Then there is a unique point p_2 on l_1 and a unique line l_2 that goes through both p_1 and p_2, such that we have the chain of incidences $p_1 \; I \; l_2 \; I \; p_2 \; I \; l_1$. Consider a second point p_3 on l_1 and a line $l_3 \neq l_1$ through p_3. Then p_1 and l_3 are not incident, because if they were, then p_3 and l_3 would be another pair of a point and a line with incidence chain $p_1 \; I \; l_3 \; I \; p_3 \; l_1$, contradicting the assumption that C is a generalized quadrangle. Therefore there exists a point p_4 on l_3 and a line l_4 and the incidence chain $p_1 \; I \; l_4 \; I \; p_4 \; I \; l_3$. Concluding, we have the incidence chain $p_1 \; I \; l_2 \; I \; p_2 \; I \; l_1 \; I \; p_3 \; I \; l_3 \; I \; p_4 \; I \; l_4 \; I \; p_1$. In other words, we have a quadrangle through the point p_1 and the line l_1. Therefore, a generalized quadrangle is a combinatorial configuration in which every non-incident point-line pair (p, l) is on a quadrangle.

2.1.4 Latin squares

Latin squares are related to the construction of both projective planes and transversal designs.

Definition 2.1.21. *Let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$ be an alphabet with n symbols. A Latin square of order n is an $n \times n$ matrix with coefficients in $\mathcal{A}$, such that every symbol from $\mathcal{A}$ appears exactly once in every row and exactly once in every column.*

For an example of a Latin square, see Figure 2.5

Definition 2.1.22. *Let A and B be two Latin squares of order n over the alphabets $\mathcal{A}_1$ and $\mathcal{A}_2$. We say that A and B are orthogonal, if, for every pair of symbols $(x, y) \in \mathcal{A}_1 \times \mathcal{A}_2$, there is only one pair of indices $1 \leq i, j \leq n$, such that $A(i, j) = x$ and $B(i, j) = y$.*

For an example of two orthogonal Latin squares of order 3, see Figure 2.6. The existence of a pair of orthogonal Latin squares is equivalent to the existence of a Graeco-Latin square. A Graeco-Latin square is an $n \times n$ matrix M with coefficients in $\mathcal{A}_1 \times \mathcal{A}_2$, such that every element $(x, y) \in \mathcal{A}_1 \times \mathcal{A}_2$ appears only once in M and such that the projection

1	2	3	4	5	6
2	3	4	5	6	1
3	4	5	6	1	2
4	5	6	1	2	3
5	6	1	2	3	4
6	1	2	3	4	5

Figure 2.5: A Latin square of order 6

1	2	3
2	3	1
3	1	2

1	3	2
2	1	3
3	2	1

Figure 2.6: Two orthogonal Latin squares of order 3

on the first and the second coordenates gives a Latin square over A_1 and A_2, respectively. For an example of a Graeco-Latin square, see Figure 2.7. In 1782 Euler stated the conjecture that no Graeco-Latin squares of order $n = 2 \pmod 4$ exists. For $n = 2$ the conjecture is very easy to check. It took 118 years until Tarry [79] proved that there are no Graeco-Latin square for $n = 6$, the year 1900. It was not until 1959-1960 that Parker, Bose, and Shrikhande finally proved that the conjecture was false for every $n > 6$.

Theorem 2.1.23. *[12] There exists no Graeco-Latin square of order n when $n = 2$ and $n = 6$. For all other $n \geq 3$ there exists at least one Graeco-Latin square.*

The Graeco-Latin squares are sometimes also called Euler squares. Because of the equivalence of the existence of Graeco-Latin squares and pairwise orthogonal Latin squares, Theorem 2.1.23 says that there is no pair of orthogonal Latin squares of order $n \in \{2, 6\}$ and also that there

$\alpha 1$	$\gamma 2$	$\beta 3$
$\beta 2$	$\alpha 3$	$\gamma 1$
$\gamma 3$	$\beta 1$	$\alpha 2$

Figure 2.7: A Graeco-Latin square that represents the two orthogonal Latin squares in Figure 2.6

is at least one pair of orthogonal Latin squares for all orders $n \geq 3$, $n \neq 6$. As a consequence of this, there is no Latin square of order 6 that is orthogonal to the Latin square in Figure 2.5.

Definition 2.1.24. *A set of mutually orthogonal Latin squares (MOLS) of order n is a set of pairwise orthogonal Latin squares of order n.*

We have the following sharp upper bound on the number of MOLS of order n.

Theorem 2.1.25. *[1] The number x of MOLS of order n satisfies $x \leq n-1$, with equality if n is a prime power.*

Therefore, the number x of MOLS of order n is known when n is a prime power and when $n \in \{2, 6\}$. For any other value of n we know that

$$2 \leq x \leq n-2.$$

2.1.5 Finite projective and affine planes

This section is a short survey on projective and affine planes. For more details, see for example [9, 44, 43, 45, 48, 75].

Definition 2.1.26. *A finite projective plane is an incidence structure in which*

- *every two points span exactly one line,*
- *every two lines meet in exactly one point and*
- *there is a quadrilateral: a set of four points such that any line contains at most two of them.*

We say that two lines are parallel if they do not intersect. The second condition in the definition of projective plane implies that a projective plane has no parallel lines. In a projective plane every 3 distinct points are either collinear or form a triangle.

Example 2.1.27. *The projective plane over a finite field $\mathbb{P}^2(\mathbb{F}_q)$ is a finite projective plane of order q.*

Examples 2.1.28. $\mathbb{P}^2(\mathbb{F}_2)$ *has 7 points, 7 lines, three points on every line and three lines through every point. It is therefore a combinatorial* $(7,7,3,3)$-*configuration. See Figure 2.8.* $\mathbb{P}^2(\mathbb{F}_3)$ *has 13 points, 13 lines, four points on every line and four lines through every point, so it is a combinatorial* $(13,13,4,4)$-*configuration. See Figure 2.9.*

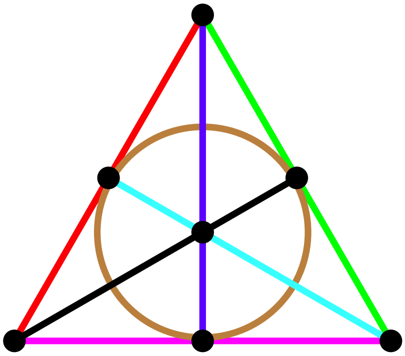

Figure 2.8: $\mathbb{P}^2(\mathbb{F}_2)$

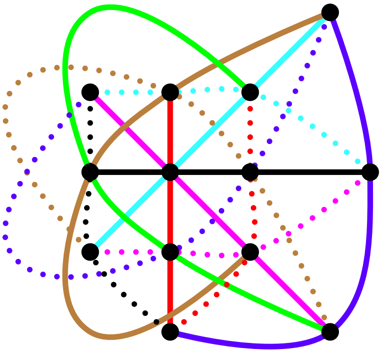

Figure 2.9: $\mathbb{P}^2(\mathbb{F}_3)$

More generally, a d-dimensional projective geometry is an incidence structure such that

- two distinct points span exactly one line,
- if a line meets two sides of a triangle, not at their intersection, then it also meets the third side,
- every line contains at least 3 points,
- the set of all points is spanned by $d+1$ points, and no fewer [87, 83].

Definition 2.1.29. *A finite affine plane is an incidence structure in which*

- *every two points span exactly one line,*
- *for every point p and line l not incident with p, there is exactly one other line $m \in \mathcal{L}$ such that p is incident with m and $l \cap m = \emptyset$,*
- *there is a quadrilateral: a set of four points such that any line contains at most two of them.*

The second condition in the definition of affine plane implies that the lines will be partioned into classes of parallel lines.

Example 2.1.30. *The affine plane over a finite field $\mathbb{A}^2(\mathbb{F}_q)$.*

Example 2.1.31. $\mathbb{A}^2(\mathbb{F}_2)$ *has 4 points, 6 lines, two points on every line and three lines through every point.*

Actually, nothing in these axiomatic definitions requires the point and line sets to be finite. Originally, these definitions were elaborated to characterize the projective and affine planes over fields and skew fields [48]. Historically, it is probably more correct to extend the infinite definitions to also include finite planes, than to extend the finite definitions to also include infinite planes. Indeed, the first planes to be studied were the planes over the real numbers, followed by the planes over the complex numbers. Therefore we should rather say: nothing in the definitions of projective and affine planes requires the point and the line sets to be infinite, they may as well be finite.

Desarguesian and Pappian planes

The axiomatic definitions for projective and affine geometries in higher dimension than 2, nicely resulted exactly in the type of spaces which were intended; projective and affine geometries over skew fields. However, in dimension 2 any intent to prove that all planes defined by the axioms are planes over skew fields, failed. The first counter-examples were found in the end of the nineteenth century.

The following two well-known theorems are crucial for the classification of projective planes. Both are true for any projective plane over a commutative field.

Theorem 2.1.32 (Pappus' theorem)**.** *Let a_1, b_1, c_1 be three points on a line l_1 and a_2, b_2, c_2 be three points on a line $l_2 \neq l_1$. Consider the three intersection points*

$$\begin{aligned} a_3 :=&\quad b_1c_2 \cap b_2c_1 \\ b_3 :=&\quad a_1c_2 \cap a_2c_1 \\ c_3 :=&\quad a_1b_2 \cap a_2b_1. \end{aligned}$$

If no three of the points a_1, b_1, a_2, b_2 are collinear, then a_3, b_3, c_3 are collinear.

Theorem 2.1.33 (Desargues' theorem)**.** *Let abc and $a^{'}b^{'}c^{'}$ be two triangles. If the three lines $aa^{'}$, $bb^{'}$ and $cc^{'}$ intersect in a point, then the three points $ab \cap a^{'}b^{'}$, $bc \cap b^{'}c^{'}$ and $ac \cap a^{'}c^{'}$ are collinear.*

As was noted by Hilbert [45], the Desargues' theorem is valid in any projective geometry of dimension at least three, and it is valid in a projective plane exactly when the plane can be embedded into a projective geometry of dimension at least three. For projective planes, it can be proved that the Desargues' theorem is true exactly in the projective planes over skew fields, and that the Pappus' theorem is true if and only if the field is commutative [48]. As a consequence of this, the projective planes over skew fields and over fields are called Desarguesian and Pappian projective planes, respectively.

For the finite case, the following famous algebraic theorem therefore has interesting geometric consequences.

Theorem 2.1.34 (Wedderburn's theorem)**.** *A finite skew field is commutative.*

As a corollary we get that in a finite projective plane the Desargues' theorem is true if and only if the Pappus' theorem is true.

Ternary rings

In [43, 44], ternary rings were introduced as the algebric structures that exactly defines coordinates for axiomatic projective planes.

Any ternary ring will give us a projective plane and any projective plane defines a ternary ring [87].

Observe that the concept of ternary ring has little to do with the concept of a ring. A ternary ring has one ternary operation, while a ring has two binary operations.

Definition 2.1.35. *A ternary ring is a set R with two distinguished elements 0,1 and a ternary operation $T : R^3 \to R$ satisfying the following conditions:*

- *(T1) $T(1, a, 0) = T(a, 1, 0) = a$ for all $a \in R$;*
- *(T2) $T(a, 0, c) = T(0, a, c) = c$ for all $a, c \in R$;*
- *(T3) If $a, b, c \in R$, the equation $T(a, b, y) = c$ has a unique solution y;*
- *(T4) If $a, a', b, b' \in R$ and $a \neq a'$, the equations $T(x, a, b) = T(x, a', b')$ have a unique solution x in R;*
- *(T5) If $a, a', b, b' \in R$ and $a \neq a'$, the equations $T(a, x, y) = b$ and $T(a', x, y) = b'$ have a unique solution x, y in R.*

If we only are interested in finite projective planes, and therefore only in finite ternary rings, we can forget about (T5). When R is finite, the condition (T5) is redundant.

Example 2.1.36. *Any field is a ternary ring with the ternary operation*

$$T(x, y, z) = xy + z.$$

Example 2.1.37. *Let J_9 be a vector space over $\mathbb{Z}/3\mathbb{Z}$ with basis $\{1, i\}$. Define $j = 1 + i$ and $k = 1 - i$ and give J_9 the multiplication defined by the quaternions $\{0, \pm 1, \pm i, \pm j, \pm k\}$, so that $\{\pm 1, \pm i, \pm j, \pm k\}$ is the quaternion group of order 8. Then J_9 is not a skew field. Instead, J_9 is what is called a near field.*

A (right) near-field is an associative ring K with 1 whose non-zero elements $K^* = K \setminus 0$ form a group under multiplication, such that:

1. *multiplication is right distributive: $(a + b)c = ac + bc$;*
2. *If $a, a', b \in K$ and $a \neq a'$, then the equation $xa - xa' = b$ has a (unique) solution x.*

If K is finite, then the axiom (2) is redundant [87].

The construction of a projective plane with coordinates in a ternary ring is analogous to the construction of a projective plane over a field.

Using projective coordinates we represent points by $(x : y : z)$ and lines by $[x : y : z]$, with x, y, z being elements of a ternary ring (R, T). Then a projective plane is a set of

- one point $(1 : 0 : 0)$,
- q points $(x : 1 : 0)$ with $x \in R$,
- q^2 points $(x : y : 1)$ with $x, y \in R$,

together with

- one line $[0 : 0 : 1]$ containing $(1 : 0 : 0)$ and $(x : 1 : 0)$,
- q lines $[0 : 1 : a]$ with $a \in R$, containing $(1 : 0 : 0)$ and $(x : -a : 1)$ for $x \in R$,
- q^2 lines $[1 : b : c]$ with $b, c \in R$, containing $(-b : 1 : 0)$ and $(x : y : 1)$ for any pair $x, y \in R$ such that $T(b, y, x) = -c$.

Properties of finite projective and affine planes

Finite projective and affine planes are projective and affine planes with a finite set of points. Every pair of points span one line, and therefore also the set of lines is finite.

For the finite projective planes, the following well-known relations are true.

Theorem 2.1.38. *[75] Let S be an incidence relation consisting of a finite number of points and a finite number of at least two blocks for which any two distinct points are on exactly one block, and there is an integer $n \geq 2$ such that any block has exactly $n + 1$ points. Then these assertions are equivalent:*

1. *S is a projective plane;*
2. *Any point of S is on at most $n + 1$ blocks;*
3. *Any point of S is on exactly $n + 1$ blocks;*
4. *There are exactly $n^2 + n + 1$ points in S;*

5. *S is a $2-(n^2+n+1, n+1, 1)$ design, or in other words, a $(n^2+n+1, n+1, 1)$-BIBD.*

As a corollary to Theorem 2.1.38 we deduce that a finite projective plane is always a combinatorial configuration. The number n is called the order of the finite projective plane.

The analogous properties for finite affine planes are stated in the following Theorem 2.1.39.

Theorem 2.1.39. *[75] Let S be a set system consisting of points and at least two blocks for which any two distinct points are on exactly one block, and there is an integer $n \geq 2$ such that any block has exactly n points. Then these assertions are equivalent:*

1. *S is an affine plane;*
2. *Any point of S is in exactly $n+1$ blocks;*
3. *There are exactly n^2 points in S;*
4. *S is a $2-(n^2, n, 1)$ design, or in other words, an $(n^2, n, 1)$-BIBD.*

Therefore any finite affine plane is also always a combinatorial configuration. The number n is called the order of the finite affine plane.

In [75], it is stated that determining which positive integers that are orders of finite projective planes is one of the most difficult questions in finite geometry. It is conjectured that the order of a finite projective or affine plane must be a power of a prime.

Conjecture 2.1.40. *The order of a finite projective plane is a power of a prime.*

At the moment the smallest order for which Conjecture 2.1.40 has not been checked is 12.

Finite projective and affine planes are not only combinatorial configurations, they are also examples of $(v, k, 1)$-BIBD. As we saw in Section 2.1.2 a $(v, k, 1)$-BIBD is a design in which every pair of points is connected by exactly one line. A third example of $(v, k, 1)$-BIBD are the unitals.

Definition 2.1.41. *A unital design is a $(n^3+1, n+1, 1)$-BIBD.*

Unital designs exist as sub-designs of projective planes of square order. A unital in a projective plane of order $n = q^2$ is a set of q^3+1 points that meets every line in either one or $q+1$ points [75]. However, any $2-(n^3+1, n+1, 1)$ design is called a unital.

Example 2.1.42. *Consider the projective plane constructed using coordinates from a finite field of square order $\mathbb{F}_{q^2}$, as described above. Then the points $(x : y : z)$ for which $xx^q + yy^q + zz^q = 0$ form a unital. This unital is a hermitian curve [75].*

Theorem 2.1.43. *[25] The following examples of infinite families of $(v, k, 1)$-BIBD (also called Steiner systems $S(2, k, v)$) are known.*

- *A finite projective geometry of order q and dimension n is a*

  $$(q^n + \ldots + q + 1, q + 1, 1)\text{-BIBD}$$

  for q a prime power and $n \geq 2$ (and it is also a $S(2, q+1, q^n+\ldots+q+1)$ Steiner system);

- *A finite affine plane of order q and dimension n is a $(q^n, q, 1)$-BIBD, for q a prime power and $n \geq 2$ (and also a $S(2, q, q^n)$ Steiner system);*

- *A unital design is a $(q^3 + 1, q + 1, 1)$-BIBD for q a prime power (and also a $S(2, q + 1, q^3 + 1)$ Steiner system);*

- *A Denniston design is a $(2^{r+s} + 2^r - 2^s, 2^r, 1)$-BIBD for $2 \leq r < s$ (and also a $S(2, 2^r, 2^{r+s} + 2^r - 2^s)$ Steiner system).*

The finite projective planes and the finite affine planes are finite projective and affine geometries of dimension 2, respectively.

2.1.6 Transversal designs

Transversal designs are block designs which are extremely “well organized”. This implies that they are very easy to represent and to construct, so they are well-suited for applications.

Definition 2.1.44. *Let $S = (\mathcal{P}, \mathcal{L}, I)$ be an incidence structure. A parallel class (or a spread) in S is a subset L of the block set $\mathcal{L}$, such that for all $p \in \mathcal{P}$ there is a unique block $l \in L$ such that $p \in l$. Hence L is a partition of the set of points $\mathcal{P}$.*

Two lines are parallel if they do not intersect and a line is parallel to itself. Parallelism is an equivalence relation and the set of lines that are parallel with l forms a class of parallel lines. If this class contains all points in the point set of the incidence structure, then it is a parallel class. Of course, not all incidence structures contain a parallel class, nor do all combinatorial configurations. For example, in a finite projective plane there are no pairs of parallel lines, since every pair of lines meet in one point. Consequently a finite projective plane has no parallel class.

Definition 2.1.45. *An incidence structure $S = (\mathcal{P}, \mathcal{L}, I)$ is a resolvable design if there exists a partition of $\mathcal{L}$*

$$\mathcal{L} = L_1 \cup \ldots \cup L_s$$

such that L_i is a parallel class.

In a resolvable design every line is in a parallel class. Example of resolvable designs are the finite affine planes.

Definition 2.1.46. *A group divisible design $(\mathcal{P}, \mathcal{L}, G)$ is an incidence structure $(\mathcal{P}, \mathcal{L}, I)$ such that*

- *G is a partition of the point set $\mathcal{P}$,*
- *every pair of points is contained either in a unique group or in a unique block, but not both.*

The parts in G are usually called groups, thereof the name group divisible design.

In a group divisible design the groups may be of different cardinality and the design is not necessarily uniform. A transversal design is a uniform group divisible design in which the group size $|G|$ equals the length of the blocks k. As a consequence, in a transversal design every block intersects every group in exactly one point.

Definition 2.1.47. *A transversal design $TD_\lambda(k, n) = (\mathcal{P}, \mathcal{L}, G)$ is a block design (X, B) such that*

- *$|X| = nk$,*
- *(X, B) is uniform of rank k,*
- *G is a partition of X in k parts (or groups) of size n,*
- *any group and any block contain exactly one common point, and*
- *every pair of points from distinct groups is contained in exactly λ blocks.*

In a transversal design (X, G, B) the set of groups G forms a partition of X, but it is not a parallel class since the elements of G do not pertain to B. On the other hand, if the block set B can be partitioned in parallel classes, then we get a resolvable transversal design.

Definition 2.1.48. *A resolvable transversal design $RTD(k, n)$ is a transversal design in which the block set B can be partitioned into k parallel classes.*

Definition 2.1.49. *We denote by $TD(k, n)$ a transversal design $TD_\lambda(k, n)$ with $\lambda = 1$.*

We have the following well-known relation between transversal designs and combinatorial configurations. For a better understanding we supply the short proof.

Proposition 2.1.50.

1. *A transversal design $TD(k, n)$ is always a combinatorial (kn, n^2, n, k)-configuration.*
2. *A transversal design $TD_\lambda(k, n)$ with $\lambda > 1$ is never a combinatorial configuration.*

Proof.

1. By definition, a $TD(k, n)$ has $v = kn$ points and the linesize is k. Any pair of points from distinct groups is contained in exactly one block and points from the same groups are not collinear. Since the group size is n, this gives $r = n$ lines through every point and we count $b = n^2$ lines. Finally, again, since any pair of points from distinc groups is on exactly one line and pairs of points from the same group are not collinear, the condition that any pair of points is on at most one line is satisfied.
2. In a $TD_\lambda(k, n)$ with $\lambda > 1$, every pair of points in distinct groups are in exactly λ blocks, so the condition that any pair of points is on at most one line is not satisfied in a $TD_\lambda(k, n)$.

□

Example 2.1.51. *The Pappus configuration in Figure 2.4 is a $TD(3, 3)$. The partition G consists of*

$$g_1 := \left\{ \begin{array}{l} \textit{the leftmost points on the red and the black lines} \\ \textit{together with the rightmost point on the lila line} \end{array} \right\};$$

$$g_2 := \{\textit{the three points in the middle}\};$$

$$g_3 := \left\{ \begin{array}{l} \textit{the rightmost points on the red and the black lines} \\ \textit{together with the leftmost point on the lila line} \end{array} \right\}.$$

Removing, for example, the three middle points from the configuration gives a $TD(2,3)$.

As is well-known, affine planes can be used to construct transversal designs as described in the following Lemma 2.1.52.

Lemma 2.1.52. *Whenever there exists a finite affine plane of order n, then for every $2 \leq k \leq n$ there exists a transversal design $T(k,n)$.*

Proof. As X, take k lines from one of the parallel classes of an affine plane of order n. As G, take the partition of the point set that is defined by these lines. As B, take the lines in the rest of the parallel classes of the affine plane, restricted to X. □

More generally, it is well-known that the existence of transversal designs $TD(k,n)$ is equivalent to the existence of a set of mutually orthogonal Latin squares (MOLS). Because of its importance for us, we here supply a proof of this result in the following Lemma 2.1.53.

Lemma 2.1.53. *[1] The existence of a set of $k-2$ MOLS of order n is equivalent to the existence of a $TD(k,n)$.*

Proof. Suppose that we have $k-2$ MOLS of order n

$$\{A_i := (x^i_{ab})\}_{i=1}^{k-2},$$

over the $k-2$ distinct alphabets $\mathcal{A}_i$ of empty intersection. The cardinality of $\mathcal{A}_i$ then equals the order of A_i, that is, it is n. We define the groups $G = \{g_i\}_{i=1}^{k}$ to be

- one group g_{k-1} containing the row indices $\{a_1, \ldots, a_n\}$,
- one group g_k containing the column indices $\{b_1, \ldots, b_n\}$ and
- $k-2$ groups g_i containing the n symbols of $\mathcal{A}^i$ for $i \in [1, \ldots, k-2]$.

Hence the point set X is defined as

$$X := \bigcup_{i=1}^{k-2} \mathcal{A}^i \cup \{a_1, \ldots, a_n\} \cup \{b_1, \ldots, b_n\}.$$

Now define the blocks as

$$\{a, b, x^1_{ab}, \ldots, x^{k-2}_{ab}\}$$

for $a \in [a_1, \ldots, a_n]$ and $b \in [b_1, \ldots, b_n]$. We can now simply affirm four of the five conditions in the definition of $TD(k,n)$:

- $|X| = kn$;
- (X, B) is a uniform block design of rank k;
- G is a partition of X in k groups of size n;
- Any group and any block contain exactly one common point.

Finally, the orthogonality of the MOLS and the two squares

$$\boxed{\begin{array}{cccc} 1 & 1 & \cdots & 1 \\ 2 & 2 & \cdots & 2 \\ \vdots & \vdots & & \vdots \\ n & n & \cdots & n \end{array}}$$

and

$$\boxed{\begin{array}{cccc} 1 & 2 & \cdots & n \\ 1 & 2 & \cdots & n \\ \vdots & \vdots & & \vdots \\ 1 & 2 & \cdots & n \end{array}}$$

implies that the constructed design satisfies the fifth condition, that every pair of points from distinct groups should be contained in exactly one block.

The other implication is obtained by reversing the process. Represent the transversal design $TD(k, n)$ in an incidence table with the lines in the rows and take two of the columns of the incidence table to be the row and column indices, respectively. Then every other column of the incidence table is a Latin square, and the set of Latin squares obtained from the $k - 2$ columns that are not used as indices, together form a set of $k - 2$ mutually orthogonal Latin squares (MOLS). □

Observe that, since they come from an affine plane, the transversal designs constructed in Lemma 2.1.52 are resolvable, while the transversal designs constructed in Lemma 2.1.53 are not necessarily resolvable. Lemma 2.1.52 can be used for the construction of at least one combinatorial (r, k)-configuration for any natural number $r, k \geq 2$.

Theorem 2.1.54. *For any pair of natural numbers $r, k \geq 2$, there always exists a combinatorial (r, k)-configuration.*

Proof. Take a transversal design $T(k, n)$ from a finite affine plane of order n as in Lemma 2.1.52, with

$$n \geq \max(r, k),$$

but as B take only r of the parallel classes. Then every line will have k points, there will be r lines through every point and through every pair of points there will go at most one line. □

The combinatorial configuration constructed in Theorem 2.1.54 has $v = nk$ points and $b = nr$ lines, and is therefore a combinatorial (nk, nr, r, k)-configuration.

2.1.7 Necessary conditions for the existence of combinatorial configurations

We have the following well-known necessary conditions for the existence of combinatorial (v, b, r, k)-configurations [39, 42].

Theorem 2.1.55. *In a combinatorial (v, b, r, k)-configuration we always have*

1. $v \geq r(k-1) + 1$ *and* $b \geq k(r-1) + 1$*;*
2. $vr = bk$*.*

Proof.

1. Take a point p. There are r lines through p with $k - 1$ more points, hence at least $r(k-1) + 1$ points. The other inequality is proved analogously.
2. There are v points in r incidence relations. And b lines in k incidence relations. Since the incidence relation is symmetric we get $vr = bk$.

□

Notation 2.1.56. *We call a quadruple satisfying the necessary conditions of Theorem 2.1.55 admissible.*

Observe that in the symmetric case ($v = b$ and $r = k$) a quadruple is admissible iff

$$v \geq r(r-1) + 1 = r^2 - r + 1.$$

2.1.8 Sufficient conditions for the existence of combinatorial configurations

We saw in Theorem 2.1.54 that for any pair of integers $r, k \geq 2$ there exists at least one combinatorial configuration. In this section we will see many other examples of theorems of existence of combinatorial configurations.

Balanced combinatorial configurations

This section will discuss some known results on the existence of balanced combinatorial configurations.

Theorem 2.1.57. *[40] Balanced combinatorial $(v, v, 3, 3)$-configurations exist if and only if the parameters are admissible, that is, if $v \geq 7$.*

Proof. Consider the incidence structure in which the point set is the set $\{1, \ldots, v\}$ and the lines are the columns of the following table. This table is called the incidence table of the incidence structure, see Section 2.1.2.

1	2	3	$\ldots$	$v-2$	$v-1$	v
2	3	4	$\ldots$	$v-1$	v	1
4	5	6	$\ldots$	1	2	3

This incidence structure is a combinatorial $(v, v, 3, 3)$-configuration, whenever $v \geq 7$. □

Example 2.1.58. *We can represent the same combinatorial configuration in several ways. The projective plane over $\mathbb{F}_2$ is the unique combinatorial $(7, 7, 3, 3)$-configuration. Below we have an incidence table of $\mathbb{F}_2$. The points are the numbers $\{1, \ldots, v\}$ and the lines are the columns of the table.*

1	1	1	2	2	3	3
2	4	6	4	5	4	5
3	5	7	6	7	7	6

Below we see the same $\mathbb{F}_2$ again, in what is called a cyclic representation.

1	2	3	4	5	6	7
2	3	4	5	6	7	1
4	5	6	7	1	2	3

In a cyclic representation the points are the elements in $\mathbb{Z}/(v)$ and the lines can be obtained from each other by applying a cyclic translation $f(x) = x + a \pmod{v}$ to all points in the line.

Theorem 2.1.59. *Balanced combinatorial $(v, v, 4, 4)$-configurations exist if and only if the parameters are admissible, that is, if $v \geq 13$.*

Proof. Consider the combinatorial configuration with point set

$$\mathcal{P} = \{1, \dots, n\}$$

and with line set the cyclic translations of the line $(1, 2, 5, 7)$. □

Example 2.1.60. *The following incidence table represents a combinatorial $(13, 13, 4, 4)$-configuration.*

1	2	3	4	5	6	7	8	9	10	11	12	13
2	3	4	5	6	7	8	9	10	11	12	13	1
5	6	7	8	9	10	11	12	13	1	2	3	4
7	8	9	10	11	12	13	1	2	3	4	5	6.

Cyclic representations are useful, because they are resource saving. Much less memory is needed to save one of the lines of a cyclic representation of a configuration than what is needed to save all the lines of the configuration. Also, the computational effort to calculate the rest of the lines is very small.

Definition 2.1.61. *The deficiency graph of a (v, b, r, k)-configuration is defined as the graph with the v points of the configuration as vertex set and an edge between two vertices when the corresponding points are not collinear.*

The following well-known lemma is very easy to prove. Just observe that a point in a combinatorial (r, k)-configuration is collinear with $r(k-1)$ other points, so that there are $v - (r(k-1)+1)$ points to which the point is not collinear.

Lemma 2.1.62. *Let $d := v - (r(k-1)+1)$. The deficiency graph is d-regular.*

The integer d is called the deficiency of the configuration. The deficiency graph can help to determine isomorphisms between combinatorial configurations, and also to prove the nonexistence of combinatorial configurations for some parameters.

Theorem 2.1.63. *Suppose that there exist a combinatorial (v, v, k, k)-configuration C, balanced and with deficiency d, and suppose that C is also a group divisible design, with $m = 1 + k(k-1)/(d+1)$ groups of common group size $n = d+1$. Define P and Q so that*

- $P = k^2 - v > 0$,
- $Q = k$.

Then

- *if m is even then P must be a square and if further $m = 4t + 2$ for some $t \in \mathbb{Z}$ and n is even, then k is a sum of two squares,*
- *if m is odd and n is even, then $Q = k$ is a square and the equation*

  $$Px^2 + (-1)^{(m-1)/2}ny^2 = z^2$$

  has a non-trivial solution in integers x, y, z, and
- *if both n and m are odd, then the two equations*

  $$(-1)^{(m-1)/2}nx^2 + Py^2 = z^2$$

  and

  $$(-1)^{(n-1)/2}nx^2 + Qy^2 = z^2$$

  both have or both have not a non-trivial solution in integers x, y, z.

As a consequence of Theorem 2.1.63 the following corollary can be obtained [41].

Corollary 2.1.64. *Each balanced combinatorial (v, v, r, r)-configuration with deficiency $d = 1$ has a 1-regular deficiency graph. This implies that it is a group divisible design and that k or $k - 2$ is a square.*

Proof. In a balanced combinatorial configuration that has parameter set (v, v, k, k) and deficiency $v - k(k-1) + 1 = 1$, given a point p there is exactly one point q such that p and q are not collinear. Hence the configuration is a group divisible design with group size $n = 2$ and number of groups $m = 1 + k(k-1)/2$, so that Theorem 2.1.63 applies. Suppose that n is even, then either m is odd so that $Q = k$ is a square, or m is even so that $P = k^2 - v = k^2 - (2 + k(k-1)) = k - 2$ is a square. □

Cyclic translations of the line $(1, 4, 5, 10, 12)$ yields a combinatorial $(21, 21, 5, 5)$-configuration. It is obvious that it works also for all $v \geq 23$. The $(21, 21, 5, 5)$-configuration is the projective plane over $\mathbb{F}_4$. There is no $(22, 22, 5, 5)$-configuration. This is because of Corollary 2.1.64. Since a $(22, 22, 5, 5)$-configuration would have deficiency 1 it can not exist, because neither $k = 5$ nor $k - 2 = 3$ is a square. Of course, this could have been deduced directly from Theorem 2.1.63, by inserting $m = 11$ and $n = 2$ and concluding that $k = 5$ should have been a square. We combine these results on the existence of combinatorial $(5, 5)$-configurations in the following Theorem 2.1.65.

Theorem 2.1.65. *Balanced combinatorial $(v, v, 5, 5)$-configurations exist if and only if the parameters are admissible and $v \neq 22$, that is, if $v = 21$ or $v \geq 23$.*

The unique $(31, 31, 6, 6)$-configuration is the projective plane over $\mathbb{F}_5$. The proof of the non-existence of a $(32, 32, 6, 6)$-configuration is the smallest which is not a consequence of the theorem by Bose and Connor. It is due to Schellenberg (1975). The proof of the non-existence of a $(33, 33, 6, 6)$-configuration is due to Kaski and Östergard (2006) who made an exhaustive computer search for such a configuration, with negative result. Therefore the existence of combinatorial $(6, 6)$-configurations is described by the following Theorem 2.1.66.

Theorem 2.1.66. *Balanced combinatorial $(v, v, 6, 6)$-configurations exist if and only if the parameters are admissible and $v \notin \{32, 33\}$, that is, if $v = 31$ or $v \geq 34$.*

Non-balanced combinatorial configurations

This section will discuss some existence results on non-balanced combinatorial configurations. Many of these results are due to Gropp [39, 40].

Theorem 2.1.67. *There exists a combinatorial $(v, b, r, 3)$-configuration if and only if the parameters are admissible, that is, whenever $v \geq 2r + 1$ and $vr = 3b$ [40].*

Theorem 2.1.68. *There exists a combinatorial $(v, b, r, 4)$-configuration for $v \equiv 4 \pmod{12}$, $v \geq 3r + 1$ and $vr = 4b$ [40].*

As the parameters increase, the knowledge of the existence of combinatorial configurations is more sparse.

There is also an asymptotic result, ensuring the existence of large configurations with what Gropp calls natural index ($t = \frac{r}{k} \in \mathbb{N}$).

Theorem 2.1.69. *For given k and r with $r = tk$, so that $t \in \mathbb{Z}$, there is a $v_0(k, t)$ such that there is a combinatorial (v, b, r, k)-configuration for all admissible parameters with $v > v_0$ [40].*

In particular the theorem can be applied for $t = 1$, that is, when the configuration is balanced. We have seen that it is true for $r = k \in \{3, 4, 5, 6\}$, but the theorem affirms that it is true for all $r = k$. Theorem 2.1.69 can not be applied when $t = \frac{r}{k} \notin \mathbb{N}$.

This short section does not in any way pretend to be a survey on previous results on the existence of combinatorial configurations. Plenty of other partial results that are not listed here can be found for example in [39, 40, 41, 42], or in their bibliography.

2.1.9 Previous results on the existence of triangle-free combinatorial configurations

Below the state of the art of the research on the existence of triangle-free configurations is explained, as far as it is known to the author.

The smallest polygon that can be contained in a triangle-free configuration is a quadrangle. As we saw in Section 2.1.3, a quadrangle in a combinatorial configuration $(\mathcal{P}, \mathcal{L}, I)$ is a set of four different lines l_1, l_2, l_3, l_4 and four different points p_1, p_2, p_3, p_4 such that the incidence relation of the configuration defines a sequence

$$l_1 \; I \; p_1 \; I \; l_2 \; I \; p_2 \; I \; l_3 \; I p_3 \; I \; l_4 \; I \; p_4 \; I \; l_1,$$

that is, a cycle of length 8 in the incidence graph. A triangle-free configuration in which every non-incident point-line pair is on a quadrangle is called a generalized quadrangle. The incidence graph of a generalized quadrangle of order $(r-1, k-1)$ is a bipartite, (r, k)-biregular graph with girth 8 and diameter 4. The following is a well-known necessary condition for a generalized quadrangle to exist.

Proposition 2.1.70. *[58] If a generalized quadrangle of order* $(r-1, k-1)$ *exists, then it has number of points*

$$v = |\mathcal{P}| = k((r-1)(k-1)+1)$$

and number of lines

$$b = |\mathcal{L}| = r((r-1)(k-1)+1).$$

There are several known families of generalized quadrangles [58]. All these families, except one, have order $(r-1, k-1)$ where $r-1$ and $k-1$ are powers of the same prime number. The exception is a family of generalized quadrangles of order $(r-1, k-1) = (q-1, q+1)$ where

q is a power of a prime number. The results on the orders of known generalized quadrangles as appearing in the book 'Finite generalized quadrangles' by Payne and Thas [58] is concluded in Proposition 2.1.71.

Proposition 2.1.71. *Let q be a power of a prime number. Then there exists a generalized quadrangle of order $(r-1, k-1)$ if*

$$(r-1, k-1) \in \{(q,1),(q,q),(q,q^2),(q^2,q^3),(q-1,q+1)\}.$$

The first question on the existence of triangle-free configurations, is answered by the following Theorem 2.1.72, but only in the balanced case. We have not found previous general results in the non-balanced case, that is, when $r \neq k$.

Theorem 2.1.72. *[42] For every integer $r \geq 2$ there exist (geometric) (r,r)-configurations that are triangle-free.*

The (r,r)-configuration used in the proof is what Pisanski calls a generalized Grey configuration [59] and Grunbaum a $LC(r)$ configuration [42]. It has r^r points and r^r lines.

The book by Grünbaum [42], which mostly treats configurations that are geometrically realizable, contains the following theorem that collects the available knowledge on the existence of triangle-free geometric $(3,3)$-configurations.

Theorem 2.1.73. *For every $v \geq 15$ except $v = 16$ and possibly $v = 23$ and $v = 27$, there are triangle-free geometric $(v,v,3,3)$-configurations.*

Theorem 2.1.73 contains the results by Betten et al. [8], who counted all triangle-free combinatorial configurations with $v \leq 21$ for $r = k = 3$. Their calculations show us that there exist triangle-free combinatorial $(3,3)$-configurations with

$$v \in \{15, 17, 18, 19, 20, 21\}.$$

The unique triangle-free $(3,3)$-configuration with $v = 15$ is the famous Cremona-Richmond configuration, which is a generalized quadrangle. In the tables of [8] it can be observed how the number of triangle-free combinatorial configurations grows very quickly with v.

Theorem 2.1.73 also contains results by Visconti [85]. Finally, the proof of Theorem 2.1.73 constructs larger configurations joining smaller ones in two different ways. This is interesting, and it is worth pointing out that both these constructions are different from the 'addition' of configurations used in this thesis. Using these constructions, starting with two triangle-free (r, r)-configurations with p and q points and p and q lines, respectively, the result is either a triangle-free (r, r)-configuration with $p + q - 1$ points and $p + q - 1$ lines, or one with $p + q + 1$ points and $p + q + 1$ lines.

Any geometric configuration is also a combinatorial configuration, and there is no triangle-free $(3, 3)$-configuration with $v = 16$ [56], so that the available knowledge on the existence of combinatorial $(3, 3)$-configurations at this moment coincides with the knowledge on the existence of geometric $(3, 3)$-configurations in Theorem 2.1.73. Observe that Theorem 2.1.73 does not count the number of triangle-free $(3, 3)$-configurations. It is not known how many of the triangle-free $(3, 3)$-configurations counted by Betten et al. are geometrically realizable [42].

Considering larger parameters, there is much less known already for triangle-free combinatorial $(4, 4)$-configurations. Recently, van Maldeghem constructed a triangle-free $(4, 4)$-configuration with $v = 40$ [11]. Since it satisfies the bound from Proposition 2.1.70, Proposition 4.3.9 says that it is a generalized quadrangle. There are also triangle-free $(4, 4)$-configurations with $v = 60$ (found by Boben), $v = 120$ and $v = 256$ [42], plus infinite families of triangle-free $(4, 4)$-configurations constructed from these using the two constructions from the proof of Theorem 2.1.73.

For triangle-free (k, k)-configurations the generalized Gray / $LC(r)$ configuration can be used to construct infinite families of triangle-free (k, k)-configurations in the same way.

Sinha constructs a family of triangle-free $(3, k)$-configurations with special parameters [69]. The Cremona-Richmond configuration appears as the smallest example of the members of this family.

Graphs and configurations are not the same thing, but some results in graph theory can be interpreted as if they treated configurations. Many proofs in this article are also expressed in the language of graphs. In particular, the following result on the existence of regular graphs, due to Sachs [66], is important and will be used later.

Theorem 2.1.74. *Let $r \geq 3$ and $g \geq 2$ be two integers. Then there always exists an r-regular graph of girth g.*

Because of Sachs' Theorem 2.1.74, there is always an r-regular graph of girth g, so it makes sense to ask for the smallest one. In graph theory an (r, g)-*cage* is an r-regular graph of girth g with the smallest possible number of vertices. It is conjectured that all cages of even girth are bipartite [60, 89]. We can identify a triangle-free (r, k)-configuration with its incidence graph, a connected, bipartite, (r, k)-biregular graph of girth at least 8. If we suppose the conjecture true, we therefore have that a triangle-free (r, r)-configuration with the smallest possible number of points and lines is exactly an $(r, 8)$-cage.

There is a well-known lower bound for the number of vertices in an (r, g)-cage [10] giving us the lower bound for the number of vertices in an $(r, 8)$-cage

$$n_0(r) = 2(1 + (r-1) + (r-1)^2 + (r-1)^3) = \frac{2(r-1)^4 - 2}{r-2}.$$

A regular cage of even girth that reaches this bound, is the incidence structure of a (balanced) generalized quadrangle [10].

In [52], Lazebnik, Ustimenko and Woldar constructed small and r-regular graphs of girth g for any $r \geq 2$ and $g \geq 3$.

Proposition 2.1.75. *[52] Let $r \geq 2$ and $g \geq 5$ be integers, and let q denote the smallest odd prime power for which $q \geq r$. Then there exists an r-regular graph of girth g and number of vertices*

$$2rq^{3g/4-a},$$

with $a = 4, 11/4, 7/2, 13/4$, for $g = 0, 1, 2, 3 \pmod 4$ respectively.

The smallest known r-regular graphs of girth 8, when r is not a power of a prime, are at the moment the ones constructed by Balbuena.

Proposition 2.1.76. *[7] Let r be an integer and q a power of a prime such that $3 \leq r \leq q$. Then there exists an r-regular bipartite graph of girth 8 with $rq^2 - q$ vertices in each bipartite set.*

The smallest known q-regular graphs when q is a power of a prime, were constructed by Gács and Héger.

Proposition 2.1.77. *[37] Let q be a power of a prime. If q is even then there exists a q-regular graph of girth 8 and with $2(q^3 - 3q - 2)$ vertices. If q is odd, then there exists a q-regular graph of girth 8 and with $2q(q^2 - 2)$ vertices.*

As recently was proved by Araujo-Pardo in [6], small odd girth g graphs can be obtained from small even girth $g + 1$ graphs. In particular, upper bounds on the number of vertices of an $(r, 7)$-cage can be obtained from the upper bound on the number of vertices of an $(r, 8)$-cage.

Proposition 2.1.78. *[6] Let $r \geq 3$ be an odd integer. If $f(r)$ is an upper bound for the number of vertices of an $(r, 8)$-cage, then an upper bound for the number of vertices of an $(r, 7)$-cage is*

$$f(r) - \frac{2(r-1)^2 - 2}{r-2}.$$

2.1.10 Numerical semigroups

Definition 2.1.79. *A numerical semigroup is a subset $S \subseteq \mathbb{N} \cup \{0\}$, such that*

- *S is closed under addition,*
- *$0 \in S$ and*
- *the complement $(\mathbb{N} \cup \{0\}) \setminus S$ is finite.*

We write $\langle a_1, \ldots, a_n \rangle$ to denote the numerical semigroup generated by the natural numbers $a_1, \ldots, a_n$ through addition. We also write $\{0, x_1, x_2, x_3, \rightarrow\}$ to denote that the numerical semigroup contains all natural numbers $n \geq x_3$.

Definition 2.1.80. *The multiplicity of a numerical semigroup is its smallest non-zero element.*

Definition 2.1.81. *The conductor of a numerical semigroup is the smallest element such that all subsequent natural numbers belong to the numerical semigroup.*

Definition 2.1.82. *Let S be a numerical semigroup. The largest element in $\mathbb{N} \cup \{0\} \setminus S$ is called the Fröbenius number.*

The conductor is then the Fröbenius number plus one.

Definition 2.1.83. *The gaps of a numerical semigroup S are the natural numbers that do not belong to S. The number of gaps is called the genus of the numerical semigroup.*

Example 2.1.84.

$$\langle 3, 7 \rangle = \{0, 3, 6, 7, 9, 10, 12, 13, 14, 15, 16, \ldots\}$$

is the numerical semigroup generated by 3 and 7. In this numerical semigroup

- *the multiplicity is 3,*

- *the Fröbenius number is 11,*
- *the conductor is 12 and*
- *the gaps are* $\{1, 2, 4, 5, 8, 11\}$.

A numerical semigroup of the form

$$\{0, a, \rightarrow\}$$

is called an ordinary numerical semigroup. Hence, in an ordinary numerical semigroup, all gaps are smaller than the multiplicity.

We say that a set of integers are coprime if the ideal they generate is $\mathbb{Z}$.

Lemma 2.1.85. *A set of integers generate a numerical semigroup if and only if they are coprime.*

The proof of this lemma can be found in the book [65], which serves as a general reference on numerical semigroups.

When the number of coprime generators is two, then it is easy to calculate the conductor of the generated numerical semigroup, with the help of the following Theorem 2.1.86

Theorem 2.1.86. *Two coprime positive integers* a, b *generate a numerical semigroup whose conductor is* $(a - 1)(b - 1)$.

When more than two generators of the numerical semigroup are involved, then the calculation of the conductor of a numerical semigroup generated by n elements is difficult [61]. However, the conductor can be bounded as a function of other properties of the numerical semigroup. For example, we have the following upper bound in terms of the genus (see Lemma 2.14 in [65].

Theorem 2.1.87. *The genus* g *and the conductor* c *of a numerical semigroup always satisfy*

$$2g \geq c.$$

2.2 Notions and definitions of privacy

As individuals in modern society we are all very well-documented. We are of course in public records like the census records, the tax records, the educational records, and perhaps also in the marriage records, the hospital records, the police records and the records of the national employment agency. There are also records which are generated from daily actions like for example the use of mobile phones, credit cards and automatic toll payment systems. These records are usually owned by companies. Indeed any company that provides a service to its customers is likely to record the preferences and habits of these customers. The access to such records is usually an important source for prosperous business.

The information collected by companies can also be important to researches and to society. For example, the collection of the location records of the mobile phones of the citizens in a geographical area provides a database of daily travelling habits that is a source of information of great importance when new infrastructure are planned. To get access to this information the politics must get (or buy) it from the phone companies.

Usually the customers are not completely aware of the fact that the company registers their actions. A clear example of this is the anger shown by customers who have bought an operative system to run on a computer or a portable device, and later discover that inside the operative system there are programs which report to the creator of the operative system the actions performed by the customer on this device.

Another way to collect information about citizens is to ask them. Most countries have a national statistics department, which collects information through surveys directed to the citizens. Usually, to the survey there is attached some information on how the citizens' private data is protected against intents of reidentification. The naive solution is to protect the database by simply removing the identifiers, like name, ID and social security number, from the tables. This solution has however shown to be far from satisfactory. In many cases it is rather easy to recover the identifier of the anonymized record [78].

Other more sofisticated solutions for protecting databases have been proposed. Examples of such solutions are methods for obtaining n-anonymity (usually called k-anonymity), rank-swapping, methods that use aggregation operators, clustering and noise addition. All these solutions are designed to be executed by the data owner, that is, by the organism or company that wants to publish the database. The data owner is then the only one responsible for the anonymization of the database. Indeed, the customers usually have no other option than to rely on the good-will of the data owner, even when the data owner is a private company without interest in preserving the privacy of their customers. As Google Sweden expressed it: "...you always have the choice not to use our services" [76].

The anonymization method treated in this document, the P2P UPIR protocol, is different in that it allows the customers, or clients, of a search engine, to anonymize the database produced by the collection of their queries even before it reaches the hands of the data owner, that is, the company behind the search engine.

The process of linking records or parts of records in a protected database to a record in the original database is called record linkage or reidentification. The model of reidentification that we present here is commonly used in research, when the researchers want to test a protection method by attacking it. The protected database is then compared to the original database, in order to deduce the quality of the protection. One could argue that a real world adversary most likely does not have access to the original database. Instead he may have access to auxiliary information, which we can suppose to be in table form, just like the original database; other databases, public census registers, etc. In this case a record linkage or reidentification process is done by combining the information in order to link a record or a part of a record in the protected database to a real individual. Observe that in order to link a protected record to a real individual the real individual must have a record in the original database. Therefore the model of a real world adversary can be representated according to the researchers' model presented here.

In data privacy, the assessment of risk is one of the elements of major importance. At present, several approaches have been studied in the literature. The major approaches are k-anonymity [67, 68, 77, 78], reidentification [32, 88] and differential privacy [33]. We will concentrate on two of these approaches: reidentification and k-anonymity, although we will denote the latter by n-anonymity, for notation reasons.

2.2.1 Reidentification

A database is a collection of records of data. We will suppose that all records correspond to distinct individuals or objects. Every record has a unique identifier and is divided into attributes. The attributes can be very specific, as the attributes "height" or "gender", or more general, as the attributes "text" or "sequence of binary numbers".

Suppose that the database can be represented as a single table. Let the records be the rows of the table and let the attributes be the columns.

The intersection of a row and an attribute is a cell in the table, and we call the data in the cells the entries of the database. Also other data structures, like for example graphs, are representable in table form. In the example of a graph the adjacency matrix is a table representation of the graph.

Let T be a table with n records and m attributes. Consider the underlying set of entries of T, $E = \bigcup\{T[i,j] : 1 \leq i \leq n, 1 \leq j \leq m\}$. We define the partition set $\mathcal{P}(T)$ of T to be the set of subsets of E.

Definition 2.2.1. *A method for anonymization of databases is any transformation or operator*

$$\begin{aligned} \rho : \quad D &\rightarrow D \\ \\ X &\mapsto Y, \end{aligned}$$

where D is a space of databases.

Then ρ, given a database X, returns a database Y. Since Y is a database, all entries in Y will correspond to a unique individual or object, which we will suppose to be the same individuals as the ones behind the records in X. Usually it is assumed that there is, in some sense, less sensible information about the individuals behind the records in X in the transformed database Y than there was in the original database X.

We propose the following formal definition of reidentification [74].

Definition 2.2.2. *Let ρ be a method for anonymization of databases, X a table with n records indexed by I in the space of tables D and $Y = \rho(X)$ the anonymization of X using ρ. Then a reidentification method is a function that given a collection of entries y in $\mathcal{P}(Y)$ and some additional information from a space of auxiliary informations A, returns the probability that y are entries from the record with index $i \in I$,*

$$\begin{aligned} r : \quad \mathcal{P}(Y) \times A &\rightarrow [0,1]^n \\ \\ (y, a) \qquad &\mapsto (P(y \in X[i]) : i \in I) \,. \end{aligned}$$

Researchers typically use parts of the original database X as auxiliary information. A common assumption is to consider that a reidentification occurs when the probability function returned by the reidentification method takes the value 1 at one index, say at i_0, and the value 0 at all the other indices. That is, given the auxiliary information a there is probability 1 that y belongs to the record with index i_0 in X.

We say that the entries $s \in \mathcal{P}(Y)$ are linked to a collection of indices $J \subseteq I$ if the probability returned by the reidentification method takes non-zero values over the indices J and is zero on the complement $I \setminus J$. Typically, a possible non-zero value for the reidentification method over J is then $1/|J|$.

2.2.2 n-anonymity

As in Section 2.2.1 we represent a database as a table and we say that the rows are the records and the columns are the attributes. We suppose that every record contains information about a unique individual. We use the notation $T(A)$ to say that T is a table with the set of attributes A. Let $B \subseteq A$ be a set of attributes of the table. We denote the projection of the table on the attributes B by $T[B]$. We suppose that every record contains information about a unique individual. An identifier I in a database is an attribute such that it uniquely identifies the individuals behind the records. In particular, any entry in $T[I]$ is unique. A quasi-identifier QI in the database is a collection of attributes $\{A_1, \ldots, A_n\}$ that belongs to the public domain (i.e. are known to an adversary), such that they in combination can uniquely, or almost uniquely, identify a record [26]. That is, the structure of the table allows for the possibility that an entry in $T[QI]$ is unique, or that there is only a small number of equal entries. In the former case the entry in $T[QI]$ uniquely identifies the individual behind the record and in the latter, the few other individuals with the same entries in $T[QI]$ may form a collusion and use secret information about themselves in order to make this identification possible.

The former case may be formalized as follows. Consider the table $\tilde{T}$ obtained by permuting randomly the records of T. Let s be an element in $\mathcal{P}(T)$ such that the entries of s all belong to the same record in $\tilde{T}$ (and therefore also in T). Then, if there is a method of reidentification $r : \tilde{T} \times A \to [0,1]^n$ such that $r(s,a)[i] = 1$ for some $a \in A$ and one index i, then s belongs to $T[QI]$.

In the latter case, an s such that $r(s,a)$ is large for a small subset J of indices and 0 for the others (so that s is linked to J) would also belong to $T[QI]$.

Example 2.2.3. *If a table contains information on students in a school class, the attributes birth data and gender could be sufficient to determine to which individual a record of the table corresponds, although it is possible that not all* *records will be uniquely identified in this way. Hence for this table, birth date and gender are an example of a quasi-identifier.*

The following definition of k-anonymity appeared for the first time in [68] (see also the articles by Samarati [67] and Sweeney [78]).

Definition 2.2.4. *A table T, that represents a database and has associated quasi-identifier QI, is k-anonymous if every sequence in $T[QI]$ appears with at least k occurrences in $T[QI]$.*

2.2.3 P2P UPIR: A peer-to-peer user-private information retrieval protocol

User-private information retrieval (UPIR) is defined as the discipline that studies how a user should retrieve an element from a database or a search engine without the system or the server being able to deduce who the retrieving user is [28, 29]. Since UPIR does not hide the content of the query for the database, but instead obstructs the possibilities for the database of profiling users, formally a UPIR protocol does not have to be a Private information retrieval (PIR) protocol (see the introduction). UPIR is also called anonymous keyword search or anonymous database search.

UPIR and mixers both deal with anonymity, but the concepts are different. As was explained in the introduction, mixers provide anonymity on the network layer, but the user can still be profiled through e.g. cookies. UPIR deals with anonymity on the application layer. In this section we will describe the UPIR protocol that will be analyzed in this thesis.

In [28, 29], a UPIR protocol was presented which was based on a peer-to-peer network, P2P UPIR. The idea behind the P2P UPIR protocol is that the clients who want to retrieve information collaborate in posting each others queries. The clients use a P2P network to interchange queries and the answers to these queries. P2P UPIR preserves the privacy of a user's query profile in front of the database and external intruders. In addition the protocol also offers privacy versus peer users. Other users see only a small part of the other user's queries. Peers can be made anonymous to each other also on the network layer by using mixers.

The communication over the P2P network should be encrypted. We will assume that the encryption is done using a symmetric encryption scheme. If the encryption is made with the same key over the entire network, then there is a high risk that the key is compromised. On the other hand, if the encryption uses different keys for every pair of clients, then this risk is low. But if the protocol prescribes one key for every pair of clients and the number of clients is large then the number of needed keys is very large, which is a problem. There are however more sophisticated ways to distribute cryptographic keys than the two trivial examples just described. The articles [28, 29] treat a version of the P2P UPIR protocol which uses combinatorial configurations (defined below) to manage the keys. The idea to use combinatorial configurations for key distributions can also be found in [53, 54]. The main problem when dealing with configurations is that they are very easy to define but not so easy to find.

The idea behind the key distribution used in [28, 29] is to represent the collaborating clients by the points of a combinatorial configuration and to use the lines to represent "communication spaces", that is, a memory sector together with a belonging cryptographic key. A client that is represented by the point p has access to the communication spaces that are represented by the lines through p and he stores the keys corresponding to these communication spaces. When the client wants to submit a query to the server, then he uploads the query to one of the communication spaces to which he has access, after encrypting it with the corresponding cryptographic key. Another client represented by the point q can read p's query on the communication space iff he has access to the corresponding cryptographic key. In other words, the client q can read p's query iff the communication space is represented by a line passing through both p and q.

Next, the client q posts the query to the server. When q receives the answer to the query he uploads it to the same communication space from where he previously read the query, after encrypting it with the corresponding cryptographic key. Subsequently p can read the answer to his query from the communication space, after decrypting it.

Below we present the configuration based P2P UPIR protocol described in [28, 29]. The precondition of the protocol is that the client or user pertains to a community of users that are mapped to the points of a combinatorial configuration and that the client or user wants to post a query to the server. The postcondition of the protocol is that the client or user obtains the answer to his query. We will abuse notation and not distinguish the points and the lines of the configuration from the clients and the communication spaces that they represent.

Protocol 1 (P2P UPIR (I)).

1. *A client or user represented by the point u selects randomly a communication space represented by a line c passing through u.*

2. *u decrypts the content on the memory sector of c using the corresponding cryptographic key of a symmetric cipher. Now the protocol ramifies into five cases depending on the outcome of the decryption.*

   (a) *The outcome is **garbage**. Then u encrypts his query and records it in c;*

   (b) *The outcome is **a query posted by another user**. Then u forwards the query to the server and awaits the answer. When u receives the answer, he encrypts it and records it in c. He then restarts the protocol with the intention to post his query;*

   (c) *The outcome is **a query posted by the user himself**. Then u does not forward the query to the server. Instead u restarts the protocol with the intention to post his query;*

   (d) *The outcome is **an answer to a query posted by another user**. Then u restarts the protocol with the intention to post his query;*

   (e) *The outcome is **an answer to a query posted by the user himself**. Then u reads the query and erases it from the communication space. Subsequently u encrypts his new query and records it in c.*

Remark 2.2.5. *We permit the users to start the protocol with a garbage query. We say that a user who starts the protocol with a query, which is either garbage or not, checks the communication spaces. We say that a user who uses his cryptographic keys to read the content on the communication spaces, without performing any other action, reads the communication spaces.*

The P2P UPIR (I) protocol is called by an initializing protocol which we call P2P UPIR INIT, which is implemented by all the community of users together. This protocol takes as parameters the combinatorial configuration to use for the distribution of communication spaces and the P2P UPIR protocol to use (later we will define P2P UPIR protocols that will differ from the P2P UPIR (I) protocol).

The precondition is here that a community of n users wants to implement a P2P UPIR protocol. The postcondition is that some user has dropped out of the protocol.

Protocol 2 (P2P UPIR INIT)**.**

1. *The points of the combinatorial configuration are mapped to the users of the community.*

2. *Then the users will repeat only one of the following steps until some user drops out of the protocol:*

    (a) *Whenever the user has a query to post, the user executes the P2P UPIR protocol.*

    (b) *After a short fixed time interval* t*, the user executes the P2P UPIR protocol. If the user has a real query to post, he will use this query when he executes the P2P UPIR, otherwise, he will use a garbage query.*

Steps 2.a and 2.b define two different strategies in the execution of the P2P UPIR protocol and it will show later, in Section 3.1, that these strategies have different consequences.

Chapter 3

Choosing configurations for P2P UPIR

3.1 Optimal combinatorial configurations for P2P UPIR and the neighborhood problem

This section is dedicated to a discussion on optimal configurations for P2P UPIR. First we will see two examples of executions of the P2P UPIR protocol that are extreme in the sense that the two types of combinatorial configurations that assign the communication spaces to the users are degenerate.

Suppose that a community of users share a communication space formed by one memory sector and one cryptographic key. The community users use the communication space to write their query requests, to read the query requests of other users and then commit these, and finally to write the answers to the queries. In that way all users collaborate for the good of the group and the server can not know who is asking what, nor elaborate any profiles, at least not more specific profiles than one describing the entire community. This system is really good if we consider the privacy against the server, but it is not so good when it comes to privacy between users. Although it is not known who made a particular query request, all requests from a certain user pass through the shared communication space.

We can think of another system where each user shares a different communication space with every other user. In that way he can spread his query requests between them. The privacy against the server is maintained. Every user reads only a portion of the query requests of a user with whom he collaborates, but on the other hand he can be certain of who of the users requested the query.

The use of combinatorial configurations to represent the distribution of the communication spaces gives us a way to parametrize intermediate solutions to the problem defined above. That is, we have a set of b communication spaces, all of them consisting of a memory sector and a cryptographic key and a set of v users, all of them having access to a subset of r communication spaces so that every communication space is shared by k users and every pair of users share at most one communication space. Then the combinatorial object that exactly represents this situation is a combinatorial (v, b, r, k)-configuration. In fact, the two extreme implementations just explained correspond to a (degenerate) combinatorial $(v, 1, 1, k)$-configuration and to a combinatorial $(v, v(v-1)/2, v-1, 2)$-configuration respectively. Given some criteria for privacy, we can ask for the optimal combinatorial configurations with respect to these criteria.

3.1.1 Optimal configurations for peer to peer private information retrieval in terms of profile diffusion

In the following discussion criteria like storage and time efficiency will be considered, alongside with the arguments on how to optimize the privacy preserving properties of the protocol.

Suppose that a community of users implement the P2P UPIR protocol with a combinatorial $(v, b, r, k)-$configuration as parameter. We next analyze its performance by means of the parameters

- $v :=$ number of users;
- $b :=$ number of communication spaces;
- $r :=$ number of communication spaces assigned to each user;
- $k :=$ number of users who share each communication space.

The following is a list of some aspects to consider in the process of choosing optimal combinatorial configurations for P2P UPIR.

1. We know from Theorem 2.1.55 that in a combinatorial configuration, with parameters (v, b, r, k), we always have

$$vr = bk. \tag{3.1}$$

This implies that the number of required keys and memory sectors is $b = vr/k$. Therefore, for a fixed v, as k/r grows, there is a reduction in the number of required keys and memory sectors (storage efficiency).

2. In addition to storage, another performance metric is how long it takes for a user to get his query submitted and answered. Clearly, the greater the number k with whom the user shares a selected communication space, the shorter the expected waiting time (time efficiency). Also, if the number of communication spaces r to which the users has access is small, then the frequency with which a user return to a particular communication space is higher. Together, this implies that the expected waiting time is shorter when k/r is large. However, if we impose on all users to check their communication spaces with a fixed frequence, as described in Protocol 2 (the P2P UPIR INIT protocol), step 2.2 and Section 3.1.4, then the expected waiting time is fixed, so that the value of k/r is no longer important;

3. The risk that a user can profile and thereby re-identify another user decreases as r increases, since the user then distributes his queries to a wider subset of communication spaces (privacy in front of other users);

4. The query profile of a particular user is diffused among the $r(k-1)$ users with whom the user shares a key and confused among the other queries submitted by those users (privacy in front of server). We deduce that the privacy of the users in front of the server is an increasing function of $r(k-1)$.

From Theorem 2.1.55 we know that

$$r(k-1) \leq v-1$$

and combining this with the fact that the privacy of the users facing the server is an increasing function of $r(k-1)$, we deduce that the parameters for an optimal configuration for the P2P UPIR, considering the privacy against the server, should satisfy the following relation:

$$r(k-1) = v-1. \tag{3.2}$$

In a combinatorial configuration satisfying Equation 3.2 a point p is collinear with all other points. In other words, any two points are connected by exactly one line. Such a combinatorial configuration is a $(v, k, 1)$-BIBD. Indeed, this property is what characterizes the $(v, k, 1)$-BIBD, and we see that the combinatorial configurations that we are looking for are exactly the $(v, k, 1)$-BIBD.

A finite affine plane of order q is a $(q^2, q, 1)$-BIBD. The following example shows the affine plane of order 3 as a combinatorial configuration for the P2P UPIR.

Example 3.1.1. *Let $v = 9$, $b = 12$, $r = 4$ and $k = 3$. Consider the following adjacency list of users and communication spaces:*

$u_1:$	c_1	c_2	c_3	c_4
$u_2:$	c_1	c_5	c_6	c_7
$u_3:$	c_1	c_8	c_9	c_{10}
$u_4:$	c_2	c_5	c_8	c_{11}
$u_5:$	c_2	c_6	c_9	c_{12}
$u_6:$	c_3	c_5	c_{10}	c_{12}
$u_7:$	c_3	c_7	c_9	c_{11}
$u_8:$	c_4	c_6	c_{10}	c_{11}
$u_9:$	c_4	c_7	c_8	c_{12}

$c_1:$	u_1	u_2	u_3
$c_2:$	u_1	u_4	u_5
$c_3:$	u_1	u_6	u_7
$c_4:$	u_1	u_8	u_9
$c_5:$	u_2	u_4	u_6
$c_6:$	u_2	u_5	u_8
$c_7:$	u_2	u_7	u_9
$c_8:$	u_3	u_4	u_9
$c_9:$	u_3	u_5	u_7
$c_{10}:$	u_3	u_6	u_8
$c_{11}:$	u_4	u_7	u_8
$c_{12}:$	u_5	u_6	u_9

The users u and the communication spaces c are then the points and the lines of the finite affine plane of order 3, respectively.

In the following lemma we see what relation r and k should keep for these combinatorial configurations.

Lemma 3.1.2. *Given a combinatorial configuration with $r(k-1) = v-1$, we always have $k \leq r$.*

Proof. Suppose that the point p_1 is incident with the lines $l_1, \ldots, l_r$. We get from the condition $r(k-1) = v-1$ that $p_2, \ldots, p_v$ are incident with one and only one of the lines $l_1, \ldots, l_r$. We also suppose, without loss of generality, that p_2 is incident with the lines l_1 and l_j with $j > r$. Then each of the other $k-1$ users on the line l_j, must be incident with a distinct line in $\{l_2, \ldots, l_r\}$. Therefore $k-1 \leq r-1$ and the result follows. □

By the arguments at the beginning of the section, larger values of k/r give better performance as for storage, and, if considered, shorter expected waiting time per query. On the other hand, larger r's give more privacy against peers. If we do not consider the privacy against other peers, then we are interested in configurations with the largest possible k/r. By Lemma 3.1.2 this means that

$$k = r. \tag{3.3}$$

Now by Equation 3.1 we also have that $v = b$ and therefore we are dealing with symmetric configurations. Define $x := b = v$ and $n := k - 1 = r - 1$. From the condition in Equation 3.2 we deduce that $x = n^2 + n + 1$. This implies also that every pair of users share one and only one communication space and that every pair of communication spaces is assigned simultaneously to one and only one user. By Theorem 2.1.38, this corresponds to a projective plane of order n.

One can also argue that once observed that the condition $r(k-1) = v - 1$ gives a $(v, k, 1)$-BIBD, it is a question of choosing optimal $(v, k, 1)$-BIBD for the P2P UPIR protocol. We have seen in Theorem 2.1.43 that examples of $(v, k, 1)$-BIBD are

- the finite projective planes, which are combinatorial configurations with parameters $(q^2 + q + 1, q^2 + q + 1, q + 1, q + 1)$,
- the finite affine planes, with parameters $(q^2, q^2 + q, q + 1, q)$, and
- the unitals with parameters $(q^3 + 1, q^4 - q^3 + q^2, q^2, q + 1)$.

In any $(v, k, 1)$-BIBD, as in every combinatorial configuration, we have that

$$k(r - 1) \leq b - 1.$$

Resource efficiency would therefore imply

$$k(r - 1) = b - 1$$

and then it is enough to observe that a $(v, k, 1)$-BIBD that satisfies this condition is a finite projective plane. We conclude that the optimal configurations for the peer to peer user private information retrieval, with respect to the privacy of the users in front of the server and also with respect to some aspects of efficiency, are, indeed, the projective planes. As described in Section 2.1.5, it is known that finite projective planes of order n exist whenever n is a power of a prime number, but when n is an integer in general the existence is not guaranteed. Actually there is not a single known example of a projective plane where n is not a power of a prime. In [75] it is specified that the existence of projective planes of arbitrary orders is one of the most difficult questions within finite geometry. These restrictions in the parameters for the existence of finite projective planes, make also other $(v, k, 1)$-BIBD interesting as combinatorial configurations for P2P UPIR.

In this discussion we did not take into account the privacy against other peers. Alternative solutions for avoiding collusions of peers are analyzed in Section 3.3.

3.1.2 A possible linkage between queries and users

The purpose with the P2P UPIR protocol is to protect the privacy of the user when retrieving information from a server. Therefore the natural starting point for the analysis is the privacy of the user in front of the server.

We will first introduce some notation.

Definition 3.1.3. *Let U be a community of users implementing an instance of the P2P UPIR protocol. The real query profile $RP(u)$ of a user $u \subseteq U$ is the temporal sequence of queries which u posts to the communication spaces.*

Definition 3.1.4. *Let U be a community of users implementing an instance of the P2P UPIR protocol. The real query profile $RP(V)$ of a set of users $V \subseteq U$ is the temporal sequence of queries which V posts to the communication spaces.*

Definition 3.1.5. *Let U be a community of users implementing an instance of the P2P UPIR protocol. The apparent query profile $AP(u)$ of a user $u \in U$ is the temporal sequence of queries which the user posts to the server.*

Definition 3.1.6. *Let U be a community of users implementing an instance of the P2P UPIR protocol. The apparent query profile $AP(V)$ of a set of users $V \subseteq U$ is the temporal sequence of queries which V posts to the server.*

The following definition is of a more combinatorial nature.

Definition 3.1.7. *We call the collection of users which are collinear with a user u but different from u the neighbors of u and denote these by $N(u)$.*

In the P2P UPIR (I) protocol the user forwards to the server only queries from collinear users different from himself. This strategy is controlled by steps 2.(b) and 2.(c) in the protocol. We will now see that this is not the perfect strategy to follow. Rather it causes the user to put his privacy at risk.

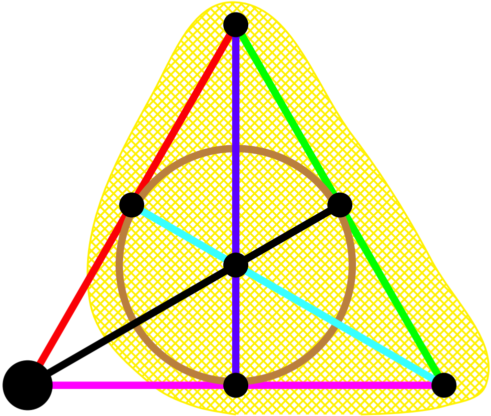

Figure 3.1: The neighborhood of a point in $\mathbb{P}(\mathbb{F}_2)$ painted in yellow

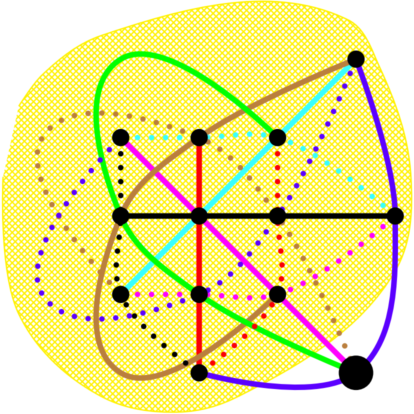

Figure 3.2: The neighborhood of a point in $\mathbb{P}(\mathbb{F}_3)$ painted in yellow

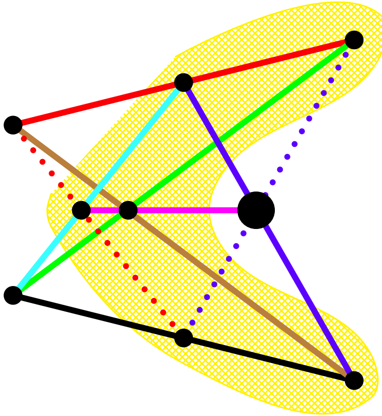

Figure 3.3: The neighborhood of a point in the Pappus configuration painted in yellow

Consider a community of users implementing the P2P UPIR (I) protocol and suppose that the initialization protocol is given a $(v,k,1)$-BIBD as parameter. The users are mapped to the points in the BIBD, and a user u will share communication spaces with the set of users $N(u)$, so that the users who post the queries in $RP(u)$ are the users in $N(u)$. In a $(v,k,1)$-BIBD every pair of points span a line. This implies that, for all points p, the neighborhood $N(p)$ is the whole set of point $\mathcal{P}$, except for the point itself. In particular, given a point p, the neighborhood $N(p) = \mathcal{P} \setminus \{p\}$ is always trivially known,

As already commented, in the P2P UPIR (I) protocol the user u is the only member of $\{u\} \cup N(u)$ who does not post the queries in $RP(u)$ to the server. We deduce that if the user community implements the P2P UPIR (I) with a $(v,k,1)$-BIBD, then the user u is the only user in the community who does not post the queries in $RP(u)$ to the server. Therefore, if a user u posts repeatedly a unique query, then the server can deduce that u is posting the query, *since u is the only user not posting the query.*

What this argumentation tells us is that if the server

- knows that a community of users is implementing the P2P UPIR (I) protocol with a combinatorial configuration of a known type,

- is interested in the real profiles of the users, but

- remains ignorant of any details on the mapping between the users and the points in the configuration,

then the $(v, k, 1)$-BIBD are indeed a bad choice of combinatorial configuration for the P2P UPIR (I). A finite projective plane is an $(n^2 + n + 1, n + 1, 1)$-BIBD. From this perspective, it is therefore reasonable to claim that a finite projective plane is a bad choice of configuration for the P2P UPIR (I). The use of a finite projective plane, or any $(v, k, 1)$-BIBD, implies that any repeated query which is odd enough to identify u can be traced back to him.

For an illustration of the relation between a point and its neighborhood in two finite projective planes and a transversal design, see Figure 3.1, Figure 3.2 and Figure 3.3. The finite projective planes are $(v, k, 1)$-BIBD, while the transversal designs are not. However, the transversal designs have an interesting property which will be used in Section 3.2. Fixed a point p, all points which are not in $N(p)$ have the same neighborhood $N(p)$.

The P2P UPIR (I) protocol is designed to protect, for example, the privacy of the users of web-based search engines. The notations regarding queries and the terms of queries are taken from [70].

Definition 3.1.8. *A term is any unbroken string of alphanumeric characters entered by a user. Terms included words, abbreviations, numbers, and logical operators (AND, OR, NOT). An URL or an e-mail address is considered to be a single term.*

Definition 3.1.9. *A query is a set of one or more search terms. It may include advanced search features, such as logical operators and modifiers.*

Definition 3.1.10. *A repeated query is a query which occurs more than once in the real profile of a user.*

The next definition is vague and ambiguous, but still useful.

Definition 3.1.11. *A repeated variation of a query is a query posted by a user which is a slight modification of a previous query posted by the same user.*

We say that a profile is rare if it contains many unique queries or unique combinations of queries and we say that it has repetition if it contains many repeated queries or repeated variations of queries. The following discussion will try to investigate if reidentification is possible considering the worst case scenario, that is, when the profile of a user is rare and has repetition. We have seen above that a user with a worst case scenario profile is vulnerable for reidentification attacks when the configuration used is a $(v, k, 1)$-BIBD. This contradicts the recommendation in the previous section to use finite projective planes for P2P UPIR (I). However, one should notice that in this analysis we assume that the mapping between the points in the combinatorial configuration and the users in the community is secret. This can of course not be assumed. Indeed one should always assume that everything in the protocol except for the cryptographic keys is public knowledge. Also, it would surely be very inefficient to construct a new configuration (the topology of the P2P network) for every collection of users that wants to implement the protocol. Finally, if it is decided that the configurations to use should be finite projective planes, then it must be taken into account that there are very few such planes for a given number of users, so in this case there is no secret at all or hardly any secret at all.

We will now assume the Kerckhoffs' principle and so we assume that both the topology of the combinatorial configuration and the mapping between the points and the users are public. Then $N(u)$ is known for all u. Two types of combinatorial configurations which are considered in this document for their otherwise good properties as combinatorial configurations for the P2P UPIR are the $(v, k, 1)$-BIBD and the triangle-free combinatorial configurations. We have seen in the previous discussion that the $(v, k, 1)$-BIBD have a property which hardly is desired in a context of privacy and anonymity. The next Theorem 3.1.12 shows that the triangle-free combinatorial configurations also have this property.

Theorem 3.1.12. *Let $C = (\mathcal{P}, \mathcal{L}, I)$ be a combinatorial (r, k)-configuration without triangles with $k > 2$ or a $(v, k, 1)$-BIBD. Then there is a bijection between the sets $\mathcal{P}$ and $\{N(p) : p \in \mathcal{P}\}$.*

Proof. A $(v, k, 1)$-BIBD is a $2-(v, k, 1)$ design, hence every pair of points is collinear. Therefore, for any point p, the neigborhood $N(p)$ is all the point set $\mathcal{P}$ except for $\{p\}$. This defines a function

$$\begin{array}{rcl} \mathcal{P} & \to & \{N(p) : p \in \mathcal{P}\} \\ p & \mapsto & \mathcal{P} \setminus \{p\} = N(p), \end{array}$$

which is obviously injective and exhaustive.

Now suppose that C is triangle-free. Fix a point $p_0 \in \mathcal{P}$ and let $p_1, p_2 \in N(p_0)$ be two points which are collinear with p_0. Let $p_3 \in \mathcal{P}$ be a point such that $N(p_0) = N(p_3)$. This implies that p_3 is collinear with p_1 and p_2, but not with p_0. so that there is no line through all the four points p_0, p_1, p_2 and p_3. Indeed it implies that p_1 and p_2 can not be collinear, because if they were, then at least one of the triples p_0, p_1, p_2 or p_1, p_2, p_3 would form a triangle. In other words, no pair of points in $N(p_0) = N(p_3)$ is collinear. Therefore the number of points on every line in C is $k = 2$, because if $k > 2$, then there would be at least a pair of collinear points $p, q \in N(p_0) = N(p_3)$. We deduce that, whenever $k > 2$, given a point $p_0 \in \mathcal{P}$ there is no point $p_3 \in \mathcal{P}$ distinct from p_0 such that $N(p_0) = N(p_3)$. Hence, for $k > 2$ the function

$$\begin{array}{rcl} \mathcal{P} & \to & \{N(p) : p \in \mathcal{P}\} \\ p & \mapsto & \mathcal{P} \setminus \{p\} = N(p) \end{array}$$

is injective. The function is obviously exhaustive and therefore a bijection. □

Theorem 3.1.12 implies that when both the combinatorial configuration and the mapping between the points and the users are known to a curious server, then both the $(v, k, 1)$-BIBD and the triangle-free combinatorial configuration have a property which implies the possibility to link parts of a diffused real profile $RP(u)$ to its owner u. In general, whenever both the combinatorial configuration and the mapping between the points and the users are known to the curious server, the P2P UPIR (I) permits an adversary to reidentify a small list of users as the possible origin of a collection of queries in the apparent profiles $AP(U)$, so that the real owner of the queries is on this list, regardless of the combinatorial configuration used for the protocol. In this context, the $(v, k, 1)$-BIBD and the triangle-free combinatorial configurations are examples of combinatorial configurations for which the length of this list of users is one, so that a reidentification is produced. In Section 3.2, we will discuss how to find combinatorial (r, k)-configurations which maximize the length of the list of possible neighbors to a given collection of users X of cardinality $|X| = r(k - 1)$.

Among the combinatorial configurations for which there is a bijective mapping between the points p and their neighborhoods $N(p)$, the finite projective planes, and in general the $(v, k, 1)$-BIBD, are however still optimal combinatorial configurations for P2P UPIR (I). The reason for this is that in the $(v, k, 1)$-BIBD, the real profiles of the users are maximally dispersed by P2P UPIR (I), that is, distributed into the apparent profiles of all the users in the community, except for the apparent profile of the user himself. Therefore, given a community of users of cardinality v, among all combinatorial configurations with v points, the number of repetitions of a query the user has to do until there is a risk for reidentification is largest exactly for the $(v, k, 1)$-BIBD. However, as we just saw, exactly because the only apparent profile of the users in which queries from the real profile of the user is missing is the apparent profile of the user himself, the dispersion of the real profile of the user performed by the P2P UPIR (I) can never be complete. Concluding, a worst case scenario real profile can be mapped to the user behind this profile, also when the configuration that is used is of a type which we previously argumented to be optimal!

One can argue that in the description of the P2P UPIR (I) in [28, 29] the protocol lets the user post his own queries if the waiting time for another user to post it is too long. Therefore it is of course possible that the user by accident is lucky enough to post the same proportion of his queries as do his neighbors, meaning that in this case the attack described above would not work. One can however not rely on such arbitrary circumstances for the protection of the privacy of the user.

3.1.3 Real examples of repeated queries

One can also ask if it is a common behavior of real users to repeatedly post a query. An interesting question is also how a typical real profile of a user looks like. In 2006 AOL released search logs that contains 20 million web queries from 658,000 AOL users posted in a period of 3 months. The released data was anonymized by replacing the identity of the users by a random index, but this quickly showed insufficient as several sequences of queries were mapped to real persons. AOL withdrew the query logs from internet, but the files were of course already downloaded by many people. The AOL search data release caused a privacy scandal which is the reason why the query logs published by AOL are practically the only material available for non-corporative research on the subject.

A quick look at the AOL query logs [5] makes it reasonable to assume that posting the same query (or a slight modification of a query) several times is a common behavior of users of web-based search engines. There seems to be at least three scenarios which can result in this behavior.

The first scenario is explained by to the way people normally use their browsers. In the common internet browsers, when the user queries a search engine the result will be presented to him as a list of links in the browser window. The user′s next step is to choose a promising link from the list and to follow it. Later he may want to return to the list of search results. Although he may still have the page with the list of search results open in the browser, this page will not be on top of the windows the user has opened. The user, being lazy, does not change to the previous window with the search results, nor does he press the return button of the browser in order to return to the previous window, but simply posts the query again. This behavior leads to many repeated queries without much or any variation.

In the released AOL query log files there are many query sequences with repeated queries which can be explained by this scenario. For example, user 1783081 has one query for ′digital camoflasges′ at 2006-03-15 12:49:29 and then 9 identical queries for ′digital camouflages′, the last one posted at 2006-03-15 13:00:09. AOL registered 7 different clicked url as a result from this sequence of queries, giving an example of a user which probably has followed a behavior similar to the one just described. Between 2006-04-18 15:14:03 and 2006-04-18 15:14:03 user 672368 posts 7 queries on ′abortion clinic charlotte′ and later between 2006-04-18 21:45:39 and 2006-04-18 21:45:49 5 queries on ′abortion clinic charlotte nc′ We observe that some users have sequences with up to 25 equal queries in very short time.

The second scenario is when the user posts a set of very similar queries in order to adjust and limit the search result so that it resembles more what the user aimed at. Misspellings are a similar scenario, but misspellings do not tend to result in multiple repetitions of a query. For example, between 2006-03-19 19:24:09 and 2006-03-19 19:30:02 the user 1783081 from the previous examples posted 3 queries concerning ′the long ranger′, 1 query on ′the legend of the long ranger′, 5 queries on ′the legend of the lone ranger′, 6 queries on ′the lone ranger theme song′ 3 queries on ′lone ranger theme′. User 1783081 generally shows a general interest for fantasy, movies and as more particular interests figures lolita porn, occult rituals, incest and young teen girls. User 672368 has a sequence of queries starting at 2006-04-18 06:50:07 with a query ′effects oon on fibriods′, then three queries on ′effects of abortion on fibroids′ and four queries on ′abortion fibroids′, with the last query at 2006-04-18 06:59:32. After this the user continues to post queries for example on the subject ′abortion′. At 2006-04-20 17:55:18 the user continues posting 11 queries on ′abortion fibroids′. Totally on this subject the user posts 19 queries on the subject ′abortion fibroids′. The user started the query sequence with 'curb morning sickness', 'get fit while pregnant', continued with 'you're pregnant he doesn't want the baby' and many queries on abortion, abortion clinics and later misscarriage. It seems likely that this user would have preferred a better privacy than AOL could offer.

The third scenario occurs when the user posts queries on something that appears in his daily life. For example, it seems to be rather common that users post a query to a search engine in order to search for the webpage of the school of their kids, or their own workplace, instead of browsing to the webpage directly. The user's workplace and the school of his kids are highly interesting information for reidentification. Considering that this kind of queries can be repeated several times a month, the risk of reidentification can not be neglected.

The AOL query logs are not well suited for finding repeated examples of the third scenario, since they only cover a time period of 3 months. However, AOL themselves and other search engine providers have of course access to query logs from much longer time periods.

A study of the query logs from the Excite search engine was published in 2000, which included 211063 users who had posted a total of 1025910 queries on 16 September 1997 [70]. Of these queries, 395461 were repeated queries, in the sense that the user had posted another query with exactly the same terms, and 531416 were unique queries which differed from the rest of the user's queries with at least one term.

In 2005 Yahoo released a study of query logs which was focused on the repetition of queries caused by the desire to return to a webpage visited before, a behaviour called re-finding [80]. The study was based on the queries posted by 114 users during one year, posting a total of 13060 queries and with a total of 21942 subsequent clicks. The researchers were not interested in short-term query repetitions, and considered all instances of the same query string that occurred within thirty minutes to be a single query. It was observed in the study that re-finding behavior is common, and it was shown that repeat clicks can often be predicted based on a users previous queries and clicks.

We conclude that repeated queries and repeated variations of queries is a frequent and common phenonemon.

It should be observed that although when the P2P UPIR (I) protocol fails to provide complete protection of the privacy of the user in front of the server in the case of many repeated queries, single queries can still not be traced back to the emitter.

The level of privacy provided by the P2P UPIR (I) protocol can be specified more exactly. The user diffuses his real profile into the apparent profiles of his $r(k-1)$ neighbors. However since he chooses communication space randomly and has no control over who will forward the query of the other $k-1$ users sharing the same communication space, it is not possible to say exactly how many repetitions of a query a user must post until his privacy is broken. Also, in general it is possible that a user may be the unique common neighbor also to sets of users of cardinality smaller than $r(k-1)$. This also affects the efficiency of the attack.

Finally it should be noticed that we below will provide a fix of the problem encountered in the P2P UPIR (I) protocol, as will be seen in the following Section 3.1.4.

3.1.4 Modifications to prevent linkages

The previous section was dedicated to a privacy analysis of the P2P UPIR (I) protocol, which is the version of the P2P UPIR protocol that appears in [28, 29]. In the P2P UPIR (I) protocol the user forwards to the server only queries from collinear users different from himself.

In this section we will discuss two variations of the P2P UPIR (I) protocol. The discussion will provide a modification of the protocol which solves the privacy flaw discussed in the previous section.

We define the P2P UPIR (II) protocol as obtained from the P2P UPIR (I) protocol by replacing step 2.(b) and 2.(c) by the single step:

2.(b) The outcome is ***a query posted either by the user himself or by another user****. Then u forwards the query to the server and awaits the answer. When u receives the answer, he encrypts it and records it in c. He then restarts the protocol with the intention to post his query;*

Hence, the only difference from the P2P UPIR (I) protocol is that in the latter the user does not forward his own queries to the server, but in the P2P UPIR (II) he does.

The following lemma implies that the users in a community that follow the P2P UPIR (II) protocol will forward more of their own queries to the server than queries of the other users. As a consequence of this, the users' real profiles can be inferred from the apparent profiles of the users.

Lemma 3.1.13. *Consider a community of users $U = \{u_i\}_I$ implementing the P2P UPIR (II) protocol. Suppose that in a fixed time interval t a user u_i posts*

q_i queries. Denote by p_{ij} the proportion of queries from the real profile of u_i on the communication space c_j. Let $\{c_{ij_n}\}$ be the set of communication spaces incident with u_i, indexed by $n \in [1, \ldots, r]$. Then the proportion of queries from the real profile of u_i in the apparent profile of u_i is

$$\sum_{n=0}^{r} \frac{p_{ij_n}}{r} q_i.$$

The proportion of queries from the real profile of u_i in the apparent profile of $u_m \neq u_i$ is

$$\begin{cases} \frac{p_{mj}}{r} q_i & \text{if } u_m \text{ is collinear with } u_i \text{ by the line/communication space } c_j; \\ 0 & \text{otherwise.} \end{cases}$$

Proof. During time t a user u_i posts a set of q_i queries to the r communication spaces to which he is connected. This set of queries is the real profile of u_i. Any communication space incident with u_i receives q_i/r queries from u_i. Let p_{ij} denote the proportion of queries from u_i on the communication space c_j, so that for all j

$$\sum_i p_{ij} = 1.$$

The P2P UPIR (II) protocol is originally designed so that a user reads the content on the communication space only when he wants to post a query (step 2.1 in Protocol 2), so that the proportion of his queries on the communication space equals the proportion of queries on the communication space that he reads. The amount of queries from the real profile of u_i which are sent to c_j and return to u_i to be forwarded by u_i to the server is therefore

$$p_{ij} \frac{q_i}{r} = \frac{p_{ij}}{r} q_i.$$

Adding over the r communication spaces incident with u_i the number of queries in the apparent profile of u_i coming from the real profile of u_i is

$$\sum_{n=0}^{r} \frac{p_{ij_n}}{r} q_i = \left(\sum_{n=0}^{r} \frac{p_{ij_n}}{r} \right) q_i.$$

Since p_{ij_n} depends on n, this expression cannot be simplified.

Consider another user $u_m \neq u_i$. Suppose that u_m and u_i share the line/communication space c_j. Then there are

$$\frac{p_{mj}}{r} q_i$$

queries from the real profile of u_i in the apparent profile of u_m. This is so, because the user u_i and the user u_m share only one communication space: c_j.

Finally, if the users u_m and u_i are not collinear, then the apparent profile of u_m will not contain any query from the real profile of u_i. □

Under particular circumstances Lemma 3.1.13 has the simpler expression given in Corollary 3.1.14.

Corollary 3.1.14. *Under the same assumptions as in Lemma 3.1.13, suppose that all users post queries with the same frequency, so that $q_i = q_j$ for all i, j. Then the proportion of queries from the real profile of u_i in the apparent profile of u_i is*

$$\frac{1}{k}.$$

The proportion of queries from the real profile of u_i in the apparent profile of another user $u_m \neq u_i$ is

$$\begin{cases} \frac{1}{rk} & \text{if } u_m \text{ is collinear with } u_i; \\ 0 & \text{otherwise.} \end{cases}$$

Proof. If we suppose that all users post queries with the same frequency q of queries, then $p_{ij} = \frac{1}{k}$, so that the amount of queries from the real profile of u_i which is sent to c_j and returns to u_i to be forwarded by u_i to the server is $\frac{1}{rk} q_i$. In this case, adding over the r communication spaces incident with u_i, the number of queries in the apparent profile of u_i coming from the real profile of u_i is

$$\sum_{n=0}^{r} \frac{p_{ij_n}}{r} q_i = \sum_{n=0}^{r} \frac{1}{rk} q = \frac{1}{k} q,$$

so the proportion of the real profile of u_i in the apparent profile of u_i is

$$\frac{1}{k}.$$

However, we see that the proportion of the real profile of u_i in the apparent profile of another user $u_m \neq u_i$, that is collinear with u_i, is only

$$\frac{1}{rk}.$$

A user u_m that is not collinear with u_i will not have any query from the real profile of u_i in his apparent profile. □

Lemma 3.1.13 has the following interpretation.

Corollary 3.1.15. *The users in a community who follow the P2P UPIR (II) protocol will forward to the server more of the queries posted by themselves than they will forward queries posted by other users.*

This corollary implies that the server can infer the real profile of a user from his apparent profile. The P2P UPIR (II) provides a partial protection of the privacy of the user in front of the server, valid for sparse use. But if we let the protocol run for a while in order to let the user post enough queries, then the users real profile will get inferable from his apparent profile.

We have seen two different strategies for how the user should treat his own queries when implementing P2P UPIR. In the first the user does not forward his own queries to the server and in the second he does. Both provide insufficient privacy protection. Now we will look at a third variation of the P2P UPIR protocol where the user adjusts the number of his own queries he should forward to the server so that his real profile results uniformly distributed over the apparent profiles of his neighbors and himself.

The protocol which we call P2P UPIR (III) differs from the P2P UPIR (I) protocol only in the steps the user follows when the decrypted content of the communication space is a query originally posted by himself which is waiting for a user to post it to the server. The P2P UPIR (III) protocol is obtained from the P2P UPIR (I) protocol by replacing step 2.(c) by:

*2.(c) If the outcome is **a query posted by the user himself**, then u forwards the query to the server with a probability to decide. If u forwards the query to the server, then u also awaits the answer. When u receives the answer, he encrypts it and records it in c. In any case then u restarts the protocol with the intention to post his query;*

The idea behind the modification of the protocol is to adjust the number of his own queries the user forwards to the server in order to obtain a smooth diffusion of his real profile over the apparent profiles of the collection of his neighbors and his own apparent profile.

We have seen before that a $(v, k, 1)$-BIBD is an optimal solution to the problem of preserving the privacy of the user in front of the server, in the sense that it maximizes the number of apparent profiles into which the real profile of a user is diffused, under the restriction to keep the size of the user community fixed. More important, it is the only type of combinatorial configuration where $N(u)$, the set of users collinear with the user u and different from u, are all the users in the configuration different from u. As already commented, the user who adopts the strategy to not forward any of his own queries (the P2P UPIR (I) protocol) as well as the user who adopts the opposite strategy to forward his own queries (the P2P UPIR (II) protocol) are both hazarding the privacy of their real profiles in front of the server, even when the used configuration is a finite projective plane. The P2P UPIR (III) is an intention to avoid these flaws in privacy by adjusting the number of own queries a user should forward to the server. The idea is to adjust so that a user forwards the correct proportion of his own queries in order for the proportion of his real profile to be constant, or at least asymptotically constant, over the apparent profiles of $\{u\} \cup N(u)$. Adjusting in this way, a $(v, k, 1)$-BIBD indeed does provide privacy for the user, since u's queries are uniformly diffused into the apparent profiles of the users $\{u\} \cup N(u)$, which in a $(v, k, 1)$-BIBD is the whole set of users.

Such an adjustment is possible if the frequencies with which the users post queries is the same for all users, as will be required for Proposition 3.1.16. What is perhaps surprising is that the adjustment is still possible when the frequency with which they post queries is not the same for all users, under the assumption that the users check the communication spaces with equal frequency. Following Remark 2.2.5, a user who checks a communication space is a user who starts the protocol with a query which may be garbage or not.

Proposition 3.1.16. *Consider a community of users implementing a P2P UPIR protocol with a combinatorial (v, b, r, k)-configuration and impose on the users to check their communication spaces with a fixed frequency higher or equal to the frequency with which they post queries. Then the user u's real profile is optimally diffused into the apparent profiles of $\{u\} \cup N(u)$ if u forwards* *a proportion of*

$$\frac{1}{r(k-1)+1}$$

of his own queries to the server.

Proof. The number of users that are collinear with u is $N(u) = r(k-1)$. The set of users who forward queries from u's real profile is $\{u\} \cup N(u)$. A homogeneous diffusion of u's real profile into the apparent profiles of $\{u\} \cup N(u)$ means that every one of the users should forward a proportion of

$$\frac{1}{r(k-1)+1}$$

of u's queries. Indeed this clearly implies that u should forward a proportion of

$$\frac{1}{r(k-1)+1}$$

of his own queries. What is perhaps not as obvious is that in order to get a homogeneous diffusion of u's real profile into the apparent profiles of $\{u\} \cup N(u)$ it is sufficient that all the users adopt this strategy.

Denote the collection of queries posted by u during the time interval t by q. Let c be one of the r communication spaces used by u. Then u sends $\frac{q}{r}$ queries to c during t. But u forwards

$$\frac{q}{r(k-1)+1}$$

of his own queries to the server. A proportion of $\frac{1}{r}$ of these comes from c and the rest from the other communication spaces used by u, so there are

$$\frac{q}{r} - \frac{1}{r} \cdot \frac{q}{r(k-1)+1}$$

of u's queries on c which are left to be forwarded to the server by the other $k-1$ users using c. Therefore each of these users forward

$$\begin{aligned}
\frac{\frac{q}{r}-\frac{q}{r(r(k-1)+1)}}{k-1} &= q\left(\frac{1-\frac{1}{r(k-1)+1}}{r(k-1)}\right) \\
&= q\left(\frac{\frac{r(k-1)}{r(k-1)+1}}{r(k-1)}\right) \\
&= q\left(\frac{1}{r(k-1)+1}\right)
\end{aligned}$$

of u's queries to the server during t, just as u does.

The frequency with which u posts queries is $\frac{q}{t}$, but it is interesting to observe that the frequency with which the other users post queries only matters for the proof because it determines the frequency with which the users check the communication spaces. If the frequency of checking the communication spaces differs among the users, then we can no longer ensure that all users forward a proportion of $\frac{1}{k-1}$ of the queries on the communication spaces to which they are connected. □

Proposition 3.1.16 therefore suggests a change in the protocol so that the users check their communication spaces with a fixed frequency. One should probably choose this frequency to be higher than or equal to the highest frequency with which any user posts queries. In this way, when a user has a query to post, he can post it to the first communication space that he checks. When the user has no query, then he checks the communication space anyway.

3.1.5 Optimal configurations for peer to peer private information retrieval in terms of non-linkability

Consider a community of users implementing an instance of the P2P UPIR (III) protocol in which the proportion of selfsubmission of queries is set to

$$\frac{1}{r(k-1)+1}$$

as recommended by Proposition 3.1.16 and suppose that the users check their communication spaces with a fixed frequency that is higher than or equal to the highest frequency with which any user posts queries. Then Proposition 3.1.16 says that the real profile of a user u is optimally diffused into the apparent profiles of $\{u\} \cup N(u)$. In the following we will always assume that the P2P UPIR (III) is implemented under these circumstances.

We observe that examples of combinatorial configurations in which the points $\{p\} \cup N(p)$ is the whole set of points are the finite projective planes, but also the finite affine planes satisfy this criteria. Indeed, as we have already seen, this property is characteristic for the $(v, k, 1)$-BIBD. This observation has two consequences.

1. The P2P UPIR (III) protocol diffuses the real profile of the user u into the apparent profiles of $\{u\} \cup N(u)$. The larger $\{u\} \cup N(u)$, the larger is the protection of the real profile of u in front of the server.

Given a set of v users such that there exists a $(v, k, 1)$-BIBD, the BIBD is a combinatorial configuration which for all users u the set $\{u\} \cup N(u)$ is the whole set of points, that is, as large as possible.

2. In many combinatorial configurations there is a bijection between the sets $\mathcal{P}$ and $\{\{p\} \cup N(p) : p \in \mathcal{P}\}$ defined by $p \mapsto \{p\} \cup N(p)$. In a $(v, k, 1)$-BIBD there is no such bijection, because $\{p\} \cup N(p)$ is the whole pointset $\mathcal{P}$ for all points $p \in \mathcal{P}$, hence equal for all points $p \in \mathcal{P}$. This property will be further studied in Section 3.2.3.

These two facts together provide an affirmation of the optimality of the finite projective planes as combinatorial configurations for P2P UPIR (III), whenever the number of users v is such that there exists a finite projective plane with v points.

There are two important differences between the arguments used in Section 3.1.1 and the arguments used in this section.

1. In this section we are considering the P2P UPIR (III) protocol, instead of the P2P UPIR (I) protocol;

2. In Section 3.1.1, arguments of efficiency were used, while the analysis in this section focused only on the privacy in front of the server.

We observe that the focus on the privacy in front of the server is logical, since the users privacy in front of the server is exactly the purpose of the protocol.

3.2 Transversal designs and n-anonymous P2P UPIR

One of the more important notions in the theory for the anonymization of databases is n-anonymity. For a short introduction to methods for the anonymization of databases and relevant formal definitions, see the preliminary Section 2.2. The aim of the following discussion is to see if the notion of n-anonymity can be useful for an analysis of the disclosure control provided by the P2P UPIR protocol. For this the rather ad-hoc context from Section 3.1 will be replaced by the more standard context of anonymization of databases.

3.2.1 Privacy notions for P2P UPIR: n-anonymity and n-confusion

Consider a community of users U who are executing an instance of the P2P UPIR protocol in order to protect their query profiles from the server, during a time t. A query profile of a user who is not implementing the P2P UPIR protocol, is a temporal sequence of queries posted by the user to the server. As we saw in Definition 3.1.9, a query is a set of one or more search terms. For users who are implementing the P2P UPIR protocol, we will follow the notation introduced in Section 3.1.2, with the addition of the temporal restriction t. Consequently, we will say that the temporal sequence of queries posted by the user u to the communication spaces during t is the real profile $RP_t(u)$ of u and that the temporal sequence of queries posted by the user u to the server during t is the apparent profile $AP_t(u)$ of u. Also, $RP_t(V)$ and $AP_t(V)$ will denote the real and the apparent profiles during t of a collection $V \subseteq U$ of users, respectively. In all this section, the word adversary will always refer to a curious server.

The P2P UPIR as a method for anonymization of databases

Because of the time constraint, the collection of real profiles of the users $RP_t(U)$ is a finite set of finite sequences of queries. The set is indexed by the users, and the sequences are ordered temporally. The set $RP_t(U)$ therefore allows an interpretation as a database, that is, a table of rows (records) and columns (attributes), in which every row is occupied by the identifier or index of a user u and the real query profile $RP_t(u)$. The attributes in this table can be defined in different ways. For example, the table may be regarded to have

1. two attributes, one for the user id and the other for the real query profile sequence, or

2. one attribute for the user id and one attribute for every query in the real query profile in temporal order: query 1, query 2, etc.

In the following discussion, the second viewpoint will be more common than the first. The collection of the apparent profiles of the community of users $AP_t(U)$ allows an analogous interpretation as a database.

In this context, the P2P UPIR protocol may be considered to be a method for anonymization of databases, see Definition 2.2.1. Then the database to protect is $RP_t(U)$, the real profiles of the community of users, collected during t. The result from applying the method for anonymization of databases to the database $RP_t(U)$ is the database $AP_t(U)$, the collection of the apparent profiles of the users during t. In this context, the P2P UPIR protocol is a transformation of the database which we will denote by

$$\begin{array}{rccc} \rho: & D & \to & D \\ & RP(U) & \mapsto & AP(U), \end{array}$$

where D is the space of all possible query databases.

Observe that there are three major differences between the P2P UPIR protocol and most methods for anonymization of databases.

- The responsible for the execution of a method for anonymization of databases is usually the data owner, that is, the entity that has collected the data. In an analysis of privacy protection provided by the P2P UPIR to the users in front of the server, the data owner is the curious server. However it is the users who are responsible for the execution of the protocol;

- The P2P UPIR protocol is executed in real-time as the users post queries to the server, that is, as the information is introduced into the database;

- A method of anonymization of databases is normally designed to preserve the utility of the anonymized database. In this case, a good method of anonymization of databases is a method which provides an anonymized database with a low risk of reidentification and a low information loss. However, for the aim of the P2P UPIR there is no need to control the utility of the transformed database. *Indeed, we assume that the users of the server have no interest in providing a useful statistical database.*

Difficulties in determining the quasi-identifiers

Methods of anonymization of databases usually assume that the entity who executes the method, normally the data owner, has complete knowledge of the content of the database. This is for example true for methods which transform a database into an n-anonymous database. If not all records or attributes of the original database are known when the method is executed, then there is always a risk that the transformed database will not be n-anonymous.

In the case of the P2P UPIR, the entities who execute the protocol are the users, and every user executes the protocol independently of the others. Keeping track of all the queries posted by all the other users in the community is costly and difficult. Also, a user can not see queries posted by users to whom he is not a neighbor. The fact that the P2P UPIR is executed in real-time implies that the users intend to protect the database $RP_{t+1}(U)$, only with (a partial) knowledge of $RP_t(U)$. Because of all these difficulties, we may assume that the knowledge the individual users have of the content of the database $RP_{t+1}(U)$ at the time t is restricted. This also restricts the knowledge the users may have of the quasi-identifiers of $RP_{t+1}(U)$ and implies that an application of the notion of n-anonymity to the P2P UPIR protocol is hard to justify, at least in this context.

Now assume that the users have complete knowledge of the content of $RP_{t+1}(U)$ before executing the protocol at time $t+1$. It is still hard for the users to foresee the auxiliary information available to the curious server, and hence also to predict how the database $AP_t(U)$ may be reorganized in order to create quasi-identifiers and perhaps reidentify one of the users. Indeed the curious server is likely to define other attributes (like for example age, dog owner, geographic names), than the attributes that we work with (query 1, query 2, etc.). This last comment is indeed valid for every method for anonymization of databases; without prior assumptions on the auxiliary information available to the adversary, it is very hard to predict how the adversary will behave. Concluding, not much can be said in advance about the nature of the quasi-identifiers of $RP_t(U)$.

The determination of the correct quasi-identifiers is crucial for a correct application of n-anonymity. The observations of the difficulties for the determination of the quasi-identifiers therefore suggests that, in the discussed context, n-anonymity is perhaps not the most adequate notion for the P2P UPIR.

Sensitive sequences and n-confusion

Regarding the P2P UPIR we observe the following. As before, the word adversary will always refer to a curious server.

- An adversary (a curious server) is given only the protected database $AP_t(U)$ and so his knowledge about $RP_t(U)$ is limited. This is indeed the purpose of the anonymization of the database. However, as stated in previous arguments, it is very hard to foresee the knowledge an adversary may have on $RP_t(U)$, considering that he may have access to important auxiliary information. Because of this, it is difficult for the user, who is the one executing the protocol, to make prior judgements about which of the queries in his real profile may cause a reidentification of him in the protected database $AP_t(U)$. The most prudent approach is therefore to assume that any collection of his queries carries a high risk of reidentification.

- Not everyone has the same requirements of privacy. What one user considers highly sensitive information, the other user may not consider sensitive at all. It is possible to design a protocol in which every user decides what to protect, but the protocol becomes more complicated and experience shows that even when similar advanced individualized features are available in software, most people do not use them. It is therefore better to assume that all subsets of queries in the real profile of the user require the same protection.

- The entries of the database $RP_t(U)$ are web search queries. According to Definition 3.1.9 a query is a set of one or more search terms. In particular, a single query may cause a reidentification and contain sensitive information, simultaneously. For example, consider the query

  $$\{\text{Anna Svensson AND stripshow}\}.$$

  More generally, a collection of queries may cause a reidentification and contain sensitive information simultaneously. Usually when a database is anonymized using the notion of n-anonymity, it is assumed that the quasi-identifier and the sensitive information are separable. As just observed, in a database of web search queries, this is not always possible.

In the following we will therefore not distinguish between quasi-identifiers and sequences of sensitive information and we will let both go under the name *sensitive sequence*. This approach is perhaps not standard, but suites well the analysis of the P2P UPIR protocol. Indeed, allowing for sensitive information to also be quasi-identifying, or a quasi-identifier to contain sensitive information, is stronger than assuming that the two types of information are independent.

Consequently, one approach for the formalization of the analysis of the P2P UPIR protocol is to assume that any sequence of queries which form part of a record of $RP_t(U)$ is a sensitive sequence to be protected. The result after applying the protocol is the database $AP_t(U)$. Consider a subset $\{s_1, \ldots, s_m\}$ of queries s_i with $m \leq t$ which forms part of a record $RP_t(u)$ of $RP_t(U)$, ordered temporally in the sequence $s = (s_1, \ldots, s_m)$. The P2P UPIR transformation ρ distributes the queries s_i over the records in $AP_t(N(u))$, that is, the records in $AP_t(U)$ which correspond to the neighbors $N(u)$ of u in the combinatorial configuration used by the protocol. For a combinatorial configuration with parameters (r, k), the number of neighbors of u is $r(k-1)$.

Suppose that s is a sensitive sequence of minimal length, so that any proper subsequence of s is not a sensitive sequence. In particular, this assumption implies that s does not contain a repetition of a sensitive subsequence.

Suppose that s is a sensitive sequence of length larger than one. Although the queries s_i are distributed uniformly over $AP_t(N(u))$, there is of course a non-zero probability that a large proportion or even all of the queries s_i will end up in the apparent profile of the user u. However, if the server is aware of the fact that the users are implementing the P2P UPIR protocol, it will not be able to tell if s pertains to $RP_t(u)$ or if s pertains to the real profile of some other user. Indeed, the curious server can not even tell if s was a sensitive sequence in $RP_t(U)$. This is due to the fact that the s_i's, which isolated do not contain any sensitive information, were distributed uniformly into $AP_t(N(u))$. Any quasi-identifying property and sensitive information in s can therefore not be traced back further than to $N(u)$.

A curious server could have interest also in analyzing the real profile of, say, a group of users who are friends or who work in the same company. It is reasonable to assume that a group of users with the same affiliation have somehow similar real query profiles. Assume that the curious server analyzes $AP_t(U)$ only with respect to the content of individual records, and that it has no interest in analyzing the real profile $RP_t(V)$ of a set of several users $V \subset U$. Also assume that the adversary does not have access to auxiliary information (see Definition 2.2.2 of reidentification) that has potential enough to link all the queries s_i to the user u. Then, since s is minimal, we may regard the sensitive information in s as destroyed. Under these circumstances we may reduce the risk analysis to the case in which s has length one.

Now assume that the length of s is one, so that the sensitive sequence is a single query. Then the information in s can not be hidden by the protocol, since the protocol does not split single queries. In this case the situation is different depending on the content of the sensitive sequence s.

- If s only contains sensitive information, then this sensitive information will be available to the server, but the server will only be able to say that the sensitive information belongs to someone in $N(u')$, where u' is the user who emits the query to the server;
- If s only contains identifying information, then the server will be able to use the information to deduce that u belongs to $N(u')$;
- If s contains both identifying information and sensitive information, as in the example with Anna Svensson and the stripshow, then the server will indeed be able to deduce that the sensitive information belongs to u.

Using this approach, what is obtained is not n-anonymity in the sense of Definition 2.2.4. Rather, in all cases but the last, what is obtained is n-confusion, with $n = r(k-1)$, that is, the P2P UPIR protocol introduces a confusion of magnitude $n = r(k-1)$ on who is the real owner of the sensitive subsequences of length one. In order to stress the distinction between the two concepts, we provide the following definition of n-confusion.

Definition 3.2.1. *Let U be a community of users implementing a P2P UPIR protocol with a combinatorial configuration C as parameter. Let s be a sensitive sequence in $AP(U)$. We say that the P2P UPIR protocol provides n-confusion, if a curious server that is given $AP(U)$, C and the mapping between the users U and the points in C, can only determine the real owner of s, that is, the user u such that $s \in RP(u)$ with a precision of n, in the sense that there are at least n users in U who could be the real owner of s.*

Observe that n-anonymity may imply n-confusion, while the opposite in general is not true.

Repetition and n-anonymous P2P UPIR

Consider a user who is still convinced that n-anonymity is the key to protect his real query profile from the curious server. An intent to make the protected database $AP_t(U)$ n-anonymous in the sense of Definition 2.2.4 would imply imposing n occurrences of the sensitive sequence s in n different records of $AP_t(U)$. Because of the structure of the protocol, the user can not assume that another user will post s. The responsability of repeating s at least n times is therefore laid on the individual user u with $s \in RP_t(u)$. All these copies of s will be located in the records of $AP_t(N(u))$, that is, the records of $AP_t(U)$ indexed by $N(u)$, with $s \in RP(u)$.

As seen in Section 3.1.2, in a quite different context it can be observed that repetition of sensible sequences of queries to web-based search engines is a common phenomenon, which appears as a natural behaviour of the users.

Whatever the reason may be for the repetition of s, the result of the P2P UPIR (I) protocol will be that several copies of s will be located in the records of $AP_t(N(u))$. If the users instead execute the P2P UPIR (III) protocol then the copies of s will be located in the records of $AP(N(u) \cup \{u\})$. The analysis in Section 3.1.2 showed that there are problems associated with the repetition of queries and the neighborhoods of the users.

The presence of a quasi-identifier

Indeed, the results from Theorem 3.1.12 suggest that the use of the protocol P2P UPIR (I) implies that there is another type of quasi-identifier present in $AP_t(U)$, namely the neighborhoods of the users $N(U)$.

With the notation from the previous discussion, let the number of copies of s in $RP(u)$ be x. Suppose that x is large enough for s to occur in all or almost all records of $N(u)$ in $AP_t(U)$ and suppose that the number of copies of s in $RP(u')$ is small for all $u' \neq u$. Also suppose that the combinatorial configuration C used by the P2P UPIR (I) protocol is triangle-free or a $(v, k, 1)$-BIBD. Then Theorem 3.1.12 implies that s can be linked to u, since the occurrences of s are linked to $N(u)$.

Therefore Theorem 3.1.12 suggests that if a community of users U implement the P2P UPIR (I) protocol with parameter the combinatorial configuration C, then the neighborhood $N(u)$ of u in $C = (U, \mathcal{L})$ is a quasi-identifier of the user $u \in U$. More formally, before transforming the database with the real profiles using the P2P UPIR (I) protocol transformation ρ, we first add the neighborhoods of the users as an attribute to the original database. That is, consider the database $RP_t(U)$ and the

P2P UPIR transformation

$$\begin{array}{rccc} \rho: & D & \to & D \\ & RP_t(U) & \mapsto & AP_t(U). \end{array}$$

The combinatorial configuration C used by the P2P UPIR protocol is a parameter given to ρ, and the properties of the points in C are assigned to the users of the P2P UPIR protocol by the mapping between the points and the users. Every point p in C has associated a neighborhood $N(p)$, and so every user u of the P2P UPIR protocol also has associated a neighborhood $N(u)$. Associated to the user is of course also his real profile $RP_t(u)$, which is his record in $RP_t(U)$. It makes sense to add the new attribute $N(u)$ to the database $RP_t(U)$, so that a record in the resulting database contains one identifier u, the attribute/ attributes of the real profile $RP_t(u)$ and the attribute $N(u)$. A record in the new database will have the following aspect:

$$[u, RP_t(u), N(u)].$$

The attribute $N(u)$ is invariant for the action of ρ, which in particular means that ρ preserves the quasi-identifying properties of $N(u)$. According to Definition 2.2.4, ensuring that $AP_t(U)$ is n-anonymous with respect to this quasi-identifier, means ensuring that every element of the family of neighborhoods $\{N(u) : u \in U\}$ occurs at least n times in $AP_t(U)$. That is, the combinatorial configuration C should be chosen so that every point shares its neighborhood with at least $n-1$ other points. Such a combinatorial configuration provides P2P UPIR (I) which is n-anonymous with respect to the quasi-identifier $N(U)$.

Definition 3.2.2. *A combinatorial configuration provides n-anonymous P2P UPIR (I) when every set of points which is a neighborhood of one point, is the neighborhood of at least n distinct points.*

The following Example 3.2.3 presents a small combinatorial configuration satisfying the condition in Definition 3.2.2.

Example 3.2.3. *An example of a combinatorial configuration which provides 3-anonymous P2P UPIR (I) is the Pappus' configuration. Figure 3.4 shows three points in the Pappus' configuration with the same neighborhood. The neighborhood is indicated in yellow and the three points with large dots.*

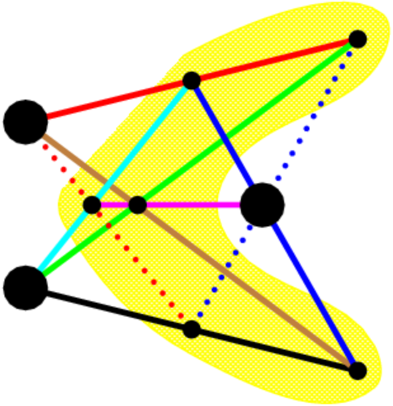

Figure 3.4: Three points in the Pappus' configuration with the same neighborhood

Section 3.2.2 will characterize and give constructions for the combinatorial configurations in Definition 3.2.2. Indeed the Pappus' configuration is only a small example of a large family of combinatorial configurations providing n-anonymous P2P UPIR (I).

3.2.2 Combinatorial configurations providing n-anonymous P2P UPIR (I)

We saw in Section 2.1.5 that a finite affine plane of order n is a combinatorial configuration A_n with parameters $(n^2, n^2+n, n+1, n)$. The line set $\mathcal{L}$ of a finite affine plane A_n is partioned into $n+1$ equivalence classes of parallel lines. Every such parallel class contains n lines. It is conjectured that a finite affine plane A_n exists if and only if n is a power of a prime. For every power of a prime q the affine plane over the finite field with q elements is a finite affine plane of order q, so at least one finite affine plane of order n exists whenever n is a power of a prime.

Theorem 3.2.4. *Consider the configuration $C = (\mathcal{P}, \mathcal{L}, I)$ obtained by taking $\mathcal{L}$ as q of the $q+1$ parallel classes of lines in a finite affine plane A_q of order q and $\mathcal{P}$ as the point set of A_q. The users of the P2P UPIR (I) protocol taking C as parameter are q-anonymous, in the sense that for every user u_i there are exactly $q-1$ other users $\{u_i\}_{i=2}^{q}$ such that $N(u_i) = N(u_j)$ for all $i, j \in \{1, \ldots, q\}$.*

Proof. The remaining parallel class of lines in A_q gives a partition $\mathcal{P} = P_1 \cup \cdots \cup P_q$ of the point set of C such that the points in any P_n are not collinear in C. Indeed the lines which make these points collinear in A_q are exactly the lines which we removed in order to construct C. In A_q every two points are collinear, implying that for any point u in C with $u \in P_n$ we have that $N(u) = \mathcal{L} \setminus P_n$. Therefore all the q points in P_n have the same neighborhood $N(u)$. □

The configuration in Theorem 3.2.4 has parameters (q^2, q^2, q, q). The use of the affine plane of order 2 gives an ordinary square with 4 points and 4 lines with 2 points on every line. The use of the affine plane of order 3 gives the Pappus configuration. We can generalize Theorem 3.2.4 by reducing the point set of C so that it contains only the points in k of the n parts of the partition of th point set of the affine plane, for $2 \leq k \leq n$. Generalizing further we see that the combinatorial configurations which we are looking for are exactly the transversal 1-designs.

Theorem 3.2.5. *The users of an instance of the P2P UPIR protocol that takes a transversal design $TD(k, n)$ as parameter are n-anonymous, in the sense that for every user u_i there are exactly $n-1$ other users $\{u_i\}_{i=2}^n$ such that $N(u_i) = N(u_j)$ for all $i, j \in \{1, \ldots, n\}$.*

Proof. The groups $G = \{g_i\}_{i=1}^k$ are a partition of the point set $\mathcal{P}$, such that the points in the same group are not collinear. Any pair of points not pertaining to the same group is contained in exactly one line. This implies that the n points inside the same group g_i all have the same neighborhood. Since G is a partition of the point set, any point $p \in \mathcal{P}$ pertains to a unique group g_i and p will share its neighborhood $N(p)$ with the n points in g_i. □

The transversal design $TD(k, n)$ in this construction is a combinatorial (nk, n^2, n, k)-configuration. Hence the construction provides an n-anonymous combinatorial configuration suitable for nk users and implies the use of n^2 communication spaces.

We have seen in Lemma 2.1.53 that the existence of a $TD(k, n)$ is equivalent to the existence of a set of $k-2$ MOLS of order n. In Theorem 2.1.25 we saw that the number x of MOLS of order n satisfies $x \leq n-1$, with equality if and only if n is a prime power.

We will now characterize the n-anonymous combinatorial configurations exactly.

Theorem 3.2.6. *An n-anonymous combinatorial (v, b, r, k)-configuration is a combinatorial configuration that satisfies the following conditions:*

- *There exists a partition $G = \{g_i\}_{i=1}^m$ of the point set such that the points in the same part are not collinear. We have $|g_i| \geq n$ for all $i \in [1, \dots, m]$;*
- *We have that $r \geq n$ and $k \leq m$;*

Proof.

- Let $C = (\mathcal{P}, \mathcal{L}, I)$ be an n-anonymous combinatorial configuration for P2P UPIR (I). Then every point $p \in \mathcal{P}$ shares its neighborhood $N(p)$ with $n - 1$ other points. "Having the same neighborhood" is a binary relation which is obviously
  - reflexive (p has the same neighborhood as p);
  - symmetric (if $N(p) = N(q)$ then $N(q) = N(p)$);
  - transitive (if $N(p_1) = N(p_2)$ and $N(p_2) = N(p_3)$, then $N(p_1) = N(p_3)$).

  Hence it is an equivalence relation and defines a partition

  $$G = \{g_1, \dots, g_m\}$$

  of the point set, in which $|g_i| \geq n$ for all $g_i \in G$. We will call the parts $g_i \in G$ groups. The neighborhood $N(p)$ of the point p is defined as the set of points which are collinear with p and different from p. In particular, if two points p and q have the same neighborhood $N(p) = N(q)$, then they can not be collinear, since if they were, then $p \in N(q)$ which would imply $p \in N(p)$. Therefore points in the same group $g \in G$ are not collinear.

- For the bound on r, consider a point $p_i \in g = \{p_1, \dots, p_n\}$ that is collinear with another point $q \in g'$, with $q \neq p_i$ (and therefore also $g' \neq g$). Then $q \in N(p_i)$, but $N(p_i) = N(p_j)$ for all $i, j \in [1, \dots, n]$, so we have $q \in N(p_i)$ for all $i \in [1, \dots, n]$. Since no line contains two points in g, we deduce that there are at least $|g| \geq n$ lines through q, so that $r \geq n$.

  Regarding the number of points on every line k, we see that, since points in the same part of G are not collinear, it is clear that any line contains k distinct points from k distinct parts of G, so that $k \leq m$.

□

If we add a restriction on regularity and maximize k when we search for n-anonymous combinatorial configurations for P2P UPIR (I), what we obtain are exactly the transversal designs.

Theorem 3.2.7. *In an n-anonymous combinatorial (v, b, r, k)-configuration C with partition $G = \{g_i\}_{i=1}^m$ and $|g_i| = n$ for all $i \in [1, \ldots, m]$, we have that*

$$r = n \text{ if and only if } m = k.$$

In this case C is a transversal design $TD(k, n)$ and $v = kn$, $b = n^2$.

Proof. We see that if $r = n$, then necessarily $k = m$ since otherwise the configuration would not be connected. On the other hand, if $k = m$, then necessarily $r = n$, since if we fix one part $g \in G$ and a point $p \in g$, then a line through g has k points through $k = m$ distinct parts $g \in G$, so the line have one point in every part in G. For any part $g' \in G$ different from g there are also a total of n lines through p. Since these lines have one point in every part of G, we get that $r = n$.

A transversal design is a uniform group divisible design in which the group size $|G|$ equals the length of the blocks k. We have seen that an n-anonymous combinatorial (v, b, n, m)-configuration such that $|g_i| = n$ and $m = k$ satisfy exactly these conditions, so it is a transversal design $TD(k, n)$. In particular, we have that

- $v = kn$,
- the line size is k,
- there is a partition G of $\mathcal{P}$ in k parts (or groups) of size n,
- any group and any block contain exactly one common point, and
- every pair of points from distinct groups is contained in exactly one block.

This is indeed exactly the definition of a transversal design $TD(k, n)$. □

There are indeed, n-anonymous combinatorial configuration which are not transversal designs.

Example 3.2.8. *Consider the combinatorial $(36, 72, 6, 3)$-configuration with point set $\mathcal{P} = \{1, \ldots, 36\}$ and line set*

$$\begin{array}{llll}
\{\{1,4,7\}, & \{4,16,19\}, & \{10,28,31\}, & \{19,22,31\}, \\
\{1,5,8\}, & \{4,17,20\}, & \{10,29,32\}, & \{19,23,32\}, \\
\{1,6,9\}, & \{4,18,21\}, & \{10,30,33\}, & \{19,24,33\}, \\
\{2,4,8\}, & \{5,16,20\}, & \{11,28,32\}, & \{20,22,32\}, \\
\{2,5,9\}, & \{5,17,21\}, & \{11,29,33\}, & \{20,23,33\}, \\
\{2,6,7\}, & \{5,18,19\}, & \{11,30,31\}, & \{20,24,31\}, \\
\{3,4,9\}, & \{6,16,21\}, & \{12,28,33\}, & \{21,22,33\}, \\
\{3,5,7\}, & \{6,17,19\}, & \{12,29,31\}, & \{21,23,31\}, \\
\{3,6,8\}, & \{6,18,20\}, & \{12,30,32\}, & \{21,24,32\}, \\
\{1,10,13\}, & \{7,22,25\}, & \{13,16,34\}, & \{25,28,34\}, \\
\{1,11,14\}, & \{7,23,26\}, & \{13,17,35\}, & \{25,29,35\}, \\
\{1,12,15\}, & \{7,24,27\}, & \{13,18,36\}, & \{25,30,36\}, \\
\{2,10,14\}, & \{8,22,26\}, & \{14,16,35\}, & \{26,28,35\}, \\
\{2,11,15\}, & \{8,23,27\}, & \{14,17,36\}, & \{26,29,36\}, \\
\{2,12,13\}, & \{8,24,25\}, & \{14,18,34\}, & \{26,30,34\}, \\
\{3,10,15\}, & \{9,22,27\}, & \{15,16,36\}, & \{27,28,36\}, \\
\{3,11,13\}, & \{9,23,25\}, & \{15,17,34\}, & \{27,29,34\}, \\
\{3,12,14\}, & \{9,24,26\}, & \{15,18,35\}, & \{27,30,35\}\}
\end{array}$$

It is clear that this combinatorial $(36, 72, 6, 3)$-configuration is 3-anonymous, but $k = 3 < 12 = m$ and $r = 6 > 3 = n$. We also observe that

$$\begin{array}{lllll}
G = & \{\{1,2,3\}, & \{4,5,6\}, & \{7,8,9\}, & \{10,11,12\}, \\
 & \{13,14,15\}, & \{16,17,18\}, & \{19,20,21\}, & \{22,23,24\}, \\
 & \{25,26,27\}, & \{28,29,30\}, & \{31,32,33\}, & \{34,35,36\}\}
\end{array}$$

and that $rk = 18$ divides $v = 36$ and $b = 72$.

3.2.3 Completely private P2P UPIR

In Section 3.1 some modifications were proposed to the P2P UPIR (I) protocol. The modified protocol was called P2P UPIR (III). Applying the modification of the protocol implies modifying the definition of n-anonymous P2P UPIR.

Definition 3.2.9. *Let $C = (\mathcal{P}, \mathcal{L}, I)$ be a combinatorial (r, k)-configuration. We say that C provides n-anonymous P2P UPIR (III) when for every point $p \in \mathcal{P}$ there are at least n distinct points $p_i \in \mathcal{P}$ for $i \in [1, \ldots, n]$ with*

$$N(p_i) \cup \{p_i\} = N(p) \cup \{p\}.$$

Then we have the following result.

Theorem 3.2.10. *A $(v, k, 1)$-BIBD provides n-anonymous P2P UPIR (III) with $n := v$. Since v is the total number of users implementing the P2P UPIR (III) protocol this is optimal.*

Proof. In a $(v, k, 1)$-BIBD any two points are connected by a line, so that the neighborhood $N(p)$ of a point p is the set of all the points in the point set except for the point p. Therefore $N(p) \cup \{p\}$ is the point set of the BIBD. From the definition of n-anonymous P2P UPIR (III) we get that the BIBD provides n-anonymous P2P UPIR with $n = v$.

More generally, in a (v, k, λ)-BIBD any two points are connected by $\lambda \geq 1$ lines, so also in this case the observation is true. However in this case the BIBD is not a combinatorial configuration. Since we want two users to share only one communication space, we are only interested in the case when $\lambda = 1$.

The n-anonymity implies better protection from reidentification for large n. The definition of combinatorial configuration that provides n-anonymous P2P UPIR says that the number of points p, for which $N(p) \cup \{p\}$ is the same, should be at least n. Therefore n can not be larger than the number of points in the configuration, so that it is optimal for $n = v$. □

Corollary 3.2.11. *[73] A finite projective plane of order m provides n-anonymous P2P UPIR (III) with $n := m^2 + m + 1$ and this is optimal for a community of $m^2 + m + 1$ users.*

In Section 3.1.2, the analysis of the P2P UPIR (I) protocol showed that it is vulnerable to attacks based on the users' repeated queries combined with knowledge of the neigborhood of the user. Combinatorial configurations which provide n-anonymity with respect to the neighborhood of the users, give risk assessment and protection. The protection coming from using n-anonymous combinatorial configurations consists in that any reidentifications process based on the users neighborhood will identify the owner of some queries only up to n other users, where n is smaller than the total number of protocol users. It should be clear to the reader that although this type of privacy protection is a result from the use of an n-anonymous combinatorial configuration, it differs from the concept of n-anonymity. We have called this notion n-confusion, see Definition 3.2.1.

The type of privacy protection provided by the P2P UPIR (III) protocol, with a $(v, k, 1)$-BIBD as parameter, is also n-confusion. In this case n equals the total number of users of the protocol. This fact justifies the use of the name complete privacy for P2P UPIR.

Definition 3.2.12. *Let U be a community of users implementing a P2P UPIR protocol with a combinatorial configuration as parameter. Let s be a sensitive sequence in $AP(U)$. We say that the P2P UPIR protocol provides complete privacy if a curious server that is given $AP(U)$, C and the mapping between the users U and the points in C, can not say to which of the real profiles $RP(U)$ of all users in U, s belongs.*

Observe that the definition of complete privacy is equivalent to the definition of n-confusion with $n = |U|$. Also observe that complete privacy is impossible for the P2P UPIR (I) protocol. Our previous results can now be restated as in the following Corollary 3.2.13.

Corollary 3.2.13. *The combinatorial configurations that offer complete privacy for P2P UPIR (III) are exactly the $(v, k, 1)$-BIBD.*

In Section 3.2.4 we will give constructions of other combinatorial configurations that provide n-anonymous P2P UPIR (III) with n smaller than v. Because of Corollary 3.2.13, we could say that the $(v, k, 1)$-BIBD offer optimal n-anonymous P2P UPIR (III) and that the question of n-anonymous combinatorial configurations for P2P UPIR (I) treated in Section 3.2.2 less interesting. However, although the modification proposed in [73] is small and easy to implement, the P2P UPIR (I) protocol is still simpler than the P2P UPIR (III) protocol. Also, giving a finite projective plane of order m as a parameter to P2P UPIR (III), implies that for a community of $m^2 + m + 1$ users the protocol needs $m^2 + m + 1$ communication spaces, that is, it will need $m^2 + m + 1$ memory sections and the same amount of cryptographic keys. The individual user has to store only $m + 1$ keys. If we use a finite affine plane of order m for a community of m^2 users the protocol needs $m^2 + m$ communication spaces. It is interesting to explore if there are solutions which require less communication spaces. Finally, finite projective and affine planes of order m are only known to exist when m is a power of a prime. There are therefore restrictions in the choice of parameters, which can be seen as a challenge to break. It is therefore interesting to search for combinatorial configurations providing n-anonymous P2P UPIR (I). It is also interesting to search for non-optimal combinatorial configurations providing n-anonymous P2P UPIR (III) which either requires less communication spaces or with more flexible parameters, or both.

3.2.4 Combinatorial configurations providing n-anonymous P2P UPIR (III)

As we saw in Theorem 3.2.10 the $(v, k, 1)$-BIBD offer n-anonymous P2P UPIR (III) with $n = v$. The $(v, k, 1)$-BIBD are therefore optimal among the n-anonymous combinatorial configurations for P2P UPIR (III) for a community of v users. In this section we will show that there are other combinatorial configurations that provide n-anonymous P2P UPIR (III), however not optimal.

In a combinatorial configuration that provides n-anonymous P2P UPIR (I) any point p shares its neighborhood $N(p)$ with at least $n-1$ other points. On the other hand, in a combinatorial configuration that provides n-anonymous P2P UPIR (III) for any point p, there are at least $n-1$ other points $(p_i)_{i=1}^{n-1}$ for which the set $N(p) \cup \{p\} = N(p_i) \cup \{p_i\}$. The resemblance of these definitions suggests that it should be possible to construct combinatorial configurations for n-anonymous P2P UPIR (III) from combinatorial configurations for n-anonymous P2P UPIR (I). Indeed this is the case, as we will see in the next Theorem 3.2.14.

Theorem 3.2.14. *Let C be a combinatorial (v, b, r, k)-configuration with $k|n$ that provides n-anonymous P2P UPIR (I) so that every point shares neighbors with exactly n more points. Then there also exists a combinatorial $(v, b+n, r+1, k)$-configuration C' that provides k-anonymous P2P UPIR (III).*

Proof. Let C be a combinatorial (v, b, r, k)-configuration with $k|n$ that provides n-anonymous P2P UPIR (I) so that every point shares neighbors with exactly n more points. Theorem 3.2.6 implies that in C there is a partition G of the point set so that points in the same partition are the points with the same neighborhood. This implies that points in the same partition are not collinear. Define C' by adding

$$k\frac{n}{k} = n$$

new lines, so that every new line contains only points from the same part of G. Let

$$P = \{p_1, \ldots, p_n\}$$

be points with $N(p_i) = N(p_j)$ in C. For any of these points p_i, in C' there will be the $k-1$ other points $(p_{i_j})_{j=1}^{k-1}$ in P, collinear with p_i by one of the new lines, such that

$$N(p_i) \cup \{p_i\} = N(p_{i_j}) \cup \{p_{i_j}\}.$$

This concludes the proof. □

As a corollary of Theorem 3.2.14 we get that an affine plane of order k is a k-anonymous combinatorial configuration for P2P UPIR (III). Just apply the construction in the proof of Theorem 3.2.14 to a transversal design $TD(k,k)$. Of course, an affine plane of order k is a $(v,k,1)$-BIBD with $v=k^2$, so we already know from Theorem 3.2.10 that it is a k^2-anonymous combinatorial configuration for P2P UPIR (III). Indeed, n-anonymity implies m-anonymity for all $m \leq n$. However, in general the combinatorial (r,k)-configuration constructed in Theorem 3.2.14 is k-anonymous but not m-anonymous for $m > k$.

Observe that not all combinatorial configurations that provide n-anonymous P2P UPIR (III) can be obtained using the construction in Theorem 3.2.14.

3.3 Collusions of users and triangle-free configurations

Consider a community of users that are implementing an instance of a P2P UPIR protocol that takes as parameter a combinatorial (r,k)-configuration C. The community of users are mapped to the points in C and they are assigned communication spaces that correpond to the lines of C. A user u_0 shares his queries with the users who are assigned the neighbor points to u_0 in C. As in Section 3.1, we call these users the neighborhood $N(u_0)$ of u_0. For u_0, to share his queries implies a privacy risk. In this section we will try to estimate how large this risk is.

3.3.1 Two different strategies for constellations of colluding users

First we observe that a user $u_1 \in N(u_0)$ who is interested in u_0's queries can choose not to follow the protocol and read all queries on the communication space c_1 that he shares with u_0, without being obliged to forward the query to the server or upload another query to the communication space. We may therefore assume that u_1 has access to all queries that u_0 uploads to c_1. These queries form a proportion of $1/r$ of the whole set of u_0's queries, that is, to the real profile $RP(u_0)$ of the user u_0. However, on the communication space c there are queries from k different users and the queries from u_0 are mixed with the other queries. Therefore u_1 does not know to whom of the $k-1$ users different from himself the queries on c belong. An adversary who owns all users on c except for u_0 will however know which of the queries belong to u_0. This observation suggests a strategy for an adversary who wants to access the real profile of u_0. This strategy consists in introducing users in the protocol such that they are collinear with u_0 and such that they are all on the same line. In order to completely control one of u_0's communication spaces, the adversary has to introduce $k-1$ users on one line which goes through u_0. In order to completely control m of u_0's communication spaces, the adversary must introduce $m(k-1)$ users on m lines which all go through u_0.

Concluding, we see that an adversary who controls $k-1$ users on the same line through u_0 has complete control over a proportion of $1/r$ of the real profile of u_0. If he controls $m(k-1)$ users on m lines through u_0, then he has complete control over a proportion of m/r of the real profile of u_0. Indeed, if the adversary controls all other users who signed up to implement the protocol, then the adversary will know the entire real profile of u_0. We will assume that it is difficult for the adversary to introduce large quantities of colluding users. If the adversary wants to have access to the largest possible proportion of the queries in the real profile of u_0, but cares less if this profile is mixed with queries from other users, then he will be more interested in introducing the colluding users such that they are collinear with u_0 by different lines. In this way the adversary will only need m users in order to have access to a proportion of m/r of the queries in the real profile of u_0, although these queries will be mixed with the queries of other users.

3.3.2 Colluding users that communicate only over channels provided by the protocol

We will first assume that the users are only able to communicate over the channels given by the protocol. In this case, any set of colluding users are forced to communicate only over the communication spaces to which they have access.

Let U be a set of users implementing an instance of a P2P UPIR protocol with a combinatorial configuration C. Consider the users u_0, u_1 and u_2 in U. Suppose that u_1 and u_2 want to form a collusion with the aim to obtain an advantage over the protocol and get access to a larger proportion of the real profile of u_0 than the protocol normally permits. If u_1 and u_2 share two different communication spaces with u_0, then the quantity of queries from the real profile of u_0 accessible to u_1 and u_2 together, is twice the quantity accessible to u_1 and u_2 on their own. In the geometric language we used before, we say that u_1 and u_2 are collinear to u_0 by two different lines, say l_1 and l_2, the lines that correspond to the two different communication spaces they share with u_0.

Since we have assumed that all communication between the users must be done over the communication channels provided by the protocol, in order for u_1 and u_2 to share their information on u_0 they must have access to a common communication space. That is, u_1 and u_2 must be collinear, say by the line l_3. We see that u_0 can not be on the line l_3. Indeed if u_0 was on l_3, then the pair of points u_0 and u_1 would be both on l_1 and l_3, so that $l_1 = l_3$. Also the pair of points u_0 and u_2 would be both on l_2 and l_3, so that $l_2 = l_3$. But we have supposed $l_1 \neq l_2$, so this is absurd. We deduce that l_1, l_2 and l_3 form a triangle in C, according to Definition 2.1.15. We see therefore that in order to avoid collusions of two users communicating over the channels provided by the P2P UPIR protocol, the configuration given as parameter to the protocol should be triangle-free.

The previous argumentation can be generalized to a set of n colluding users. Suppose that a set of n users want to form a collusion to spy on u_0 and that they only have access to the communication channels provided by the P2P UPIR, that is, to the communication spaces. From the previous discussion it is clear that the n users should all be collinear to u_0. We also previously saw that the users can be either

1. collinear with u_0 on the same line,
2. collinear with u_0 by different lines and finally, for $n > 2$ users,
3. both of the previous situations can occur.

Suppose that the adversary introduces n colluding users in the protocol. Then he obtains access to the largest proportion of the real profile if the colluding users are introduced so that they are collinear with u_0 by different lines. On the other hand if the colluding users are introduced on the same line, then the adversary obtains better control of which queries on the communication space that pertain to the real profile of u_0. Suppose that the former type of control is more interesting to the adversary than the latter. That is, suppose that the adversary wants to introduce the colluding users so that they are collinear with u_0 by different lines.

In order for these users to communicate they need to share communication spaces, that is, they need to be collinear. The best communication is obtained if they are pairwise collinear, that is, if every pair of users in the set of colluding users shares a communication space. Following the same arguments as in the case of two colluding users, it is easy to see that this requires the existence of a triangle through every triple of points u_0, u_i, u_j where u_i and u_j are colluding users. A simple counting argument then shows that the number of required triangles through u_0 is $n^2/2$. The highest proportion of the real profile of u_0 which can be read in this way requires a set of r colluding users, one sitting on every line through u_0. The number of triangles through u_0 required in this case is $r^2/2$. In this constellation the r colluding users can indeed read the entire real profile of u_0, although it will be mixed with queries from other users. One type of combinatorial configuration which permits this attack are the finite projective planes, in which every three points are on a triangle.

One can imagine a more sparse constellation of colluding users that may require less triangles. For example, n colluding users $\{u_i\}_{i=1}^{n}$ may be located so that they are all collinear with u_0 and connected in between the collusion only by, say, one path of lines $\{l_i\}_{i=1}^{n-1}$, so that the line l_i is spanned by the points assigned to the users u_i and u_{i+1}. In any case, all these constellations of colluding users are avoided if the combinatorial configuration used in the P2P UPIR protocol is triangle-free. In Section 4.3, we will treat results regarding existence and construction of triangle-free combinatorial configurations.

3.3.3 Colluding users that use external channels of communication

In the previous discussion we assumed that the colluding users only had access to the channels of communication provided by the protocol. However, for colluding users who are controlled centrally by an adversary, it is reasonable to assume that they can communicate also over channels which are not controlled by the protocol.

In a combinatorial configuration any pair of lines meet in at most one point. Therefore, if two lines are not parallel, then they identify their point of intersection. Indeed this is a property which is familiar from any linear space. For example, everyone knows that in the Euclidean plane, two intersecting lines determine a point. What this simple geometric observation suggests is that whoever controls the intersecting lines that determine the point, also controls the point. This argument was used also in the previous section, but it is important to stress that if communication between colluding users is permitted also on channels not controlled by the protocol, then the argument is independent of the existence of triangles. That is, if two users u_1 and u_2 colluding on u_0 can share their information on a channel external to the protocol, then it does not matter whether they are on a triangle or not. The only important requirement is for u_1 and u_2 to be collinear with u_0, on two different lines.

In general, for n users colluding on u_0 to obtain access to n times the information on u_0 as permitted by the protocol, they must be collinear with u_0 by n different lines.

As we already know, the assignment of communication spaces to the users follows the structure of a combinatorial configuration. We will assume that this assignment is done in the following way.

The users sign up as interested in participating in the protocol. The initialization algorithm P2P UPIR INIT (Protocol 2) is then executed by a dealer. The dealer chooses a combinatorial configuration C and a set of users with cardinality equal to the number of points in C. Then the dealer assigns the points of C (randomly) to the users, and subsequently distributes the communication spaces among the users according to the geometry of the combinatorial configuration.

Say that we want to estimate how hard it is for an adversary to introduce n colluding users so that $m \leq n$ of these are neighbors to u_0 and span m different lines with u_0. The total number of lines through u_0 is r, so we have $m \leq r$. Since we have supposed that the adversary has no control of the assignment of points to the users, the estimation should be expressed as the probability that n users owned by the adversary will be assigned points on m different lines intersecting in the same point u_0.

Proposition 3.3.1. *Consider a combinatorial (v, b, r, k)-configuration C and fix a point p_0 in C. Let an adversary A choose randomly a set $\{p_i\}_{i=1}^{n}$ of n other points from C. The probability that exactly m of the points p_i are collinear with p_0 and that the m lines p_ip_0 are all different in C, for $m \leq r$ and $m \leq n \leq v$ is*

$$\frac{n!}{v!}\sum_{s=0}^{n\text{-}m\text{-}1}\sum_{i_s=s}^{n\text{-}m\text{-}1}\prod_{j=0}^{m\text{-}1}((r-j)(k-1))\prod_{t=0}^{n\text{-}m\text{-}1}(v-r(k-1)+i_t(k-2)-t-1).$$

Proof. The number of points on a line through p_0 different from p_0 is $k-1$, the number of points which are collinear with p_0 but different from p_0 is $r(k-1)$ and the total number of points in C is v. For the sake of simplicity, define $a = r(k-1)$. Define

- the event $\oplus_i$ as the event that the adversary A introduces a point p_i so that p_i is collinear with p_0 but not on any of the lines p_0p_j for $1 \leq j \leq i-1$;

- the event $\ominus_i$ as the event that A introduces the point p_i so that p_i is not collinear with p_0 or on some of the lines p_0p_j for $1 \leq j \leq i-1$.

Observe that the events $\oplus_i$ and $\ominus_i$ are complementary for the introduction of p_i. We will see now that the probability of $\oplus_i$ or $\ominus_i$ depends on how A has chosen the points $p_1, \ldots, p_{i-1}$. That is, the probability for the events $\oplus_i$ and $\ominus_i$ depends on the sequence of previous events $(x_j)_{j=1}^{i-1}$ in which the elements x_j take values in $\{\oplus_j, \ominus_j\}$. Using the complementary property, the probabilities $P(\oplus_i|(x_j)_{j=1}^{i-1})$ and $P(\ominus_i|(x_j)_{j=1}^{i-1})$ can be represented in a binary directed tree of height i, so that the vertices are assigned these probabilities as weights.

The weights on level 1 in this tree are the probabilities $P(\oplus_1)$ and $P(\ominus_1)$. For the first point p_1 introduced by A, these probabilities are

- $P(\oplus_1) = \frac{a}{v-1}$: the probability that p_1 will be collinear with p_0;

- $P(\ominus_1) = 1 - \frac{a}{v-1} = \frac{v-a-1}{v-1}$: the probability that p_1 will *not* be collinear with p_0.

The weights on level 2 in the tree are the probabilities $P(\oplus_2|x_1)$ and $P(\ominus_2|x_1)$, with $x_1 \in \{\oplus_1, \ominus_1\}$. Suppose that p_1 was chosen collinear with p_0 ($x_1 = \oplus_1$). For the second point p_2 introduced by A, these probabilities are then

- $P(\oplus_2|\oplus_1) = \frac{a-(k-1)}{v-2}$: the probability that p_2 will be collinear with p_0 but not on the line p_0p_1;

- $P(\ominus_2|\oplus_1) = \frac{v-a-1+(k-2)}{v-2}$: the probability that p_2 will *not* be collinear with p_0 or that p_2 will be on the line p_0p_1.

Now suppose that p_1 was *not* chosen collinear with p_0 ($x_1 = \ominus_1$). Then, the probabilities are

- $P(\oplus_2|\ominus_1) = \frac{a}{v-2}$: the probabilities that p_2 will be collinear with p_0 but not on the lines p_0p_1;

- $P(\ominus_2|\ominus_1) = \frac{v-a-2}{v-2}$: the probabilities that p_2 will *not* be collinear with p_0 or that p_2 will be on the line p_0p_1.

Observe that the fact that p_1 is chosen collinear with p_0 ($\oplus_1$) implies that the possible points for the event $\oplus_2$ to occur is lowered by all the $k-1$ points on the line p_0p_1 different from p_0, while the possible points for the event $\ominus_2$ to occur is augmented by the $k-2$ points on the line p_0p_1 different from p_0 and p_1 (a point can only be chosen once).

On the other hand, if p_1 is chosen *not* collinear with p_0 ($\ominus_1$) then the possible points for the event $\oplus_2$ to occur is not lowered at all (since the number of non-assigned collinear points stay the same) and the number of possible points for the event $\ominus_2$ is lowered by one (corresponding to the assigned non-collinear point p_1).

In general, the weights on level i in the tree are the probabilities $P(\oplus_i)$ and $P(\ominus_i)$. For the ith point p_i introduced by A, assuming that the sequence of previous events is $(x_j)_{j=1}^{i-1}$ and the number of $x_j = \oplus_j$ is t and the number of $x_j = \ominus_j$ is s (we have $i-1 = s+t$), the probability is

- $P(\oplus_i|(x_j)_{j=1}^{i-1}) = \frac{a-t(k-1)}{v-i}$ that p_2 will be collinear with p_0 but not on any of the lines p_0p_j for $1 \leq j \leq i-1$;
- $P(\ominus_i|(x_j)_{j=1}^{i-1}) = \frac{v-a-1+t(k-2)-s}{v-i}$ that p_i will *not* be collinear with p_0 or that p_i will be on some of the lines p_0p_j for $1 \leq j \leq i-1$.

The probability that A introduces n different points $(p_i)_{i=1}^{n}$ in C so that m of these are collinear with p_0 and so that they span m different lines with p_0, for $m \leq r$ and $m \leq n \leq v$ is

$$\sum_{x \,\in\, X(n,\, m)} P((x_i)_{i=1}^{n}) = \sum_{x \,\in\, X(n,\, m)} \prod_{\mathrm{i=1}}^{\mathrm{n}} P\left(x_i|(x_j)_{j=1}^{i-1}\right), \tag{3.4}$$

where $X(n,m)$ is the set of sequences

$$X(n,m) = \{x := (x_i)_{i=1}^{n} : \sharp\{x_i = \oplus_i\} = m\}\,.$$

That is, it is the sum of the probabilites obtained by multiplying the weights on the vertices of the paths that contain m events $\oplus_j$ and $n-m$ events $\ominus_j$. When $m > r$ then the correponding probability is 0, since there are only r lines through p_0.

In any of the terms of the sum in Equation 3.4 the m probabilities

$$P(\oplus_i|(x_j)_{j=1}^{i-1})$$

will take the values

$$\frac{a-t(k-1)}{v-i}$$

for $0 \leq t \leq m-1$, but the probabilites

$$P(\ominus_i|(x_j)_{j=1}^{i-1})$$

depend on for which of the $n-m$ different values of i the event $\ominus_i$ occurs. The order of the events $\oplus_i$ and $\ominus_i$ is therefore important only for the value of the probability $P(\ominus_i|(x_j)_{j=1}^{i-1})$. In particular, the number of terms in the sum is

$$\left(\begin{array}{c} n \\ m \end{array}\right) = \left(\begin{array}{c} n \\ n-m \end{array}\right).$$

The denominator is always the same in all terms of the sum in Equation 3.4:

$$\prod_{i=0}^{n-1}(v-i) = \frac{v!}{n!}.$$

We deduce that the probability that A introduces n different points $(p_i)_{i=1}^{n}$ in C so that m of these are collinear with p_0 and so that they span m different lines with p_0, for $m \leq r$ and $m \leq n \leq v$ is

$$\sum_{s\,=\,0}^{\text{n-m-1}}\sum_{i_s\,=\,s}^{\text{n-m-1}}\prod_{j\,=\,0}^{\text{m-1}}((r(k-1)-j(k-1))\prod_{t\,=\,0}^{\text{n-m-1}}(v-r(k-1)+i_t(k-2)-t-1),$$

divided by

$$\frac{v!}{n!}.$$

□

Proposition 3.3.2. *Consider a combinatorial (v,b,r,k)-configuration C and fix a point p in C. Let a dealer choose a set $\{p_i\}_{i=1}^{n}$ of n other points from C. The probability that exactly m of the points p_i are collinear with p, for $m \leq r$ and $m \leq n$ is*

$$\left(\begin{array}{c} n \\ m \end{array}\right)\frac{r(k-1)!}{(r(k-1)-m)!}\,\frac{(v-r(k-1))!}{(v-r(k-1)-(n-m))!}\,\frac{(v-n)!}{v!}.$$

Proof. The proof is as in Proposition 3.3.1 but simpler, since in this case when p_i is chosen collinear with p_0 ($\oplus_i$) the possible points for the event $\oplus_{i+1}$ to occur is lowered only by the point p_i, and that the number of possible points for the event $\ominus_{i+1}$ to occur stays the same.

As before, the weights on level 1 in the tree are the probabilities $P(\oplus_1)$ and $P(\ominus_1)$. Also this time, for the first point p_1 introduced by A, these probability are

- $P(\oplus_1) = \frac{a}{v-1}$: the probability that p_1 will be collinear with p_0;
- $P(\ominus_1) = 1 - \frac{a}{v-1} = \frac{v-a-1}{v-1}$: the probability that p_1 will *not* be collinear with p_0.

The weights on level 2 in the tree are the probabilities $P(\oplus_2 : x_1))$ and $P(\ominus_2 : x_1)$, with $x_1 \in \{\oplus_1, \ominus_1\}$. Suppose that p_1 was chosen collinear with p_0 ($x_1 = \oplus_1$). For the second point p_2 introduced by A, in this case these probabilities are then

- $P(\oplus_2|\oplus_1) = \frac{a-1}{v-2}$: the probability that p_2 will be collinear with p_0;
- $P(\ominus_2|\oplus_1) = \frac{v-a-1}{v-2}$: the probability that p_2 will *not* be collinear with p_0.

Now suppose that p_1 was *not* chosen collinear with p_0 ($x_1 = \ominus_1$). Then, the probabilities are

- $P(\oplus_2|\ominus_1) = \frac{a}{v-2}$: the probabilities that p_2 will be collinear with p_0;
- $P(\ominus_2|\ominus_1) = \frac{v-a-2}{v-2}$: the probabilities that p_2 will *not* be collinear with p_0.

If we draw this tree until level n, and investigate all paths containing m events $\oplus_i$, then we see that all the probabilities in these paths have the same factors. The product of these factors is

$$\frac{(\prod_{i=1}^{m}(a-i))(\prod_{i=1}^{n-m}(v-a-i))}{\prod_{i=1}^{n}(v-i)}$$

and there are

$$\left(\begin{array}{c} n \\ m \end{array}\right)$$

such paths so that the probability that exactly m of the points p_i are collinear with p, for $m \leq r$ and $m \leq n$ is

$$\left(\begin{array}{c} n \\ m \end{array}\right) \frac{r(k-1)!}{(r(k-1)-m)!} \frac{(v-r(k-1))!}{(v-r(k-1)-(n-m))!} \frac{(v-n)!}{v!}.$$

□

Chapter 4

Constructing configurations

4.1 Constructions of finite projective planes

A standard construction of a finite projective plane of order q uses homogeneous coordinates over the finite field of order q. This method gives us one plane for every order q: $\mathbb{P}(\mathbb{F}_q)$. Many more projective planes can be constructed using algebraic structures with less postulates than fields. M. Hall defined the concept of *ternary ring* (see below) as the algebraic structure that exactly corresponds to the structure needed in the construction of projective planes [43, 44].

Another very simple construction is given by the existence in some projective planes of the so-called Singer cycles. The projective planes that have Singer cycles can be constructed by defining the lines of the plane as the successive translations of a difference set. Apart from being a very simple construction (once given a difference set), it gives a compact way of representing the projective plane, since defining one of the lines is enough to have constructed the entire plane. However, not all planes can be constructed in this way and, more important, constructing a difference set is equivalent to constructing a projective plane.

When q is prime there is a rather straight forward algorithm for constructing $\mathbb{P}(\mathbb{F}_q)$. We will here give one variant of this straight forward algorithm an efficient and explicit expression, at the same time generalizing it in order to be able to efficiently construct any projective plane (whenever it exists). In particular we construct projective planes of order a power of a prime number, using an algorithm that is efficient and of easy implementation. The generalization uses the mathematical structure ternary ring, which was first introduced by M. Hall as a tool for his construction and classification of projective planes [43]. For the definition and a short introduction on ternary rings, see Section 2.1.5. Ternary rings make possible a general formulation of the algorithm, and simplify the proof.

In the following we will represent a finite projective plane using an adjacency list, i.e. a list of the subsets defining the lines of the plane, using the set of integers $\{1,\dots,n\}$ to represent the points.

The following defines a general and efficient algorithm for the construction of finite projective planes.

Proposition 4.1.1. *Let R be a ternary ring with q elements*

$$R_0 = \{0,\dots,R_{q-1}\}$$

and ternary operation T. Let $\iota(R_i) = i$. Consider the matrix $A = (a_{i,j})$ defined by

$$a_{i,j} = 2 + iq + j,$$

with $i \in \{0,\dots,q\}$ and $j \in \{0,\dots,q-1\}$, and the q matrices $B^0 = (b^0_{i,j}), B^1 = (b^1_{i,j}),\dots,B^{q-1} = (b^{q-1}_{i,j})$ defined by

$$b^k_{i,j} = 2 + (j+1)q + \iota(T(R_j, R_k, R_i)),$$

with $i \in \{0,\dots,q-1\}$ and $j \in \{0,\dots,q-1\}$. The following matrix gives us an adjacency list defining a projective plane of order q:

$$C_q = \left(\begin{array}{c|c}
1 & \\
1 & A \\
\vdots & \\
1 & \\
\hline
2 & \\
2 & B^0 \\
\vdots & \\
2 & \\
\hline
3 & \\
3 & B^1 \\
\vdots & \\
3 & \\
\hline
\vdots & \vdots \\
\vdots & \vdots \\
\hline
q+1 & \\
q+1 & B^{q-1} \\
\vdots & \\
q+1 &
\end{array}\right)$$

Proof. First we observe that in the $(q+1) \times q$ matrix A, all numbers in $\{2, \ldots, q^2+q+1\}$ appear once.

Now let A' be the matrix A, dropping the first row, i.e. the row containing the numbers $\{2, \ldots, p+1\}$. We observe that by property T2 from Definition 2.1.35 the matrix B^0 is A' transposed, and therefore no two numbers appearing in the same row in A' (and hence in A) can appear in the same row in B^0.

Consider now the construction of the matrices B^k. In every column b_j of B^k, the elements

$$b_{i,j}^k = 2 + (j+1)q + \iota(T(R_j, R_k, R_i))$$

by property T3 take all the values in

$$\{2 + (j+1)q, \ldots, 1 + (j+2)q\}, \tag{4.1}$$

as i goes through $\{0, \ldots, q-1\}$. We will call (4.1) the element sequence of column b_j.

Each of these sequences corresponds to the elements in a row of A', so if two numbers appear in the same row in A they can not appear in the same row in B^k for any k.

On the other hand, all the sequences are disjoint, so all numbers in $\{q+2, \ldots, q^2+q+1\}$ appear exactly once in each B^k. Consequently, the only case in which two elements may still coincide in more than two rows is when the two rows belong to two different minors

$$\left(\begin{array}{c|c} \vdots & \\ k+2 & B^k \\ \vdots & \end{array}\right)$$

and

$$\left(\begin{array}{c|c} \vdots & \\ k'+2 & B^{k'} \\ \vdots & \end{array}\right)$$

with $k \neq k'$.

Suppose that these two rows are the ith row in $(k+2|B^k)$ and the i'th row in $(k'+2|B^{k'})$. Let the two repeated elements be b^k_{ij} and $b^k_{ij'}$, with $j \neq j'$. Since the set of elements of the jth column in B^k is the same as the set of elements of the jth column in $B^{k'}$, and the analogous case is true for the j'th column, we must have $b^k_{ij} = b^{k'}_{i'j}$ and $b^k_{ij'} = b^{k'}_{i'j'}$, i.e. $T(R_j, R_k, R_i) = T(R_j, R'_k, R'_i)$. By condition T4 it must be $j = j'$, a contradiction. □

Proposition 4.1.2. *The computational cost of the algorithm given by Proposition 4.1.1 is $O(q^3)$, provided that q is prime and that the ternary ring we use corresponds to the arithmetic of $\mathbb{F}_q$. In particular the number of operations used in each case is*

- *Additions: $q + q^2 + 2q^3$;*
- *Multiplications: $q + q^2 + 2q^3$;*
- *Modulo operations: q^3.*

Proof. The matrix A needs $q(q+1)$ multiplications and $2q(q+1)$ additions. If we do not count addition of the constant 2 we get $q(q+1)$ additions. The matrix B^k needs $2q^2$ multiplications, q^2 modulo operations and $4q^2$ additions. If we continue not counting addition of the constants

1 and 2 we get $2q^2$ additions. Considering that $k \in \{0, \ldots, q-1\}$ we get $q(q+1) + q(2q^2) = q + q^2 + 2q^3$ multiplications, $q(q+1) + q(2q^2) = q + q^2 + 2q^3$ additions and q^3 modulo operations. □

In order to clarify how to implement the algorithm, we will now see some examples. In the following, if nothing else is said, the operations $+$ and $\cdot$ stand for ordinary integer sum and product.

Example 4.1.3. *$(\mathbb{Z}/p\mathbb{Z}, +, \cdot)$, the integers modulo a prime number p, give rise to a ternary ring with ternary operation $T(x, y, z) = xy + z \pmod{p}$. With the notation from Proposition 4.1.1 we calculate $A = (a_{i,j})$ using*

$$a_{i,j} = 2 + ip + j,$$

and for $k \in \{0, \ldots, p-1\}$ we calculate $B^k = (b^k_{i,j})$ using

$$b^k_{i,j} = 2 + (j+1)p + [i + jk \pmod{p}],$$

with $i \in \{0, \ldots, p-1\}$ and $j \in \{0, \ldots, p-1\}$.

We now present the results from this construction for $p = 2, p = 3$.

$$C_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 4 & 5 \\ 1 & 6 & 7 \\ 2 & 4 & 6 \\ 2 & 5 & 7 \\ 3 & 4 & 7 \\ 3 & 5 & 6 \end{pmatrix} \qquad C_3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 5 & 6 & 7 \\ 1 & 8 & 9 & 10 \\ 1 & 11 & 12 & 13 \\ 2 & 5 & 8 & 11 \\ 2 & 6 & 9 & 12 \\ 2 & 7 & 10 & 13 \\ 3 & 5 & 9 & 13 \\ 3 & 6 & 10 & 11 \\ 3 & 7 & 8 & 12 \\ 4 & 5 & 10 & 12 \\ 4 & 6 & 8 & 13 \\ 4 & 7 & 9 & 11 \end{pmatrix}$$

Example 4.1.4. *A finite field $\mathbb{F}_q$ defines a ternary ring with ternary operation $T(x, y, z) = xy + z$, where the sum and the product follow the arithmetic rules of $\mathbb{F}_q$. With the notation from Proposition 4.1.1 we calculate $A = (a_{i,j})$ using*

$$a_{i,j} = 2 + iq + j.$$

We represent the elements of $\mathbb{F}_q$ *with the integers in the array* $F = (0, \dots, q-1)$. *With the zero of the field represented as* 0 *and the unit represented as* 1, *the elements of the matrices* B^k, *for* $k \in \{0, \dots, q-1\}$, *can now be calculated as*

$$b^k_{i,j} = 2 + (j+1)q + [F_i + F_{k-1}F_j],$$

with $i \in \{0, \dots, q-1\}$ *and* $j \in \{0, \dots, q-1\}$. *The arithmetic of the elements of the array* F *must follow the arithmetic rules of* $\mathbb{F}_q$.

Observe that Example 4.1.3 is a special case of this one.

We now present the result from this construction for $q = 4$.

$$C_4 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 6 & 7 & 8 & 9 \\ 1 & 10 & 11 & 12 & 13 \\ 1 & 14 & 15 & 16 & 17 \\ 1 & 18 & 19 & 20 & 21 \\ 2 & 6 & 10 & 14 & 18 \\ 2 & 7 & 11 & 15 & 19 \\ 2 & 8 & 12 & 16 & 20 \\ 2 & 9 & 13 & 17 & 21 \\ 3 & 6 & 11 & 16 & 21 \\ 3 & 7 & 10 & 17 & 20 \\ 3 & 8 & 13 & 14 & 19 \\ 3 & 9 & 12 & 15 & 18 \\ 4 & 6 & 12 & 17 & 19 \\ 4 & 7 & 13 & 16 & 18 \\ 4 & 8 & 10 & 15 & 21 \\ 5 & 6 & 13 & 15 & 20 \\ 5 & 7 & 12 & 14 & 21 \\ 5 & 8 & 11 & 17 & 18 \\ 5 & 9 & 10 & 16 & 19 \end{pmatrix}$$

In this thesis the focus is on constructing optimal combinatorial configurations for P2P UPIR, and it is probably enough with one combinatorial configuration for a given $d = q - 1$. It is conjectured that all finite projective planes have order a power of a prime number. Therefore it is highly probable that all projective planes constructed by our algorithm will have this property. The finite projective planes constructed using finite fields constitute a subset of all finite projective planes, with the particularity that they satisfy the theorem of Desargues. Although the existence of finite fields is restricted to q a power of a prime number, since there always exists one finite field for every q, we will always get at least one projective plane of order q, using Example 4.1.4. It is therefore of little interest to continue the examples further.

Observe though that some projective planes constructed using less 'regular' (i.e. satisfying less axioms) ternary rings could be interesting when some properties associated to the theorem of Desargues are to be avoided.

4.2 The numerical semigroup associated to the existence of combinatorial configurations

4.2.1 The set of $(r, k)-$configurable tuples

Definition 4.2.1. *We say that the tuple of parameters (v, b, r, k) is configurable if there exists a (v, b, r, k)-configuration.*

As we saw in Theorem 2.1.55, if (v, b, r, k) is configurable, then $vr = bk$. Consequently there exists d such that

$$v = \frac{bk}{r} = d\frac{k}{\gcd(r,k)}$$

and symmetrically

$$b = \frac{vr}{k} = d'\frac{r}{\gcd(r,k)}.$$

Since v and b are integers, so are d and d'. We also have

$$\begin{aligned} d &= \tfrac{v\gcd(r,k)}{k} \\ &= \tfrac{bk\gcd(r,k)}{rk} \\ &= \tfrac{b\gcd(r,k)}{r} = d' \end{aligned}.$$

Therefore, to each configurable tuple (v, b, r, k) we can associate an integer d. On the other hand, given r and k, any $d \in \mathbb{N}$ determines two integers v and b, perhaps corresponding to the number of points and lines of a combinatorial configuration.

In the following we will consider the set of integers such that they may be associated to a combinatorial (r,k)-configuration, for r and k fixed. That is, consider the set of natural numbers

$$D_{(r,k)} = \left\{ d \in \mathbb{N} \cup \{0\} : \left(d\frac{k}{\gcd(r,k)}, d\frac{r}{\gcd(r,k)}, r, k \right) \text{ is configurable} \right\}.$$

The aim here is to study the set $D_{(r,k)}$. A first observation shows that the duality of combinatorial configurations (see Section 2.1.2) implies that $D_{(r,k)} = D_{(k,r)}$.

By convention we will say that the tuple $(0,0,r,k)$ is configurable for any pair r,k, although it represents the empty combinatorial configuration, and we associate the integer 0 to the configurable tuple $(0,0,r,k)$. As a consequence we get $0 \in D_{(r,k)}$, for any pair of r,k.

4.2.2 The numerical semigroup $D_{(x,2)} = D_{(2,x)}$

As we saw previously, in Section 2.1.2 and Example 2.1.14, the combinatorial $(v,b,r,2)$-configurations are r-regular undirected connected graphs with v vertices and b edges. The vertices in the graph correspond to the points of the configuration and the lines, which have only $k=2$ points, correspond to the edges.

Therefore the following Lemma 4.2.2 and Lemma 4.2.4 on the existence of regular graphs provide the key results for describing the set $D_{(2,k)}$. Because of duality, if we determine $D_{(r,2)}$, then we also determine $D_{(2,r)}$.

Although the results in Lemma 4.2.2 and Lemma 4.2.4 are well-known, for the sake of completeness we will provide the proofs.

Lemma 4.2.2. *Let r be an even positive integer. A connected r-regular graph with v vertices exists if and only if $v \geq r+1$.*

Proof. By definition, any r-regular graph must have a number of vertices at least $r+1$.

Conversely, suppose $v \geq r+1$. Consider a set of vertices $x_1, \ldots, x_v$. Put an edge between x_i and x_j, with $i \leq j$, if $j-i \leq r/2$ or $i+v-j \leq r/2$. This gives a connected r-regular graph with v vertices. □

The construction in this last proof is illustrated in Figure 4.1.

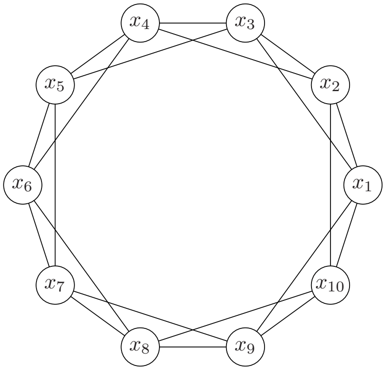

Figure 4.1: Construction of a connected 4-regular graph with 10 vertices

Since the $(v, b, r, 2)$-configurations are r-regular connected graphs with v vertices and b edges, we get the following Corollary 4.2.3. Remember that we write $\langle a_1, \ldots, a_n \rangle$ to denote the numerical semigroup generated by $a_1, \ldots, a_n$.

Corollary 4.2.3. *If r is an even positive integer then*

$$D_{(2,r)} = D_{(r,2)} = \{0, r+1, \rightarrow\}.$$

Lemma 4.2.4. *Let r be an odd positive integer. A connected r-regular graph with v vertices exists if and only if v is even and $v \geq r+1$.*

Proof. By definition, any r-regular graph must have a number of vertices at least $r+1$. Now, since the number of edges is $b = vr/2$, then vr must be even and since r is odd, then v must be even. Conversely, suppose that v is even and that $v \geq r+1$. Consider a set of vertices $x_1, \ldots, x_v$. Put an edge between x_i and x_j, with $i \leq j$, if $j - i \leq (r-1)/2$ or $i + v - j \leq (r-1)/2$. Put also edges between x_i and $x_{i+v/2}$ for i from 1 to $v/2$. This gives a connected r-regular graph with v vertices. □

The construction in this last proof is illustrated in Figure 4.2.

Now, the fact that the $(v, b, r, 2)$-configurations are r-regular connected graphs with v vertices and b edges, implies the following corollary.

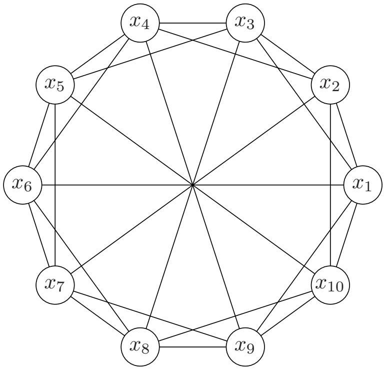

Figure 4.2: Construction of a connected 5-regular graph with 10 vertices

Corollary 4.2.5. *If r is an odd positive integer then*

$$D_{(2,r)} = D_{(r,2)} = \{0, \frac{r+1}{2}, \rightarrow\}.$$

4.2.3 The numerical semigroup $D_{(x,3)} = D_{(3,x)}$

We can use Theorem 2.1.67 to completely describe the numerical semigroup $D_{(3,x)} = D_{(x,3)}$ for all integers $x \geq 3$. The case $r = 3$ is the dual of $k = 3$.

Theorem 4.2.6. *Suppose $k > 1$ then*

$$D_{(3,k)} = \begin{cases} \{0, 2k+1, 2k+2, \ldots\} & \text{if } k \equiv 0 \pmod 3 \\ \{0, \frac{2k+1}{3}, \frac{2k+1}{3}+1, \frac{2k+1}{3}+2, \ldots\} & \text{if } k \equiv 1 \pmod 3 \\ \{0, \frac{2k+2}{3}, \frac{2k+2}{3}+1, \frac{2k+2}{3}+2, \ldots\} & \text{if } k \equiv 2 \pmod 3 \end{cases}$$

Proof. Dually, by Theorem 2.1.67 we know that any tuple $(v, b, 3, k)$ with $b \neq 0$ is configurable if and only if $3v = bk$ and $b \geq k(3-1)+1 = 2k+1$. In particular, the non-zero values b for which there exists a configurable tuple $(v, b, 3, k)$ are exactly those integers $b \geq 2k+1$ such that $\frac{bk}{3}$ is an integer.

If $k \equiv 0 \pmod 3$ then the only condition is $b \geq 2k+1$ which results in

$$d = \frac{b \gcd(3,k)}{3} = \frac{3b}{3} = b \geq 2k+1$$

and this proves the result in this case.

Otherwise, we need $b \geq 2k+1$ and b be a multiple of 3. If $k \equiv 1 \pmod 3$ this is equivalent to $b \in \{2k+1, 2k+4, 2k+7, \ldots\}$ and so $d = \frac{b \gcd(3,k)}{3} = \frac{b}{3}$ is in

$$\left\{ \frac{2k+1}{3}, \frac{2k+1}{3}+1, \frac{2k+1}{3}+2, \ldots \right\}.$$

If $k \equiv 2 \pmod 3$ this is equivalent to $b \in \{2k+2, 2k+5, 2k+8, \ldots\}$ and so $d = \frac{b \gcd(3,k)}{3} = \frac{b}{3}$ is in

$$\left\{ \frac{2k+2}{3}, \frac{2k+2}{3}+1, \frac{2k+2}{3}+2, \ldots \right\}.$$

□

4.2.4 The set of integers associated to the combinatorial (r,k)-configurations forms a numerical semigroup

We want to prove that $D_{(r,k)} \subset \mathbb{N} \cup \{0\}$ is a numerical semigroup. As we saw in Section 2.1.10, a numerical semigroup is a subset $S \subset \mathbb{N} \cup \{0\}$, so that S is closed under addition, $0 \in S$ and the complement $(\mathbb{N} \cup \{0\}) \setminus S$ is finite.

Lemma 2.1.85 says that in order to prove that a set is a numerical semigroup it is enough to prove that the set is a submonoid of the natural numbers with coprime elements. In particular it is enough to prove that

- $0 \in D_{(r,k)}$,
- $D_{(r,k)}$ is closed under addition,
- at least two elements of $D_{(r,k)}$ are coprime.

The two first conditions ensure that the subset $D_{(r,k)}$ of the natural numbers is a monoid. The operation of the monoid is addition. The last condition ensures that the monoid contains the numerical semigroup generated by the two coprime elements. The complement of this numerical semigroup is finite, therefore also the complement of the monoid, and we deduce that it is a numerical semigroup.

The set of configurable tuples is a submonoid in the natural numbers

As already commented, by convention we consider the empty combinatorial configuration to be a combinatorial $(0, 0, r, k)$-configuration for every pair of r, k, so that we have $0 \in D_{(r,k)}$.

We will now prove that the set $D_{(r,k)}$ is closed under addition.

Lemma 4.2.7. *If (v, b, r, k) and (v', b', r, k) are configurable tuples, so is the tuple $(v + v', b + b', r, k)$.*

Proof. Suppose that we have a (v, b, r, k)-configuration $C = (\mathcal{P}, \mathcal{L}, I)$ with points

$$\mathcal{P} = \{x_1, \ldots, x_v\}$$

and lines

$$\mathcal{L} = \{y_1, \ldots, y_b\}$$

and another (v', b', r, k)-configuration $C' = (\mathcal{P}', \mathcal{L}', I')$ with points

$$\mathcal{P}' = \{x'_1, \ldots, x'_{v'}\}$$

and lines

$$\mathcal{L}' = \{y'_1, \ldots, y'_{b'}\}.$$

Consider the incidence graphs G and G' of C and C' respectively. Then the graph G is bipartite with partitioned vertex set $\mathcal{P} \cup \mathcal{L}$ and the edges are the elements in I. Analogously, the graph G' is bipartite with partitioned vertex set $\mathcal{P}' \cup \mathcal{L}'$ and edge set I'.

Define another graph $\tilde{G}$ with vertices $\mathcal{P} \cup \mathcal{P}' \cup \mathcal{L} \cup \mathcal{L}'$ and the edges $I \cup I'$. Then this graph is also bipartite. We can assume without loss of generality that the edges x_1y_1, x_vy_b,$x'_1y'_1$, $x'_{v'}y'_{b'}$ belong to the original configurations.

Replace the edges x_vy_b and $x'_1y'_1$ by the edges $x_vy'_1$ and x'_1y_b. This is then the incidence graph of a $(v + v', b + b', r, k)$ configuration [29]. An example of this construction is illustrated in Figure 4.3. □

Let

$$d = v \gcd(r, k)/k = b \gcd(r, k)/r$$

and

$$d' = v' \gcd(r, k)/k = b' \gcd(r, k)/r$$

be the two integers associated to the two configurable tuples (v, b, r, k) and (v', b', r, k). Then

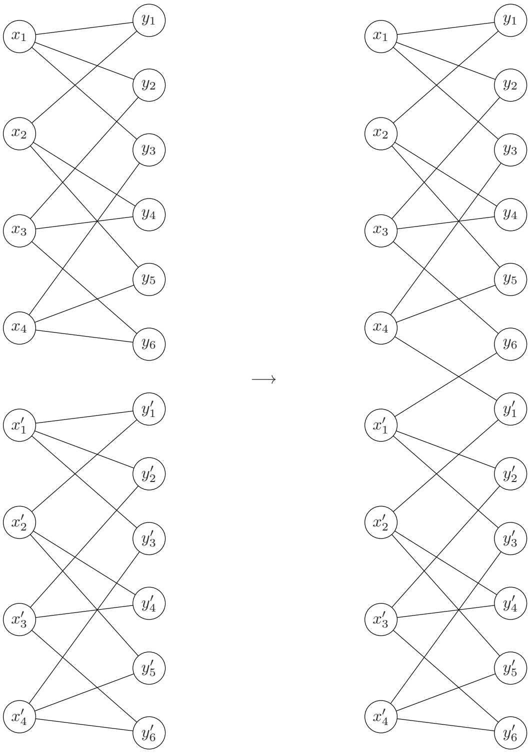

Figure 4.3: Construction of a $(v+v', b+b', r, k)$ configuration from a (v, b, r, k) configuration and a (v', b', r, k) configuration.

$$d'' = (v+v')\gcd(r,k)/k = (b+b')\gcd(r,k)/r = d + d'.$$

is the integer associated to the configurable tuple $(v+v', b+b', r, k)$. Hence if $d, d' \in D_{(r,k)}$, then also $d + d' \in D_{(r,k)}$. In other words $D_{(r,k)} \subset \mathbb{N} \cup \{0\}$ is closed under addition. Together with the fact that $0 \in D_{(r,k)}$ we get the result we were looking for.

Proposition 4.2.8. *$D_{(r,k)}$ is a submonoid of the natural numbers.*

The submonoid contains two coprime elements

In Section 4.2.2 we determined exactly the sets $D_{(r,2)}$ and $D_{(2,k)}$. We may therefore assume that $r, k \geq 3$.

From Theorem 2.1.54, we know that for any pair of natural number $r, k \geq 3$, there exists a combinatorial (r, k)-configuration.

We will now construct a second element of $D_{(r,k)}$, such that the element we already have and the new one are coprime. In order to do so we need the following lemma.

Lemma 4.2.9. *Suppose we have a (v, b, r, k)-configuration with $r \geq 3$ and incidence graph G. There exist three edges in G such that the six ends are all different.*

Proof. It is easy to see, by the property that no cycle of length 4 exists, that there exists a path with four edges with the five ends being different. Three of these ends will be in one partition of the graph while the other two will be in the other partition. Take the vertex at the end of the path. It must be one of the three in the same partition. Since its degree is at least 3, then it will have one neighbor not in the path. So, by adding the edge from the end of the path to this additional vertex, we obtain a new path with 5 edges with all its vertices being different. By taking the first, third, and fifth edges of this new path we obtain the result. □

Lemma 4.2.9 tells us that the vertices $\{x_1, \ldots, x_v\}$, $\{y_1, \ldots, y_b\}$ in the incidence graph of a combinatorial (v, b, r, k)-configuration with $r \geq 3$ can be indexed so that the edges (x_1, y_1), (x_2, y_2) and (x_v, y_b) belong to the edge set.

We are now ready to prove the existence of two coprime elements of $D_{(r,k)}$.

Proposition 4.2.10. *$D_{(r,k)}$ contains two elements $m \neq 0$ and $sm + 1$, with $s = rk/\gcd(r, k)$, so that the two are coprime.*

Proof. Remember that we have assumed that r and k are larger than 3. Because of Theorem 2.1.54 and since $D_{(r,k)} \subseteq \mathbb{N} \cup \{0\}$, there is a minimal non-zero element m in $D_{(r,k)}$. Let us call

$$v = mk/\gcd(r, k)$$

and

$$b = mr/\gcd(r, k).$$

Select a (v, b, r, k) configuration. Take

$$s = rk/\gcd(r, k)$$

copies of this configuration. Let us call the vertices in the incidence graph of the ith copy

$$x_1^{(i)}, \dots, x_v^{(i)}, y_1^{(i)}, \dots, y_b^{(i)}.$$

By Lemma 4.2.9 we can assume that

$$x_1^{(i)}y_1^{(i)}, x_2^{(i)}y_2^{(i)} \text{ and } x_v^{(i)}y_b^{(i)}$$

belong to the ith copy. Consider $\alpha := k/\gcd(r,k)$ further vertices

$$x'_1, \dots, x'_\alpha$$

and $\beta := r/\gcd(r,k)$ further vertices

$$y'_1, \dots, y'_\beta.$$

Now perform the following changes to the edge set of the graph defined by the union of all parts previously mentioned. It may be clarifying to contemplate Figure 4.4. In the figure the edges to be removed are dashed, while the edges to add are thick lines.

- For all $2 \leq i \leq s$ replace the edges

  $$x_v^{(i)}y_b^{(i)} \text{ and } x_1^{(i-1)}y_1^{(i-1)}$$

  by

  $$x_v^{(i)}y_1^{(i-1)} \text{ and } x_1^{(i-1)}y_b^{(i)}.$$

- Also, remove the edges $x_2^{(i)}y_2^{(i)}$ for all $2 \leq i \leq s$.
- Add the edges

  $$\begin{aligned} &x'_1y_2^{(1)}, x'_1y_2^{(2)}, \dots, x'_1y_2^{(r)}, \\ &x'_2y_2^{(r+1)}, x'_2y_2^{(r+2)}, \dots, x'_2y_2^{(2r)}, \\ &\vdots \\ &x'_\alpha y_2^{(s-r+1)}, \dots, x'_\alpha y_2^{(s)} \end{aligned}$$

and
$$\begin{aligned}
&x_2^{(1)}y'_1, x_2^{(2)}y'_1, \ldots, x_2^{(k)}y'_1,\\
&x_2^{(k+1)}y'_2, x_2^{(k+2)}y'_2, \ldots, x_2^{(2k)}y'_2,\\
&\vdots\\
&x_2^{(s-k+1)}y'_\beta, \ldots, x_2^{(s)}y'_\beta.
\end{aligned}$$

As can be verified, the construction gives a new configuration with parameters

$$\begin{aligned}
(v', b', r, k) &= (sv+\alpha, sb+\beta, r, k)\\
&= (sv + \tfrac{k}{\gcd(r,k)}, sb + \tfrac{r}{\gcd(r,k)}, r, k)\\
&= \left(\tfrac{smk}{\gcd(r,k)} + \tfrac{k}{\gcd(r,k)}, \tfrac{smr}{\gcd(r,k)} + \tfrac{r}{\gcd(r,k)}, r, k\right)\\
&= (\tfrac{(sm+1)k}{\gcd(r,k)}, \tfrac{(sm+1)r}{\gcd(r,k)}, r, k)
\end{aligned}$$

and so $sm+1 \in D_{(r,k)}$. □

From Proposition 4.2.10 we deduce that $D_{(r,k)}$ contains two coprime elements, so that they generate a numerical semigroup and this semigroup is contained in $D_{(r,k)}$. So the complement of $D_{(r,k)}$ in $\mathbb{N}_0$ is finite and $D_{(r,k)}$ is a numerical semigroup.

We have seen in Section 4.2.2 that $D_{(r,2)} = D_{(2,r)}$ is a set of the form

$$\{0, r+1, \rightarrow\}$$

if r is even, and of the form

$$\{0, \frac{r+1}{2}, \rightarrow\},$$

when r is odd. Therefore these sets consist of integers in $\mathbb{N} \cup \{0\}$

Sets of this form satisfy the conditions for being a numerical semigroup and in Section 2.1.10 we saw that they are called ordinary numerical semigroups. Concluding, we obtain the following Theorem 4.2.11.

Theorem 4.2.11. *For every pair of integers $r, k \geq 2$, $D_{(r,k)}$ is a numerical semigroup.*

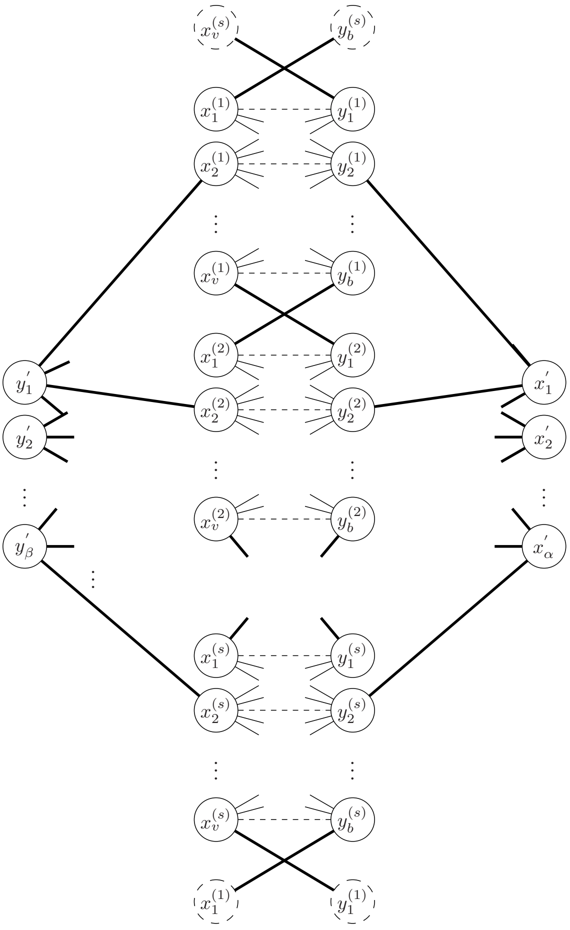

Figure 4.4: The construction of a $(sv+k/\gcd(r,k), sb+r/\gcd(r,k), r, k)$-configuration from the number of s (v,b,r,k)-configurations and $\alpha + \beta = k/\gcd(r,k) + r/\gcd(r,k)$ extra vertices.

So far, all examples that we have seen of numerical semigroups associated to configurable tuples of parameters for combinatorial (r,k)-configurations, have been ordinary. However, there are pairs r,k for which $D_{(r,k)}$ is not ordinary. For example the multiplicity of $D_{(5,5)}$ is 21, but as we saw in Theorem 2.1.65, 22 is a gap of $D_{(5,5)}$. Also, Theorem 2.1.66 says that although the multiplicity of $D_{(6,6)}$ is 31, the integers 32 and 33 do not pertain to $D_{(6,6)}$, so in this case there are two gaps that are larger than the multiplicity.

4.2.5 Bounds on the existence of combinatorial configurations in terms of bounds on the multiplicity and conductor of the associated numerical semigroup

Given the numerical semigroup structure of $D_{(r,k)}$ it is natural to formulate the following questions.

- Which is the smallest non-zero element in $D_{(r,k)}$?
- Since the complement $\mathbb{N} \setminus D_{(r,k)}$ is finite, which is the largest element in the complement?

Remember from Definitions 2.1.81, 2.1.82 and 2.1.80 that the smallest non-zero element of a numerical semigroup is its multiplicity, the smallest element of the semigroup such that all subsequent natural numbers belong to the semigroup is its conductor and the largest element in the complement is called the Fröbenius number. Therefore the conductor is the Fröbenius number plus one.

Lower bounds on the existence of combinatorial configurations

As we saw in Section 2.1.7, given parameters $r, k \geq 2$, a combinatorial (v,b,r,k)-configuration with the smallest possible numbers of points and lines satisfies the necessary conditions

$$v \geq r(k-1)+1$$

and

$$b \geq k(r-1)+1$$

from Theorem 2.1.55. As a consequence we get a lower bound for the multiplicity of the numerical semigroup $D_{(r,k)}$.

Proposition 4.2.12. *Given $r, k \geq 2$, a lower bound for the multiplicity m of the numerical semigroup $D_{(r,k)}$ is*

$$m \geq \max\left(\frac{(r(k-1)+1)\gcd(r,k)}{k}, \frac{(k(r-1)+1)\gcd(r,k)}{r}\right)$$

Proof. Just apply the definition of the integer associated to the parameter tuple of a combinatorial configuration to the bounds for the number of points and lines of a combinatorial (r,k)-configuration from Theorem 2.1.55. □

In the balanced case, that is, when $r = k$, the combinatorial (r,r)-configuration with the smallest number of points and lines is necessarily a finite projective plane, should it exist (see Section 2.1.5). Then we have $\gcd(r,k) = r = k$ so we get the following corollary.

Corollary 4.2.13. *Given $r \geq 2$, a lower bound for the multiplicity m of the numerical semigroup $D_{(r,r)}$ is*

$$m \geq r(r-1)+1.$$

If $r-1$ is a power of a prime, then equality holds.

Upper bounds on the existence of combinatorial configurations

In this section we consider two different upper bounds on the existence of combinatorial configurations. The first bound is an upper bound on the multiplicity of the numerical semigroup $D_{(r,k)}$, hence an upper bound on the size of the smallest existing combinatorial configuration for fixed r and k. The second bound is an upper bound on the conductor of the numerical semigroup $D_{(r,k)}$, hence an upper bound on the parameters v and b such that there exists at least one combinatorial (v,b,r,k)-configuration for all admissible v and b that are larger than this bound. Remember that a tuple (v,b,r,k) is admissible if the necessary conditions of Theorem 2.1.55 are satisfied.

Upper bounds on the existence of combinatorial configurations based on the multiplicity of $D_{(r,k)}$

The lower bound on the multiplicity was deduced from the definition of combinatorial configurations, but upper bounds will rely on their explicit constructions.

The numerical semigroups $D_{(r,2)} = D_{(2,r)}$, as well as the numerical semigroups $D_{(r,3)} = D_{(3,r)}$, were completely described for any $r \geq 2$ in Section 4.2.2 and Section 4.2.3). Therefore we may assume that $r, k \geq 3$ (and even that $r, k \geq 4$).

From Theorem 2.1.54, we know that for any pair of natural numbers $r, k \geq 3$, there exists a combinatorial (r, k)-configuration. Theorem 2.1.54 is constructive; the construction starts with an affine plane of order n such that $n \geq \max(r, k)$ and the result is a combinatorial (v, b, r, k)-configuration with parameters $v = nk$ and $b = nr$. The integer that we associate to such a combinatorial configuration is

$$d = \frac{v \gcd(r, k)}{k} = \frac{nk \gcd(r, k)}{k} = n \gcd(r, k).$$

In Section 2.1.5 we saw that it is conjectured that the order n of an affine plane is always a power of a prime. We deduce the following bound for the multiplicity of the numerical semigroup.

Theorem 4.2.14. *The multiplicity m of the numerical semigroup $D_{(r,k)}$ satisfies*

$$m \leq q \gcd(r, k),$$

where q is the smallest prime power such that $q \geq \max(r, k)$.

Remember that in general there may be more than one affine plane of order n. Therefore, for any pair of integers $r, k \geq 3$ there are combinatorial (v, b, r, k)-configurations; at least one for every integer $d := q \gcd(r, k)$ with $q \geq \max(r, k)$ and q a power of a prime. From this we deduce the following Theorem 4.2.15.

Theorem 4.2.15. *Let $\gcd(r, k) = 1$. Then every prime power $q \geq \max(r, k)$ belongs to $D_{(r,k)}$.*

Upper bounds on the existence of combinatorial configurations based on the conductor of $D_{(r,k)}$

Using our construction of a second element in $D_{(r,k)}$, coprime with the first, it is easy to construct bounds on the conductor using Theorem 2.1.86. As we saw in Theorem 2.1.86, the conductor of the numerical semigroup generated by a and b is $(a - 1)(b - 1)$.

In Theorem 4.2.14 we saw that an upper bound for the multiplicity of $D_{(r,k)}$ was $m \leq q \gcd(r, k)$, for the smallest prime power $q \geq$

$\max(r,k)$, and in Proposition 4.2.10 we saw that the natural number $sm+1$ for $s = rk/\gcd(r,k)$ belongs to $D_{(r,k)}$.

The numerical semigroup generated by the elements m and $sm+1$ has conductor

$$\begin{aligned} c &= (m-1)(sm+1-1) \\ &= (m-1)sm \\ &= (m-1)mrk/\gcd(r,k), \end{aligned}$$

and by replacing m by $q\gcd(r,k)$, we get that the conductor $c_{(r,k)}$ of $D_{(r,k)}$ is bounded by

$$\begin{aligned} c_{(r,k)} &\leq (q\gcd(r,k)-1)q\gcd(r,k)rk/\gcd(r,k) \\ &= (q\gcd(r,k)-1)rkq. \end{aligned}$$

Observe that when r and k are coprime, so that $\gcd(r,k)=1$, then Theorem 4.2.15 tells us that every prime power $q \geq \max(r,k)$ belongs to $D_{(r,k)}$. Since primes are always coprime, in this case Theorem 4.2.10 is not necessary in order to prove that there are at least two coprime elements in $D_{(r,k)}$. Indeed, in this case Theorem 4.2.15 implies that all prime powers larger than $\max(r,k)$ belong to $D_{(r,k)}$, so that the numerical semigroup generated by these prime powers is contained in $D_{(r,k)}$.

When more than two generators of the numerical semigroup are involved, then the calculation of the conductor of a numerical semigroup generated by n elements is difficult [61]. Therefore it is not immediate how to determine the conductor of this numerical semigroup. It is possible that the calculation of a conductor of a numerical semigroup generated by a sequence of successive prime powers is more easy to calculate than the conductor in the general case, but this is an open question.

However, for us the numerical semigroup generated by the prime powers is only a tool to prove that the set $D_{(r,k)}$ is a numerical semigroup and to give an upper bound of its conductor. For this purpose it is enough to use an upper bound of the conductor of the numerical semigroup generated by a sequence of successive prime powers. We provide the following upper bound for this conductor.

Theorem 4.2.16. *Let c be the conductor of a numerical semigroup that contains all prime powers larger than or equal to a given integer n. Then this* *conductor satisfies*

$$c \leq 2 \prod_{p\ prime,\ p<n} (\lfloor \log_p(n-1) \rfloor + 1),$$

and also

$$c \leq \prod_{p\ prime,\ p<n} p^{(\lfloor \log_p(n-1) \rfloor)} + 1.$$

Proof. As we saw in Section 2.1.10, the genus of a numerical semigroup is the number of gaps of the numerical semigroup. In Theorem 2.1.87 we saw that the conductor of a numerical semigroup is smaller or equal to two times the genus.

Suppose that Λ is a numerical semigroup that contains all prime powers larger than or equal to a given integer n. We want to estimate the genus of Λ. Then any gap x can be expressed as a product

$$x = p_1^{n_1} \cdots p_k^{n_k}$$

with n_i integers such that $1 \leq n_i \leq \log_{p_i}(n-1)$ for all i. In particular $p_1, \ldots, p_k$ are prime numbers smaller than n.

Indeed, decompose x as a product of powers of different primes $x = p_1^{n_1} \cdots p_k^{n_k}$. If $n_i > \log_{p_i}(n-1)$ for some i then $p_i^{n_i}$ is a prime power larger than or equal to n and so it belongs to Λ and so does any multiple of it, like x.

Therefore the genus, that is, the number of gaps of Λ, is at most

$$\prod_{p\ prime,\ p<n} (\lfloor \log_{p_i}(n-1) \rfloor + 1),$$

so that the conductor of Λ is at most

$$2 \prod_{p\ prime,\ p<n} (\lfloor \log_{p_i}(n-1) \rfloor + 1).$$

The second inequality is deduced from the fact that the Frobenius number (the largest gap) must be smaller than

$$\prod_{p\ prime,\ p<n} p^{(\lfloor \log_p(n-1) \rfloor)},$$

and the fact that the conductor is the Frobenius number plus one. □

In the particular case of the numerical semigroups $D_{(r,k)}$ associated to the existence of combinatorial (r,k)-configurations with $\gcd(r,k) = 1$, from Theorem 4.2.15 we know that every prime power $q \geq \max(r,k)$ belongs to $D_{(r,k)}$. We therefore deduce the following bound on the conductor of the numerical semigroup $D_{(r,k)}$.

Corollary 4.2.17. *If* $\gcd(r,k) = 1$*, then the conductor* $c_{(r,k)}$ *of the numerical semigroup* $D_{(r,k)}$ *satisfies*

$$c_{(r,k)} \leq 2 \prod_{p\ prime,\ p<\max(r,k)} (\lfloor \log_p(\max(r,k)-1)\rfloor + 1),$$

and also

$$c_{(r,k)} \leq \prod_{p\ prime,\ p<\max(r,k)} p^{(\lfloor \log_p(\max(r,k)-1)\rfloor)} + 1.$$

4.3 The numerical semigroup associated to the existence of triangle-free combinatorial configurations

As we saw in Section 3.3, one problem that the UPIR system could have is that two adversary users connected to a third user through two different communication spaces, could communicate themselves through a third communication space and infer some joint information. This can be avoided by simply avoiding circuits of length 6 in the bipartite incidence graph that represents the combinatorial configuration. Avoiding circuits of length 6 in this graph means avoiding triangles in the configuration. In another context, triangle-free configurations are also called (0,1)-geometries, see Definition 2.1.18 and [24, 75].

Using the existence of regular graphs of girth 8 and any degree [66] we demonstrate in this section the existence of triangle-free (r,k)-configurations for every pair $r,k \geq 2$. Composing triangle-free configurations we deduce that the subset of the natural numbers that is associated to the triangle-free (r,k)-configurations forms a submonoid of the non-negative integers and through constructions of triangle-free combinatorial configurations, analogous to the constructions in Section 4.2, we prove that this submonoid is in fact a numerical semigroup. This will imply, for example, that there exist infinitely many triangle-free (r,k)-configuration for any pair $r,k \geq 2$.

4.3.1 Associating a set of integers to the existence of triangle-free (r, k)-configurations

We saw in Section 4.2.1 that to any tuple of parameters (v, b, r, k) that is admissible for combinatorial configurations, we can associate an integer d.

To remind the reader, this was due to the fact that for any (v, b, r, k)-configuration we have the following expressions for v and b:

$$v = \frac{bk}{r} = d\frac{k}{\gcd(r, k)}$$

and symmetrically

$$b = \frac{vr}{k} = d'\frac{r}{\gcd(r, k)}.$$

Since v and b are integers, so are d and d'. We also have

$$\begin{array}{rl} d & = \frac{v\gcd(r,k)}{k} \\ \\ & = \frac{bk\gcd(r,k)}{rk} \\ \\ & = \frac{b\gcd(r,k)}{r} = d' \end{array}.$$

To any (r, k)-configuration, with or without triangles, we can therefore associate the integer d. On the other hand, given r and k, any $d \in \mathbb{N}$ determines two integers v and b, perhaps corresponding to the number of points and lines of a configuration. For some $d \in \mathbb{N}$ there is no triangle-free combinatorial configuration with parameter set

$$\left(d\frac{k}{\gcd(r, k)}, d\frac{r}{\gcd(r, k)}, r, k\right).$$

Determining for which d there exist triangle-free combinatorial configurations is the problem of existence of triangle-free combinatorial configurations.

In the following we will consider the set of integers such that they may be associated to a triangle-free configuration. The aim is here to prove that this set is a numerical semigroup.

Definition 4.3.1. *For $r, k \in \mathbb{N}$, $r, k \geq 2$ we define*

$$D^{\bigtriangledown}_{(r,k)} := \{d \in \mathbb{N} : \exists \textit{ triangle-free } (v, b, r, k) - \textit{configuration and}$$
$$v = d\tfrac{k}{\gcd(r,k)}, b = d\tfrac{r}{\gcd(r,k)}\}.$$

4.3.2 The set of integers associated to the triangle-free (r, k)-configurations forms a numerical semigroup

In this section we will prove that $D^{\triangledown}_{(r,k)}$ is a numerical semigroup. The proof is similar to the proof we used in Section 4.2 to show that $D_{(r,k)}$ is a numerical semigroup.

Just as in Section 4.2 we will use that it is enough to prove that

- $0 \in D^{\triangledown}_{(r,k)}$,
- $D^{\triangledown}_{(r,k)}$ is closed under addition,
- at least two elements of $D^{\triangledown}_{(r,k)}$ are coprime.

Again, the two first conditions ensure that the subset $D^{\triangledown}_{(r,k)}$ of the natural numbers is a monoid. The operation of the monoid is addition. The last condition ensures that the monoid contains the numerical semigroup generated by the two coprime elements. The complement of this numerical semigroup is finite, therefore also the complement of the monoid, and we deduce that it is a numerical semigroup.

The set of integers associated to the triangle-free (r, k)-configurations is a submonoid of the natural numbers

We first observe that since we consider that the empty set is a triangle-free (r, k)-configuration, we have $0 \in D^{\triangledown}_{(r,k)}$.

We will now prove that the set $D^{\triangledown}_{(r,k)}$ is closed under addition.

Lemma 4.3.2. *If there exist two triangle-free (r, k)-configurations*

$$S_1 = (\mathcal{P}_1, \mathcal{L}_1, I_1)$$

and

$$S_2 = (\mathcal{P}_2, \mathcal{L}_2, I_2)$$

with mutually disjoint point and line sets, then there also exists a triangle-free (r, k)-configuration

$$S_1 \oplus S_2 = (\mathcal{P}_1 \cup \mathcal{P}_2, \mathcal{L}_1 \cup \mathcal{L}_2, I).$$

Proof. First observe that if we use $\emptyset$ to denote the empty triangle-free (r, k)-configuration for any r and k, then, for any triangle-free (r, k)-configuration S, we have, in a natural way,

$$S \oplus \emptyset = S$$

and

$$\emptyset \oplus S = S.$$

Now suppose we have a nonempty triangle-free (r, k)-configuration S_1 with vertices $\mathcal{P}_1 = \{p^1_1, \ldots, p^1_{v_1}\}$ and $\mathcal{L}_1 = \{l^1_1, \ldots, l^1_{b_1}\}$ and another nonempty triangle-free (r, k)-configuration S_2 with vertices $\mathcal{P}_2 = \{p^2_1, \ldots, p^2_{v_2}\}$, $\mathcal{L}_2 = \{l^2_1, \ldots, l^2_{b_2}\}$. Consider the graph with vertices

$$\mathcal{P}_1 \cup \mathcal{L}_1 \cup \mathcal{P}_2 \cup \mathcal{L}_2$$

and edges

$$I_1 \cup I_2.$$

Observe that by definition we have $r, k \geq 2$, so we can assume without loss of generality that

$$(p^1_{v_1}, l^1_{b_1}), (p^2_1, l^2_1) \in I_1 \cup I_2.$$

Replace the relations $(p^1_{v_1}, l^1_{b_1})$ and (p^2_1, l^2_1) by $(p^1_{v_1}, l^2_1)$ and $(p^2_1, l^1_{b_1})$ and consider the resulting incidence relation I. We want to prove that the incidence graph of the incidence structure $(\mathcal{P}_1 \cup \mathcal{P}_2, \mathcal{L}_1 \cup \mathcal{L}_2, I)$ is a connected, bipartite, $(r, k)-$biregular graph of girth at least 8, hence an incidence graph of a triangle-free combinatorial (r, k)-configuration.

But that this graph is connected, bipartite and $(r, k)-$biregular is obvious, so we only need to prove that the girth is at least 8. Now almost all incidence relations in I are the same as in $I_1 \cup I_2$, so the only delicate part of the graph is where the two original graphs were connected, that is, we need to check that the vertices $p^1_{v_1}, p^2_1, l^1_{b_1}, l^2_1$ are not on any cycle of length less than 8.

Now S_1 and S_2 have girth at least 8, so the shortest path between $p^1_{v_1}$ and $l^1_{b_1}$ inside S_1 other than $(p^1_{v_1}, l^1_{b_1})$ (which we have removed) has length at least 7, and the shortest path between p^2_1 and l^2_1 inside S_2 other than (p^2_1, l^2_1) (which we have removed) also has length at least 7. Therefore, in $(\mathcal{P}_1 \cup \mathcal{P}_2, \mathcal{L}_1 \cup \mathcal{L}_2, I)$ the vertices $p^1_{v_1}, p^2_1, l^1_{b_1}, l^2_1$ can not be on a cycle of length less than 8. We get that $(\mathcal{P}_1 \cup \mathcal{P}_2, \mathcal{L}_1 \cup \mathcal{L}_2, I)$ is a triangle-free (r, k)-configuration. □

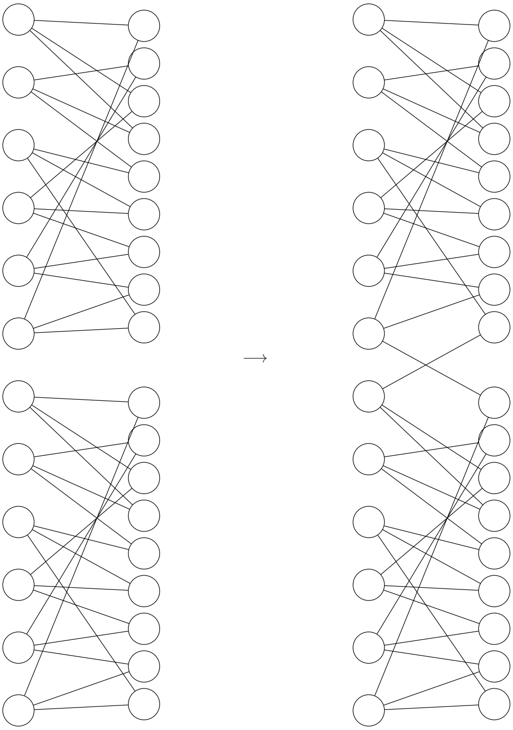

Figure 4.5: Two triangle-free (r, k)-configurations are combined so that the integer associated to the resulting (r, k)-configuration is the sum of the integers associated to the original configurations.

Proposition 4.3.3. *$D^{\bigtriangledown}_{(r,k)}$ is a submonoid of the natural numbers.*

The proof of this proposition is completely analogous to the proof of Proposition 4.2.8, although using Lemma 4.3.2 instead of Lemma 4.2.7 and will therefore not be repeated here.

The submonoid contains two coprime elements

We start by proving that given any pair of natural numbers $r, k \geq 2$, there exists at least one element in $D^{\bigtriangledown}_{(r,k)}$ different from 0. We do this by constructing the incidence graph of a nonempty triangle-free (r, k)-configuration.

For the construction we use a regular graph of girth at least 8.

Theorem 2.1.74 says that for any $n \geq 3$ and $g \geq 2$ there exists an $n-$regular graph of girth g. In particular for any $n \geq 3$ there exists an $n-$regular graph of girth at least 8. We will use one of these graphs to construct a connected, bipartite, $(r, k)-$biregular graph of girth at least 8, defining a triangle-free (r, k)-configuration.

Proposition 4.3.4. *For any pair of integers $r, k \geq 2$, there exists at least one non-zero integer in $D^{\bigtriangledown}_{(r,k)}$.*

Proof. Consider the complete bipartite graph $K_{r,k}$, with edge set E and vertex set V. We consider one spanning tree $T_{r,k}$ of $K_{r,k}$. Then $T_{r,k}$ has the same vertex set V as $K_{r,k}$, but its edge set $E' \subset E$ is smaller. We have

$$|E'| = r + k - 1.$$

The number of edges in $K_{r,k}$ outside $T_{r,k}$, that is, in $E - E'$, is

$$n = rk - r - k + 1 = (r - 1)(k - 1).$$

Suppose $n \geq 3$. (This excludes the cases $(r, k) \in \{(2, 2), (2, 3), (3, 2)\}$, which must be treated separately and will be so, at the end of this proof.)

From Theorem 2.1.74 we know that there exists at least one n-regular graph of girth at least 8. Take one of these graphs and call it G. Associate to each of the vertices of G a copy of the complete bipartite graph $K_{r,k}$. For all edges ab in G, consider its end vertices a and b and let A and B be the copies of $K_{r,k}$ associated to these vertices. Also let T_A and T_B be the corresponding spanning trees in A and B. Now choose one edge $x_A y_A$ in A, but not in T_A and one edge $x_B y_B$ in B, but not in T_B and swap them so that we instead get two edges $x_A y_B$ and $x_B y_A$. Since G is $n-$regular and n is the number of edges in $K_{r,k}$ that are not in its spanning tree, we can choose different edges $x_A y_A$ and $x_B y_B$ for every edge in G.

In this way we get a bipartite, $(r,k)-$biregular graph of girth at least 8, from a $n-$regular graph of girth at least 8, with $n = (r-1)(k-1)$.

The resulting graph may not be connected. If this is the case, we can proceed in two ways.

- We can choose any of the connected subgraphs, and consider that graph to be the incidence graph of the triangle-free configuration we want to construct. If we choose the smallest connected subgraph, then we minimize the size of the smallest known triangle-free (r,k)-configuration proved to exist in this manner;
- We can use the 'addition' law from Lemma 4.3.2 to connect all the connected subgraphs.

In any case we get a connected, bipartite, $(r,k)-$biregular graph of girth at least 8, that is, the incidence graph of a triangle-free (r,k)-configuration.

We still must treat the cases $(r,k) \in \{(2,2),(2,3),(3,2)\}$.

- When $(r,k) = (2,2)$, the connected graph with 8 vertices of degree 2 is a connected, bipartite, $(2,2)$-biregular graph of girth 8, so it is the incidence graph of the smallest nonempty triangle-free $(2,2)$-configuration. It has parameters $d = v = b = 4$;
- When $(r,k) = (2,3)$, the following is an incidence list of a triangle-free $(2,3)$-configuration. We have represented the points as $\mathcal{P} = \{1,\ldots,9\}$ and the lines as $\mathcal{L} = \{A,\ldots,F\}$. Consequently $v = 9$, $b = 6$ and $d = 3$.

A	1	2	9
B	2	3	8
C	3	4	7
D	4	5	1
E	5	6	8
F	6	7	9

- When $(r,k) = (3,2)$, we can consider the dual triangle-free configuration of the previous example.

This concludes the proof. □

Remark 4.3.5. *Observe that since a bipartite graph always has even girth, even if we start with a n-regular graph of girth at least 7, the result will be a graph with girth at least 8. This is interesting if we want the corresponding triangle-free configuration to be as small as possible.*

We will now construct a second element of $D^{\bigtriangledown}_{(r,k)}$, such that the element of Proposition 4.3.4 and the new one are coprime. In order to do so we need the following lemma.

Lemma 4.3.6. *Suppose that $r \geq 3$ and $k \geq 3$. Consider a nonempty triangle-free (r, k)-configuration $(\mathcal{P}, \mathcal{L}, I)$. Then there exist three different points p_1,p_2 and p_3 and three different lines l_1, l_2 and l_3, such that (p_1, l_1), (p_2, l_2) and (p_3, l_3) are in I, but (p_i, l_j) is not in I if $i \neq j$.*

Proof. Since the girth of the incidence graph is at least 8, no cycle of length 7 exists. The graph is connected and has at least 8 edges. It therefore exists a path of length 6 not passing through the same vertex twice. Without loss of generality we may suppose that if $r \geq 3$, then the path starts with a vertex representing a point and ends with a vertex representing a point, and if $r < 3$ but $k \geq 3$, then the path starts with a vertex representing a line and ends with a vertex representing a line. (Remember that the graph is bipartite, with the points on one side and the lines on the other.)

Take the first and the fourth edge of this path. The ends of these edges are separated by paths of length at least two. Also take the seventh (the last) vertex of the path. It is separated from the first and the forth edge by paths of length at least two. If $r \geq 3$, then we have chosen the path so that the seventh vertex represents a point, so it has degree at least 3. If $r < 3$ but $k \geq 3$, then we have chosen the path so that the seventh vertex represents a line, so also in this case it has degree at least 3. Therefore it will have at least two neighbors not in the path. Since the girth of the graph is larger than 4, these two neighbors can not be simultaneously neighbors of the first vertex of the path. Moreover, since the girth is larger than 7, if we choose a vertex, neighbor to the seventh vertex, but not to the first vertex, it will be separated from all first six vertices on the path by paths of at least length 2. We take the edge between the seventh vertex an this vertex. Together with the two edges selected before, they constitute a set of three edges where the ends are all different and ends of different edges are not neighbors.

Consequently we obtain three edges (p_1, l_1), (p_2, l_2) and (p_3, l_3), so that the three points and the three lines are all different and such that $(p_i, l_j) \notin I$ if $i \neq j$. □

We are now ready to prove the existence of two coprime elements of $D^{\bigtriangledown}_{(r,k)}$.

Proposition 4.3.7. *$D^{\bigtriangledown}_{(r,k)}$ contains two elements $m \neq 0$ and $am+1$, with $a \in \mathbb{N}$, so that the two elements are coprime.*

Proof. Consider first the case $(r,k) = (2,2)$. We saw in the proof of Proposition 4.3.4 that the connected graph with 8 vertices of degree 2 is a connected, bipartite, $(2,2)$-biregular graph of girth 8, so it is the incidence graph of the smallest nonempty triangle-free $(2,2)$-configuration. The parameters of this triangle-free configuration were $v = b = d = 4$.

Actually, for any integer $d \geq 4$, the connected graph with $2d$ vertices of degree 2 gives us a triangle-free $(2,2)$-configuration with associated integer $d = v = b$. Therefore we have $D^{\bigtriangledown}_{2,2} = \mathbb{N} \cup \{0\} \setminus \{1,2,3\}$. This proves that $D^{\bigtriangledown}_{2,2}$ is a numerical semigroup and also reveals completely the structure of $D^{\bigtriangledown}_{2,2}$.

Now we may suppose $r \geq 3$ or $k \geq 3$. By Proposition 4.3.4 and since $D^{\bigtriangledown}_{(r,k)} \subseteq \mathbb{N} \cup \{0\}$, there is a minimal non-zero element m in $D^{\bigtriangledown}_{(r,k)}$.

Select a triangle-free (r,k)-configuration S with

$$v = m\tfrac{k}{\gcd(r,k)}$$

and

$$b = m\tfrac{r}{\gcd(r,k)}.$$

Take

$$a = \tfrac{rk}{\gcd(r,k)}$$

copies of S. Let us call the vertices of the ith copy

$$p^{(i)}_1, \ldots, p^{(i)}_v, l^{(i)}_1, \ldots, l^{(i)}_b.$$

By Lemma 4.3.6 we can assume that

$$(p^{(i)}_1, l^{(i)}_1), (p^{(i)}_2, l^{(i)}_2) \text{ and } (p^{(i)}_v, l^{(i)}_b)$$

are edges of the ith copy and that all other combinations

$$(p^{(i)}_a, l^{(i)}_b)$$

with $a \in \{1,2,v\}$ and $b \in \{1,2,b\}$, are not edges of the ith copy.

Consider $\alpha := k/\gcd(r,k)$ further vertices

$$p'_1, \ldots, p'_\alpha$$

and $\beta := r/\gcd(r,k)$ further vertices

$$l'_1, \ldots, l'_\beta.$$

Now perform the following changes to the edge set of the graph defined by the union of all parts previously mentioned. It may be clarifying to contemplate Figure 4.6. In the figure the edges to be removed are dashed, while the edges to add are thick lines.

- "Add" together the a copies of the original configurations. That is, for all $1 \leq i \leq a-1$ replace the edges

  $$(p_v^{(i)}, l_b^{(i)}) \text{ and } (p_1^{(i+1)}, l_1^{(i+1)})$$

  by

  $$(p_v^{(i)}, l_1^{(i+1)}) \text{ and } (p_1^{(i+1)}, l_b^{(i)}).$$

- Also, remove the edges $(p_2^{(i)}, l_2^{(i)})$ for all $1 \leq i \leq a$.

- Add the edges

  $$(p'_1, l_2^{(1)}), (p'_1, l_2^{(2)}), \ldots, (p'_1, l_2^{(r)}),$$
  $$(p'_2, l_2^{(r+1)}), (p'_2, l_2^{(r+2)}), \ldots, (p'_2, l_2^{(2r)}),$$
  $$\vdots$$
  $$(p'_\alpha, l_2^{(a-r+1)}), (p'_\alpha, l_2^{(a-r+2)}), \ldots, (p'_\alpha, l_2^{(a)})$$

  and

  $$(p_2^{(1)}, l'_1), (p_2^{(2)}, l'_1), \ldots, (p_2^{(k)}, l'_1),$$
  $$(p_2^{(k+1)}, l'_2), (p_2^{(k+2)}, l'_2), \ldots, (p_2^{(2k)}, l'_2),$$
  $$\vdots$$
  $$(p_2^{(a-k+1)}, l'_\beta), (p_2^{(a-k+2)}, l'_\beta), \ldots, (p_2^{(a)}, l'_\beta).$$

The constructed graph is connected, bipartite, and $(r, k)-$biregular. The edges $(p_1^{(i)}, l_1^{(i)})$, $(p_2^{(i)}, l_2^{(i)})$ and $(p_v^{(i)}, l_b^{(i)})$ were chosen as permitted by Lemma 4.3.6 and the copies of the original graph are bipartite. Therefore we get that a cycle passing through two different copies has length at least 8. Together with the fact that the girth of the a original copies was at least 8, this implies that the girth of the resulting graph also must be at least 8. So we constructed an incidence graph of a triangle-free (r, k)-configuration, which we may call S'.

We have

$$\begin{aligned} v' = |\mathcal{P}'| &= a|\mathcal{P}| + \alpha \\ &= a|\mathcal{P}| + \tfrac{k}{\gcd(r,k)} \\ &= \tfrac{amk}{\gcd(r,k)} + \tfrac{k}{\gcd(r,k)} \\ &= (am+1)\tfrac{k}{\gcd(r,k)} \end{aligned}$$

and

$$\begin{aligned} b' = |\mathcal{L}'| &= a|\mathcal{L}| + \beta \\ &= a|\mathcal{L}| + \tfrac{r}{\gcd(r,k)} \\ &= \tfrac{amr}{\gcd(r,k)} + \tfrac{r}{\gcd(r,k)} \\ &= (am+1)\tfrac{r}{\gcd(r,k)} \end{aligned}$$

and so $am + 1 \in D^{\triangledown}_{(r,k)}$. □

From Proposition 4.3.7 we deduce that $D^{\triangledown}_{(r,k)}$ contains two coprime elements, so that they generate a numerical semigroup and this semigroup is contained in $D^{\triangledown}_{(r,k)}$. So the complement of $D^{\triangledown}_{(r,k)}$ in $\mathbb{N}_0$ is finite and $D^{\triangledown}_{(r,k)}$ is a numerical semigroup.

4.3.3 Bounds on the existence of triangle-free configurations in terms of bounds on the multiplicity and conductor of the associated numerical semigroup

As in the case with the numerical semigroup associated to the configurable tuples, we will now ask the following questions.

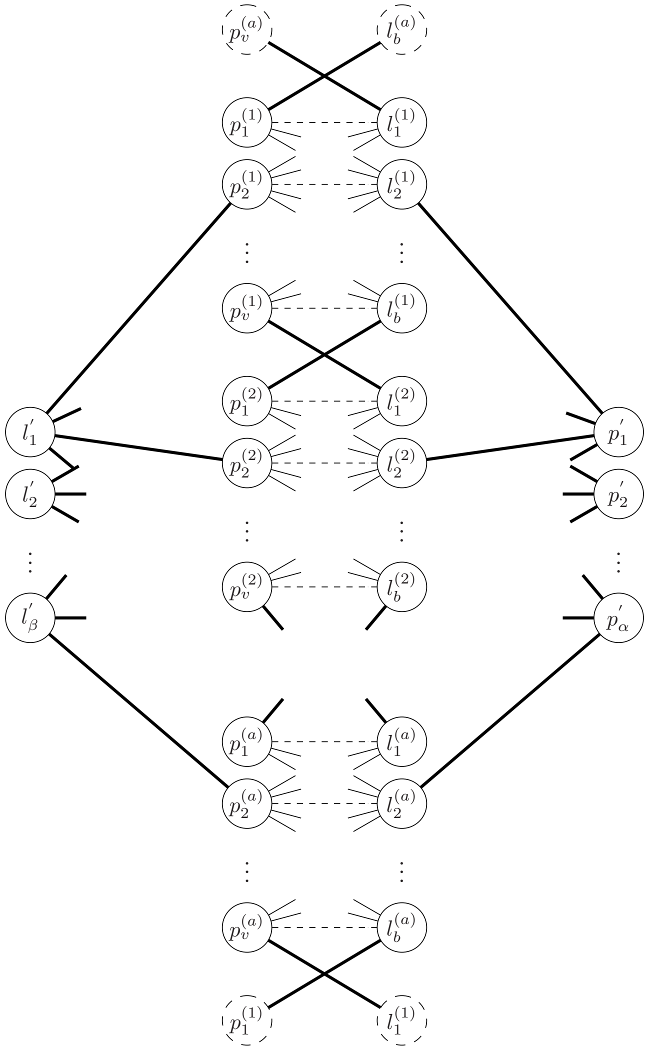

Figure 4.6: Construction of a triangle-free (r, k)-configuration with associated integer $am + 1$ from a smaller triangle-free (r, k)-configuration with associated integer m using $\alpha + \beta$ extra vertices.

- Which is the smallest non-zero element in $D^{\bigtriangledown}_{(r,k)}$?

- Since the complement $\mathbb{N} \setminus D^{\bigtriangledown}_{(r,k)}$ is finite, which is the largest element in the complement?

These smallest elements correspond to the multiplicity and the Frobenius element respectively. The conductor is one plus the Frobenius element.

Lower bounds on the existence of triangle-free configurations

The smallest number of points and lines of a triangle-free (r, k)-configuration is necessarily the number of points and lines of a generalized quadrangle of order $(r-1, k-1)$, should it exist (see Section 2.1.9).

Proposition 4.3.8. *A triangle-free (v, b, r, k)-configuration satisfies*

$$v \geq k((r-1)(k-1)+1)$$

and

$$b \geq r((r-1)(k-1)+1).$$

Proof. Consider a line $l \in \mathcal{L}$ and the two sets

$$A = \{x : x \in \mathcal{P} \text{ and } (x, l) \notin I\}$$

and

$$B = \{x : x \in \mathcal{P} \text{ and } (x, l) \notin I \text{ and } \exists M \in \mathcal{L}, y \in \mathcal{P} \text{ such that } x\, I\, M\, I\, y\, I\, l\}.$$

The number of points not on l is $|A| = |\mathcal{P}| - k$. The number of lines concurrent with l is $k(r-1)$ and these lines have together $|B| = k(r-1)(k-1)$ points which are not their intersection points with l. Obviously $A \supset B$, so

$$|\mathcal{P}| - k = |A| \geq |B| = k(r-1)(k-1),$$

that is,

$$v = |\mathcal{P}| \geq k((r-1)(k-1)+1).$$

Dually

$$b = |\mathcal{L}| \geq r((r-1)(k-1)+1).$$

□

Proposition 4.3.9. *If the bounds in Proposition 4.3.8 are attained, then the triangle-free (v, b, r, k)-configuration is a generalized quadrangle.*

Proof. If the bounds in Proposition 4.3.8 are attained, then the sets A and B have the same cardinality, and since $B \subset A$, they are equal. Therefore the points x not incident with l but connected to l through a pair $(M, y) \in \mathcal{L} \times \mathcal{P}$ are all points of $\mathcal{P}$ except those incident with l. In other words, for every point x not incident with l there is a pair $(M, y) \in \mathcal{L} \times \mathcal{P}$ for which $x\ I\ M\ I\ y\ I\ l$.

On the other hand, a triangle-free combinatorial configuration is a $(0, 1)$-geometry (see Section 2.1.3), so for any point x not incident with l, there can be at most one pair (M, y) such that $x\ I\ M\ I\ y\ I\ l$. Since the existence of a unique pair of such (M, y) is exactly the definition of a generalized quadrangle [58], this proves the statement. □

Remark 4.3.10. *The proof of Proposition 4.3.8 is a simple generalization of the proof of Proposition 2.1.70 as it appears in [58].*

Remark 4.3.11. *Proposition 4.3.8 gives a lower bound on the multiplicity of the numerical semigroup $D^{\bigtriangledown}_{(r,k)}$.*

Upper bounds on the existence of triangle-free configurations

In this section we consider two different upper bounds on the existence of triangle-free configurations.

The first bound is an upper bound on the multiplicity of the numerical semigroup $D^{\bigtriangledown}_{(r,k)}$, hence an upper bound on the size of the smallest existing triangle-free configuration for fixed r and k.

The second bound is an upper bound on the conductor of the numerical semigroup $D^{\bigtriangledown}_{(r,k)}$, hence an upper bound on the size of configuration from which there exists at least one configuration for all admissible sizes which are larger than this bound.

Upper bounds on the existence of triangle-free configurations based on the multiplicity of $D^{\bigtriangledown}_{(r,k)}$

If the lower bound on the multiplicity is deduced from the definition of triangle-free configurations, the upper bound on the other hand relies on their explicit constructions. Expressed in terms of graphs, in order to prove that there always exists a triangle-free (r, k)-configuration it is necessary to prove that for every pair of natural numbers $r, k \geq 2$ there exists a connected, bipartite, (r, k)-biregular graph of girth at least 8. This was proved in Proposition 4.3.4.

In the proof of Proposition 4.3.4 we needed the existence of an r-regular graph of girth 8. For this we used Theorem 2.1.74 due to Sachs. The graphs that Sachs used to prove Theorem 2.1.74 are constructed recursively. As the parameters grow they get large quickly. In order to obtain smaller, general, upper bounds on the multiplicity of $D^{\triangledown}_{(r,k)}$, the n-regular graphs from Propositions 2.1.76, 2.1.77 and 2.1.78 are better suited. From Proposition 2.1.75 we get that there exists an n-regular graph of girth 7 and $2nq^2$ vertices, for a prime power $q \geq n$. Replacing each vertex of this graph with the vertices of the complete, bipartite (r, k)-regular graph on $r + k$ vertices, means multiplying the number of vertices by $r + k$. So the resulting incidence graph has

$$2nq^2(r + k) = 2(r - 1)(k - 1)(r + k)q^2$$

vertices.

We get the following:

Proposition 4.3.12. *For any integers $r, k \geq 2$*

1. *there exists a triangle-free (r, k)-configuration with $2(r - 1)(k - 1)kq^2$ points and $2(r - 1)(k - 1)rq^2$ lines, for $q \geq (r - 1)(k - 1)$ a prime power;*

2. *$D^{\triangledown}_{(r,k)}$ has multiplicity at most $2(r - 1)(k - 1)q^2 \gcd(r, k)$, where q is as before.*

If n is odd, that is, if both r and k are even, then instead of Proposition 2.1.75 we can use the graphs of girth 8 from Proposition 2.1.76 together with the result from Proposition 2.1.78 to deduce the existence of an n-regular graph of girth 7 with

$$2(nq^2 - q) - \frac{2(n - 1)^2 - 2}{n - 2}$$

vertices, so the resulting incidence graph will have

$$\begin{aligned}
&(r + k)\left(2(nq^2 - q) - \tfrac{2(n-1)^2-2}{n-2}\right) \\
&= (r + k)\left(2((r - 1)(k - 1)q^2 - q) - \tfrac{2((r-1)(k-1)-1)^2-2}{(r-1)(k-1)-2}\right),
\end{aligned}$$

for a prime power $q \geq n = (r-1)(k-1)$.

When $n = (r-1)(k-1)$ is a power of a prime Proposition 2.1.77 can be used together with Proposition 2.1.78 to improve further and then if n is odd we get an incidence graph with

$$(r+k)\left(2q(q^2-2) - \frac{2((r-1)(k-1)-1)^2-2}{(r-1)(k-1)-2}\right)$$

vertices.

These results can now be combined with results on the distribution of primes to express the number of points and lines of the constructed configuration as a function of r and k.

Upper bounds on the existence of triangle-free configurations based on the multiplicity of $D^{\triangledown}_{(r,k)}$ for special parameters

When the configuration is balanced, so that $r = k$, and if we suppose that the conjecture that all cages of even girth are bipartite is true [89], then the upper bound on the multiplicity of $D^{\triangledown}_{(r,r)}$ is given by an upper bound on the existence of a $(r, 8)$-cage.

If r is a power of a prime, then Proposition 2.1.77 implies that there exists a triangle-free (r, r)-configuration with

$$v \leq r(r^2-2)$$

$$b \leq r(r^2-2).$$

If r is not a power of a prime, then Proposition 2.1.76 implies that, if q is a power of a prime such that $3 \leq r \leq q-1$, then there exists a triangle-free (r, r)-configuration with

$$v \leq rq^2 - q$$

$$b \leq rq^2 - q.$$

Whenever a generalized quadrangle exists, it is the smallest triangle-free (r, k)-configuration that exists. Then the bound in Proposition 4.3.8 is reached:

$$v = k((r-1)(k-1)+1)$$

and

$$b = r((r-1)(k-1)+1).$$

For example, when $r = k$ and $r-1$ is a power of a prime, it is proved that there is a generalized quadrangle of order $(r-1, r-1)$, with

$$v = b = r((r-1)^2 + 1)$$

(see [58] or Proposition 2.1.71 and 2.1.70).

We also repeat the results stated in Proposition 2.1.71: Let q be a power of a prime. Then there exists a generalized quadrangle of order $(r-1, k-1)$ if

$$(r-1, k-1) \in \{(q,1), (q,q), (q,q^2), (q^2,q^3), (q-1,q+1)\}.$$

The Cremona-Richmond configuration is a famous combinatorial configurations, with parameters $(15_3, 15_3)$ is an example of a smallest triangle-free combinatorial configuration for its parameters, so the multiplicity of $D^{\triangledown}_{(3,3)}$ is 15.

Upper bounds on the existence of triangle-free configurations based on the conductor of $D^{\triangledown}_{(r,k)}$

Using our construction of a second element in $D^{\triangledown}_{(r,k)}$, coprime with the first, it is easy to construct bounds on the conductor using Theorem 2.1.86, which tells us that the conductor of the numerical semigroup generated by the a and b has conductor $(a-1)(b-1)$.

When more than two generators of the numerical semigroup are involved, then the calculation of the conductor of a numerical semigroup generated by n elements is difficult [61].

Regarding the case $r = k = 4$, as we saw in Section 2.1.9, we have that $40, 60, 120 \in D^{\triangledown}_{(4,4)}$ and applying the two constructions from the proof of Theorem 2.1.73 together with the addition of the numerical semigroup it can be calculated that the numerical semigroup generated by these elements has conductor 411. More specifically:

$$
\begin{array}{ll}
\{ & 40, 60, 79, 80, 81, 99, 100, 101, 118, 119, 120, 121, 122, \\
 & 138, 139, 140, 141, 142, 157, 158, 159, 160, 161, 162, 163, \\
 & 177, 178, 179, 180, 181, 182, 183, 196, 197, 198, 199, 200, \\
 & 201, 202, 203, 204, 216, 217, 218, 219, 220, 221, 222, 223, \\
 & 224, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, \\
 & 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 274, \\
 & 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, \\
 & 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, \\
 & 306, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, \\
 & 324, 325, 326, 327, 333, 334, 335, 336, 337, 338, 339, 340, \\
 & 341, 342, 343, 344, 345, 346, 347, 352, 353, 354, 355, 356, \\
 & 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, \\
 & 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, \\
 & 384, 385, 386, 387, 388, 391, 392, 393, 394, 395, 396, 397, \\
 & 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, \\
 & 411, \rightarrow\} \subset D^{\bigtriangledown}_{(4,4)},
\end{array}
$$

and all natural numbers larger than 411 are also contained in $D^{\bigtriangledown}_{(4,4)}$.

Chapter 5

More on the numerical semigroup associated to the existence of combinatorial configurations

This chapter collects various observations regarding the numerical semigroup associated to the configurable tuples $D_{(r,k)}$.

5.1 Another necessary condition for the existence of combinatorial configurations

Consider a (v, b, r, k)-configuration and choose two points p and q that are on a line l. The number of points that are collinear with p but not on l is $(r-1)(k-1)$ and the number of points that are collinear with q but not on l is $(r-1)(k-1)$. We have then counted at most $(r-1)^2$ points twice, one for every intersection of a line through p and a line through q. Adding the k points on the line l we obtain that

$$v \geq 2(r-1)(k-1) - (r-1)^2 + k.$$

Dually we obtain that

$$b \geq 2(k-1)(r-1) - (k-1)^2 + r.$$

Rewriting these expressions we get the following result:

Proposition 5.1.1. *In a (v, b, r, k)-configuration we always have that*

$$v \geq r(k-1) + 1 + (k-r)(r-1)$$

and

$$b \geq k(r-1) + 1 + (r-k)(k-1).$$

However, as we will see in Proposition 5.1.3 the inequalities from Theorem 2.1.55 are stronger than the inequalities in Proposition 5.1.1.

We have seen that in a combinatorial configuration necessarily $vr = bk$ and that the number of points v and the number of lines b therefore are given by:

$$v = \frac{bk}{r} = d\frac{k}{\gcd(r,k)}$$

and symmetrically

$$b = \frac{vr}{k} = d'\frac{r}{\gcd(r,k)}.$$

Since v and b are integers, so are d and d'. Also

$$\begin{array}{rl} d & = \frac{v\gcd(r,k)}{k} \\ & = \frac{bk\gcd(r,k)}{rk} \\ & = \frac{b\gcd(r,k)}{r} = d' \end{array}.$$

We have therefore associated the integer d to the configuration. If one prefers, one can also express this integer as

$$d = \frac{vr}{\mathrm{lcm}(r,k)} = \frac{bk}{\mathrm{lcm}(r,k)}.$$

Observe that using $v = \frac{dk}{\gcd(r,k)}$ and $b = \frac{dr}{\gcd(r,k)}$ the inequalities from Theorem 2.1.55 can be written as in the following Corollary.

Corollary 5.1.2. *In a combinatorial (v, b, r, k)-configuration we always have*

$$r \leq \frac{\frac{dk}{\gcd(r,k)} - 1}{k-1}$$

and

$$k \leq \frac{\frac{dr}{\gcd(r,k)} - 1}{r-1},$$

The same inequalities can be rewritten as

$$r(k-1)+1-k=(r-1)(k-1)\leq k\left(\frac{d}{\gcd(r,k)}-1\right)$$

and

$$k(r-1)+1-r=(r-1)(k-1)\leq r\left(\frac{d}{\gcd(r,k)}-1\right),$$

and combining these last two we get

$$(r-1)(k-1)\leq \min(r,k)\left(\frac{d}{\gcd(r,k)}-1\right). \tag{5.1}$$

We can also rewrite the inequalities from Proposition 5.1.1 in an analogous way. Add $(r-1)(k-1)-k$ to both sides of the first inequality and
$(k-r)(r-1)-r$ to both sides of the second inequality. We then obtain

$$\begin{aligned}
&r(k-1)+1+(k-r)(r-1)+(r-k)(k-1)-k\\
&=(r-1)(k-1)+(k-r)((r-1)-(k-1))\\
&=(r-1)(k-1)+(k-r)(r-k)\\
&\leq \tfrac{dk}{\gcd(r,k)}+(r-k)(k-1)-k\\
&=k\left(\tfrac{d}{\gcd(r,k)}+(r-k)\tfrac{k-1}{k}-1\right)\\
&\leq k\left(\tfrac{d}{\gcd(r,k)}+(r-k)-1\right),
\end{aligned}$$

and dually

$$(r-1)(k-1)+(k-r)(r-k)\leq r\left(\frac{d}{\gcd(r,k)}+(k-r)-1\right)$$

in other words

$$\begin{aligned}
&(r-1)(k-1)+(k-r)(r-k)\leq\\
&\qquad\min\left(k\left(\tfrac{d}{\gcd(r,k)}+(r-k)-1\right),r\left(\tfrac{d}{\gcd(r,k)}+(k-r)-1\right)\right).
\end{aligned}$$

We will now compare the two inequalities.

Proposition 5.1.3. *The inequalities from Theorem 2.1.55:*

$$r(k-1)+1 \leq v \tag{5.2}$$

and

$$k(r-1)+1 \leq b \tag{5.3}$$

are always sharper together than the inequalities from Proposition 5.1.1:

$$r(k-1)+1+(k-r)(r-1) \leq v \tag{5.4}$$

and

$$k(r-1)+1+(r-k)(k-1) \leq b \tag{5.5}$$

are together.

Proof. First we observe that when $r = k$, then the two pairs of inequalities are equivalent. Because of the symmetry of the problem we may therefore assume that $k > r$. The function $f(x) = \frac{x}{x-1} = \frac{1}{1-1/x}$ satisfies $f(2) = 2$ and as x grows the function decreases towards 1. Therefore $\frac{r}{r-1} > \frac{k}{k-1}$ or equivalently $r(k-1) > k(r-1)$. Of the two inequalities (5.2) and (5.3) it is therefore the first one which has the largest lefthand expression. Proposition 5.1.1 bounds v by inequality 5.4. One could think that since $k > r$ inequality (5.4) would be sharper than inequality (5.2). But $vr = bk$ so that $v = bk/r$ and using inequality (5.3) we get

$$(k(r-1)+1)\frac{k}{r} \leq v. \tag{5.6}$$

Since $k > r$, $\frac{k}{r}$ is larger than 1 and can indeed get very large. Write (5.4) as

$$k(r-1)+1+r(k-r) \leq v.$$

Suppose that (5.4) is sharper than the inequalities (5.2) and (5.3) for some r and k so that

$$(k(r-1)+1)\frac{k}{r} < k(r-1)+1+r(k-r),$$

implying

$$(k(r-1)+1)(\frac{k}{r}-1) < r(k-r).$$

Since $\frac{k}{r} > 1$ we can divide by $\frac{k}{r} - 1$ and we get

$$k(r-1)+1 < \frac{r(k-r)}{\frac{k}{r}-1} = \frac{r^2(k-r)}{k-r} = r^2. \tag{5.7}$$

Again, we have that $k > r$, so that

$$k(r-1)+1 \geq (r+1)(r-1)+1 = r^2. \tag{5.8}$$

But there are no r and k such that both (5.7) and (5.8) are satisfied simultaneously. We have therefore proved that (5.2) and (5.3) combined are always sharper than (5.4).

Finally, combining $v = bk/r$ and (5.5) as we did with (5.3) obtaining (5.6) gives

$$(k(r-1)+1+(r-k)(k-1))\frac{k}{r} \leq v.$$

But we have assumed $k > r$, so the term $(r-k)(k-1)$ is negative, implying that

$$(k(r-1)+1+(r-k)(k-1))\frac{k}{r} < (k(r-1)+1)\frac{k}{r}$$

so that it is always sharper to use the rightmost expression, that is, the bound in (5.3). □

5.2 Small configurations and their numerical semigroups

In the previous Section 4.2 we proved that the set of integers associated to the (r,k)-configurable tuples form a numerical semigroup. We have seen examples of these numerical semigroups for small parameters, that is, for $r \leq 3$ or $k \leq 3$. We do not know what numerical semigroups appear as associated to (r,k)-configurations for some pair of integers $r, k \geq 4$. In general, given a pair of integers $r, k \geq 2$ we ask for the natural numbers $d \in \mathbb{N}$ that belong to the numerical semigroup $D_{(r,k)}$. This question is equivalent to the existence problem for combinatorial (r,k)-configurations.

Fixing r and k, the first necessary conditions from Theorem 2.1.55 restrict which integers d can appear as an integer attached to an (r,k)-configuration. Since the set of these integers form the numerical semigroup $D_{(r,k)}$, it is an infinite set, and the restrictions take the form of a bound for the smallest non-zero element in this set, that is, the multiplicity of the numerical semigroup. If we instead fix the integer d, the same inequalities give restrictions on the pairs of integers (r, k) such that $d \in D_{(r,k)}$.

Definition 5.2.1. *For $d \in \mathbb{N}$ we denote*

$$R_d = \{(r, k) \in \mathbb{N}^2_{\geq 2} : r(k-1)+1 \leq \frac{dk}{\gcd(r, k)} \text{ and } k(r-1)+1 \leq \frac{dr}{\gcd(r, k)}\}.$$

In the appendix the reader can find diagrams that show admissible pairs of integers (r, k) for some small integers d. For example, for $d = 2$ we have

$$R_2 = \{(2, 3), (3, 2)\}.$$

For $d = 3$ the admissible pairs of natural numbers (r, k) are

$$R_3 = \{(2, 2), (2, 3), (3, 2), (2, 5), (5, 2), (3, 4), (4, 3)\}.$$

In general, the set R_d has the following properties.

Proposition 5.2.2. 1. *The set R_d is finite.*

2. *The set R_d is symmetric, in the sense that if $(r, k) \in R_d$ then $(k, r) \in R_d$.*

3. *If $d < d'$ then $R_d \subseteq R_{d'}$.*

4. *The l_1 norm of a point $P = (r, k) \in R_d$ satisfies $l_1(P) = |r|+|k| = r+k \leq 2d+2$ and when $r \neq k$ we have $l_1(P) = |r|+|k| = r+k \leq 2d+1$.*

Proof. 1. This can (for example) be deduced from point 4 of this same proposition.

2. This follows from the symmetry of the definition of R_d.

3. We have that if $d' > d$ then

$$\frac{d'k}{\gcd(r, k)} > \frac{dk}{\gcd(r, k)} \geq r(k-1)+1$$

and

$$\frac{d'r}{\gcd(r, k)} > \frac{dr}{\gcd(r, k)} \geq k(r-1)+1$$

so that if the inequalities with d are satisfied, so are the inequalities with d'. The statement follows.

4. Fix d and $g = \gcd(r, k)$. Because of the symmetry of the problem we may suppose that $k \geq r$. From Theorem 2.1.55 we have

$$k(r-1)+1 \leq b = \frac{dr}{g},$$

or the equivalent expression from Corollary 5.1.2

$$k \leq \frac{\frac{dr}{g}-1}{r-1}.$$

Consider the upper bound of k expressed as a real function in r

$$K = K(r) = \frac{\frac{dr}{g}-1}{r-1}.$$

We have

$$\frac{\partial K(r)}{\partial r} = \frac{\frac{d}{g}}{r-1} - \frac{\frac{dr}{g}-1}{(r-1)^2} = \frac{\frac{d}{g}(r-1)-(\frac{d}{g}r-1)}{(r-1)^2} = \frac{1-\frac{d}{g}}{(r-1)^2}.$$

Using Theorem 2.1.55 again we get

$$\begin{aligned} \tfrac{d}{g} &\geq \tfrac{k(r-1)+1}{r} \\ &= \tfrac{kr-k+1}{r} \\ &= k - \tfrac{k-1}{r} \\ &= 1+(k-1)-\tfrac{k-1}{r} \\ &= 1+\tfrac{r(k-1)-(k-1)}{r} \\ &= 1+\tfrac{(r-1)(k-1)}{r} > 1. \end{aligned}$$

In the cases that we consider, $\frac{\partial K}{\partial r}$ is therefore negative and strictly increasing so that $K(r)$ is a convex function. Consider two points on the graph of K: $P_1 = (2, K(2))$ and $P_2 = (r_2, K(r_2))$ with r_2 such that $r_2 = K(r_2)$. P_1 is the intersection point of the graph of K with the line $r = 2$ and P_2 is the intersection point of the graph of K with the line $r = k$. Note that $l_1(P_1) = 2 + K(2) = 2\frac{d}{g}+1$. Since K is a convex function, its graph drawn over the interval $[2, r_2]$

will be situated below the straight line $\overline{P_1P_2}$ drawn between the end points P_1 and P_2 of the graph over the interval, giving an upper bound on $K(r)$ and hence on $r + K(r)$. We will now see that we can bound the line $\overline{P_1P_2}$ from above with the line

$$\left\{P = (r, k) \in \mathbb{R} \times \mathbb{R} : l_1(P) = |r| + |k| = 2\frac{d}{g} + 2\right\}, \tag{5.9}$$

giving the upper bound

$$r + K(r) \leq 2\frac{d}{g} + 2.$$

Consider the points $P_3 = (2, 2\frac{d}{g})$ and $P_4 = (\frac{d}{g} + 1, \frac{d}{g} + 1)$. We have that both P_1 and P_3 are on the line $r = 2$, but P_3 is farther away from origo than P_1. Both P_2 and P_4 are points on the line $r = k$. We will see that either $P_2 = P_4$ or P_4 is farther away from origo than P_2. We have

$$\frac{\frac{dr_2}{g} - 1}{r_2 - 1} = K(r_2) = r_2,$$

$$r_2^2 - (\tfrac{d}{g} + 1)r_2 + 1 = 0,$$

$$r_2 = \tfrac{d}{g} + 1 - \tfrac{1}{r_2},$$

so that

$$r_2 \leq \frac{d}{g} + 1, \tag{5.10}$$

so that

$$|r_2| + |K(r_2)| = 2r_2 \leq 2\frac{d}{g} + 2. \tag{5.11}$$

Therefore either $P_2 = P_4$ or P_4 is farther away from origo than P_2. We deduce that the straight line drawn between the two points P_1 and P_2 on the graph of $K(r)$ is situated below the straight line (5.9) drawn between the points P_3 and P_4. Since $K(r)$ is the bound for admissible parameters of combinatorial configurations, we deduce that the parameters for any combinatorial configuration with $2 \leq r \leq k$ satisfy

$$|r| + |k| \leq 2\frac{d}{g} + 2.$$

By symmetry this is also true for any combinatorial configuration with $2 \leq k \leq r$, that is, for the parameters of any combinatorial configuration.

Now consider the point $P_5 = (\frac{d}{g} + \frac{1}{2}, \frac{d}{g} + \frac{1}{2})$. Then P_2, P_4 and P_5 are points on the line $r = k$. In a combinatorial configuration we always assume $r \geq 2$ so that $-\frac{1}{r} \geq -\frac{1}{2}$ and in particular $-\frac{1}{r_2} \geq -\frac{1}{2}$. Therefore (5.10) gives

$$r_2 \geq \frac{d}{g} + \frac{1}{2},$$

implying that either $P_2 = P_5$ or that P_5 is closer to origo than P_2. We have that P_2 is a point on $r = k$ between the points P_4 and P_5 and that

$$l_2(P_2) = |r_2| + |K(r_2)| = 2r_2 \geq 2\frac{d}{g} + 1,$$

so we can not use $k \leq K(r)$ to ensure that the parameters of any combinatorial configuration satisfy

$$|r| + |k| \leq 2\frac{d}{g} + 1. \tag{5.12}$$

However, as we will now see, (5.12) is valid whenever $r \neq k$. Let $P_6 = (r_6, K(r_6))$ be the intersection point of the graph of $K(r)$ with the line $k = r + 1$, so that r_6 is such that $K(r_6) = r_6 + 1$. As before, because of the convexity of $K(r)$ we know that, if drawn over any interval, the graph of $K(r)$ will always be situated below the straight line drawn between the two end points of the drawn graph. Therefore, over the interval $[2, r_6]$ the graph of $K(r)$ will be situated below the straight line drawn between the points P_1 and P_6.

The point r_6 is the solution to the equation

$$K(r_6) = \frac{\frac{dr_6}{g} - 1}{r_6 - 1} = r_6 + 1,$$

in other words

$$\frac{dr_6}{g} - 1 = (r_6 - 1)(r_6 + 1) = r_6^2 - 1$$

so that

$$\frac{dr_6}{g} = r_6^2$$

or, since $r_6 \neq 0$,

$$\frac{d}{g} = r_6.$$

The point P_6 is therefore

$$P_6 = (r_6, K(r_6)) = (r_6, r_6 + 1) = \left(\frac{d}{g}, \frac{d}{g} + 1\right)$$

so that the l_1-norm of P_6 is

$$|r_6| + |K(r_6)| = r_6 + K(r_6) = 2\frac{d}{g} + 1.$$

Indeed the line between P_1 and P_6 is the line defined by

$$\left\{P = (r, k) \in \mathbb{R} \times \mathbb{R} : l_1(P) = |r| + |k| = 2\frac{d}{g} + 1\right\}.$$

Since $K(r)$ is the bound for admissible parameters of combinatorial configurations, we deduce that the parameters for any combinatorial configuration with $2 \leq r \leq k - 1$ satisfy

$$|r| + |k| \leq 2\frac{d}{g} + 1.$$

By symmetry, this is also true for any combinatorial configuration with $2 \leq k \leq r - 1$, so that it is true for any combinatorial configuration with $r \neq k$.

□

Corollary 5.2.3. *The parameters r and k of a combinatorial (r, k)-configuration with associated integer d satisfy*

$$r + k \leq 2d + 1.$$

Proof. This can be seen by observing that the only case in which the first of the inequalities

$$r + k \leq 2\frac{d}{\gcd(r, k)} + 1 \leq 2d + 1$$

fails, is when $r = k$. But when $r = k$, then $\gcd(r, k) = r = k$ so we can use the inequality

$$r + k \leq 2\frac{d}{\gcd(r, k)} + 2,$$

which is valid for all parameters of combinatorial configurations, to see that since $d \geq 2$ and $\gcd(r, k) \geq 2$, we have that

$$r + k \leq 2\frac{d}{\gcd(r, k)} + 2 \leq \frac{2d}{2} + 2 = d + 2 \leq d + d + 1 = 2d + 1.$$

□

Observe that the fact that $(r, k) \in R_d$ does not imply that $D_{(r,k)}$ actually contains d. For example, define $X = \{x \in \mathbb{N} : x \geq 2\}$, then

$$\begin{aligned} R_{43} = \{(r, k) \geq \mathbb{N}^2 : \quad & r(k-1) + 1 \geq \tfrac{43k}{\gcd(r,k)}, \\ & k(r-1) + 1 \geq \tfrac{43r}{\gcd(r,k)} \text{ and } r, k \geq 2\}, \end{aligned}$$

which means that $(7, 7) \in R_{43}$, but if 43 was in $D_{(7,7)}$, then there would be a $(43, 43, 7, 7)$-configuration and this configuration would be a finite projective plane of order 6. But there is no finite projective plane of order 6, so 43 can not be in $D_{(7,7)}$. This fact is a consequence of Theorem 2.1.63.

Since the finite set R_d is small when d is small, it is a possible and interesting task to list all numerical semigroups $D_{(r,k)}$ containing d. This will be done in this section.

5.2.1 Numerical semigroups $D_{r,k}$ with the integer 2

Since $R_2 = \{(2, 3), (3, 2)\}$ we only have to analyze $D_{2,3}$ and $D_{3,2}$. Because of duality, we always have $D_{(r,k)} = D_{(k,r)}$, so it is enough if we analyze one of these two. But we saw in Section 4.2.2 that $(3, 2)$-configurations exist for all $d \geq 2$. Indeed, k is the number of points on the lines and if $k = 2$ then the configurations are graphs and the lines are edges in the graph, see Section 2.1.2. As a consequence of Lemma 4.2.4, there is a 3-regular graph on $2d$ vertices for every integer $d \geq 2$, that is, for every integer $d \geq 2$ there is a $(2d, 3d, 3, 2)$-configuration.

The smallest $(3, 2)$-configuration is the affine plane $\mathbb{A}(\mathbb{F}_2)$, which is the complete graph on 4 vertices and it consequently has 6 edges. It is the unique $(4, 6, 3, 2)$-configuration and $d = \frac{v \gcd(3,2)}{k} = \frac{4}{2} = 2$ or equivalently $d = \frac{b \gcd(r,k)}{r} = \frac{6}{3} = 2$. One should note that the fact that $R_d = \{(2, 3), (3, 2)\}$ implies that the only configurations with associated integer 2 are $\mathbb{A}(\mathbb{F}_2)$ and its dual configuration, which consists in the 6 points $1, 2, 3, 4, 5, 6$ on the 4 lines a, b, c, d of linesize $k = 3$ listed here below.

a	b	c	d
1	1	2	3
2	4	4	5
3	5	6	6

There is therefore only one numerical semigroup $D_{(r,k)}$ containing 2, namely

$$D_{(2,3)} = D_{(3,2)} = \langle 2, 3\rangle = \{0, 2, 3, 4, 5, \rightarrow\},$$

which is the numerical semigroup with multiplicity 2 and conductor 2. Hence none of the numerical semigroups with multiplicity 2 and conductor larger than 2 can appear as the numerical semigroup attached to the existence of (r, k)-configurations.

5.2.2 Numerical semigroups $D_{r,k}$ with the integer 3

We have

$$R_3 = \{(2, 2), (2, 3), (3, 2), (2, 5), (5, 2), (3, 4), (4, 3)\}.$$

We know that $D_{(r,k)} = D_{(k,r)}$, so we only have to analyze one of the pairs (r, k) and (k, r). From Section 5.2.1 we know that

$$D_{(2,3)} = D_{(3,2)} = \langle 2, 3\rangle = \{0, 2, 3, 4, 5 \rightarrow\}.$$

Again, we know that whenever $k = 2$, then the configurations are graphs, so that $D_{(2,2)}$ and $D_{(2,5)} = D_{(5,2)}$ represent the existence of 2-regular and 5-regular graphs, respectively. According to Corollary 4.2.3,

$$D_{(2,2)} = \langle 3, 4, 5\rangle = \{0, 3, 4, 5, 6, \rightarrow\},$$

and Corollary 4.2.5 gives

$$D_{(5,2)} = D_{(2,5)} = \langle 3, 4, 5\rangle = \{0, 3, 4, 5, 6, \rightarrow\}.$$

Observe that we have $D_{(5,2)} = D_{(2,5)} = D_{(2,2)}$, which means that the same numerical semigroup can appear as $D_{(r,k)}$ for distinct parameter pairs (r, k).

Now the only case left is $D_{(3,4)} = D_{(4,3)}$. Since $k > 2$, neither the $(3,4)$-configurations nor the $(4,3)$-configurations are graphs. $(3,4)$-configurations exists for $3d = v \geq r(k-1)+1 = 9$, that is, for $d \geq 3$. For $d = 3$ the $(12,9,3,4)$-configuration is given by the following table,

a	b	c	d	e	f	g	h	i
1	1	1	11	11	11	12	12	12
2	7	10	2	7	10	2	7	10 .
3	6	9	9	3	6	6	9	3
4	5	8	5	8	4	8	4	5

This is the dual of the affine plane of order 3, $\mathbb{A}(\mathbb{F}_3)$, which has 9 points, 12 lines, 3 points on any line and 4 lines through any point. The dual has 12 points, 9 lines, 4 points on any line and 3 lines through any point. For $d = 4$ one can take the affine plane of order 4, $\mathbb{A}(\mathbb{F}_4)$, which has $4^2 = 16$ points and $4^2 + 4 = 20$ lines, then choose 3 of the 5 parallel classes of lines. This gives a $(16,12,3,4)$-configuration, with associated integer $d = 4$. For $d = 5$ one can take the affine plane of order 5, $\mathbb{A}(\mathbb{F}_5)$, which has $5^2 = 25$ points and $5^2 + 5 = 30$ lines, then choose a parallel class of lines and take the $4 \cdot 5 = 20$ points on 4 of the lines in this class. An affine plane of order n has $n^2 + n$ lines partitioned into $n+1$ parallel classes. The affine plane of order 5 has its 30 lines partitioned into 6 classes, of which we have already used one class. Take the lines in $r = 3$ of the other 5 parallel classes and consider their restriction to the set of 20 points which we chose before. Then this point set together with these restricted lines form a $(20,15,3,4)$-configuration, with associated integer $d = 5$. To see this, just remember that the affine plane is a configuration, which means that there can be at most one line through every pair of points and observe that, by restricting the lines to the chosen point set, any line will contain $k = 4$ points and, by taking $r = 3$ parallel classes of lines, any point will be on $r = 3$ different lines.

Theorem 4.2.11 says that $D_{(3,4)}$ is a numerical semigroup and now we know that $3, 4, 5 \in D_{(3,4)}$. But the numerical semigroup generated by $3, 4, 5$ contains every integer $d \geq 3$, so there is a $(3,4)$-configuration for all integers $d \geq 3$. Therefore,

$$D_{(3,4)} = D_{(4,3)} = \langle 3, 4, 5 \rangle = \{0, 3, 4, 5, 6, \rightarrow\}.$$

Observe that this means that all the numerical semigroups $D_{(r,k)}$ for $(r,k) \in R_3 \setminus R_2$ are the same. These numerical semigroups is the set of $D_{(r,k)}$ with multiplicity 3.

5.2.3 Numerical semigroups $D_{r,k}$ with the integer 4

We have

$$\begin{aligned} R_4 = \quad & \{(2,2),(2,3),(3,2),(2,5),(5,2),(3,4), \\ & (4,3),(2,7),(7,2),(3,5),(5,3),(4,5),(5,4)\}. \end{aligned}$$

Of these, only the last six elements are in $R_4 \setminus R_3$. Therefore the rest of the elements were treated already in the previous section. Also, just as in the previous section, since $D_{(r,k)} = D_{(k,r)}$, we only have to analyze one of these two. The question is therefore: which numerical semigroups are attached to the (r,k)-configurations with $(r,k) \in \{(7,2),(3,5),(4,5)\}$?

- Again, we know that whenever $k = 2$, then the configurations are graphs, so that $D_{(7,2)}$ represents the existence of 7-regular graphs. Since 7 is an odd integer, Corollary 4.2.5 gives

$$D_{(2,7)} = D_{(7,2)} = \{0, 4, \rightarrow\}.$$

- According to Gropp [40], $(v,b,r,3)$-configurations exists for all admissible parameters. Therefore

$$D_{(3,5)} = D_{(5,3)} = \{0, 4 \rightarrow\}.$$

- Only $D_{(4,5)} = D_{(5,4)}$ remains. The affine plane of order 4 has 4 points on every line, 5 lines through every point, $4^2 = 16$ points and $4^2 + 4 = 20$ lines, and is indeed a $(16,20,5,4)$-configuration with associated integer

$$d = \frac{v \gcd(r,k)}{k} = 4 \in D_{(5,4)}.$$

The dual configuration, with parameters $(20,16,4,5)$ is a configuration with associated integer

$$d = \frac{v \gcd(r,k)}{k} = 4 \in D_{(4,5)}.$$

Following the examples from the previous section we can construct more $(4,5)$-configurations using parallel classes of affine planes. Such configurations are called transversal designs. Indeed we have the following more general result in Theorem 4.2.15. From Theorem 4.2.15 we deduce that $(4,5)$-configurations with associated integer q exist for all prime power $q \geq \max(4,5) = 5$. Indeed

$$5, 7, 8, 9, 11, 13, 16, \ldots \in D_{(4,5)}.$$

The question is now if $6 \in D_{(4,5)}$. There is the following result by Gropp:

Proposition 5.2.4. *[40] There exists a combinatorial $(v, b, r, 4)$-configuration for every $v \equiv 0 \pmod{12}$ whenever the parameters are admissible and v is not in the set*

$$\begin{aligned} E = \{ \quad & 84, 120, 132, 180, 216, 264, 312, 324, \\ & 372, 456, 552, 648, 660, 804, 852, 888\,\}. \end{aligned}$$

Indeed the proposition implies that since $v = 24 \notin E$ we have the existence of a $(24, 30, 5, 4)$-configuration and therefore also a $(30, 24, 4, 5)$- configuration, so that $6 \in D_{(4,5)}$. We have proved that $D_{(4,5)} = D_{(5,4)} = \{0, 4, \rightarrow\}$, and so all the numerical semigroups associated to (r, k)-configurations with $(r, k) \in R_4 \setminus R_3$ are the same, just as was the case for $R_3 \setminus R_2$.

5.2.4 Numerical semigroups $D_{r,k}$ with the integer 5

The region of $\mathbb{N} \times \mathbb{N}$ which contains the pairs (r, k) such that 5 is an admissible associated integer is

$$R_5 = R_4 \cup \{(2, 4), (4, 2), (2, 9), (9, 2), (3, 7), (7, 3), (5, 6)(6, 5)\}.$$

As before, because of the symmetry we only need to study

$$\{(4, 2), (9, 2), (3, 7), (5, 6)\}.$$

Of these the two pairs with $k = 2$ correspond to 4- and 9-regular graphs respectively, which according to Corollary 4.2.3 and Corollary 4.2.5 exist for all admissible integers d so that

$$D_{(2,4)} = D_{(4,2)} = D_{(2,9)} = D_{(9,2)} = \{0, 5, \rightarrow\}.$$

For $D_{(3,7)}$ we can use the dual theorem of Gropp, [40], which says that they exists for all admissible integers d, hence also

$$D_{(3,7)} = \{0, 5, \rightarrow\}.$$

For $D_{(5,6)}$, first observe that there exists an affine plane of order 5. The affine plane of order 5 is a $(25, 30, 6, 5)$-configuration and it has a dual $(30, 25, 5, 6)$-configuration. Both have associated integer $d = 5$ so that $5 \in D_{(5,6)}$. As before, using transversal designs we can obtain many more $(5, 6)$-configurations. Indeed using Theorem 4.2.15 we have that there are $(5, 6)$-configurations with associated integer d for all prime powers d that satisfy $d \geq \max(5, 6) = 6$, so that

$$7, 8, 9, 11, 13, 16, \ldots \in D_{(5,6)}.$$

We have therefore that

$$\langle 5, 7, 8, 9, 11 \rangle \subseteq D_{(5,6)}.$$

This does however not resolve the question if $6 \in D_{(5,6)}$.

5.2.5 Numerical semigroups $D_{r,k}$ with the integer 6

The region of $\mathbb{N} \times \mathbb{N}$ which contains the pairs (r, k) such that 6 is an admissible associated integer is

$$\begin{aligned} R_6 = R_5 \cup \quad & \{(2, 11), (11, 2), (3, 8), (8, 3), (4, 7), \\ & (7, 4), (5, 7), (7, 5), (6, 7), (7, 6)\}. \end{aligned}$$

We observe that this set contains the interesting element $(7, 6)$. A $(7, 6)$-configuration with associated integer 6 would have the parameter set $(36, 42, 7, 6)$ and would therefore be an affine plane of order 6. Since there is no affine plane of order 6 we know that $6 \notin D_{(7,6)} = D_{(6,7)}$ Again we can use transversal designs to construct more $(7, 6)$-configurations. Indeed Theorem 4.2.15 implies that there exists a $(7, 6)$-configuration with associated integer d for all prime powers $d \geq \max(7, 6) = 7$, so that

$$7, 8, 9, 11, 13, \ldots \in D_{(7,6)}.$$

5.2.6 Non-ordinary numerical semigroups associated to (r,k)-configurable tuples

We have investigated the numerical semigroups associated to the (r, k)-configurable tuples that contain small integers. These numerical semigroups all have small multiplicity, or more precisely, the numerical semigroups $D_{(r,k)}$ that contain an integer d all have multiplicity smaller than or equal to d. We have seen that such numerical semigroups are, mostly, ordinary. In general this is however not true.

A major obstacle in the execution of the analysis is the sparse available knowledge when the parameters r and k are large. In this context, large means larger than 5. There are more examples of balanced combinatorial configurations than there are in the non-balanced case. The reason for this is that more research has been made in the balanced case. From what we saw in Section 2.1.8 we can deduce that the first non-ordinary numerical semigroup associated to the (r,r)-configurable tuples, when ordered by size of the multiplicity, is $D_{(5,5)}$. As we saw there, the next example is $D_{(6,6)}$, which has two gaps that are larger than the multiplicity

The numerical semigroups associated to balanced (r,r)-configurable tuples have larger multiplicity than the numerical semigroups associated to non-balanced (r,k)-configurable tuples, as can be observed in the figures of the region R_d in the appendix. There it can be observed that the smaller $\gcd(r,k)$ gets, the smaller is the multiplicity of $D_{(r,k)}$, so that the smallest multiplicities are observed for pairs r,k with $\gcd(r,k)=1$ and the largest multiplicities can be found in the balanced case, when $\gcd(r,k)=r=k$. The multiplicity of $D_{(5,5)}$ is 21, see Theorem 2.1.65, and we do not have access to sufficient material for an analysis of all the numerical semigroups $D_{(r,k)}$ with multiplicity 21.

If there are non-ordinary numerical semigroups associated to (r,k)-configurable tuples with multiplicity smaller than 5, then the tuples are non-balanced ($r \neq k$). The largest number of parameters (r,k) with numerical semigroups $D_{(r,k)}$ of small multiplicity are parameters with $\gcd(r,k)=1$. However, it is still an open question whether there are parameters (r,k) with $\gcd(r,k)=1$ such that the numerical semigroup $D_{(r,k)}$ is non-ordinary.

Chapter 6

Conclusions

6.1 Results

On one hand this thesis has treated some questions about the existence and construction of combinatorial configurations, on the other, it has treated an application of combinatorial configurations to user-private information retrieval or anonymous database search. The results that have been presented can therefore be divided into two categories, mathematical results on combinatorial configurations and applied results in computer science on the P2P UPIR protocol. Some of the results may of course belong to both categories.

P2P UPIR using combinatorial configurations

The content of Chapter 3 was mainly of applied nature and contained analyses of different modalities and scenarios for the execution of P2P UPIR using combinatorial configurations. Three types of combinatorial configurations were recognized as useful for P2P UPIR, because of the properties that define them.

The $(v,k,1)$-BIBD, or with another name, the $S(2,k,v)$ Steiner systems, were recognized as optimal combinatorial configurations for P2P UPIR with respect to the diffusion of the real profile of the protocol user. As a special case of $(v,k,1)$-BIBD, the finite projective planes were identified as the optimal combinatorial configurations for P2P UPIR, with respect to criteria as privacy in front of the server and storage efficiency.

The neighborhood of a point in a combinatorial configuration was recognized as a quasi-identifier of that point and it was explained that users of the P2P UPIR protocol who combine repeated queries with a unique neigborhood are not protected by the protocol. We analyzed the query behaviour of some users in the AOL query log files, and with the additional statistical information from other query logs we concluded that repetition of queries is a common and frequent phenomenon. Since repeated queries are hard to avoid, we therefore justified why it is important to avoid combinatorial configurations that give unique neighborhoods to the users.

The theory of n-anonymity was applied to the neighborhood problem and the use of transversal designs for n-anonymous P2P UPIR was proposed. In general, we characterized the combinatorial configurations that provide n-anonymous P2P UPIR.

As an alternative solution to the neighborhood problem we proposed a modification of the P2P UPIR protocol. The $(v, k, 1)$-BIBD were recognized as optimal combinatorial configurations for this modified P2P UPIR protocol. Indeed they were recognized as exactly the only combinatorial configurations capable of providing complete privacy for any P2P UPIR protocol. We explained how modified P2P UPIR with a $(v, k, 1)$-BIBD can also be understood as v-anonymous modified P2P UPIR. We also defined the concept of n-confusion for P2P UPIR.

We proved that when k divides n, then it is possible to construct a combinatorial (r, k)-configuration that provides k-anonymous modified P2P UPIR from a combinatorial (r, k)-configuration that provides n-anonymous P2P UPIR.

Collusions of adversary protocol users communicating only over the channels provided by the protocol were recognized as a privacy risk. Several users can collude in order to get advantage over the protocol and obtain more information on the query profiles of their neighbors than expected. Analyses of different scenarios of colluding users were provided. Triangle-free combinatorial configurations were proposed in order to avoid the privacy risk caused by collusions of users communicating over channels provided by the protocol. For collusions of users that communicate also over external channels, a calculation of the magnitude of the privacy risk was presented.

Combinatorial configurations and numerical semigroups

In Section 4.2 we associated a subset of the natural numbers $D_{(r,k)}$ to the combinatorial (r,k)-configurations and the following theorem was proved.

Theorem 6.1.1. *For every pair of integers $r, k \geq 2$, $D_{(r,k)}$ is a numerical semigroup.*

Theorem 6.1.1 has several corollaries.

Corollary 6.1.2. *For any pair of integers $r, k \geq 2$ there exist infinitely many combinatorial (r,k)-configurations.*

A numerical semigroup has a conductor, a smallest number from which all larger integers belong to the numerical semigroup. Using this we get Corollary 6.1.3

Corollary 6.1.3. *Given a pair of integers $r, k \geq 2$ there exists a positive number N such that for all integers $n \geq N$ there exists at least one combinatorial configuration with parameters*

$$\left(\left(n\frac{k}{\gcd(r,k)}\right), \left(n\frac{r}{\gcd(r,k)}\right), r, k\right),$$

that is, when the number of points (and lines) is big enough, then there is at least one configuration for any admissible parameters.

The number N in Corollary 6.1.3 is the conductor of the numerical semigroup $D_{(r,k)}$. We have bounded this conductor with the upper bound

$$(q\gcd(r,k)-1)rkq,$$

where q is the smallest prime power such that $q \geq \max(r,k)$.

Using the construction and the combination of combinatorial configurations in the proof of Theorem 6.1.1, it is possible to explicitly construct combinatorial (r,k)-configurations with associated integers in the numerical semigroup generated by the coprime integers m and $am+1$, where m is the associated integer of an existing combinatorial (r,k)-configuration and $a = rk/\gcd(r,k)$. Theorem 2.1.54 gives the example $m = q\gcd(r,k)$ where $q \geq \max(r,k)$ is a prime power, but any other integer m associated to a (r,k)-configuration can be used.

In the special case $\gcd(r,k) = 1$ it was proved that all prime powers q that satisfy $q \geq \max(r,k)$ belong to the numerical semigroup $D_{(r,k)}$

and it was deduced that the conductor of $D_{(r,k)}$ in this case is bounded by

$$2 \prod_{p\ prime,\ p<\max(r,k)} (\lfloor \log_p(\max(r,k)-1) \rfloor + 1).$$

Triangle-free combinatorial configurations

In Section 4.3 we associated a subset of the natural numbers $D^{\bigtriangledown}_{(r,k)}$ to the triangle-free (r,k)-configurations and then we proved the following theorem.

Theorem 6.1.4. *For every pair of integers $r, k \geq 2$, $D^{\bigtriangledown}_{(r,k)}$ is a numerical semigroup.*

In particular, we proved that for every pair of natural numbers r and k, larger than 2, the set $D^{\bigtriangledown}_{(r,k)}$ contains at least one non-zero element m. This integer m corresponds to a triangle-free (v,b,r,k)-configuration with number of points $v = |\mathcal{P}| = m\frac{k}{\gcd(r,k)}$ and number of lines $b = |\mathcal{L}| = m\frac{r}{\gcd(r,k)}$. We get the following result (Proposition 4.3.12).

Corollary 6.1.5. *For any pair of integers $r, k \geq 2$ and a prime power q that satisfies $q \geq (r-1)(k-1)$ there exists a triangle-free (r,k)-configuration with*

$$2(r-1)(k-1)kq^2$$

points and

$$2(r-1)(k-1)rq^2$$

lines.

As we saw in Section 4.3.3, for many cases there are much smaller triangle-free configurations. These results should be compared with the previous bound for the smallest *balanced* triangle-free (r,r)-configuration that was given in Theorem 2.1.72, which uses the generalized Gray / $LC(r)$ configuration with r^r points and r^r lines. Beside the fact that this bound was of exponential size, while our bound is polynomial, our bound is more general, since we also treat unbalanced configurations.

The proof of Corollary 6.1.5 is constructive and can be used as an algorithm to construct a triangle-free (r,k)-configuration. The construction can be found in Proposition 4.3.4. Further constructions are given in Lemma 4.3.2 and Proposition 4.3.7.

From Theorem 6.1.4 we can also deduce the existence of infinite families of triangle-free (r, k)-configurations. These families are different from the families which can be constructed from the results presented in [42], and also in this case it should be noticed that we treat both balanced and unbalanced configurations.

Corollary 6.1.6. *For any pair of integers $r, k \geq 2$ there exist infinitely many triangle-free (r, k)-configurations.*

A numerical semigroup has a conductor, a smallest number from which all larger integers belong to the numerical semigroup. Using this we get the following result.

Corollary 6.1.7. *Given a pair of integers $r, k \geq 2$ there exists a positive number N such that for all integers $n \geq N$ there exists at least one triangle-free configuration with parameters*

$$\left(\left(n\frac{k}{\gcd(r,k)}\right), \left(n\frac{r}{\gcd(r,k)}\right), r, k\right),$$

that is, when the number of points (and lines) is big enough, there is at least one triangle-free configuration for any admissible set of parameters.

Using the construction and the combination of triangle-free combinatorial configurations in the proof of Theorem 6.1.4, it is possible to explicitly construct triangle-free combinatorial (r, k)-configurations with associated integers in the numerical semigroup generated by the coprime integers m and $am + 1$, where m is the associated integer of an existing triangle-free combinatorial (r, k)-configuration (for example the one from Corollary 6.1.5) and $a = rk/\gcd(r, k)$.

We have described an application of configurations to P2P UPIR. In Section 3.3 it is justified that in order to avoid collusions of two users that are spying on a third, configurations without triangles should be used.

This is of course not the only application of triangle-free configurations. Configurations have been used in coding theory, for example in the construction of LDPC codes [34, 35, 57, 82], where a large girth is important. In this context a girth which is at least 8 may be considered to be large. The results on the existence and the explicit constructions of triangle-free configurations presented in this article have therefore applications to coding theory. One should also notice that the arguments used in this thesis in general work for configurations with an incidence graph of girth at least $n \in \mathbb{N}$, for $n \geq 6$.

Another example of an application of configurations is the deterministic key distribution scheme for distributed sensor networks that was described in [53, 54].

For all applications it is obviously useful to know that there are plenty of triangle-free configurations, so that it is possible to find one for any specified parameters. Finding means explicit construction, which we provide. Also, the addition in $D^{\triangledown}_{(r,k)}$ comes from combining two triangle-free configurations, so we can construct larger (r, k)-configurations from smaller ones, which is very useful in many applications. We do not call it addition of configurations, since it is not well-defined. Indeed we can combine the same two configurations in many ways, by choosing different vertices in the combination process.

6.2 Open problems

In the elaboration of this thesis open problems have been formulated. Some of these have been solved, others are still without answering. Some of the problems that are still open are listed below.

- In many combinatorial configurations there is a bijection between the sets $\mathcal{P}$ and $\{\{p\} \cup N(p) : p \in \mathcal{P}\}$ defined by $p \mapsto \{p\} \cup N(p)$. We have seen that this is the case for triangle-free combinatorial configurations and for $(v, k, 1)$-BIBD. In general, we still do not know exactly for which combinatorial configurations this is true.

- We propose as an open problem the calculation of the conductor of a numerical semigroup generated by a sequence of consecutive prime powers.

- The question if there are parameters (r, k) with $r \neq k$ such that the numerical semigroup $D_{(r,k)}$ is non-ordinary is also still open.

- This thesis has provided some contributions to our knowledge on the existence of combinatorial configurations. However, in general the existence problem for combinatorial configurations remains unanswered.

Bibliography

[1] R. J. R. Abel, C.J. Colbourn, J.H. Dinitz (2007) Mutually Orthogonal Latin Squares (MOLS). In the Second Edition of C.J. Colbourn, J.H. Dinitz (Eds.) "The CRC Handbook Of Combinatorial Designs." CRC Press, Boca Raton, FL, pp. 160–193.

[2] R.J.R. Abel and M. Greig (2007) BIBDs with Small Block Size. In the Second Edition of C.J. Colbourn, J.H. Dinitz (Eds.) "The CRC Handbook Of Combinatorial Designs." CRC Press, Boca Raton, FL, pp. 72–79.

[3] C. Aguilar-Melchor and P. Gaborit (2007) *Single-database private information retrieval protocols: Overview, usability and trends.* Research Report.

[4] C. Aguilar-Melchor and Y. Deswarte (2009) *Trustable relays for anonymous communication.* Transactions on Data Privacy, 2:2, pp. 101–130.

[5] AOL query logs, *www.aolstalker.com*.

[6] G. Araujo-Pardo (2010) *On upper bounds of odd girth cages.* Discrete Mathematics, 310:10-11, pp. 1622 – 1626.

[7] C. Balbuena (2009) *A construction of small regular bipartite graphs of girth 8.* Discrete Mathematics & Theoretical Computer Science, 11:2, pp. 33–46.

[8] A. Betten, G. Brinkmann and T. Pisanski (2000) *Counting symmetric configurations v_3.* Discrete Applied Mathematics, 99:1-3, pp. 331–338.

[9] A. Beutelspacher and U. Rosenbaum "Projective Geometry, from Foundations to Applications." Cambridge University Press, Cambridge, 1998.

[10] N. Biggs, "Algebraic Graph Theory." Cambridge University Press, Cambridge, second edition, 1993.

[11] J. G. Bokowski, "Computational Oriented Matroids." Cambridge University Press, Cambridge, 2006.

[12] R. C. Bose, S.S. Shrikhande and E.T. Parker (1960) *Further Results on the Construction of Mutually Orthogonal Latin Squares and the Falsity of Euler's Conjecture.* Canad. J. Math., 12, pp. 189–203.

[13] M. Boben, B. Grünbaum, T. Pisanski and A. Zitnik (2006) *Small triangle-free configurations of points and lines.* Discrete and Computational Geometry, 35, pp. 405–427.

[14] M. Bras-Amorós, J. Domingo-Ferrer and K. Stokes (2009) *Configuraciones combinatóricas y recuperación privada de información por pares.* In Congreso de la Real Sociedad Matemática Española-RSME 2009, Oviedo, Spain.

[15] M. Bras-Amorós and K. Stokes (2011) *The semigroup of combinatorial configurations.* Semigroup Forum, to appear (also available as preprint, arXiv: 0907.4230v3).

[16] M. Bras-Amorós and K. Stokes (2010) *On the existence of combinatorial configurations.* In Proceedings of the 3rd International Workshop on Optimal Networks Topologies (IWONT), 9-11 June 2010, Barcelona, pp. 145–168.

[17] M. Bras-Amorós, K. Stokes and M. Greferath (2010) *Using (0,1)-geometries for collusion-free P2P user private information retrieval.* In Proceedings of The 19th International Symposium on Mathematical Theory of Networks and Systems, Budapest.

[18] R.H. Bruck and H.J. Ryser (1949) *The nonexistence of certain finite projective planes.* Canadian J. Math. 1, pp. 88–93.

[19] Y. Chang and M. Mitzenmacher (2005) *Privacy Preserving Keyword Searches on Remote Encrypted Data.* In Proceedings of Applied Cryptography and Network Security (ACNS 2005), pp. 442–455.

[20] B. Chor, O. Goldreich, E. Kushilevitz and M. Sudan (1995) *Private information retrieval.* In IEEE Symposium on Foundations of Computer Science (FOCS), pp. 41–50.

[21] B. Chor and N. Gilboa (1997) *Computationally private information retrieval.* In Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC'97), El Paso, Texas, May 4–6, 1997 ACM Press, New York, pp. 304–313.

[22] B. Chor, O. Goldreich, E. Kushilevitz and M. Sudan (1998) *Private information retrieval.* Journal of the ACM. 45, pp. 965–981.

[23] B. Chor, N. Gilboa, and M. Naor. *Private information retrieval by keywords.* Technical Report TR-CS0917, Dept. of Computer Science, Technion, 1997.

[24] F. De Clerck, J.A. Thas and H. Van Maldeghem (1996) *Generalized polygons and semipartial geometries.* EIDMA minicourse.

[25] C.J. Colbourn and R. Mathon (2007) Steiner Systems. In the Second Edition of C.J. Colbourn, J.H. Dinitz (Eds.) "The CRC Handbook Of Combinatorial Designs," CRC Press, Boca Raton, FL,pp. 102-110.

[26] Dalenius, T. (1986) Finding a needle in a haystack. Journal of official statistics, 2:3, pp. 329–336.

[27] M. DeSoete, K. Vedder, M. Walker (1990) *Cartesian authentication schemes.* Advances in Cryptology – Eurocrypt 89, Lecture Notes in Computer Science 434, Springer-Verlag 1990, 476–490.

[28] J. Domingo-Ferrer and M. Bras-Amorós (2008) *Peer-to-peer private information retrieval.* Privacy in Statistical Databases, Lecture Notes in Computer Science, pp. 315–323.

[29] J. Domingo-Ferrer, M. Bras-Amorós, Q. Wu and J. Manjón (2009) *User-private information retrieval based on a peer-to-peer community.* Data Knowl. Eng., 68:11, pp. 1237–1252.

[30] J. Domingo-Ferrer and U. González-Nicolás (2011) *Rational Behavior in Peer-to-Peer Profile Obfuscation for Anonymous Keyword Search.* Information Sciences, to appear.

[31] J. Domingo-Ferrer, A. Solanas and J. Castellà-Roca (2009) *h(k)-private information retrieval from privacy-uncooperative queryable databases.* Online Information Review, 33:4, pp. 720–744.

[32] J. Domingo-Ferrer and V. Torra (2001) A quantitative comparison of disclosure control methods for microdata, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. Zayatz (eds.) "Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies", North-Holland, pp. 111–134.

[33] C. Dwork (2006) *Differential privacy.* Proc. ICALP 2006, Lecture Notes in Computer Science 4052, pp. 1-12.

[34] M. Flanagan, M. Greferath and C. Roessing (2007) *On LDPC codes from (0, 1)-geometries induced by finite inversive spaces of even order.* Workshop on Coding and Cryptography 2007, WCC '07, Versailles, France.

[35] M. Flanagan, M. Greferath and C. Roessing (2007) *An encoding scheme, and a decoding scheme using a series of LDPC codes based on finite inversive spaces.* Technical Publication.

[36] M.J. Freedman and R. Morris (2002) *Tarzan: A peer-to-peer anonymizing network layer.* In Proceedings of ACM Conference on Computer and Communications Security, CCS02, ACM Press, Washington, DC, USA, pp. 193–206.

[37] A. Gács and T. Héger (2008) *On geometric constructions of (k, g)-graphs.* Contributions to Discrete Mathematics, **3:1**, pp. 63–80.

[38] W. Gasarch and A. Yerukhimovich (2006) *Computational inexpensive PIR.* http://www.cs.umd.edu/arkady/papers/pirlattice.pdf.

[39] H. Gropp (2007) Configurations. In the Second Edition of C.J. Colbourn, J.H. Dinitz (Eds.) "The CRC Handbook Of Combinatorial Designs", CRC Press, Boca Raton, FL, pp. 352–355.

[40] H. Gropp (1994) *Nonsymmetric configurations with natural index. Discrete Math.*, 124:1-3, pp. 87–98. Graphs and combinatorics (Qawra, 1990).

[41] H. Gropp (1992) *Non-symmetric configurations with deficiencies 1 and 2. Combinatorics '90 - Recent Trends and Applications,* In Proceedings of the Conference on Combinatorics, Gaeta, Italy, 20-27 May 1990, pp. 227–240.

[42] B. Grünbaum, "Configurations of Points and Lines." American Mathematical Society, Providence, RI, 2009.

[43] M. Hall (1943) *Projective planes.* Transactions of the American Mathematical Society 54:2, pp. 229–277.

[44] M. Hall, "The Theory of Groups." The Macmillan Co., New York, 1959.

[45] D. Hilbert, "The Foundations of Geometry." The Open Court Publishing Company, La Salle, Illinois, 1950.

[46] D. Hilbert and S. Cohn-Vossen, "Anschauliche Geometrie", 1932. English translation: "Geometry and the imagination". Chelsea, New York, 1952. Second Ed., Springer, Berlin, 1996.

[47] D.C. Howe and H. Nissenbaum (2009) *Trackmenot: Resisting surveillance in web search.* In Lessons from the Identity Trail: Privacy, Anonymity and Identity in a Networked Society, I. Kerr, C. Lucock and V. Steeves (Eds.), Oxford University Press, Oxford, UK, 2009. Software downloadable from: http://www.mrl.nyu.edu/dhowe/trackmenot/.

[48] D.R Hughes and F.C Piper, "Projective Planes." Springer-Verlag, New York, 1973.

[49] P. Kaski and P.R.J. Östergård (2007) *There exists no symmetric configuration with 33 points and line size 6.* Australasian Journal of Combinatorics 38, pp. 273–277.

[50] A. Kersten (1992) *Secret Sharing Schemes aus geometrisher Sicht.* Mitt. Math. Sem. Univ. Giessen 208.

[51] E. Kushilevitz and E. Ostrovsky (1997) *Replication is not needed: Single database, computationally-private information retrieval.* In Proceedings of the 38th Annual IEEE Symposium on Foundations of Computer Science, pp. 364–373.

[52] F. Lazebnik, V. A. Ustimenko and A. J. Woldar (1997) *New upper bounds on the order of cages.* Electronic Journal of Combinatorics, 4:2, Research Paper 13, approx. 11 pp. (electronic).

[53] J. Lee and D.R. Stinson (2005) *A combinatorial approach to key predistribution for distributed sensor networks.* IEEE Wireless Communications and Networking Conference, CD-ROM, paper PHY53-06, 6 pp.

[54] J. Lee and D.R. Stinson (2008) *On the construction of practical key predistribution schemes for distributed sensor networks using combinatorial designs.* ACM Trans. Inf. Syst. Secur., 11:2, pp. 1:1–1:35.

[55] F. Levi, "Geometrische Konfigurationen." Hirzel, Leipzig, 1929.

[56] V. Martinetti (1886) *Sopra alcune configurazioni piane.* Annali di Matematica Pura ed Applicata (1867 - 1897), 14, pp. 161–192.

[57] J.M.F Moura, J. Lu and H. Zhang (2004) *Structured LDPC codes with large girth.* IEEE Signal Processing Magazine, Included in Special Issue on Iterative Signal Processing for Communications, 21:1, pp. 42–55.

[58] S.E. Payne and J.A. Thas, "Finite Generalized Quadrangles." European Mathematical Society (EMS), Zürich, 2009.

[59] T. Pisanski (2007) *Yet another look at the Gray graph.* New Zealand Journal of Mathematics, 36, pp. 85–92.

[60] T. Pisanski, M. Boben, D. Marušič, A. Orbanić and A. Graovac (2004) *The 10-cages and derived configurations.* Discrete Mathematics, 275:1-3, pp. 265–276.

[61] J.L. Ramírez Alfonsín, "The Diophantine Frobenius Problem." Oxford University Press, Oxford, 2005.

[62] M. Raykova, B. Vo, S. Bellovin and Tal Malkin (2009) *Secure Anonymous Database Search.* In Proceedings of CCSW09, Chicago, Illinois, USA.

[63] M. Rennhard and B. Plattner (2004) *Practical anonymity for the masses with MorphMix.* Financial Cryptography, pp. 233–250.

[64] T. Reye, "Geometrie der Lage." I. Second Edition, Alfred Krumloner Verlag, Leipzig, 1876.

[65] J.C. Rosales and P.A. García-Sánchez, "Numerical Semigroups." Springer, New York, 2009.

[66] H. Sachs (1963) *Regular graphs with given girth and restricted circuits.* J. London Math. Soc., 38, pp. 423–429.

[67] P. Samarati (2001) *Protecting Respondents' Identities in Microdata Release.* IEEE Trans. on Knowledge and Data Engineering, 13:6,pp. 1010-1027.

[68] P. Samarati and L. Sweeney (1998) *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* SRI Intl. Tech. Rep.

[69] K. Sinha (1978) *A triangle free configuration.* Československá Akademie Věd. Časopis Pro Pěstování Matematiky, 103:2, pp. 147–148.

[70] A. Spink, D. Wolfram, M. B. J. Jansen, T. Saracevic (2001) *Searching the web: The public and their queries.* Journal of the American Society for Information Science and Technology, 52:3, pp. 226-234.

[71] K. Stokes and M. Bras-Amorós (2010) *Optimal configurations for peer-to-peer user-private information retrieval.* Computers & Mathematics with Applications, 59:4, pp. 1568 – 1577.

[72] K. Stokes and M. Bras-Amorós (2011) *Associating a numerical semigroup to the triangle-free configurations.* Advances in Mathematics of Communications, 5:2, pp. 351 – 371.

[73] K. Stokes and M. Bras-Amorós (2011) *On query self-submission in peer-to-peer user-private information retrieval.* In Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society (PAIS '11), Uppsala, Sweden.

[74] K. Stokes and V. Torra, *Reidentification and k-anonymity: a model for disclosure risk in graphs.* Submitted.

[75] L. Storme (2007) *Finite geometry.* In the Second Edition of C.J. Colbourn, J.H. Dinitz (Eds.) "The CRC Handbook Of Combinatorial Designs", CRC Press, Boca Raton, FL, 2007, pp. 702–729.

[76] SVT Kulturnyheterna, 16 May 2011, *Så kartlägger Google ditt liv.* http://svt.se/2.27170/1.2427089/sa_kartlagger_google_ditt_liv

[77] L. Sweeney (2002) *Achieving k-anonymity privacy protection using generalization and suppression.* Int. J. of Unc., Fuzz. and Knowledge Based Systems, 10:5, pp. 571–588.

[78] L. Sweeney (2002) *k-anonymity: a model for protecting privacy.* Int. J. of Unc., Fuzz. and Knowledge Based Systems, 10:5, 557–570.

[79] G. Tarry (1900) *Le probleme des 36 officiers.* Comptes-Rendus de l'Association Française pour I'Avancement des Sciences, Congrès de Paris 1900, pp. 170–204.

[80] J. Teevan, E. Adar, R. Jones, M. Potts (2005) *History repeats itself: Repeat Queries in Yahoo's query logs.* In Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '06). pp. 703-704.

[81] The Tor project, inc. Tor: Overview. http://torproject.org/overview.html.en.

[82] B. Vasic and O. Milenkovic (2004) *Combinatorial constructions of low-density parity-check codes for iterative decoding.* IEEE Transactions on Information Theory, 50:6, pp. 1156–1176.

[83] O. Veblen and J. W. Young, "Projective geometry." Ginn and company, Boston, New York, 1910.

[84] A. Viejo and J. Castellà-Roca (2010) *Using social networks to distort users' profiles generated by web search engines.* Computer Networks, 54:9, pp. 1343 – 1357.

[85] E. Visconti (1916) *Sulle configurazioni piane atrigone.* Giornale di Matematiche di Battaglini, 54, pp. 27–41.

[86] S. Wang, X. Ding, R. Deng, F. Bao, *Private Information Retrieval Using Trusted Hardware,* In Lecture Notes in Computer Science, "Computer Security ESORICS 2006", Springer Berlin / Heidelberg, 2006.

[87] C. Weibel (2007) *Survey of non-desarguesian planes-* Notices of the AMS, 54:10.

[88] W.E. Winkler (2004) *Re-identification methods for masked microdata.* PSD 2004, Lecture Notes in Computer Science 3050, pp. 216–230.

[89] P.K. Wong (1982) *Cages–a survey*, Journal of Graph Theory, 6:1, pp. 1–22.

Appendix

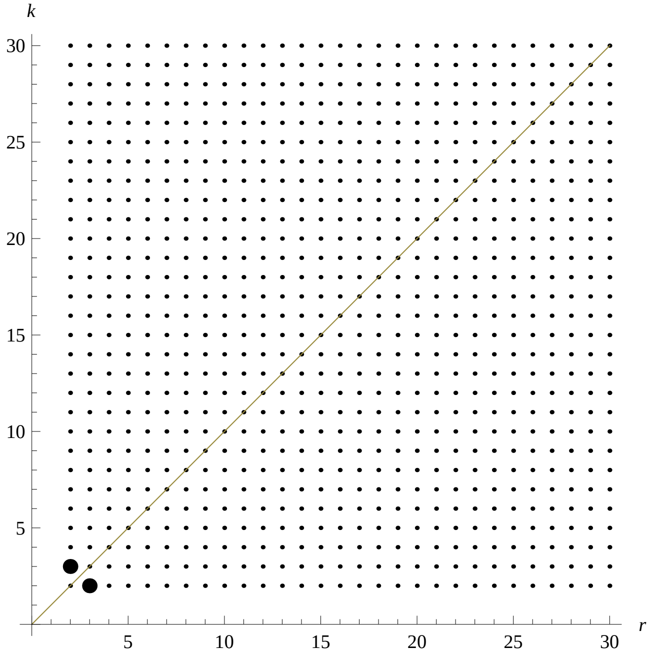

Figure 1: Admissible (r, k) for $d = 2$

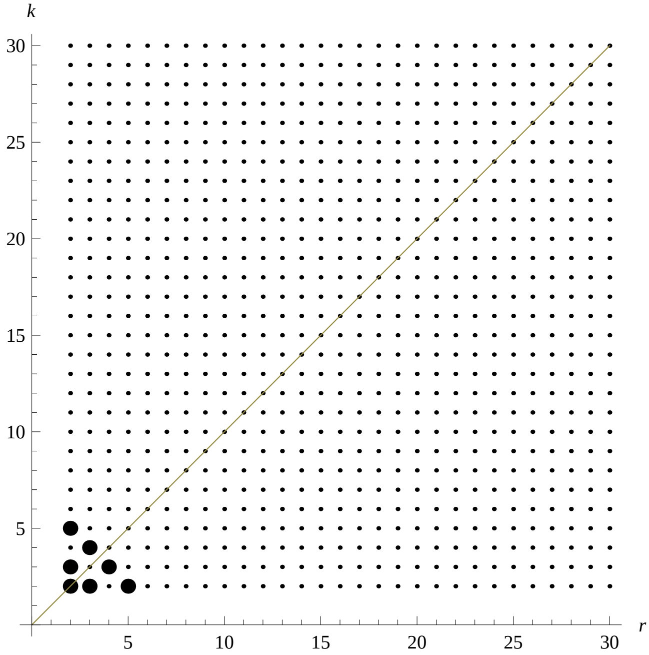

Figure 2: Admissible (r, k) for $d = 3$

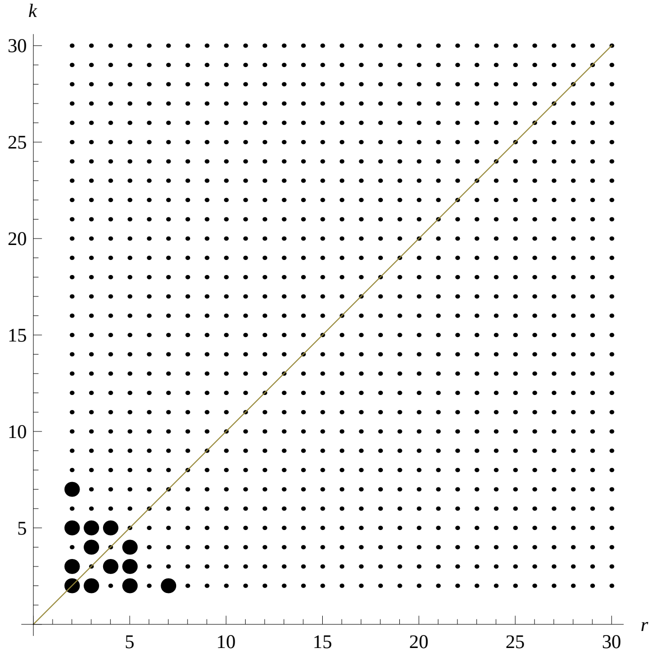

Figure 3: Admissible (r, k) for $d = 4$

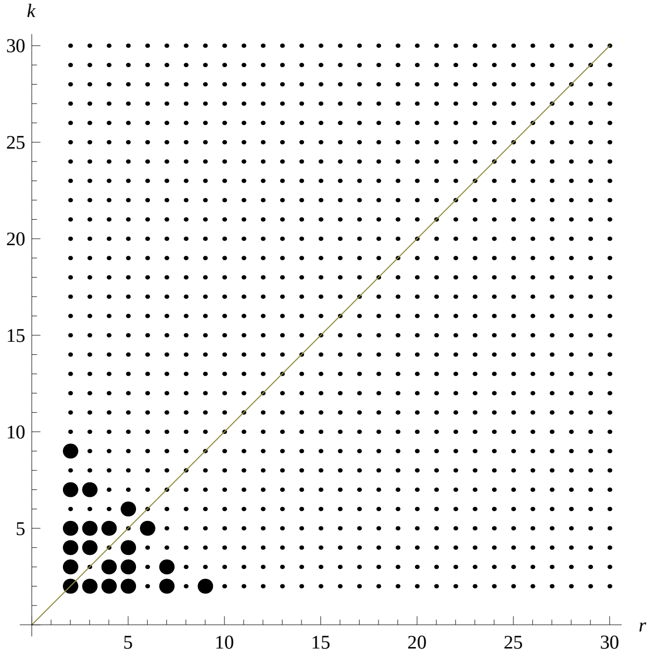

Figure 4: Admissible (r,k) for $d=5$

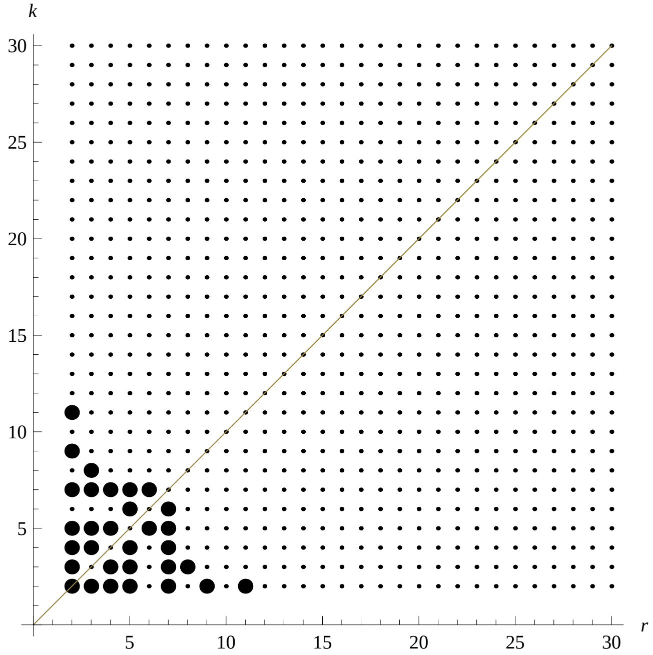

Figure 5: Admissible (r, k) for $d = 6$

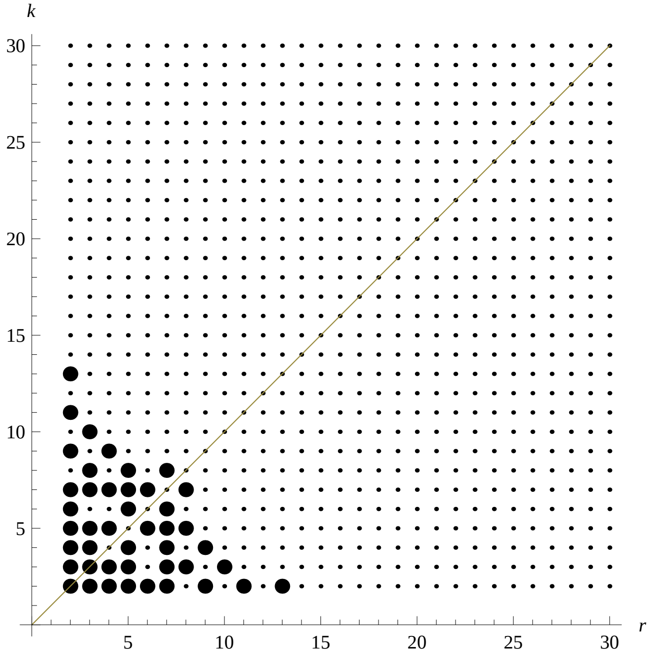

Figure 6: Admissible (r, k) for $d = 7$

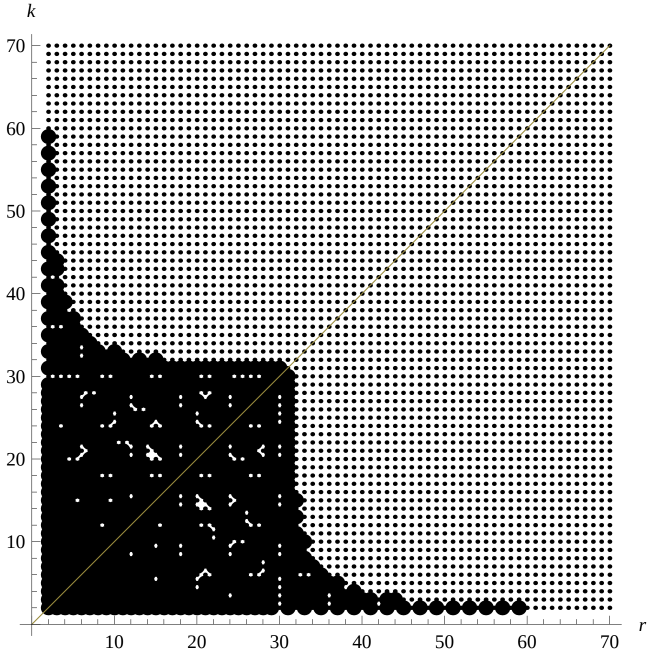

Figure 7: Admissible (r, k) for $d = 30$